UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

Inclusión de una perspectiva sintáctica a la obtención de pares semánticos

T E S I S

que para obtener el título de
Licenciada en Lengua y Literaturas Hispánicas

p r e s e n t a

Sonia Elisa Morett Álvarez

Asesor: Dr. Gerardo Sierra Martínez

México, D. F. Ciudad Universitaria 2006

A la generación de estudiantes que entre 1999 y 2000, con esfuerzo y organización, permitió que la UNAM siga siendo pública

PREFACIO

La presente tesis ha sido diseñada para optar por el grado de licenciada en lengua y literaturas hispánicas. La investigación que le dio origen fue financiada por el proyecto CONACYT R37712A.

Agradecimientos:

A Gerardo Sierra, que con su experiencia supo darme las pautas para encaminar este proyecto, por su interés y dedicación en la supervisión de todas las etapas de la tesis y por su paciencia.

A Gabriel Castillo, Javier Cuétara, Alfonso Medina y Ramón Zacarías, quienes desde diferentes ámbitos mostraron la mejor disposición para hacer aportaciones que enriquecieron mi trabajo.

Al Grupo de Ingeniería Lingüística. A los lingüistas, especialmente a Antonio y a César, siempre dispuestos a escucharme y a leerme, por ayudar a resolver mis dudas unas veces y por ampliármelas otras, acercándome al conocimiento de textos y teorías. Y sobre todo a los computólogos, que dedicaron mucho de su tiempo a permitirme salir bien librada de los misterios de la informática.

A Chuy, de quien aprendí a luchar; a María, que siempre me ha apoyado en todo lo posible y en todo lo imposible; a Lucía, que me ha enseñado importantes lecciones de vida, por haberme acompañado con decisión en este proceso. A toda mi familia.

A Lolín, por alimentar este trabajo ocupándose tan deliciosamente de mí.

A mis compañeros de estos años por el entusiasmo compartido y su solidaridad. A Alberto, con quien coincidí en el camino y decidimos abonar rumbos juntos, por esa pasión por la vida que contagia.

ÍNDICE GENERAL

CAPÍTULO 2. ANÁLISIS DEL FUNCIONAMIENTO DE LA HERRAMIENTA

2.1. ***Delimitación del objeto de estudio***

2.2. ***Pruebas***

2.2.1. Obtención de pares semánticos manualmente

2.2.2. Pruebas realizadas modificando los criterios de búsqueda del Algoritmo flexibilizado de alineamiento semántico

2.2.3. Resultados cuantitativos de las pruebas

2.2.4. Interpretación global de los resultados

2.3. ***Evaluación de cada eje del algoritmo flexibilizado***

2.3.1. Pares semi iguales

2.3.2. Pares semi nulos

2.3.3. Distinguir mayúsculas y minúsculas

2.3.4. Intercambio

2.3.4.1. **Conjuntivo**

2.3.4.2. **No conjuntivo**

2.4. ***Conclusiones***

2.5. ***Recapitulación***

CAPÍTULO 3. ANÁLISIS SINTÁCTICO Y PROPUESTAS

3.1 ***Perífrasis gramaticales***

3.1.1. Locuciones

3.1.2. Términos compuestos

3.1.3. Perífrasis verbales

3.2. ***Nexos***

3.2.1. Conjunciones

3.2.2. Preposiciones

3.2.2.1. **Verbos que rigen preposición**

3.2.2.2. **Locuciones prepositivas**

3.2.3. Verbos copulativos

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El tema que nos proponemos desarrollar se inscribe en un área interdisciplinaria de investigación aplicada que es la ingeniería lingüística, la cual aprovecha los conocimientos de la lengua obtenidos mediante la investigación para mejorar el funcionamiento de los sistemas informáticos, perfeccionando la interacción hombre-computadora con la aportación de medios para la interpretación y generación de lenguaje natural.

Al observar que, en las últimas décadas, la implementación de mecanismos para el almacenamiento de gigantescas colecciones de texto en soporte electrónico, a la que asistimos con la comercialización de internet, no se ha visto correspondida con el desarrollo de las herramientas suficientes para su procesamiento, hemos decidido realizar nuestra tesis de licenciatura en esta área, con la intención de aportar algo que pueda servir a otras investigaciones en la materia.

Concretamente, nuestro trabajo consistirá en el análisis del funcionamiento de un programa computacional denominado Clustering, que genera automáticamente agrupamientos semánticos[1] y en la elaboración de una propuesta para mejorar algunos aspectos deficientes durante el proceso que pueden ser abordados desde la sintaxis.

Dicho programa se basa en la comparación de las palabras que componen parejas de definiciones sobre un mismo término que se encuentran contenidas en un banco terminológico[2]. El mismo fue diseñado conjuntamente por el Grupo de Ingeniería Lingüística (GIL) de la UNAM y el Instituto de Ciencia y Tecnología (UMIST) de Manchester en 1999. Tres años después el sistema fue objeto de una revisión minuciosa.

Con miras a la elaboración de esta tesis, nos apoyamos en los materiales que el GIL ha producido en relación con Clustering para realizar una investigación preliminar que consistió en la aplicación aleatoria de pruebas al sistema sobre la base del banco de términos y definiciones en el área de Física. Del ejercicio surgieron temas de interés para abordarse desde una perspectiva lingüística (que se incorporaron a los ya propuestos en los textos estudiados), así como el proyecto para el presente trabajo.

En el marco de las líneas elegidas para desarrollar la investigación cabe preguntar: ¿Es posible subsanar las deficiencias de Clustering relativas a la

1 Grupos de palabras que en determinados contextos pueden funcionar como sinónimos.
2 Base de datos conformada por términos esenciales para algunas áreas del conocimiento y determinada información vinculada a cada uno, como enunciados, extraídos de diferentes contextos, que sirven para definirlos.

lengua desde una perspectiva integradora, que no atienda particularidades sino que busque acercarse a una solución considerando previamente el funcionamiento de nuestro idioma a nivel sintáctico de manera global? ¿Podrían, las propuestas resultantes de una investigación así, ser fácilmente aprovechables para su formalización matemática?

Si hiciéramos un etiquetado sintáctico a los actuales bancos terminológicos obtendríamos una solución temporal: la desambiguación dentro de un sistema de alineamiento de pares de definiciones sólo para un conjunto finito. A medida que se agregaran nuevas definiciones a la base de datos se tendría que anexar su análisis manual, lo que no resulta funcional; así que sólo consideramos la posibilidad de desarrollar propuestas automatizables.

La importancia del motivo elegido para esta investigación reside en que la herramienta que se somete a revisión es pieza fundamental de un diccionario onomasiológico[3], pues el programa de agrupamiento semántico obtiene automáticamente las imprescindibles palabras clave para la búsqueda de términos a partir de un incontable número de definiciones. La utilidad final de la tesis se verá en el conjunto del diccionario sobre el que se trabaja —con un esfuerzo colectivo— en el GIL.

Se pretende que, a través de internet, un estudiante (en el caso concreto del diccionario de Física) pueda con sus propias palabras explicar aquello que no sabe cómo nombrar pero conoce que existe y que el sistema le proporcione el término apropiado para ese referente. Clustering, con los mismos principios, puede ser aplicado a cualquier colección de términos y definiciones.

Nuestro objetivo con la presente tesis es hacer uso del análisis sintáctico para sentar las bases de modificaciones que redunden en un mejor funcionamiento del programa de agrupamiento de pares semánticos. Es decir, aportar información lingüística al proceso de recuperación de la información (dentro de un banco terminológico) que se requiere para preestablecer modelos de búsqueda para un diccionario como el que se ha descrito.

Clustering presenta una serie de deficiencias derivadas de que fue diseñado originalmente para la lengua inglesa; los puntos vulnerables para el español no fueron reforzados con las modificaciones al algoritmo llevadas a cabo posteriormente porque el trabajo que les dio origen fue de carácter matemático y no lingüístico. Nuestra hipótesis de trabajo es que la incorporación de información gramatical servirá para mejorar el proceso de recuperación de pares semánticos.

La metodología que se seguirá en la tesis es la de la lingüística de corpus. Con base en la clasificación formal de las palabras, y a partir de la teoría gramatical de Alarcos Llorach y de Rojo y Jiménez Juliá se procederá a realizar un análisis de tipo funcional. El análisis sintáctico tiene como propósito llegar a deducir

3 Tipo de diccionario en que, partiendo de una definición, se busca acceder al término que le corresponde. En el GIL se encuentra en desarrollo un diccionario con esta característica. En ello profundizaremos en el Capítulo 1.

reglas lo más generales posibles para posteriormente poderlas aplicar a nuestro corpus de manera semiautomática.

En el capítulo 1 se presentará el programa y se expondrán los conceptos básicos asociados al funcionamiento de éste; además se hablará sobre el *Diccionario Electrónico de Búsqueda Onomasiológica* y su relación con Clustering.

En el capítulo 2 describiremos las pruebas que deberán realizarse al sistema en su estado actual y definiremos el corpus al que se le aplicarán dichas pruebas (que será el mismo que se empleará en los diferentes análisis de la tesis). Además registraremos los resultados de las pruebas y haremos las evaluaciones pertinentes.

El capítulo 3 se destinará a explicar detalladamente los aspectos de interés sintáctico observables en las definiciones con las que trabaja el programa que —consideramos— nos pueden llevar a configurar cambios para incorporar al sistema. Se hará una propuesta específica para cada eje de análisis y se aportarán ejemplos que la sustenten.

El capítulo 4 registrará el procedimiento aplicado, los resultados obtenidos y nuestra evaluación a un experimento que derivará de las propuestas que hayan sido vertidas en el capítulo 3.

Por último, se extraerán conclusiones generales y se aportarán ideas para trabajos futuros.

Al concluir esta tesis se habrán definido las líneas para abordar, desde una perspectiva lingüística, las deficiencias que todavía tiene Clustering y se habrán desarrollado propuestas metodológicas para superarlas.

1. INTRODUCCIÓN A CLUSTERING

Debido a que este trabajo consiste en una propuesta de mejoramiento para el programa de agrupamiento semántico empezaremos por conocerlo; con tal propósito presentaremos, en primer lugar, el proyecto del que forma parte, el *Diccionario Electrónico de Búsqueda Onomasiológica* (1.1); contar con este antecedente es necesario para entender la utilidad de Clustering. El apartado 1.2 se destinará a describir el programa, que cuenta ya con dos versiones: el Algoritmo básico de agrupamiento semántico (1.2.1) y el Algoritmo flexibilizado de agrupamiento semántico (1.2.3). Terminaremos este capítulo señalando algunas limitaciones de Clustering (1.3).

1.1. ¿Qué es el DEBO?

Hasta este momento, el *Diccionario Electrónico de Búsqueda Onomasiológica* (DEBO)[1] constituye el desarrollo más importante del GIL. Sus características principales están contenidas en el propio nombre: electrónico y onomasiológico.

Con *electrónico* se quiere decir simplemente que la colección de voces y sus definiciones se presenta en un formato digital y que la información que ésta contiene se recupera mediante un sistema computacional, en este caso, disponible en internet. Las ventajas de los diccionarios respaldados electrónicamente son que se pueden consultar en la misma computadora en la que se está trabajando; el sistema corrige los errores ortográficos del usuario y la búsqueda se realiza de forma automática (al teclear la información que nos puede llevar a descubrir lo que deseamos, sin necesidad de buscarla entre miles de posibilidades conforme a ciertos criterios de ordenación, generalmente alfabéticos). Además, se pueden consultar simultáneamente varios diccionarios. En síntesis, con los diccionarios electrónicos se obtiene una ganancia en tiempo y una disminución de errores.

Lo *onomasiológico* se refiere a que el diccionario se enfoca a la búsqueda de términos a partir de la descripción del concepto[2]. La existencia de este tipo de diccionarios ofrece muchos beneficios a cualquier usuario, porque coadyuva a incrementar su léxico activo, pero particularmente a los especialistas en determinada materia porque agiliza su trabajo y la comunicación entre colegas.

---

1 *Vid* Gerardo Sierra, *Avances en el desarrollo del Diccionario Electrónico de Búsqueda Onomasiológica.*

2 Una de las clasificaciones de los diccionarios se debe al modo de búsqueda. Bajo este criterio se distinguen los semasiológicos y los onomasiológicos. Los diccionarios semasiológicos se apoyan en un término para acceder a los significados que le corresponden; mientras que, en los onomasiológicos, la incógnita es el término, pues se puede describir el concepto pero no se recuerda el nombre (significante) que lo denota.

Los diccionarios onomasiológicos y electrónicos, en nuestros días, tienen una estrecha relación entre sí, que podría calificarse como *lógica.* Cuando la tendencia a respaldar electrónicamente todo tipo de documentos alcanzó a los diccionarios, lo primero que se hizo fue recuperar en formato digital algunos ya existentes. En seguida, se advirtió que contar con una base de datos computarizada formada por términos y sus correspondientes definiciones (sea cual fuere su propósito original) admite dos formas de búsqueda, desde el punto de vista semántico: la semasiológica, en donde el usuario teclea un término para que el sistema recupere el o los significados vinculados a éste, y la onomasiológica, en donde se inserta una definición con la intención de obtener el término que se desconoce o no se recuerda.

Como es notorio, en la era digital, los diccionarios onomasiológicos han adquirido mayor relevancia sobre las posibilidades del estático libro impreso, gracias a la enorme capacidad de almacenamiento de los sistemas computacionales y a la flexibilidad que ofrece la automatización de procesos para el acceso a la información contenida en ellos. Tal como se señala en el artículo "Criterios para una fraseografía onomasiológica automatizable"[3]:

> Una de las novedades más interesantes de la lexicografía en soporte electrónico fue el que la ordenación onomasiológica comenzara a ser una alternativa realmente operativa. […] Una de las razones de que el desarrollo de los diccionarios onomasiológicos [sin soporte electrónico] haya sido menor en la práctica de lo que su evidente necesidad hacía suponer [es] de tipo técnico: aun suponiendo que el lexicógrafo fuese tan ingenioso como para lograr una ordenación nocional "indiscutible", seguiría siendo muy trabajoso, incierto y lento para el usuario encontrar lo que busca en semejantes inventarios léxicos, ya que éste no tiene por qué conocer o compartir los criterios del autor.
>
> […] La irrupción del soporte informático permitió en principio aplicar los dos sistemas (semasiológico y onomasiológico) a una misma colección, tanto de palabras aisladas como de unidades complejas. En informática es sabido que lo que hoy es posible, mañana será obligatorio, y este tándem pronto se ha de convertir en norma.

El hecho de consultar un diccionario a través de una computadora ofrece los beneficios que ya hemos mencionado, pero no hace más eficientes las búsquedas onomasiológicas automáticamente; esto se debe a que no se libera al usuario de la dificultad de situarse en el lugar del lexicógrafo, ya que debe insertar una descripción al sistema que coincida plenamente con los criterios y con las *palabras clave*[4] que este último empleó durante la realización del diccionario. Para que las posibilidades de éxito en estas búsquedas se multipliquen, es necesario que las definiciones contenidas en el diccionario reciban algún tipo de preprocesamiento y así puedan ser explotadas con diferentes herramientas para el tratamiento automático de bases de datos léxicos.

En el diseño del DEBO —como un diccionario electrónico y onomasiológico— se ha considerado todo lo anterior y el resultado es un sistema de recuperación de información que se encuentra en desarrollo. Actualmente, abarca las áreas

3 Antonio Pamies et.al., *Criterios para una fraseografía onomasiológica automatizable*, p.1.
4 Una palabra clave es cualquier voz que pueda considerarse importante dentro de una definición.

temáticas de *desastres, Física* y *Lingüística*; se nutre con la información léxica contenida en un banco terminológico implementado en el GIL para este propósito, el cual contiene definiciones (en español y, en algunas áreas, también en inglés) provenientes de diferentes fuentes para cada uno de los términos que maneja el mismo banco. Se puede consultar por internet en la página: http://iling.torreingenieria.unam.mx/diccionarios.

La primera versión del DEBO se implementó para una base de datos que contenía definiciones para 33 términos en el área de desastres. En esta versión se sentaron las bases del diccionario, cuyo mecanismo consiste en que el usuario introduce los datos de su búsqueda en lenguaje natural por medio de una redacción libre y sin restricciones de sintaxis; a partir de esa descripción se extraen las palabras clave, las que se contrastan con la información asociada a cada término de la base de datos. Finalmente, el sistema proporciona el o los términos más probables, acompañados de su definición, con la finalidad de que el usuario pueda comprobar sí efectivamente encontró el término que buscaba[5]. Veamos lo que obtendríamos de una búsqueda ideal:[6]

> The ideal onomasiological search must allow writers to input the concept to be searched through the ideas they may have, using any words in any order. The system must be so constructed that it accepts a wide range of words which it then analyses in order to point the user to the word that most closely approaches the concept he had in mind when he started the search.

Para acercarnos a esta búsqueda es necesario expandir las palabras clave que emplea el usuario. Una alternativa para tal propósito consiste en la determinación de *paradigmas semánticos*, formados por "el conjunto de palabras clave que presentan rasgos comunes y que pueden ser utilizadas con el mismo sentido en el contexto del término final al que corresponden, esto es, puede ser sustituido cualquier miembro del paradigma en los elementos correspondientes del sintagma sin cambiar el significado del mismo."[7]

Así por ejemplo, *realizar, producir, hacer* y *efectuar* constituyen un paradigma semántico, pues en las siguientes definiciones de *energía* tomadas del mencionado banco terminológico podemos sustituir la palabra correspondiente del paradigma por cualquiera de las otras tres sin alterar el sentido de la frase:

5 En la siguiente dirección electrónica se puede acceder a una versión de muestra del programa: http://tabasco.torreingenieria.unam.mx/debo/ventana.html.
6 Gerardo Sierra, *Design of a concept-oriented tool for terminology*, p. 34.
7 Gerardo Sierra, *Avances en,* p. 2.

- Capacidad de un sistema físico para **realizar** trabajo.[8]
- Capacidad de un sistema material para **producir** trabajo, con las propiedades de la conservación y la interconvertibilidad.[9]
- Propiedad de un sistema —su capacidad para **hacer** un trabajo[10].
- Es la capacidad de **efectuar** trabajo.[11]

A partir de estas sustituciones podríamos formar doce nuevas frases, todas referidas a energía:

- Capacidad de un sistema material para **realizar** trabajo, con las propiedades de la conservación y la interconvertibilidad.
- Propiedad de un sistema —su capacidad para **realizar** un trabajo.
- Es la capacidad de **realizar** trabajo.
- Capacidad de un sistema físico para **producir** trabajo.
- Propiedad de un sistema —su capacidad para **producir** un trabajo.
- Es la capacidad de **producir** trabajo.
- Capacidad de un sistema físico para **hacer** trabajo.
- Capacidad de un sistema material para **hacer** trabajo, con las propiedades de la conservación y la interconvertibilidad.
- Es la capacidad de **hacer** trabajo.
- Capacidad de un sistema físico para **efectuar** trabajo.
- Capacidad de un sistema material para **efectuar** trabajo, con las propiedades de la conservación y la interconvertibilidad.
- Propiedad de un sistema —su capacidad para **efectuar** un trabajo.

Al considerar el paradigma {capacidad, propiedad} para nuevas sustituciones a este grupo de definiciones, las expresiones resultantes se duplicarían. La idea es que alguien que ha olvidado el término energía pueda acceder al diccionario para explicar el concepto que quiere nombrar. Si el sistema no incluyera la expansión de la búsqueda a paradigmas semánticos, las descripciones del usuario que no se correspondieran con alguna de las definiciones obtenidas del banco terminológico no conducirían a la obtención de dicho término. Ahora bien, si a la memoria del diccionario se incorpora los paradigmas {realizar, producir, hacer, efectuar} y

8 Encarta 98.
9 Océano Uno Color.
10 Definición traducida de *Dictionary of physics*.
11 Física Weber.

{capacidad, propiedad}, el usuario puede decir que lo que busca “es la propiedad de realizar trabajo”, o “es la capacidad de producir trabajo” o “es la propiedad de hacer trabajo” y el resultado de su búsqueda será el mismo que si inserta la definición que el diccionario contiene: “Es la capacidad de efectuar trabajo”.

Entonces, la ampliación de los criterios de búsqueda en un diccionario onomasiológico a la consideración de paradigmas semánticos multiplica (por seis, en este sencillo ejemplo) las probabilidades del potencial usuario para acceder al término *energía*.

Para obtener automáticamente estos paradigmas o agrupamientos semánticos se emplea el programa Clustering.

1.2. ¿Qué es Clustering?

Clustering es una palabra inglesa que deriva de *cluster*, la cual puede definirse como “a number of things of the same kind, growing or held together.”[12] Coincidiendo con esto, en el diccionario bilingüe, por ejemplo, encontramos como sustantivo las palabras “grupo, racimo, enjambre, manada”, y como verbo los verbos “agruparse” o “arracimarse”.[13]

Las ciencias de la computación adoptaron este término con dos significados distintos. Uno de los cuales se refiere al “proceso de agrupar datos en clases o clusters de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y baja (sean muy diferentes) con objetos de otros clusters. […] La medida de similaridad está basada en los atributos que describen a los objetos.”[14]

Por tanto, un *cluster* puede ser un agrupamiento de formas lingüísticas en torno a una característica que denote similitud y ésta, a su vez, puede ser un aspecto semántico común a sus respectivos significados, pero sin tratarse de una relación de sinonimia plena.

Con la idea anterior, se desarrolló el programa de agrupamiento semántico denominado Clustering[15], el cual se orienta al trabajo del DEBO. Como se mencionó en la Introducción, este programa es un sistema de recuperación de información que tiene como finalidad agrupar formas lingüísticas con características semánticas comunes.

En una primera etapa, el sistema alinea dos definiciones de un mismo término; el procedimiento se repite hasta agotar todas las posibles combinaciones de pares de

12 *Webster’s Universal College Dictionary*, p.152.
13 Ramón García-Pelayo y Gross et. al., *Gran Diccionario Larousse Español –Inglés*, p. 834.
14 Eduardo F. Morales, *Descubrimiento de Conocimiento en Bases de Datos*, 23
15 En este trabajo hablaremos de agrupamientos semánticos y no de clusters, pero llamaremos Clustering al programa de alineamiento de pares semánticos, por ser el nombre propio que sus creadores le dieron.

definiciones para cada término. Una vez con los alineamientos llevados a cabo, se localizan los pares iguales[16], los pares nulos[17] y los pares correspondientes[18].

Siguiendo nuestro ejemplo, contamos con cuatro definiciones de energía; de la comparación de la secuencia paralela de sus formas obtendremos seis alineamientos; tomemos uno de ellos:

Def. 1 Energía	Capacidad de		un	sistema	material	para	producir	trabajo	con	las	…	
Def. 2 Energía	Capacidad de		un	sistema	físico	para	realizar	trabajo				
Tipo	par igual	par igual	par igual	par igual	par igual	par corres-pondiente	par igual	par corres-pondiente	par igual	par nulo	par nulo	pares nulos

Posteriormente y mediante el cálculo automático de un coeficiente de similitud denominado LCC (*Longest Collocation Couple*), se determina si los pares correspondientes se promueven a *pares vinculados.* Éstos son los que el sistema identifica automáticamente como posibles *pares semánticos* (pares de formas que en un contexto determinado pueden intercambiarse sin alterar el significado de la estructura que los engloba). Cuando se han establecido los pares vinculados, el proceso se repite iterativamente utilizando las definiciones resultantes del proceso de sustitución de pares vinculados hasta que no se generan nuevos.

En cada ciclo, el sistema detecta las palabras que aparecen en más de un par vinculado; de la fusión de los pares que comparten palabras se obtienen los agrupamientos semánticos. El procesamiento está descrito en forma muy clara en el artículo "Algoritmo flexibilizado de agrupamiento semántico"[19]:

> Los agrupamientos se generan a través de la siguiente *regla de transitividad entre pares-vinculados*:
>
> Sean (*a*,*b*) y (*b*,*c*) dos pares-vinculados formados por las palabras *a*, *b* y *c* ; además, dado que *a* mantiene una relación semántica con *b*, y a su vez *b* mantiene una relación semántica con *c,* entonces se puede afirmar que *a* mantiene una relación semántica con *c*.
>
> Con base en la regla de transitividad de pares-vinculados podemos afirmar que el conjunto {*a*, *b*, *c*} forma un agrupamiento semántico.

Así es como, a partir de los pares {realizar, producir}, {realizar, hacer} y {hacer, efectuar}, se consiguió el agrupamiento {realizar, producir, hacer, efectuar}.

Como ya se mencionó, las definiciones con las que trabaja Clustering previamente fueron reunidas en un banco terminológico; originalmente provienen de diccionarios y textos especializados en la materia a la que se refieren, así como de diccionarios de lengua general. Entonces, el usuario potencial del DEBO no

16 Pares formados por palabras que comparten posición y lema —no necesariamente flexiones.
17 Se consideran así las palabras que dentro de un alineamiento no tienen una paralela en la otra definición o, dicho de otro modo, su par es una palabra vacía.
18 Parejas de palabras diferentes que ocupan una posición paralela dentro del alineamiento.
19 Gabriel Castillo y Gerardo Sierra, *Algoritmo flexibilizado de agrupamiento semántico*, p.76.

consultará estos materiales directamente, sino que accederá a una información ya procesada. Las definiciones contenidas en los diccionarios serán de mayor utilidad para el investigador que trabaja sobre este sistema que para el usuario final del DEBO, tal como lo explica Antonio Zampolli: “Los productos de la tradición lexicográfica, o sea, los diccionarios impresos, son reconocidos ahora no solamente como una de las principales fuentes de datos y de información en torno al lenguaje, sino también como sólidos bancos de conocimiento general, con un papel cognitivo de importancia.”[20]

En este sentido, Clustering tiene la ventaja de que puede aplicarse a otros diccionarios (y no sólo a la base del DEBO) sin la necesidad de efectuar cambios en el sistema.

Además de la información contenida en textos especializados y en diccionarios, para los propósitos de una búsqueda onomasiológica no debe descuidarse la conceptualización de los hablantes a los que se pretende dirigir el DEBO. Por esto, encontramos en el banco terminológico muchas definiciones cuya fuente es *personal* (resultado de una investigación de campo), las cuales enriquecen de manera importante la información léxica en que se sustenta Clustering.

En resumen, la finalidad de Clustering es expandir las posibilidades de la búsqueda onomasiológica: Si la definición que el usuario inserta en su búsqueda no se encuentra en el banco terminológico, el sistema acudirá a detectar las palabras clave para esa búsqueda y remplazará cada una por todas las que componen el agrupamiento semántico del que ésta forma parte; así, se consideran todas las posibles definiciones resultantes y no únicamente las establecidas.

El éxito del DEBO dependerá, en gran medida, de que los agrupamientos que genere Clustering sean precisos. Es por esto que el programa no se da por terminado y en el GIL se continúa trabajando para mejorarlo: El algoritmo original (1.2.1) fue sometido a un análisis profundo, de cuya evaluación bien vale la pena dar a conocer los resultados (1.2.2); a partir de dicho examen se realizaron las modificaciones que dieron lugar a la versión actual del programa (1.2.3).

1.2.1. Algoritmo básico de agrupamiento semántico

Bajo el título de Clustering, Gerardo Sierra y John McNaught desarrollaron entre 1999 y 2000 el programa de alineamiento de pares semánticos que, a partir de una revisión posterior, se identificó con el nombre de *Algoritmo básico de agrupamiento semántico*.

Para el desarrollo de nuestro trabajo resulta necesario describir algunas operaciones que el programa realiza; dejamos sin exponer el cálculo de la distancia de Levenshtein y la asignación de costos, entre otras, porque, aunque son fundamentales para el sistema, quedan fuera de nuestro campo de conocimiento.

---

20 Antonio Zampolli, *Los bancos de datos léxicos: bases multifuncionales de datos léxicos*, p.130.

1.2.1.1. Lematización

Por lematización, en lexicografía, se entiende la eliminación de las formas flexivas de una palabra para obtener su lema, que es "la forma canónica a la que se reduce todo un paradigma flexivo y que se toma como representante de todas las variantes morfológicas de la palabra"[21].

Para la finalidad del programa de agrupamiento semántico y acorde con los sistemas de recuperación de información, la lematización es un proceso automático en el cual se conserva sólo la parte común de diferentes formas flexivas o derivadas, aunque esta última no necesariamente constituya un lexema. Por ejemplo, de *cálculo*, *calculista*, *calculemos*, y *calculadora*, el lema obtenido es *calcul*.

En el agrupamiento semántico es importante saber si dos formas comparten un mismo lema o si se trata de lemas distintos, pues de no aplicarse la lematización el sistema consideraría un par correspondiente a dos palabras que comparten base léxica, siendo que semánticamente son equivalentes. Por esta razón, el sistema incorpora una herramienta que lematiza automáticamente las palabras de las definiciones que se someten a alineamiento. Así, un par como {rayo, rayos} es considerado par igual y no correspondiente.

La lematización se hace con el algoritmo de Porter[22] y sirve sólo para el proceso interno del sistema, pues las definiciones se incorporan en su forma no lematizada y el *output* también se presenta sin lematizar.

1.2.1.2. LCC

El LCC es el coeficiente que determina automáticamente el grado de similitud de un par correspondiente. Se basa en el algoritmo de Wagner and Fisher[23]. Su valor se obtiene al contabilizar el número de pares iguales inmediatamente anteriores o posteriores a un par correspondiente dentro de un alineamiento (hasta la aparición de un par nulo u otro par correspondiente), más una unidad que corresponde al par correspondiente en examen. Además, debe cumplirse con la *condición de frontera*, que es una restricción que indica que se requiere, al menos, un par igual antes y uno después del par correspondiente para que las palabras que lo forman sean susceptibles de considerarse intercambiables**.**

Con base en este índice, el par correspondiente se promueve o no a par vinculado; mientras mayor sea el número, habrá mayor probabilidad de que se trate de un par semántico. Se considera que un par con un LCC igual a 5 que cumple con la condición de frontera es suficientemente bueno para ser un par vinculado, pero también aparecen buenos candidatos con valores de 4 y 3. Clustering ofrece al usuario la

21 Enrique Alcaraz Varó y Ma. Antonia Martínez, *Diccionario de lingüística moderna*, p. 321.
22 Para el español se utiliza una adaptación del algoritmo. Porter, M.F. "An algorithm for suffix stripping", publicado originalmente en *Program*, Vol. 14 no. 3, pp 130-137, July 1980.
23 Wagner, RA and Fisher, MJ. "The String To String Correction Problem", *Journal Association for Computing Machinery*, Vol.21, No.1, 1974, pp. 168-173.

posibilidad de determinar por sí mismo el valor del coeficiente de similitud que considere relevante para la obtención de nuevos pares vinculados.

A continuación mostraremos los índices de LCC para el alineamiento de nuestro ejemplo anterior, considerando pares vinculados aquellos correspondientes cuyo valor sea igual o mayor a 5 (cifra que sugieren los diseñadores del programa):

Def. 1	**Energía**	Capacidad de	un	sistema	material	para	producir	trabajo	con	las	…
Def. 2	**Energía**	Capacidad de	un	sistema	Físico	para	realizar	trabajo			
Tipo	par igual	par igual	par igual	par igual	par igual	par correspondiente	par igual	par correspondiente	par igual	par nulo	par nulo	pares nulos
LCC	0	0	0	0	0	7	0	3	0	0	0	0
Par vinculado	no	no	no	no	no	sí	no	no	no	no	no	no

Como se puede observar, el sistema asigna un LCC igual a 7 al par correspondiente {material, físico}, número que indica que se trata de un posible par semántico —lo que coincide con el análisis de un especialista en Física. El par {producir, realizar} que —conforme a nuestro conocimiento de la lengua como hablantes— es un par semántico, no es promovido a par vinculado porque su LCC es de 3, que, como habíamos mencionado, en ocasiones es suficiente, pero para los valores de esta búsqueda no fue relevante al determinar candidatos a pares semánticos.

Lo que debe buscarse con este coeficiente es que los pares vinculados coincidan con los pares semánticos que una revisión humana del alineamiento identifica. Como el único factor que el sistema considera para proponer pares semánticos es el valor de LCC, resulta muy importante revisar y mejorar constantemente los criterios para calcularlo.

1.2.1.3. Palabras irrelevantes

Para fines de recuperación de información, a las palabras clave de un registro o de una búsqueda se oponen las llamadas *palabras irrelevantes,* que son palabras que previamente han sido analizadas y se han considerado como funcionales, esto es, son elementos que unifican el discurso y le dan coherencia pero semánticamente no son significativas para el propósito de recuperación de información que se persigue.

Al algoritmo que conforma el motor del programa aparece incorporada una lista de palabras irrelevantes, *stoplist.* Cuando el programa identifica un par vinculado en donde una de las palabras que lo forman o ambas aparecen en la *stoplist*, el par es rechazado por el sistema y no se le considera para la integración de agrupamientos semánticos.

Para la primera versión del algoritmo, la incorporación de la *stoplist* fue un acierto que eliminó errores en la obtención de pares semánticos automáticamente, según se consigna en la tesis doctoral de Gerardo Sierra[24]:

24 Gerardo Sierra, *Design of,* p.178.

> Functional words can also be clustered. As an example, from table 6.5 we can observe that one of the highest lcc scores corresponds to “any-an”, and other pairs are “any-a” and “an-the”. However, in general, stop words interfere in the identification of clusters and can give more wrong than good results.

A continuación, reproducimos una parte de la tabla mencionada en el texto anterior[25]:

ff_i	ff_j	lcc_{ij}	ff_i	ff_j	lcc_{ij}
any	an	9	limits	field	6
determining	measuring	9	measuring	ascertaining	6
celestial	heavenly	8	measuring	taking	6
intensity	amount	8	method	system	6
one	that	8	sunlight	day	6
swinging	turning	8	testing	measuring	6
that	which	8	accurate	precise	5
determining	which	7	an	the	5
inclination	direction	7	analyse	recording	5
instrument	telescope	7	any	a	5
that	for	7	apparatus	instrument	5
amount	percentage	6	concentration	amount	5
determining	ascertaining	6		…	

Además, menciona otra razón significativa para filtrar los pares arrojados por el sistema mediante una *stoplist*: “It is easier to use such a stoplist than to engage in very sophisticated and hard to program methods that might not do any better.”[26]

1.2.2. Evaluación del algoritmo básico de agrupamiento semántico

En 2002, los creadores de Clustering junto con Gabriel Castillo hicieron una revisión exhaustiva del algoritmo y los resultados obtenidos de su aplicación a un corpus de metrología en inglés. De este análisis derivaron 19 líneas de trabajo, cuatro de ellas fueron abordadas en la tesis de Maestría en Ciencias de la Computación de este último[27].

Las líneas de trabajo propuestas se insertan en diferentes etapas del proceso: preprocesamiento (1.2.2.1), modificaciones al algoritmo de distancia de edición (1.2.2.2), cálculo de LCC (1.2.2.3) y lista de palabras irrelevantes (1.2.2.4). Si bien se documenta la necesidad de realizar cambios en la estructura del algoritmo, esto escapa del alcance de la tesis.

1.2.2.1. Preprocesamiento

25 No pudimos obtener ejemplos de “pares irrelevantes” para el español, ya que en la versión más reciente de Clustering la *stoplist* aparece incorporada al sistema y no como una opción que el usuario pueda elegir o rechazar.
26 Gerardo Sierra, *Design of a,* p.178.
27 Gabriel Castillo, *Algoritmo revisado para la extracción automática de agrupamientos semánticos*.

El algoritmo original no realiza ningún tipo de preprocesamiento. Se propone adaptar el lematizador a la lengua española; eliminar signos de puntuación que puedan alterar el funcionamiento del algoritmo, siempre y cuando éstos no separen ideas; identificar frases yuxtapuestas mediante conjunciones copulativas y disyuntivas para determinar cuántas definiciones existen en un enunciado; identificar las unidades léxicas formadas por colocaciones o términos compuestos para su correcto tratamiento; identificar las categorías gramaticales de las palabras que componen las definiciones para alinear sólo las que comparten categoría o cuyas categorías son afines. Se pretende que todo lo anterior pueda realizarse automáticamente, identificando y adecuando los algoritmos necesarios o, en caso de ser necesario, desarrollando nuevos.

1.2.2.2. Modificaciones al algoritmo de distancia de edición

Este eje reúne algunas líneas propias del trabajo informático, entre ellas la que se refiere a la operación de inversión de dos palabras, que sí nos atañe, pero la describiremos en el siguiente apartado porque fue una de las líneas trabajadas por Castillo.

1.2.2.3. Cálculo de LCC

Se proponen varias modificaciones para asociar un valor a los pares correspondientes, y así pueda ser un mejor indicador de si efectivamente éstos guardan relación semántica, de forma que, consecuentemente, los agrupamientos semánticos se incrementen en número y en calidad.

1.2.2.4. Lista de palabras irrelevantes

Se sugiere determinar una *stoplist* para el español y una adicional para cada dominio del banco terminológico, así como que la consideración de cada palabra de la lista como irrelevante esté en manos del usuario.

1.2.3. Algoritmo flexibilizado de agrupamiento semántico

Con el fin de mejorar el número de pares vinculados y de que éstos coincidieran con pares semánticos, Gabriel Castillo implementó seis algoritmos que son resultado del desarrollo de cuatro líneas de investigación de las 19 surgidas de la revisión de 2002 —las cuales se eligieron con base en la experiencia de los autores del algoritmo básico. De las cuatro líneas analizadas derivaron las propuestas específicas para el intercambio de palabras (1.2.3.1), pares semi nulos (1.2.3.2) y semi iguales (1.2.3.3). Asimismo, resultaron la modificación de costos y la incorporación de rutas múltiples de alineamiento (que no analizaremos).

Los algoritmos resultantes se integraron en el *Algoritmo flexibilizado de agrupamiento semántico*. Sus nuevas aplicaciones no se incorporan automáticamente a las búsquedas dentro de Clustering, sino que se presentan como

opcionales. En la siguiente página, mostramos la interfaz tal como se puede consultar en la dirección de internet
http://tabasco.torreingenieria.unam.mx/scripts/clusters.exe/Alineamiento :

“Base de datos a utilizar Fisica

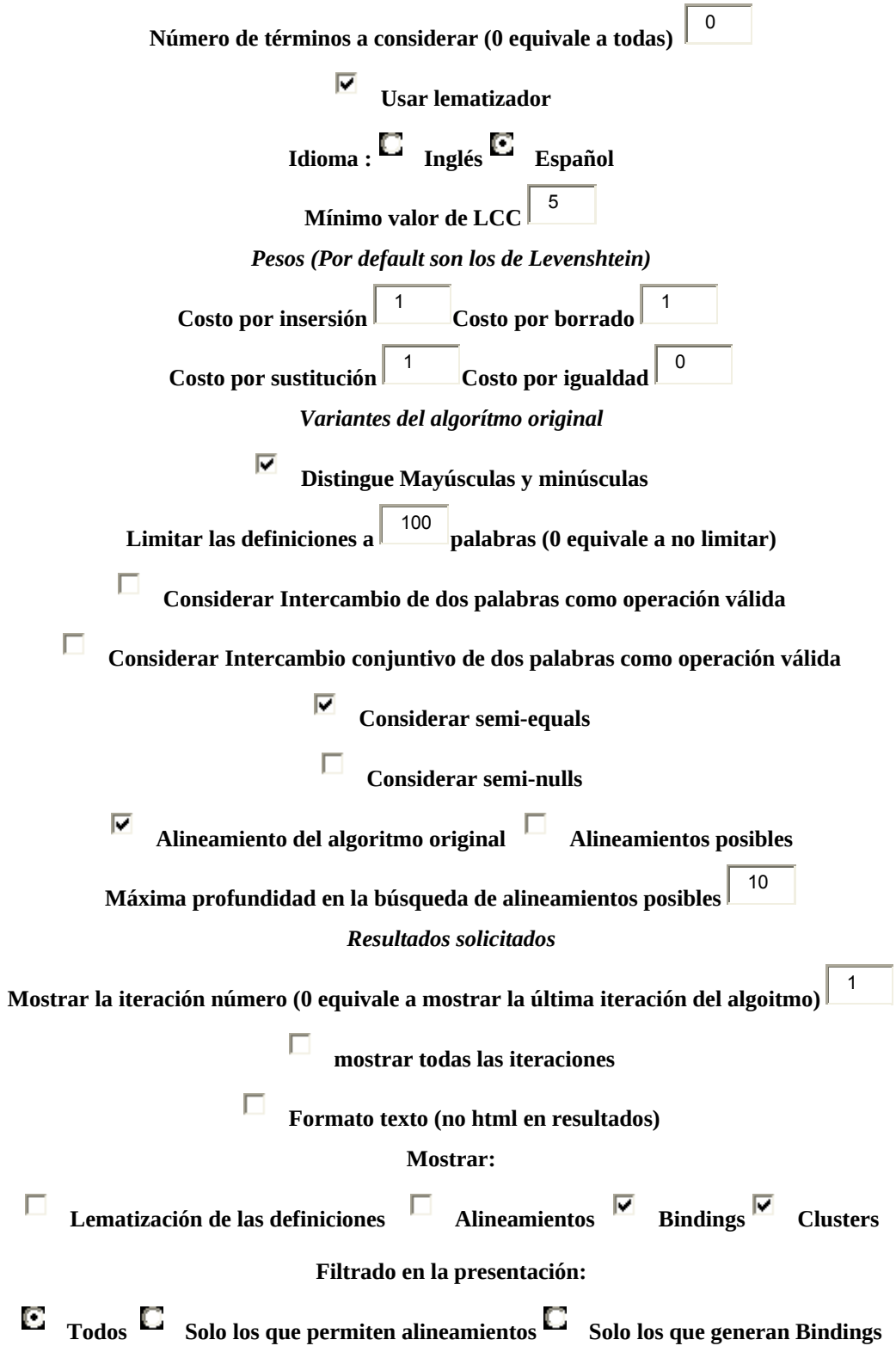

Figura 1: Interfaz para la realización de búsquedas en Clustering.

Como vemos, en primer lugar, el usuario puede elegir la base de datos, el idioma y el número de términos de los disponibles en el banco sobre los que va a realizar su

búsqueda; al igual que la presentación de exclusivamente los resultados que requiere y la forma en que éstos le interesan. La aplicación del lematizador y los valores de LCC también están a su consideración, así como las variantes del algoritmo original contenidas en el flexibilizado, las cuales se seleccionan de forma independiente conforme a las necesidades de cada búsqueda.

El diseñador del nuevo algoritmo aplicó diferentes combinaciones de búsqueda a un banco terminológico en el área de metrología con el fin de evaluar sus resultados. El inventario contiene 342 términos, cuyas definiciones se obtuvieron de dos diccionarios (el *Collins English Dictionary* y el *Oxford English Dictionary*). Tomaremos en cuenta los resultados de dicha evaluación al momento de describir las operaciones que incorpora el Algoritmo flexibilizado de agrupamiento semántico.

1.2.3.1. Operación de inversión de dos palabras

Con base en la revisión de la que hemos hablado se consideró que incorporar al sistema la inversión de dos palabras como una operación válida podía mejorar la identificación de pares semánticos. La finalidad es realizar intercambios de formas consecutivas dentro de una definición, sin que se modifique el sentido de la frase, para que sea posible promover a vinculados pares correspondientes que no alcanzaban el valor de LCC necesario en el algoritmo original.

Se preveía que esta operación resultara exitosa para nuestro idioma, pues "en español es común el intercambio del sustantivo y adjetivo cuando éstos se expresan en conjunto, así por ejemplo 'velocidad alta' y 'alta velocidad' hacen referencia al mismo concepto y con las mismas características."[28]

Veamos el ejemplo que propone Gabriel Castillo[29]:

Def. 1	**Estroboscopio**	Aparato	de	alta	velocidad	que	permite	congelar	…
Def. 2	**Estroboscopio**	Instrumento	de	velocidad	alta	que	permite		…
LCC	0	3	0	2	3	0	0	0	
Tipo*	I	C	I	C	C	I	I	N	

* Tipos de pares identificados: I = par igual; C = par correspondiente; N = par nulo.

> Como puede observarse a simple vista el segundo par (*aparato, instrumento*) tienen un valor de LCC = 3, y se observa además que *alta velocidad* y *velocidad alta* se alinearon en las parejas (*alta, velocidad*) y (*velocidad, alta*), que fueron identificadas como pares correspondientes. Si se dispusiera el algoritmo de intercambio de palabras el resultado debería identificar a las expresiones *alta velocidad* y *velocidad alta* como la misma expresión y elevar de esta manera el valor de LCC de la pareja (*aparato, instrumento*) a 7.

La operación necesaria para considerar estructuras paralelas al momento de un alineamiento a *velocidad alta* y *alta velocidad* se denominó *intercambio no conjuntivo.*

---

28 Gabriel Castillo, *op. cit*., pág. 28.
29 *Ibid*., p.29.

Además, en esta nueva versión del algoritmo se experimentó con una operación derivada de la anterior que no se había desarrollado previamente en otros sistemas, el *intercambio conjuntivo*, es decir, el de palabras no contiguas vinculadas mediante una conjunción. Sin embargo, Castillo concluye en su evaluación que la aplicación de los diferentes tipos de intercambio no cubre las expectativas previstas.

1.2.3.2. Pares semi nulos

Se denomina par *semi nulo* a un par nulo formado por una palabra funcional y un espacio vacío en la cadena, debido a que:

> Las palabras contenidas en la lista de palabras irrelevantes no aportan información relevante durante el proceso de identificación de pares vinculados, por lo tanto, los pares nulos que agrupen palabras de esta lista, y sólo de esta lista, pueden ser considerados pares iguales, bajo la óptica de que alinear palabras irrelevantes con la cadena vacía (ε) puede ser equivalente, en este contexto, a insertar la palabra involucrada en lugar de ε.[30]

El propósito de establecer este nuevo tipo de pares es aumentar el valor de LCC de los pares correspondientes que aparezcan —en los alineamientos— cerca de los semi nulos, ya que éstos, al ser considerados iguales, también lo son para el cálculo de LCC.

Tomemos el alineamiento que utilizamos para ejemplificar los tipos de pares y el coeficiente LCC. La primera tabla corresponde a la información arrojada por la versión del algoritmo original y la segunda es el alineamiento que la versión del algoritmo flexibilizado muestra para el mismo par de definiciones:

Def. 1	**Energía**	Capacidad	de	un	sistema	material	para	**producir**	trabajo	con	las	…
Def. 2	**Energía**	Capacidad	de	un	sistema	físico	para	**realizar**	trabajo			
Tipo	par igual	par igual	par igual	par igual	par igual	par correspondiente	par igual	**par correspondiente**	par igual	**par nulo**	**par nulo**	pares nulos
LCC	0	0	0	0	0	7	0	**3**	0	0	0	0
Par vinculado	no	no	no	no	no	sí	no	**no**	no	no	no	no

Resultados obtenidos con la aplicación del algoritmo original

Def. 1	**Energía**	Capacidad	de	un	sistema	material	para	**producir**	trabajo	con	las	…
Def. 2	**Energía**	Capacidad	de	un	sistema	físico	para	**realizar**	trabajo			
Tipo de par	par igual	par igual	par igual	par igual	par igual	par correspondiente	par igual	**par correspondiente**	par igual	**par semi nulo**	**par semi nulo**	pares nulos
LCC	0	0	0	0	0	7	0	**5**	0	0	0	0
Par vinculado	no	no	no	no	no	sí	no	**sí**	no	no	no	no

Resultados obtenidos con la aplicación del algoritmo flexibilizado

30 *Ibid*., p. 53.

Obsérvese el par {producir, realizar} de la novena columna. La versión original de Clustering le asigna un LCC = 3, por lo que no es promovido a vinculado siendo que se trata de un par semántico; en cambio, la versión flexibilizada del algoritmo sí lo identifica como par vinculado. Esto se debe a que los pares {con, ε } y {las, ε } que eran nulos, ahora son considerados semi nulos (pues la preposición *con* y el artículo *las* forman parte de la *stoplist)* y conforme a esto cada uno suma un punto al valor de LCC de {producir, realizar}, tal como si se tratara de un par igual.

Aunque para este ejemplo, de la consideración de pares semi nulos se obtiene un par semántico acertado, no siempre sucede así. Las pruebas realizadas por Castillo demostraron que la aplicación de este criterio aumenta significativamente el número de pares vinculados que el sistema obtiene, pero disminuye la relación entre pares vinculados y pares semánticos acertados.

1.2.3.3. Pares semi iguales

Los pares semi iguales se sustentan en la misma premisa de los pares semi nulos. Un par semi igual es aquél que está formado por dos palabras "irrelevantes" (incluidas en la *stoplist)* y que, para efectos de cómputo de LCC, se considera equivalente a un par igual.

Veamos cómo funciona en un alineamiento:

Def. 1*	**Cinemática**	Parte	de	la	mecánica	que	**describe**	un	movimiento				
Def. 2**	**Cinemática**	Parte	de	la	mecánica	que	**estudia**	el	movimiento	prescindiendo	de	…	
Tipo ***	I	I	I	I	I	I	**C**	C	I	N	N		
LCC	0	0	0	0	0	0	**7**	2	0	0	0	0	
Par vinculado	no	no	no	no	no	no	**no**	no	no	no	no	no	

Resultados obtenidos con la aplicación del algoritmo original

Def. 1*	**Cinemática**	Parte	de	la	mecánica	que	**describe**	un	movimiento				
Def. 2**	**Cinemática**	Parte	de	la	mecánica	que	**estudia**	el	movimiento	prescindiendo	de	…	
Tipo ***	I	I	I	I	I	I	**C**	SI	I	N	N	N	
LCC	0	0	0	0	0	0	**9**	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	**sí**	no	no	no	no	no	

Resultados obtenidos con la aplicación del algoritmo flexibilizado

*CIm1

** RAE 92

***Tipos de pares: I = par igual; C = par correspondiente; N = par nulo; SI = par semi igual.

Como puede verse, al par {describe, estudia} le corresponde un valor de 7 para el algoritmo original, pero no es promovido a par vinculado porque no cumple con la condición de frontera que establece que todo par correspondiente requiere de un par igual a la izquierda y otro a la derecha para ser considerado vinculado. Con la opción de semi iguales, el par {un, el} se trata como igual y el LCC del par en estudio ya cumple con la condición de frontera y se incrementa a 9. También en este caso, la

aplicación surgida de la revisión del algoritmo redunda en la identificación de un par semántico que anteriormente el sistema no registraba.

La conclusión del análisis del funcionamiento global de esta operación aplicada al corpus de metrología es la misma que para los semi nulos: incrementa la obtención de pares vinculados pero éstos disminuyen en calidad, es decir, no todos son semánticos.

Ahora veamos un ejemplo en donde se combinan las opciones de semi iguales y semi nulos:

Def. 1*	**Difracción**	Desviación	de	los	rayos	luminosos	cuando	pasan	por	los	bordes	de	un	cuerpo	opaco	
Def. 2**	**Difracción**	Desviación	del		rayo	luminoso	al	rozar	el		borde	de	un	cuerpo	opaco	
Tipo***	I	I	C	N	I	I	C	C	C	N	I	I	I	I	I	
LCC	0	0	3	0	0	0	3	1	1	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Resultados obtenidos con la aplicación del algoritmo original

Def. 1*	**Difracción**	Desviación	de	los	rayos	luminosos	cuando	pasan	por	los	bordes	de	un	cuerpo	opaco	
Def. 2**	**Difracción**	Desviación	del		rayo	luminoso	al	rozar	el		borde	de	un	cuerpo	opaco	
Tipo***	I	I	SI	SN	I	I	SI	C	C	SN	I	I	I	I	I	
LCC	0	0	3	0	0	0	0	15	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	sí	no	no	no	no	no	no	no	

Resultados obtenidos con la aplicación del algoritmo flexibilizado

* Vox

** RAE 92

*** Tipos de pares: I = par igual; C = par correspondiente; N = par nulo; SI = par semi igual; SN = par semi nulo.

Al margen de algunas precisiones que haremos en su momento (3.11.3), la aplicación de las dos opciones en forma conjunta es acertada. Pero repetimos, ésta no es la regla; los alineamientos que aquí mostramos fueron seleccionados para explicar cuáles beneficios tienen estas operaciones y, por ello, cubren la expectativa de incrementar los pares semánticos obtenidos automáticamente.

1.2.3.4. Conclusiones

Las conclusiones que sobre el algoritmo flexibilizado hace Gabriel Castillo señalan que la aplicación de las alternativas que él introduce relaja las restricciones del algoritmo original con resultados que incrementan considerablemente la cantidad de pares vinculados que el sistema identifica pero que hacen disminuir la correspondencia entre pares vinculados y semánticos. Específicamente[31]:

> El algoritmo de alineamiento semántico básico mejora notablemente al incluir la variante de par semi-igual o la variante de par semi-nulo; en particular, la primera

31 *Ibid*., pág. 70.

> ofrece mejores resultados que la segunda. La inclusión de alguna otra variante (alineamientos de costos múltiples, modificación de costos o intercambios) no aporta mejora alguna del algoritmo básico.

Además, se demostró que los programas computacionales requieren de una evaluación sistemática de los resultados que ofrecen, pues en este caso comprobamos que ésta "ayudó a evitar consideraciones cualitativas que eventualmente podrían sesgar los juicios respecto a las bondades de las variantes del algoritmo propuesto."[32]

1.3. ¿Qué falta por hacer?

En cuanto a nuestra área de estudio —la lengua española— hay mucho por hacer, sobre todo si se considera que Clustering fue diseñado para aplicarse a un corpus en inglés. En un artículo sobre el algoritmo flexibilizado, Castillo y Sierra señalan que en su trabajo "no se evaluaron los resultados que los algoritmos ofrecen cuando se aplican a un corpus en español. Como parte de los trabajos futuros deberán analizarse las modificaciones y adecuaciones necesarias para el idioma español." Y proponen incluir un etiquetador[33] de las partes de la oración que "posiblemente mejore los resultados obtenidos."[34]

La consideración anterior orientó las primeras ideas sobre esta tesis, las cuales se han ido enriqueciendo con las observaciones de John McNaught, quien diseñó una herramienta que corre el programa de Clustering con definiciones previamente etiquetadas siguiendo este planteamiento[35]. Además nos hemos valido de las líneas de investigación contenidas en el eje de preprocesamiento de la revisión de 2002 y de los resultados de nuestras propias pruebas al algoritmo.

En atención a que las modificaciones realizadas hace tres años surgieron de un trabajo basado en criterios matemáticos y no lingüísticos, como señala Castillo en las conclusiones de su trabajo[36], pretendemos aplicar un análisis sintáctico a las definiciones del banco terminológico para posteriormente incorporar la información gramatical al proceso de extracción automática de pares semánticos.

---

32 *Ibid*., pág. 79.

33 En ingeniería lingüística se entiende por etiquetador una herramienta aplicable a un corpus para adjuntar la información de tipo gramatical que corresponde a sus unidades. Existen diferentes tipos de etiquetado que se aplican según las necesidades de cada caso.

34 Gabriel Castillo y Gerardo Sierra, *op. cit.*, p.80.

35 Sierra y McNaught, *Serendipitous Wording of POS Tags to Extract Semantic Pairs of Words from Dictionary Definitions*.

36 "El algoritmo de alineamiento semántico es un método de comparación de dos definiciones que se basa en la comparación de la secuencia de las palabras que las constituyen. Las palabras son analizadas desde un punto de vista tal, que su semántica no es incluida en el análisis. Esta pérdida de información conduce a que eventualmente se agrupan palabras sin ninguna relación semántica.
[…] "Mientras se siga visualizando los textos como una secuencia de símbolos sin información adicional, los resultados no mejoraran en cuanto a los índices de evaluación."
G. Castillo, *op. cit*., pp. 76-77.

1.4. Recapitulación

En este capítulo hicimos una introducción a lo que son los programas de Clustering y expusimos qué es y en qué consiste el propuesto por Gerardo Sierra y John McNaught, en su versión original, y en la desarrollada a partir de ésta con Gabriel Castillo. Para explicar la utilidad de este programa, tuvimos que describir lo qué es el *Diccionario Electrónico de Búsqueda Onomasiológica* y, para que se comprendiera su mecanismo, introdujimos conceptos —como el de *pares correspondientes* o el de *pares semi iguales*— que serán manejados a lo largo de la tesis.

En el siguiente capítulo expondremos nuestra evaluación al funcionamiento de Clustering con la finalidad de que precisemos en qué espacios podemos incidir. La evaluación se sustentará en 32 pruebas realizadas al sistema, las que también describiremos.

2. ANÁLISIS DEL FUNCIONAMIENTO DE LA HERRAMIENTA

Conforme a la intención central de esta investigación —que es, como se ha dicho, aportar elementos que redunden en el mejoramiento de Clustering— y una vez que hemos explicado para qué sirve este programa y descrito algunos conceptos ligados a él, así como aportado información general sobre su funcionamiento, procederemos a reportar las conclusiones obtenidas del examen del estado actual de la herramienta, lo que nos llevará a determinar en qué espacios podemos incidir.

De acuerdo con las evaluaciones anteriores, Clustering recupera pares y agrupamientos semánticos de manera automática pero pasa por alto otros y recupera errores, también automáticamente; así que debe modificarse el sistema que hace posible este proceso para que su eficiencia aumente. Las modificaciones a un programa computacional sólo son posibles si se parte de un análisis que detecte las debilidades del sistema; una vez ubicadas, se elaboran propuestas que más tarde se materializarán y ya implementadas, deberá evaluarse si resultan funcionales o no. Éste es el procedimiento que nosotros seguiremos.

En este capítulo nos encargaremos de analizar en qué medida los algoritmos básico y flexibilizado (con sus diferentes variantes) obtienen pares semánticos. Para ello, describiremos nuestro objeto de estudio (2.1) y las pruebas realizadas al sistema (2.2); finalmente, haremos una evaluación del funcionamiento de los ejes del Algoritmo flexibilizado de agrupamiento semántico que son de interés lingüístico (2.3).

2.1. Delimitación del objeto de estudio

Conforme a la metodología de la lingüística de corpus, lo primero que debemos hacer es definir nuestro corpus de trabajo. Para los propósitos de esta investigación, tomaremos una muestra representativa de los alineamientos que se desprenden de todas las posibles combinaciones binarias de las definiciones que contiene el banco terminológico (que será el soporte del diccionario). Sobre este corpus se aplicará el procedimiento anteriormente descrito para implementar modificaciones a un programa computacional, pues trabajar con todas las definiciones del banco llevaría mucho tiempo —ya que la primera etapa de nuestro trabajo implica un análisis manual— y además, resulta innecesario si se considera que los aspectos de interés que buscamos en las definiciones son modelos lingüísticos que bien pueden aparecer en cualquier definición.

En primer lugar, nuestro corpus se compondrá sólo de definiciones en español pues buscamos evaluar el funcionamiento del programa para nuestro idioma. De las áreas que actualmente contiene el banco terminológico, hemos decidido trabajar con una recopilación de definiciones sobre Física.

El diccionario de Física está pensado como un material didáctico de apoyo para los alumnos de bachillerato que deben compartir un grado de conocimientos sobre esta ciencia (muy diferente al de un estudiante universitario de la especialidad, un investigador, un profesor o un niño de primaria). Por tanto, quienes participaron en la construcción del banco terminológico escogieron las definiciones que se ajustaran al propósito del DEBO, las cuales provienen de diversas fuentes: libros de texto, diccionarios de lengua general, diccionarios especializados, enciclopedias y definiciones personales.

Originalmente, se pensó trabajar con toda la base de Física, pero a medida que avanzamos en el análisis manual se vio que era excesivamente lento y para esta investigación decidimos reducir el corpus a la mitad de los alineamientos exactamente; no obstante, se intentó que éste estuviera balanceado: contiene definiciones sobre conceptos, instrumentos, experimentos, así como referentes a científicos.

Finalmente, el corpus quedó integrado por las definiciones que —en número variable— corresponden a 111 términos, las cuales suman 687. De las combinaciones binarias de estas definiciones se obtienen 2480 alineamientos. El esquema del corpus puede observarse en el Apéndice A.

No hay que perder de vista que las conclusiones que surjan de nuestro trabajo deberán ser aplicables para todos los dominios del banco terminológico y para cualquier par de definiciones que se inserte a Clustering, puesto que consideramos las definiciones del corpus como manifestaciones lingüísticas que incorporan fenómenos presentes en nuestra lengua —o, cuando menos, en el ámbito de la lexicografía.

2.2. Pruebas

Para poder evaluar Clustering, debemos, en primer lugar, conocer los métodos de evaluación para este tipo de sistemas. Existen diferentes criterios, dependiendo de los aspectos que se desee calificar.

Nosotros emplearemos el método seguido por Castillo que consiste en asignar a cada prueba valores de *recall* y *precision:* “Recall is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database. […] Precision is the ratio of the number of relevant documents retrieved over the total number of documents retrieved. Both recall and precision take

on values between 0 and 1."[1] Para nuestro trabajo, los documentos relevantes recuperados serán los pares vinculados acertados; los documentos relevantes en la base de datos, los pares semánticos ubicados en el corpus, y los documentos recuperados, los pares vinculados que el sistema recupera para una búsqueda.

Entonces, para establecer estos índices de evaluación debemos relacionar dos factores: los pares vinculados que el sistema obtiene para cada búsqueda y los pares semánticos existentes en el corpus. El primer factor se obtiene automáticamente y el segundo requiere de una mirada humana que, a partir del estudio de los alineamientos, determine cuáles son los pares semánticos dentro del corpus.

2.2.1. Obtención de pares semánticos manualmente

Con el propósito de establecer los índices de *precision* y *recall* y, con ellos, determinar cuáles son las variantes de Clustering que ofrecen los mejores resultados y cuáles las más imprecisas, revisamos manualmente los 2480 alineamientos del corpus y, con base en nuestro propio criterio, obtuvimos 460 pares semánticos (independientemente de su posición al interior de los alineamientos, es decir, de si aparecían como pares correspondientes o si se encontraban distantes). Cuando finalmente contamos con los pares, dividimos éstos en dos grupos: el primero quedó conformado por 196 que, a juzgar por nuestro conocimiento de la lengua, son incuestionablemente semánticos; mientras que el segundo grupo fue integrado por 264 pares que probablemente lo eran, pero no pudimos emitir una opinión final nosotros mismos, ya que involucraban conceptos científicos que desconocíamos.

Para determinar cuáles de estos 264 pares eran semánticos, pusimos a consideración de una profesora de Física[2] cada par en su contexto (alineamiento). La respuesta que ella nos dio es que 168 de los pares que le mostramos sí pueden ser considerados semánticos (aunque en la mayor parte de los casos no son estrictamente equivalentes para los criterios de la Física) y los restantes 96 definitivamente no son semánticos; nos explicó el motivo según cada caso.

En total, obtuvimos 364 pares semánticos[3]. Muchos de éstos (197) sólo pueden ser semánticos si se les trata como *pares semánticos compuestos,* es decir, cuando uno o ambos elementos de los que se componen están formados por estructuras sintácticas superiores a la palabra (por ejemplo *cargas eléctricas*[4] integra un par semántico con *campo eléctrico*[5], pero el par {cargas, campo} es erróneo). Por el

1 William B. Frakes, *Introduction to Information Storage and Retrieval Systems*, p. 10.
2 María Álvarez Moctezuma: Profesora de educación primaria por la Escuela Nacional de Maestros (1960) y Maestra de educación secundaria por la Escuela Normal Superior (1965). De 1965 a 2004 se desempeñó como profesora de las asignaturas de Biología, Ciencias Naturales, Física, Química e Introducción a la Física y Química en escuelas secundarias oficiales del Distrito Federal.
3 En el Apéndice B mostramos estos pares semánticos.
4 Larousse (definición de energía eléctrica).
5 Física de emergencia (definición de energía eléctrica).

contrario, los otros 167 son *pares semánticos simples*, es decir, sus elementos están compuestos por sólo una palabra.

2.2.2. Pruebas realizadas modificando los criterios de búsqueda del Algoritmo flexibilizado de agrupamiento semántico

De las modificaciones que Castillo realizó al algoritmo original, retomamos para nuestras pruebas las agrupadas en torno a cinco aspectos: distinguir mayúsculas y minúsculas, considerar intercambio de dos palabras, considerar intercambio conjuntivo de dos palabras, considerar pares semi iguales y considerar pares semi nulos.

Con la intención de tener un registro de la eficiencia actual de Clustering y de determinar con cuál de las opciones del algoritmo flexibilizado se recuperan los pares semánticos más coincidentes con los obtenidos manualmente, se corrió el programa una vez por cada opción resultante de modificar esos cinco parámetros; se consideró, igualmente, la opción que no incorpora ninguna de las modificaciones, es decir, el algoritmo original (prueba 01). Obtuvimos 32 variantes del algoritmo.

En seguida registramos los pares diferentes ofrecidos por el sistema para cada prueba y, en cada caso, comparamos los resultados con la lista de pares semánticos identificados manualmente; así, pudimos obtener los valores de *precision* y *recall* para cada prueba. Al índice de *recall* le asignamos dos valores, uno que considera como los posibles pares semánticos de nuestro corpus sólo los pares simples —pues ni el algoritmo original ni el flexibilizado recuperan pares compuestos— y otro que considera todos los pares semánticos identificados manualmente, ya sean simples o compuestos. El primer valor de *recall* servirá para evaluar el funcionamiento del programa, tal como está diseñado, mientras que del segundo se obtendrá una medida que relaciona los resultados generados con lo que se espera de la herramienta, pero que aún no está programada para realizar.

En la siguientes páginas se puede observar la tabla que registra en qué consistió cada prueba y qué resultados se obtuvieron, así como una gráfica que ilustra la intersección de los índices de *precision* y *recall* para dichas pruebas, la cual considera dos valores de *recall*, uno para los pares semánticos simples y otro para el total de los pares semánticos identificados manualmente.

Número de la prueba	M, m.[6]	I.[7]	I. C.[8]	S. I.[9]	S. N.[10]	Pares generados	Pares semánticos acertados	*Precision*	Pares semánticos simples identificados manualmente	*Recall*	Pares semánticos identificados manualmente	*Recall*
01	no	no	no	no	no	53	30	0.5660	167	0.1796	364	0.0824
02	sí	sí	sí	sí	sí	376	46	0.1223	167	0.2754	364	0.1263
03	sí	no	no	no	sí	141	38	0.2695	167	0.2275	364	0.1043
04	sí	no	no	no	no	53	30	0.5660	167	0.1796	364	0.0824
05	sí	no	no	sí	sí	374	46	0.123	167	0.2754	364	0.1263
06	sí	no	no	sí	no	193	42	0.2176	167	0.2514	364	0.1153
07	sí	no	sí	no	sí	141	38	0.2695	167	0.2275	364	0.1043
08	sí	no	sí	no	no	53	30	0.5660	167	0.1796	364	0.0824
09	sí	no	sí	sí	sí	375	46	0.1227	167	0.2754	364	0.1263
10	sí	no	sí	sí	no	194	42	0.2165	167	0.2514	364	0.1153
11	sí	sí	no	no	sí	141	38	0.2695	167	0.2275	364	0.1043
12	sí	sí	no	no	no	53	30	0.5660	167	0.1796	364	0.0824
13	sí	sí	no	sí	sí	375	46	0.1227	167	0.2754	364	0.1263
14	sí	sí	no	sí	no	194	42	0.2165	167	0.2514	364	0.1153
15	sí	sí	sí	no	sí	141	38	0.2695	167	0.2275	364	0.1043
16	sí	sí	sí	no	no	53	30	0.5660	167	0.1796	364	0.0824
17	sí	sí	no	sí	no	195	42	0.2154	167	0.2514	364	0.1153
18	no	no	no	no	sí	141	38	0.2695	167	0.2275	364	0.1043
19	no	no	no	sí	sí	374	46	0.123	167	0.2754	364	0.1263
20	no	no	no	sí	no	193	42	0.2176	167	0.2514	364	0.1153
21	no	no	sí	no	sí	141	38	0.2695	167	0.2275	364	0.1043
22	no	no	sí	no	no	53	30	0.5660	167	0.1796	364	0.0824

6 Distingue mayúsculas y minúsculas.
7 Considera intercambio de dos palabras.
8 Considera intercambio conjuntivo de dos palabras.
9 Considera pares semi iguales.
10 Considera pares semi nulos.

Número de la prueba	M, m.[11]	I.[12]	I. C.[13]	S. I.[14]	S. N.[15]	Pares generados	Pares semánticos acertados	*Precision*	Pares semánticos simples identificados manualmente	*Recall*	Pares semánticos identificados manualmente	*Recall*
23	no	no	sí	sí	sí	375	46	0.1227	167	0.2754	364	0.1263
24	no	no	sí	sí	no	194	42	0.2165	167	0.2514	364	0.1153
25	no	sí	no	no	sí	141	38	0.2695	167	0.2275	364	0.1043
26	no	sí	no	no	no	53	30	0.5660	167	0.1796	364	0.0824
27	no	sí	no	sí	sí	375	46	0.1227	167	0.2754	364	0.1263
28	no	sí	no	sí	no	194	42	0.2165	167	0.2514	364	0.1153
29	no	sí	sí	no	sí	141	38	0.2695	167	0.2275	364	0.1043
30	no	sí	sí	no	no	53	30	0.5660	167	0.1796	364	0.0824
31	no	sí	sí	sí	sí	376	46	0.1223	167	0.2754	364	0.1263
32	no	sí	sí	sí	no	195	42	0.2154	167	0.2514	364	0.1153

Tabla 1: Pruebas realizadas al sistema, resultados cuantitativos y valores de *precision* y *recall.*

11 Distingue mayúsculas y minúsculas.
12 Considera intercambio de dos palabras.
13 Considera intercambio conjuntivo de dos palabras.
14 Considera pares semi iguales.
15 Considera pares semi nulos.

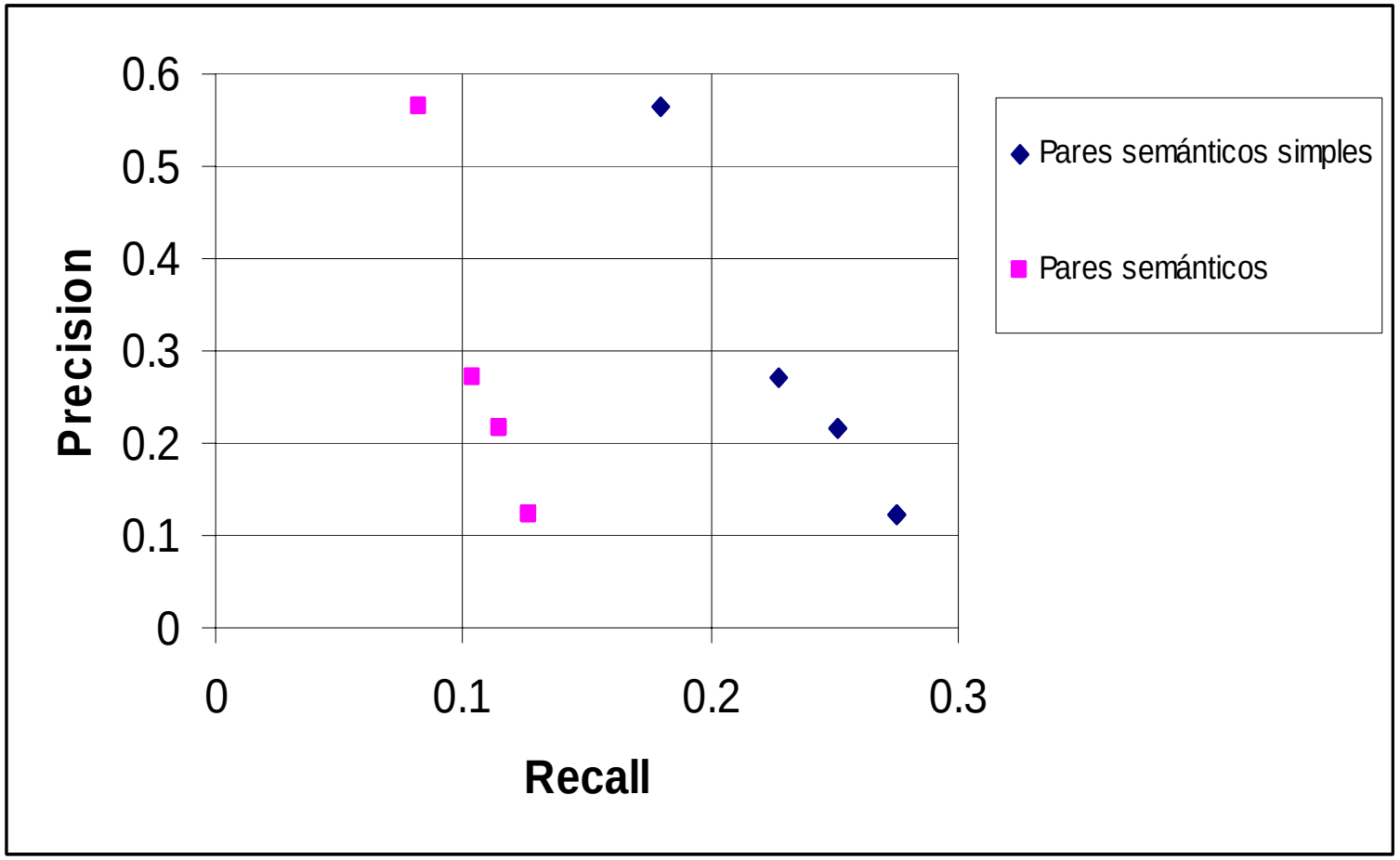

Figura 2: Índices de *recall* y *precision* correspondientes a los pares semánticos simples y al total los pares semánticos identificados manualmente para las pruebas realizadas al agoritmo flexibilizado.

2.2.3. Resultados cuantitativos de las pruebas

Observando la tabla que registra las 32 pruebas realizadas al sistema, pudimos detectar coincidencias numéricas (en cuanto a la generación de pares semánticos y, consecuentemente, en *precision* y *recall)* que nos llevaron a hacer una primera clasificación de los datos en torno a cuatro grandes grupos; tal orden es claramente perceptible en los puntos de la gráfica. Cada grupo contiene los resultados de ocho combinaciones de parámetros.

Una vez que contamos con estos grupos, comparamos los pares generados por las pruebas que arrojaron resultados similares, con el fin de determinar si dos corridas diferentes de las que se obtiene igual número de pares recuperan los mismos o la coincidencia numérica es dada por casualidad.

Tomando en cuenta lo anterior y los parámetros de búsqueda para cada prueba, sacamos inferencias generales sobre el funcionamiento de las aplicaciones del algoritmo flexibilizado que complementan la información que se desprende de la comparación de los valores de *precision* y *recall* para cada caso. Contar con esta información estructurada facilitará los análisis posteriores.

A continuación mostraremos y describiremos los grupos, considerando los siguientes parámetros de búsqueda:

Clave	**Parámetro**
A	Distingue mayúsculas y minúsculas
B	Considera intercambio de dos palabras
C	Considera intercambio conjuntivo de dos palabras
D	Considera pares semi iguales
E	Considera pares semi nulos

Primer grupo

Prueba	**A**	**B**	**C**	**D**	**E**	**Pares generados**	**Pares semánticos acertados**	**Precision**	**Recall (pares simples)**	**Recall**
01	no	no	no	**no**	**no**	53	30	0.5660	0.1796	0.0824
04	sí	no	no	**no**	**no**	53	30	0.5660	0.1796	0.0824
08	sí	no	sí	**no**	**no**	53	30	0.5660	0.1796	0.0824
12	sí	sí	no	**no**	**no**	53	30	0.5660	0.1796	0.0824
16	sí	sí	sí	**no**	**no**	53	30	0.5660	0.1796	0.0824
22	no	no	sí	**no**	**no**	53	30	0.5660	0.1796	0.0824
26	no	sí	no	**no**	**no**	53	30	0.5660	0.1796	0.0824
30	no	sí	sí	**no**	**no**	53	30	0.5660	0.1796	0.0824

Tabla 2: Primer grupo.

En este grupo incluimos las pruebas que dieron como resultado los mismos 53 pares generados (ya fueran acertados o erróneos). Observamos que la característica común a sus parámetros de búsqueda —que no está presente en las demás pruebas— consiste en que no se contempla ninguna de las opciones de pares semi

iguales y semi nulos. Lo anterior nos lleva a deducir que estos aspectos son los que concentran las diferencias relevantes del algoritmo flexibilizado con respecto al algoritmo original.

Segundo grupo

Prueba	A	B	C	D	E	Pares generados	Pares semánticos acertados	Precision	Recall (pares simples)	Recall
03	sí	**no**	no	no	**sí**	141	38	0.2695	0.2275	0.1043
07	sí	**no**	sí	no	**sí**	141	38	0.2695	0.2275	0.1043
18	no	**no**	no	no	**sí**	141	38	0.2695	0.2275	0.1043
21	no	**no**	sí	no	**sí**	141	38	0.2695	0.2275	0.1043
11	sí	**sí**	no	no	**sí**	141	38	0.2695	0.2275	0.1043
15	sí	**sí**	sí	no	**sí**	141	38	0.2695	0.2275	0.1043
25	no	**sí**	no	no	**sí**	141	38	0.2695	0.2275	0.1043
29	no	**sí**	sí	no	**sí**	141	38	0.2695	0.2275	0.1043

Tabla 3: Segundo grupo.

El segundo grupo es similar al primero en lo que se refiere a la combinación de parámetros de búsqueda; la diferencia consiste en que aquí se están considerando semi nulos (para todas las pruebas).

Decidimos agrupar estas ocho combinaciones de parámetros debido a que hay una coincidencia en el número de pares generados por el sistema (141). Sin embargo, estos pares no son exactamente los mismos, así que dividimos el grupo en dos subgrupos (pruebas 03, 07, 18, 21, para el primero, y 11, 15, 25 y 29 para el segundo), cuyas respectivas pruebas coinciden en los pares generados por el sistema. Ambos conjuntos comparten 140 pares y se diferencian en uno.

Vemos que la diferencia entre los subgrupos es mínima, pero no sólo eso; el factor que marca esta diferencia es el intercambio no conjuntivo. Las pruebas en donde se contempló esta opción generan, para el término *energía*, el par vinculado {producir, transformarse}, que no se obtuvo con las que no la consideran.

Por otra parte, todas las pruebas del subgrupo 1 generan el par {fricción, resistencia} para el término *fuerza de fricción,* mientras que para el mismo término no se encuentra el par en el subgrupo 2. La causa es igualmente la aplicación del intercambio no conjuntivo.

Destacamos que para el primer grupo no se observaron modificaciones relacionadas con la aplicación de intercambios y que, hasta el momento, la operación de intercambio no conjuntivo sólo incide en la obtención de pares en combinación con la opción de semi nulos.

Tercer grupo

Prueba	A	B	C	D	E	Pares generados	Pares semánticos acertados	Precision	Recall (pares simples)	Recall
06	sí	no	no	**sí**	**no**	193	42	0.2176	0.2514	0.1153
20	no	no	no	**sí**	**no**	193	42	0.2176	0.2514	0.1153
10	sí	no	sí	**sí**	**no**	194	42	0.2165	0.2514	0.1153
24	no	no	sí	**sí**	**no**	194	42	0.2165	0.2514	0.1153
14	sí	sí	no	**sí**	**no**	194	42	0.2165	0.2514	0.1153
28	no	sí	no	**sí**	**no**	194	42	0.2165	0.2514	0.1153
17	sí	sí	sí	**sí**	**no**	195	42	0.2154	0.2514	0.1153
32	no	sí	sí	**sí**	**no**	195	42	0.2154	0.2514	0.1153

Tabla 4: Tercer grupo.

El tercer grupo está formado por todas las pruebas que tienen como parámetros comunes la aplicación de semi iguales y la no consideración de semi nulos. La mínima diferencia que existe entre el número de pares generados por el sistema se debe a las opciones de intercambio. A las pruebas 06 y 20 no se les aplicó ninguna de estas dos opciones y el número de pares obtenidos fue de 193; mientras que con las pruebas que consideraron un tipo de intercambio se generó un par más, que es diferente para el intercambio conjuntivo y el no conjuntivo. Por último, con las pruebas 17 y 32, que corresponden a la aplicación simultánea de los dos tipos de intercambio, obtuvimos dos pares más (la suma de los pares conseguidos por cada una de las opciones de intercambio) que con las pruebas que no realizan intercambios. De cualquier forma, es de notar que los pares adicionales obtenidos no son pares semánticos acertados.

Cuarto grupo

Prueba	A	B	C	D	E	Pares generados	Pares semánticos acertados	Precision	Recall (pares simples)	Recall
5	sí	no	no	**sí**	**sí**	374	46	0.123	0.2754	0.1263
19	no	no	no	**sí**	**sí**	374	46	0.123	0.2754	0.1263
9	sí	no	sí	**sí**	**sí**	375	46	0.1227	0.2754	0.1263
23	no	no	sí	**sí**	**sí**	375	46	0.1227	0.2754	0.1263
13	sí	sí	no	**sí**	**sí**	375	46	0.1227	0.2754	0.1263
27	no	sí	no	**sí**	**sí**	375	46	0.1227	0.2754	0.1263
2	sí	sí	sí	**sí**	**sí**	376	46	0.1223	0.2754	0.1263
31	no	sí	sí	**sí**	**sí**	376	46	0.1223	0.2754	0.1263

Tabla 5: Cuarto grupo.

La explicación de esta tabla es parecida a la de la anterior. Los pares generados por el sistema se incrementan considerablemente con relación al tercer grupo (esto se debe a que aquí se están considerando, además de los pares semi iguales, los semi nulos); pero el fenómeno es el mismo: el menor número de pares vinculados

corresponde a las pruebas que no consideran intercambios; cada opción de intercambio suma sólo un par a los anteriores, y la combinación del intercambio conjuntivo y no conjuntivo aporta dos pares.

2.2.4. Interpretación global de los resultados

Los grupos mencionados se formaron a partir de la similitud que encontramos en los valores de *precision* y *recall* para determinadas pruebas. Teniendo agrupadas, de esta manera, las corridas que hicimos del algoritmo, pudimos advertir que la aplicación o no de los pares semi iguales y semi nulos es el factor que cohesiona los grupos internamente y suscita las diferencias entre uno y los demás. Son también estos ejes del algoritmo flexibilizado los que marcan un cambio importante con relación al algoritmo básico, como decíamos al inicio del apartado anterior.

Como sabemos, el primer grupo se corresponde prácticamente con el algoritmo básico; si bien es cierto que en los criterios de las diferentes pruebas se combinan tres ejes del algoritmo flexibilizado, éstas no implican ninguna modificación en cuanto a la generación de pares semánticos con respecto a los obtenidos con el algoritmo original. El segundo grupo —cuyo aspecto más relevante para la recuperación de pares semánticos es la consideración de pares semi nulos para todas las pruebas que lo componen— se considerará simplemente como la aplicación de semi nulos; lo mismo ocurrirá con el grupo 3 en relación con pares semi iguales. El cuarto grupo se considerará como la aplicación que combina los parámetros de semi iguales y semi nulos. Simplificar los resultados no sólo se hace para disminuir el trabajo, sino porque estos criterios coinciden con los que aplicó Gabriel Castillo para la evaluación del algoritmo flexibilizado, lo que nos permitirá comparar sus resultados con los nuestros.

La interpretación que aquí hacemos coincide con la de Castillo para el corpus en inglés: con las modificaciones del algoritmo flexibilizado aumenta el *recall* (es decir, se incrementa la relación entre los pares semánticos recuperados y los pares semánticos posibles dentro del universo que se estudia), pero no la *precision* (esto es, disminuye el número de pares semánticos entre los pares vinculados). En la siguiente tabla se registran los resultados de ambos análisis:

Inglés Español

Grupo	Recall (pares simples)	Precision	Recall [16] (pares simples)	Recall	Precision
1 = Algoritmo básico	0.1026	0.9375	0.1796	0.0824	0.5660
2 = Semi nulos	0.1578	0.6338	0.2275	0.1043	0.2695
3 = Semi iguales	0.2175	0.6392	0.2514	0.1153	0.2176 – 0.2154
4= Semi nulos + semi iguales	0.3157	0.4865	0.2754	0.1263	0.123 – 0.1223

Tabla 6: Pruebas a los algoritmos básico y flexibilizado para el inglés y para el español.

Como puede verse, los resultados de las dos revisiones siguen la misma tendencia: la opción que genera pares mayoritariamente semánticos es la versión original del algoritmo, pero éstos, aunque aceptables, son insuficientes en número (un 16.2% de los identificados manualmente en nuestro corpus). Por otra parte, con la prueba que involucra semi nulos y semi iguales se obtiene el mayor número de pares semánticos; al mismo tiempo, esa prueba es la que genera el mayor número de pares erróneos. Castillo señala en su evaluación que las alternativas incorporadas con su tesis "permiten relajar las restricciones del algoritmo original, incrementando el índice de *recall* pero disminuyendo en consecuencia el índice de *precision*." [17]

La diferencia notoria entre los resultados de las dos revisiones la constituyen los números de *precision* y *recall.* De acuerdo con este método de evaluación, el sistema funcionará mejor en la medida en que los índices se acerquen al número uno. Por tanto, Clustering cumple mejor su función (la de generar agrupamientos semánticos) para el inglés que para el español.

2.3. Evaluación de cada eje del algoritmo flexibilizado

Ahora, señalaremos nuestras observaciones a cada una de las modificaciones al algoritmo original a partir del análisis de la misma tabla.

2.3.1. Pares semi iguales

Las aplicaciones de semi iguales y semi nulos, como hemos visto, sí marcan una diferencia importante en cuanto a la cantidad de pares obtenidos. La opción que combina ambas proporciona casi diez veces más pares que la que no considera ninguno, pero la relación entre pares vinculados y semánticos se aleja en la medida en que se consiguen los nuevos pares.

16 Para este trabajo consideraremos en todo momento al hablar de pares semánticos la unión de los simples y los complejos. Hacemos aquí una excepción porque Gabriel Castillo sólo trabajó con pares semánticos simples, por lo que consideramos mas coherente para las comparaciones tomar datos que sean lo más próximos posible.

17 Gabriel Castillo, *op. cit.*, p. 77.

En el apartado 1.2.3.3 vimos un ejemplo donde la consideración de pares semi iguales concuerda con lo que se esperaba de esta operación, pero la disminución de los valores de *precision* muestra que la aplicación, ya implementada, no funciona siempre así. Decidimos revisar los alineamientos de nuestro corpus considerando sólo esta opción del algoritmo flexibilizado para intentar explicar por qué entre los nuevos pares que se generan predominan los que no son semánticos.

Ahora veremos un caso en donde se obtienen pares vinculados erróneamente debido a la aplicación de semi iguales:

Def. 1*	**Dinámica**	Parte	de	la	mecánica	que	**relaciona**	el	movimiento	con	la	fuerzas	asociadas	con	…
Def. 2**	**Dinámica**	Es	el	estudio	De	las	**causas**	del	movimiento	de	las	fuerzas	que	provocan	…
Tipo	I	C	SI	C	C	**SI**	**C**	**SI**	I	**SI**	**SI**	I	C	C	
LCC	0	3	0	2	2	0	**7**	0	0	0	0	0	6	1	
Par vinculado	no	no	no	no	No	no	**sí**	no	no	no	no	no	no	no	

* CIm1
** Física para ciencias e Ingeniería

El sistema asigna la categoría de vinculado al par {relaciona, causas}, que no es semántico, mientras que el algoritmo original otorga un LCC de 1 para el mismo par. El drástico incremento en los valores de LCC en el nuevo algoritmo se debe precisamente a la aplicación de semi iguales. Esto no sucedería si Clustering considerara categorías gramaticales e incorporara una restricción que impidiera que se vincularan palabras de diferente clase.

Podemos concluir que el empleo de esta opción del algoritmo flexibilizado resulta arriesgado, pues, en algunos casos, es un buen auxiliar para la obtención de pares semánticos, pero, en mayor medida, genera pares equivocados. Sin embargo, no debe eliminarse del programa la opción de semi iguales, pues el criterio que la sustenta —dos palabras funcionales son equivalentes en cuanto no aportan una carga semántica fuerte a las definiciones— es muy atinado para la finalidad de Clustering. En este sentido, proponemos acotar esta premisa a que dos palabras funcionales diferentes son equivalentes cuando cumplen una misma función en un enunciado (en este caso, definición), entendiendo que las palabras funcionales no son irrelevantes. Identificar las categorías funcionales en las definiciones permitiría impedir que se formaran pares semi iguales con palabras cuya categoría funcional es diferente.

2.3.2. Pares semi nulos

Como ya se ha mencionado, la aplicación de pares semi nulos incrementa notablemente el número de pares vinculados, pero éstos disminuyen en calidad. Gabriel Castillo identificó en su análisis al corpus en inglés algo parecido, por lo que sugiere no considerar esta opción al correr el algoritmo flexibilizado. Sin embargo, insistimos, al analizar nosotros los alineamientos con esta opción consideramos que

el problema no está en la aplicación sino en situaciones que pueden ser mejoradas. Veamos un alineamiento:

Def. 1*	**Período**	Intervalo	de	tiempo	**en**		que	se	desintegran	la	mitad	…
Def. 2**	**Período**	Espacio	de	tiempo	**especialmente**	el	que	comprende		la	duración	…
Tipo	I	C	I	I	**C**	**SN**	I	C	N	I	C	
LCC	0	4	0	0	**5**	0	0	3	0	0	2	
Par vinculado	no	no	no	no	**sí**	no	no	no	no	no	no	

* María Moliner
** Clave

El error que promueve a vinculado el par {en, especialmente} consiste en otorgarle un punto a un par que es casi nulo.

Si bien, como hemos visto, se justifica plenamente que palabras funcionales sean alineadas con conjuntos vacíos, los pares que se generan a partir de esta vinculación carecen del peso semántico de los pares iguales formados sustantivos o verbos, por lo que no debe aplicarse el mismo criterio para efectos de LCC. Además, su incidencia es tan alta que prácticamente en todas las definiciones donde se encuentran pares semi nulos, éstos ocasionan que se generen pares vinculados. Tomando en cuenta lo anterior, proponemos que, en lo que se refiere al cómputo de LCC, los pares semi nulos no corten las cadenas de palabras, pero tampoco sumen puntos a un par correspondiente.

Asimismo, el identificar la categoría gramatical de las palabras (tal como lo proponen Sierra y McNaught) evitaría que se consideraran par semántico una preposición y un adverbio.

2.3.3. Distinguir mayúsculas y minúsculas

La aplicación de la opción *Distinguir mayúsculas y minúsculas* no redunda en la obtención de pares semánticos. Para todas las combinaciones de parámetros trabajadas en estas pruebas —que son todas las posibles— no existe un sólo par que se genere debido a la aplicación de este criterio. Podría sencillamente eliminarse de Clustering (al menos en la forma en como funciona actualmente).

2.3.4. Intercambio

Las opciones de intercambio (conjuntivo y no conjuntivo) no son cuantitativamente significativas, pero esto no basta para establecer un juicio sobre su incorporación al programa; debemos conocer si los intercambios responden al propósito con el que se establecieron, es decir, si inciden en la identificación de nuevos pares semánticos y, de no ser así, cuáles son los motivos. Aquí encontramos una diferencia con el inglés, en donde sí hay un incremento importante de pares vinculados debido a la incorporación de intercambios: “Las alternativas de intercambio de palabras […] no

ofrecieron los resultados esperados, pues aumentan **fuertemente el número de pares-vinculados** sin incrementar de manera importante el número de pares semánticos.” [18]

2.3.4.1. Conjuntivo

De entre 2480 alineamientos, el intercambio conjuntivo sólo está incidiendo en la obtención de un nuevo par vinculado y esto, cuando se aplica simultáneamente con la opción de pares semi iguales o la que combina éstos y los pares semi nulos. Veamos este caso…

18 Gabriel Castillo, *Ibid*., p. 79. Las **negritas** son nuestras.

… cuando se aplica sólo la opción de semi iguales:

Def. 1*	**Física**	Ciencia	que	estudia	la	materia	la	energía	sus	propiedades	y	los	fenómenos	y	leyes	que	las	rigen	o	caracterizan	
Def. 2**	**Física**	Ciencia	que	estudia	las					propiedades	leyes	y	fenómenos			de	la	materia	y	la	energía
Tipo de par	I	I	I	I	I	N	N	N	N	I	C	C	I	N	N	SI	SI	C	SI	C	N
LCC	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	4	0	2	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

… cuando se aplica el intercambio conjuntivo:

Def. 1*	**Física**	Ciencia	que	estudia	la	materia	la	energía	sus	propiedades	y	los	fenómenos	y	leyes	que	las	**rigen**	o	caracterizan	
Def. 2**	**Física**	Ciencia	que	estudia	las					propiedades			leyes	y	fenómenos	de	la	**materia**	y	la	energía
Tipo de par	I	I	I	I	I	N	N	N	N	I	N	N	**IC*** **	**IC**	**IC**	SI	SI	**C**	SI	C	N
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**	0	2	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	**sí**	no	no	no

* Clave
** Larousse
******* Intercambio conjuntivo

El par vinculado {rigen, materia} que el sistema identifica para este alineamiento aumenta su LCC de 4 a 7 debido a la aplicación que considera un intercambio conjuntivo. Sin embargo, el par no es semántico. En él se vinculan incorrectamente dos palabras con diferente categoría gramatical; entonces, aquí ubicamos otra debilidad del sistema que podría ser eliminada con la identificación de categorías gramaticales.

2.3.4.2. No conjuntivo

Las aplicaciones del intercambio de palabras consecutivas tienen mayor incidencia en nuestro corpus que la opción anterior, pero aún ésta es muy modesta y casi nunca redunda en la generación de nuevos pares semánticos.

2.4. Conclusiones

La primera conclusión, después de haber realizado y comparado las 32 pruebas, es que nuestro análisis de los alineamientos y las propuestas que resulten de éste se harán sobre el algoritmo original debido a que su *precision* es la mejor; pero cuando sea necesario acudiremos a los alineamientos de las versiones del flexibilizado porque, a nuestro juicio, las iniciativas incluidas en los ejes de este algoritmo son realmente buenas y no deben desecharse del sistema. Si el algoritmo flexibilizado no proporciona una mejoría significativa en la obtención de pares semánticos —como se suponía— es porque hace falta incluir criterios lingüísticos al tratamiento de las definiciones, tal como señala Castillo: “mientras no se incorpore información semántica a las definiciones un incremento del índice *recall* tendrá por consecuencia una disminución del índice *precision*.”[19]

Por otra parte, al comparar los resultados de nuestra revisión con los de la anterior, queda confirmado que Clustering funciona mejor para corpus en inglés, lo cual resulta lógico puesto que fue diseñado para esta lengua. Esto reafirma la necesidad de hacer modificaciones al programa que se basen en la gramática de nuestro idioma. Nosotros trabajaremos en el preprocesamiento sintáctico de las definiciones (el procesamiento deberá realizarse computacionalmente) para así allanar el camino para que el sistema pueda analizar las definiciones y realizar los alineamientos considerando sus estructuras sintácticas, pero nuestro trabajo no llegará a este punto pues sabemos que “el desarrollo de los analizadores sintácticos […] todavía es un problema, especialmente para los idiomas que no tienen un orden de palabras fijo, como en el español (en inglés, el orden de las palabras es fijo. Por eso las teorías basadas en inglés no son fáciles de adoptar al español).”[20]

Recordemos que Clustering es un programa para la obtención de agrupamientos semánticos (conjuntos de palabras que pueden ser mutuamente sustituibles en determinados contextos). La primera etapa del proceso consiste

19 Gabriel Castillo, *Ibid*., p. 77.
20 Grigory Sidorov, *Problemas actuales de lingüística computacional*.

en la comparación de pares de definiciones; las similitudes de sentido se infieren del parecido en las relaciones sintácticas que establecen dos palabras. Es por esto que la información gramatical prioritaria para el funcionamiento del programa es la sintáctica (al menos en esta etapa, pues para otras resulta más necesaria la morfológica —que ya aparece incorporada con éxito al algoritmo original, ver 1.2.1.1).

2.5. Recapitulación

En este capítulo se presentaron y analizaron los resultados obtenidos con la aplicación individual y combinada de cinco alternativas del Algoritmo flexibilizado de agrupamiento semántico: la que distingue mayúsculas y minúsculas, las que consideran el intercambio de palabras consecutivas, el intercambio conjuntivo de dos palabras, pares semi iguales y pares semi nulos, así como la opción que no considera ninguna de éstas (el algoritmo original).

Las posibles combinaciones dan un total de 32 pruebas que dividimos en cuatro grupos, de acuerdo con sus correspondientes índices de *precision* y *recall*; la comparación de los parámetros mostró que la consideración de los pares semi iguales y semi nulos determina las mayores diferencias. Posteriormente, se compararon nuestros resultados con los del análisis del algoritmo flexibilizado hecho por Gabriel Castillo.

Para poder calificar, mediante los índices de *precision* y *recall*, la calidad de los pares generados en cada prueba, tuvimos que determinar manualmente los pares semánticos existentes en la colección de definiciones a la que se aplicaron, y por ello comenzamos por definir nuestro corpus y el procedimiento seguido para la identificación de los pares semánticos.

Finalmente se hizo una evaluación del funcionamiento de cada uno de los ejes del algoritmo flexibilizado que consideramos relevantes para nuestro trabajo. Se concluyó que la determinación de las categorías gramaticales beneficiaría a Clustering, al impedir que se vincularan palabras con distinta categoría que, por ese motivo, no pueden formar un par semántico.

En el siguiente capítulo describiremos el análisis sintáctico de las definiciones y con base en éste definiremos las propuestas de modificación a los alineamientos y cuáles probaremos en el capítulo 4.

3. ANÁLISIS SINTÁCTICO Y PROPUESTAS

En el capítulo anterior se presentaron las pruebas realizadas a Clustering considerando el algoritmo básico y las diferentes combinaciones de los criterios del flexibilizado. También se hizo una evaluación de cada eje del algoritmo flexibilizado.

En el presente capítulo expondremos los resultados de aplicar un análisis sintáctico a las definiciones del corpus. Se presentarán cada uno de los aspectos que, según nuestro criterio, deben incorporarse (además de la categoría gramatical de las palabras) como información sintáctica a Clustering, ya que podrían incidir favorablemente en la obtención de pares semánticos. Esta apreciación se apoyará en ejemplos de los casos registrados en nuestra exploración.

En este trabajo, con *análisis sintáctico* nos referimos al estudio de las relaciones y funciones de los constituyentes de la oración. Por sintagma entenderemos todo enunciado, según la acepción de De Saussure, definida en el *Diccionario de lingüística moderna* como "cualquier combinación de dos o más elementos, sea cual sea la complejidad y la estructura de éstos, que contraen relaciones gramaticales."[1] La nomenclatura a emplear para definir las clases de sintagmas que encontramos (*palabra, frase, oración, oración compleja, oración transpuesta,…*), así como para llamar a las categorías funcionales (*suplemento, implemento, atributo, aditamento*), es de Alarcos Llorach.

Cabe señalar que, para nuestra propuesta, tanto la sintaxis como el resto de los niveles de análisis lingüístico son fundamentales para una explicación más fina de los fenómenos que hemos localizado y su consecuente implementación al sistema.

3.1. Perífrasis gramaticales

Definir qué significa *palabra* ha sido un problema filosófico, al menos, desde Aristóteles. La palabra es para la morfosintaxis el signo lingüístico entre el morfema y la frase. Tomamos por válida una de las acepciones que ofrece el *Diccionario del español actual*[2]: "Conjunto fijo de fonemas que constituye una unidad indivisible dotada de significado y función estables". Así, este conjunto fijo de fonemas habitualmente se presenta en la escritura entre dos blancos, pero no se trata de una característica definitoria de la palabra.

1 Enrique Alcaraz y Ma. Antonia Martínez, *Diccionario de lingüística moderna*, p. 526.
2 Manuel Seco, Olimpia Andrés y Gabino Ramos.

Por perífrasis gramatical se entiende una construcción gramatical fija en la lengua "que procede de la falta de una voz única para expresar un concepto único"[3]. No debemos perder de vista que uno de los mecanismos en la formación de neologismos en español es la sintagmación, que consiste en la lexicalización de una estructura sintáctica (por este motivo se habla de palabras compuestas, que no son sino aquéllas que integran dos o más morfemas léxicos con un significado unitario).

En tanto "conjunto fijo de fonemas que constituye una unidad indivisible dotada de significado y función estables", entendemos que las perífrasis gramaticales son funcionalmente palabras más que frases y, por esto, deben ser consideradas como tales por el programa de agrupamiento semántico, Clustering.

Existen tres tipos de perífrasis gramaticales que vamos a considerar: locuciones, términos compuestos y perífrasis verbales.

3.1.1. Locuciones

En este trabajo llamaremos *locuciones* a estructuras que semántica y sintácticamente forman una unidad pero aparecen separadas en la escritura. Julio Casares agrega que su "sentido unitario, familiar a la comunidad lingüística, no se justifica, sin más, como una suma del significado normal de los componentes."[4]

En el alineamiento 1, podemos ver que *por encima de* y *sobre* son equivalentes, ya que introducen la misma relación semántica; por tanto, no debe alinearse *por*, *encima* o *de* con *sobre,* sino la estructura *por encima de* en su conjunto, la cual se comporta en la definición como una preposición análoga a *sobre*.

Un indicio revelador de la presencia de locuciones son las equivocaciones ortográficas cometidas por los capturistas que asentaron en el banco terminológico las definiciones que copiaban de sus fuentes originales, pues estos "errores" confirman que ciertas expresiones sufren un proceso de lexicalización: El hablante las escribe sin espacios porque las identifica como una unidad. El alineamiento 2 muestra un ejemplo de este fenómeno: El sistema arroja un par correspondiente ({*a*, *acabo*}, séptima columna) y otro nulo ({*cabo*, ε}, octava columna), donde debería aparecer un solo par igual. Proponemos que las construcciones que han mostrado inestabilidad en la escritura —*a cabo, a partir*,…— sean tratadas como unidades léxicas al momento de realizar los alineamientos (independientemente de la forma externa que presenten).

Gabriel Castillo señala que "el algoritmo no realiza identificación de unidades léxicas, por lo que expresiones tales como: 'de un' [sic.], 'por lo tanto', 'por ejemplo', etc. son identificadas manualmente y se unen a estas expresiones mediante un guión bajo ('de_un', 'por_lo_tanto', etc.) de modo que el algoritmo considere estas expresiones como una sola palabra." Y en este sentido, propone "identificar los algoritmos

3 Enrique Alcaraz y Ma. Antonia Martínez, *op. cit*., p. 432.
4 Julio Casares, *Introducción a la lexicografía moderna*, § 73.

existentes y aplicarlos aquí para una correcta identificación de los términos compuestos y colocaciones, para su correcto tratamiento en el algoritmo de alineamiento semántico considerando estos términos como una sola una unidad. En caso de no existir algoritmos adecuados, desarrollarlos."[5]

De lo anterior, se infiere que computacionalmente es posible automatizar el proceso de identificación de locuciones, así que suscribimos la propuesta. De no ser viable esto a corto plazo, provisionalmente, se podría tomar una solución —intermedia entre lo que Castillo propone y el mecanismo que teóricamente utiliza el programa— consistente en agregar una lista con las locuciones más frecuentes, las cuales deberá tratar el sistema como una sola palabra. Con ello se aceleraría el procedimiento, al no tener que marcar cada una de estas expresiones manualmente, lo que, por otra parte, no se aplica, pues de todos los alineamientos estudiados no hubo uno solo en donde el algoritmo considerara expresiones léxicas como palabras.

5 Gabriel Castillo, *op. cit.*, p. 26.

Número de alineamiento: 1
Término: Energía potencial gravitacional
Fuente de la definición 1: Física aplicada
Fuente de la definición 2: Física general Bueche-Schaum

Def. 1	**E**	**p**	**g**	La	ep	gravitacional	de	un	cuerpo																	…
Def. 2	**E**	**p**	**g**	La	epg		de	un	objeto	se	define	con	respecto	a	un	nivel	arbitrario	cero	el	cual	a	menudo	es	la	superficie	…
Tipo de par	I	I	I	I	C	N	I	I	C	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
LCC	0	0	0	0	5	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Def. 1	…	de	masa	m	y	que		está	a	una	altura	h	**por**	**encima**	**de**	un	determinado	nivel	de		referencia	...
Def. 2	…	de	la	tierra	Si	un	objeto	está	a	una	altura	h	**sobre**	el				nivel	cero	o	referencia	...
Tipo de par		I	C	C	C	C	N	I	I	I	I	I	C	C	N	N	N	I	C	N	I	
LCC		0	2	1	1	1	0	0	0	0	0	0	6	1	0	0	0	0	2	0	0	
Par Vinculado		no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 2
Término: Observador
Fuente de la definición 1: Diccionario de Física Vox
Fuente de la definición 2: Diccionario de Física. Gran Vox

Def. 1	**O**	Toda	persona	que	lleva	**a**	**cabo**	una	observación	adquiere	un	papel	muy	importante	…
Def. 2	**O**	Toda	persona	que	lleva	**acabo**		una	observación	adquiere	un	papel	muy	importante	…
Tipo de par	I	I	I	I	I	**C**	**N**	I	I	I	I	I	I	I	
LCC	0	0	0	0	0	**6**	**0**	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	**no**	**no**	no	no	no	no	no	no	no	

3.1.2. Términos compuestos

En este apartado, al igual que en el conjunto de la tesis, debemos considerar que nuestro corpus está formado por definiciones y que éstas corresponden a un lenguaje especializado, el cual se define así por conformar una terminología[6].

Las definiciones con las que trabajamos provienen tanto de fuentes acotadas al ámbito de la Física como de diccionarios de lengua general, en los cuales se busca emplear un léxico neutro que pueda ser asequible a una amplia gama de usuarios. Dentro de los materiales especializados, muchos son libros de texto dirigidos a estudiantes no familiarizados con la terminología de la materia y, en consecuencia, el grado de especialización es menor. Por lo anterior y porque los lenguajes de especialidad "tienden (**solo tienden**[7]) a disponer de una denominación para cada concepto, lo que les lleva a presentar un cierto grado de sinonimia no controlada"[8], es viable la identificación de pares semánticos formados por un término y una palabra —o locución— de uso común, o bien, por dos términos especializados diferentes.

Las locuciones que se encuentran entre los términos especializados se llaman términos compuestos. Éstos constituyen unidades léxicas con un significado unitario, y las relaciones sintácticas que pueden establecer son las mismas que establece cualquier palabra. Tal como señala Cabré[9]:

> Si se acepta la propuesta de que la terminología es una parte integrante del sistema léxico de una «gramática», los términos se revelan como unidades sígnicas que pueden ser analizadas lingüísticamente desde tres puntos de vista diferentes: formal (la denominación), semántico (el concepto) y funcional (la categoría y la distribución).

En nuestro corpus son claramente identificables como compuestos una alta proporción de los 111 términos que trabajamos[10], los cuales no ofrecen problemas a los alineamientos (pues siempre obtendremos pares iguales en los espacios correspondientes a los términos). Resulta menos sencillo identificar los términos compuestos que se encuentran insertos en las definiciones; los que marcamos como tales fueron proporcionados por la especialista que nos ayudó a establecer manualmente los pares semánticos.

---

6 "Las comunicaciones especializadas, además de contener un determinado número de palabras funcionales y léxicas propias del lenguaje común, contienen términos peculiares propios de la temática de la que tratan. El conjunto de los términos de un campo, es decir su terminología, representa la estructura conceptual de esa materia, y cada uno de los términos denomina un concepto de la red estructurada de la materia en cuestión." Ma. Teresa Cabré, *La terminología, teoría, metodología y aplicaciones*, pp. 166-167.
7 Las **negritas** son nuestras.
8 Ma. Teresa Cabré, *op.cit*., p. 131.
9 *Ibid*., pp.170-171.
10 *Vid*. Apéndice A.

En el alineamiento 3, encontramos que el término *movimiento rectilíneo uniforme* que corresponde al ámbito de la Física aparece en la primera definición; mientras que en la segunda, la opción elegida para denotar la misma realidad a la que se refiere el término compuesto es una frase construida libremente por su autor ("movimiento uniforme en línea recta"), la que, como frase, implica relaciones gramaticales más complejas que las que contiene un término.

Cada una de las definiciones del alineamiento 4 contiene un término compuesto: *cargas eléctricas* y *campo eléctrico,* respectivamente*.* Aunque no son sinónimos, pueden ser intercambiables en el contexto en que se presentan, siempre que se consideren los términos en su totalidad.

En el alineamiento 5 podemos identificar el par semántico {rayos luminosos, luz}. La acepción de *luz* que se considera para la determinación de pares semánticos es la que se restringe al dominio de la Física; lo mismo sucede con *rayos* (cuyo significado coloquial se identifica con otro fenómeno físico), pero en este caso, el término requiere precisarse. Aunque no es muy común en Física manejar la expresión *rayo(s) luminoso(s),* si ésta se emplea no debe verse como la unión circunstancial de dos palabras aisladas, sino como una unidad terminológica utilizada para referirse a un concepto cuyos constituyentes inmediatos no pueden aportar el significado aisladamente. Sólo si se trata como un elemento, es posible el alineamiento con *luz.*

La solución que se propone para disminuir los errores en la obtención de pares semánticos es localizar los términos compuestos insertos en las definiciones que alimentan Clustering*.* Esta identificación podrá realizarse automáticamente en un corto plazo.

> [La] necesidad por extraer términos de manera automática surge a finales de los años ochenta; la aparición del programa TERMINO en 1990, marca un avance significativo en el área de extracción terminológica vía procesos automáticos y muestra resultados alentadores (Cabré, María Teresa, Rosa Estopà y Jordi Vivaldi 2001). Muchos grupos se encuentran en el desarrollo de recursos para aprovechar los estudios, programas y la infraestructura existentes, y dar a conocer nuevas técnicas que redunden en el beneficio de los terminólogos[11].

Actualmente y como parte de un proyecto de extracción de patrones definitorios, el GIL desarrolla un programa para la obtención de términos compuestos o multipalabra en textos de carácter científico o técnico para el español[12], el cual combina criterios sintácticos y estadísticos en la aplicación del algoritmo C-value-E.

Una condición previa para la aplicación del algoritmo es que los textos a los cuales se va a aplicar hayan recibido previamente un etiquetado con POST. La detección de candidatos a términos se basa en un primer momento en asociar palabras contiguas que presentan patrones sintácticos previamente registrados como propios

11 Antonio Reyes Pérez, *Extracción automática de terminología en el léxico de Física*, p. 49.
12 Vid. Alberto Barrón *et al*., *C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español.*

de los términos. Posteriormente, el algoritmo determina la posibilidad de que un sintagma candidato constituya un término con base en los siguientes parámetros:

1) Frecuencia total de ocurrencia del sintagma candidato en el corpus.
2) Frecuencia total de ocurrencia del sintagma candidato dentro de candidatos de mayor longitud.
3) Número de ocurrencias de dichos candidatos de mayor longitud.
4) Longitud del sintagma candidato.

Las primeras pruebas a este método se aplicaron a un corpus de Ingeniería; la evaluación cualitativa y cuantitativa de los términos recuperados con relación a los posibles en el universo registra índices de *precision* y *recall* de 0.47 y 0.69 respectivamente, lo que, sin lugar a dudas, es muy positivo. Se espera obtener mejores resultados cuando el programa se haya concluido.

Con la finalidad de avanzar en la identificación de términos compuestos, sugerimos anexar para cada base de definiciones con que trabaje el programa una lista con los términos compuestos que sean entrada de diccionarios especializados en la materia respectiva. Así por ejemplo, la base de Física incorporaría, entre muchas otras, las siguientes entradas del diccionario *Física de emergencia*: movimiento acelerado, movimiento circular uniforme, movimiento de los fluidos, movimiento de un cuerpo rígido, movimiento parabólico, movimiento rectilíneo uniforme, movimiento sin fuerzas, movimiento uniforme, movimiento uniformemente acelerado.

La aplicación combinada de las listas de palabras y la extracción terminológica a partir del algoritmo C-value-E aseguraría una importante recuperación de términos compuestos. Con esto se eliminaría el problema que conlleva la desproporción entre el número de palabras que componen los diferentes términos, lo que desencadena un desajuste entre las estructuras que le siguen en los alineamientos y, por tanto, se mejoraría el desempeño general del programa.

Número de alineamiento: 3
Término: Primera ley de Newton
Fuente de la definición 1: Física universitaria Sears
Fuente de la definición 2: El mundo de la Física 1

Def. 1	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o		**movimiento**	**rectilíneo**	**uniforme**				a	menos	que	sea	impelido	a	cambiar	dicho	estado	…
Def. 2	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o	de	**movimiento**		**uniforme**	**en**	**línea**	**recta**	a	menos	que	sea	obligado	a	cambiar	este	estado	…
Tipo de par	I	I	I	I	I	I	I	I	I	I	I	I	I	N	I	N	I	N	N	N	I	I	I	I	C	I	I	C	I	
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	6	0	
Par Vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	sí	no	no	sí	no	

Número de alineamiento: 4
Término: Energía eléctrica
Fuente de la definición 1: Larousse
Fuente de la definición 2: Física de emergencia

Def. 1	**E**	**e**		La	producida	por	las	**cargas**	**eléctricas**	
Def. 2	**E**	**e**	Es	la	energía	debida	a	un	**campo**	**eléctrico**
Tipo de par	I	I	N	I	C	C	C	C	C	N
LCC	0	0	0	0	2	1	1	1	1	0
Par vinculado	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 5
Término: Difracción
Fuente de la definición 1: Vox
Fuente de la definición 2: Física Weber

Def. 1	**D**	Desviación	de	los	**rayos**	**luminosos**	cuando	pasan	por	los	bordes	de	un	cuerpo	opaco
Def. 2	**D**	Es	la	dispersión	de	la	**luz**	en	una	región	situada	tras	un	obstáculo	
Tipo de par	I	C	C	C	C	C	C	C	C	C	C	C	I	C	N
LCC	0	2	1	1	1	1	1	1	1	1	1	2	0	2	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

3.1.3. Perífrasis verbales

Por perífrasis verbal se entiende la unión de morfemas que en conjunto refieren a un solo proceso o estado; dicha perífrasis "consiste en el empleo de un verbo auxiliar conjugado, seguido del infinitivo, del gerundio o del participio. Entre el auxiliar y el infinitivo se interpone *que* o una preposición."[13]

La anterior es una explicación básica de las perífrasis, pues no contempla la combinación de dos o más de las fórmulas descritas. Tampoco es referente suficiente para reconocerlas en una oración donde entre las formas verbales aparece un adverbio (o frase adverbial), como en la siguiente definición de *eje de rotación*:

- Recta alrededor de la cual un cuerpo libre, no solicitado por fuerza alguna, **continuará** indefinidamente **girando** si en un instante dado girase alrededor de ella.[14]

Nuestro corpus es rico es construcciones perifrásticas, como son abundantes en español. Observamos que la presencia de éstas dentro de los alineamientos desencadena el que aparezcan como pares correspondientes palabras con diferente categoría gramatical. Resulta lógico inferir que si consideramos verbo a toda estructura gramatical que funcionalmente actúa como tal, se obtendrán alineamientos más precisos, lo que redundará en mejores pares semánticos. Por tanto, proponemos que se marquen como unidades léxicas no sólo aquellos verbos en los que el proceso que describen se manifiesta con más de una palabra (los tiempos compuestos de la conjugación y los verbos pronominales en donde el pronombre *se* aparece en posición proclítica); sino también las perífrasis verbales.

En el caso de los tiempos compuestos y de los verbos pronominales, éstos podrían identificarse a partir de cierta información gramatical que se insertara al motor de Clustering. Pero para las perífrasis, el reconocimiento automático se torna más complicado, pues la superposición de sus fórmulas básicas multiplica las expresiones que encontramos en la lengua; además "existen combinaciones de una forma verbal y un derivado que no han de interpretarse como perífrasis: no actúan como segmentos unitarios nucleares, sino como reunión de núcleo y adyacente."[15] He aquí un ejemplo que encontramos en una definición de *cinemática*:

- Se ocupa del estudio del movimiento de partículas y cuerpos rígidos, sin atender a las causas que **producen dichos movimientos**.[16]

En este caso *dichos* funciona como adjetivo de *movimientos* y esta estructura (*dichos movimientos*) es un adyacente de *producir.*

13 Samuel Gili Gaya, *Curso superior de sintaxis española*, p. 105.
14 Gran enciclopedia Larousse.
15 Emilio Alarcos, *Gramática de la lengua española*, p. 259.
16 Cinemática y dinámica básica.

Como puede observarse, no existe ninguna diferencia formal que indique cuándo la unión de un verbo conjugado y una forma no personal implica una perífrasis y cuándo no la forma.

Por otra parte, debemos considerar que alinear frases verbales que actúan como núcleo de oraciones pasivas con verbos en voz activa podría conducir a equiparar otros elementos oracionales que semánticamente son diferentes (principalmente el sujeto paciente con el sujeto agente, ya que gramaticalmente son idénticos, a diferencia del complemento y el complemento agente). Incluir una restricción en este sentido haría más complejo el procedimiento y resulta innecesario, pues estos casos son poco frecuentes.

No obstante los riesgos previstos, consideramos que debe incorporarse al programa un mecanismo que ayude a identificar las perífrasis.

Ahora veamos los siguientes alineamientos:

En el 6, la definición número 1 se compone por dos oraciones, como núcleo del segundo enunciado tenemos una perífrasis, *se puede transformar*, que en su conjunto es funcionalmente equivalente al núcleo oracional de la segunda definición, también una perífrasis, *puede ser transformada.* Ambas frases verbales cumplen la misma función, la de predicado de oraciones en voz pasiva; las diferencias de forma se deben a que la primera prefiere la pasiva refleja a la marcada con el verbo *ser* como auxiliar de la segunda.

La primera definición del alineamiento 7 sugiere, mediante una combinación de dos perífrasis básicas (verbo conjugado + infinitivo + participio), un proceso equiparable al expresado por el verbo de la segunda. Aunque entre ambas expresiones existe una diferencia semántica (ya que en la segunda definición la forma verbal *cuelga* señala que el acontecimiento existe, mientras que la primera manifiesta la posibilidad de que se dé la misma situación, *puede oscilar suspendido*), ésta no altera el que su funcionamiento sea análogo. Alinear estas formas verbales en un mismo campo impediría que el programa arroje pares correspondientes del tipo {oscilar, de} y {suspendido, un}, lo que consecuentemente redundaría en un mejor alineamiento de las definiciones en su conjunto.

Por último, vemos el número 8, en donde considerar una sola forma (y no cuatro) las perífrasis, no redundaría en beneficios para el resto del alineamiento, pues las definiciones que lo componen muestran la misma estructura sintáctica, incluyendo la estructura interna de la perífrasis. Aquí, el sistema está asignando un LCC = 7 al par correspondiente {impelido, obligado}, que efectivamente es semántico. Si se incorporara al algoritmo la modificación que proponemos en este apartado, cada uno de los participios que forman el par aparecería integrado a una perífrasis verbal, cuyo LCC disminuiría a 4 y no cumpliría la condición de frontera, lo que hace pensar que debe considerarse que, en un segundo momento, se realicen alineamientos dentro de las perífrasis a fin de recoger nuevos pares semánticos.

Por otra parte, la primera definición muestra, además, un caso de reunión de formas verbales que no constituyen perífrasis. Aquí, la identificación automática de unidades perifrásticas podría conducir a que erróneamente el sistema incorporara *dicho* a la perífrasis que le antecede, lo que confirma que las perífrasis verbales no pueden esquematizarse en un lenguaje matemático por completo. Para este alineamiento, su identificación automática afectaría negativamente la obtención de pares semánticos; pero, como ya hemos dicho, en términos generales, consideramos que el desempeño del algoritmo mejorará con tal modificación.

Número de alineamiento: 6
Término: Ley de la conservación de la energía
Fuente de la definición 1: CIm1
Fuente de la definición 2: Físíca Resnick

Def. 1	**L**	**d**	**l**	**c**	**d**	**l**	**e**	En	cualquier	proceso	la	energía	total	ni	aumenta	ni	disminuye	La	energía	**se**	**puede**		**transformar**	de	una	forma	a	otra	…
Def. 2	**L**	**d**	**l**	**c**	**d**	**l**	**e**	En	un	sistema	aislado							la	energía		**puede**	**ser**	**transformada**	de	una	clase	a	otra	…
Tipo de par	I	I	I	I	I	I	I	I	C	C	C	N	N	N	N	N	N	I	I	N	I	N	I	I	I	C	I	I	
LCC	0	0	0	0	0	0	0	0	9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	sí	no	no	

Número de alineamiento: 7
Término: Péndulo
Fuente de la definición 1: DRAE 92
Fuente de la definición 2: Física de emergencia

Def. 1	**P**	Cuerpo	grave	que	**puede**	**oscilar**	**suspendido**	de				un	punto					por	un	hilo	o	varilla	
Def. 2	**P**	Cuerpo	rígido	que	**cuelga**	de	un	hilo	delgado	sujeto	a	un	soporte	fijo	que	oscila	libremente	por	la	acción	de	su	peso
Tipo de par	I	I	C	I	C	C	C	C	N	N	N	I	C	N	N	N	N	I	C	C	C	C	N
LCC	0	0	4	0	2	1	1	1	0	0	0	0	2	0	0	0	0	0	2	1	1	1	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 8
Término: Primera ley de Newton
Fuente de la definición 1: Física universitaria Sears
Fuente de la definición 2: El mundo de la Física 1

Def. 1	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o		movimiento	rectilíneo	uniforme			a	menos	que	**sea**	**impelido**	**a**	**cambiar**	dicho	estado	…
Def. 2	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o	de	movimiento			uniforme	en	línea	recta	a	menos	que	**sea**	**obligado**	**a**	**cambiar**	este	estado	...
Tipo de par	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	N	I	N	I	N	N	N	I	I	I	I	C	I	I	C	I	I
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	6	0	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	sí	no	no	sí	no	no

3.2. Nexos

Un nexo es "cualquier elemento lingüístico que sirve para unir a otros dos, sobre todo en el plano sintagmático."[17] Las clases de palabras que funcionalmente actúan como nexos son las conjunciones, las preposiciones y también los verbos copulativos. En este apartado veremos las dificultades que presentan para los alineamientos y nuestras propuestas de solución.

3.2.1. Conjunciones

La conjunción es una "parte invariable de la oración, que denota la relación que existe entre dos oraciones o entre miembros y vocablos de una de ellas, juntándolos o enlazándolos siempre gramaticalmente, aunque a veces signifique contrariedad o separación de sentido entre unos y otros."[18]

En el caso de Clustering, uno de los factores que provoca que los alineamientos no vinculen palabras funcionalmente equivalentes es que el programa no distingue si las conjunciones afectan a palabras, frases u oraciones, o bien a unidades de diferente tipo. Aunque el corpus en estudio contiene varias clases de conjunciones, las cuales indican una amplia gama de relaciones semánticas, nuestra propuesta se limitará a las relaciones de coordinación (sólo copulativa y disyuntiva), así como a las de yuxtaposición observables en los signos de puntuación.

El término yuxtaposición designa la reunión de "dos o más unidades (no solo oracionales) que desempeñan en conjunto la misma función que cumpliría cada una de ellas aisladamente. […] Los rasgos fónicos que distinguen a las unidades yuxtapuestas se reducen al carácter descendente de la entonación en cada una de ellas, que en la ortografía queda señalada por las comas."[19] Por su parte, la coordinación es un procedimiento que sirve para enlazar segmentos yuxtapuestos del discurso mediante una conjunción.

Lo que nos interesa resaltar en las definiciones que alimentan Clustering es qué tipo de sintagmas está afectando el enlace. Vayamos a los ejemplos:

El alineamiento número 9 compara las definiciones:

- La Física puede definirse como la ciencia que investiga los conceptos fundamentales de materia**,** energía y espacio...

- Ciencia que estudia las propiedades de la materia **y** de la energía...

17 Fernando Lázaro Carreter, *Diccionario de términos filológicos* p. 293.
18 DRAE 92.
19 Emilio Alarcos, *op. cit*., p. 316.

La primera definición se expresa mediante una oración compleja, cuya oración transpuesta adverbial, introducida por *como*, presenta una estructura sintáctica semejante a la de la oración que constituye la segunda definición: Dos oraciones de relativo que adjetivan a *ciencia,* y dentro de ellas, una frase preposicional en función de complemento adnominal para el primer caso y dos frases del mismo tipo para la segunda definición.

Las frases preposicionales modifican respectivamente *conceptos* y *propiedades* (sustantivos que, en este contexto, pueden intercambiarse sin que por ello se altere el sentido de las oraciones originales). El complemento adnominal de la primera definición reúne tres elementos que en conjunto funcionan como término del enlace preposicional, pero cada uno de ellos, precedido de *de*, podría desempeñar, en forma independiente, la misma función, que es lo que sucede en la segunda definición; aquí, las dos frases preposicionales tienen, por separado, los términos enlazados mediante la coma en la frase de la primera definición.

Entonces, tenemos dos definiciones que, ya sea con una coma —a nivel de palabra—, ya sea con la conjunción *y* —a nivel de frase—, expresan semánticamente la misma adición. Vale la pena subrayar que la coma cumple una función gramatical que no es perceptible en el alineamiento, pues el programa de alineamiento semántico elimina los signos de puntuación.

En el alineamiento 10 encontramos un fenómeno similar al del alineamiento anterior. En la primera definición, la *o* está enlazando los términos *reposo* y *movimiento rectilíneo uniforme,* que (mediante una frase preposicional) califican a *estado* y, a su vez, la frase que forman con este referente (*su estado de reposo o movimiento rectilíneo uniforme*) funciona como suplemento de la oración en la que se inserta. Mientras que en la segunda definición, la locución *o bien* denota una relación de exclusión entre los dos implementos de la oración con que se define la primera ley de Newton.

En el alineamiento 11, la segunda definición presenta, a partir de *la energía,* la misma estructura que la primera definición, **Artículo + sustantivo + adverbio de negación + verbo**[20] **+ conjunción + verbo + adverbio + verbo**, y a juzgar por las conjunciones, una diferencia semántica absoluta (las cuales imprimirían un sentido disyuntivo a la primera y uno copulativo a la segunda). Para la primera definición, en sentido estricto entenderíamos, dado el carácter disyuntivo, que existen dos opciones de algo que no puede sucederle a la energía: una es *crearse* y otra es *destruirse;* simultáneamente, la presencia de cualquiera de estas cualidades estaría negando la posibilidad de existencia a la otra. Sin embargo, la conjunción *o* no pretende expresar eso, sino sumar un elemento al anteriormente negado, con lo que estaría invadiendo el terreno de *ni*. No es conveniente incorporar ninguna información en este sentido a Clustering, pues el programa carece del criterio que nos ayudó a percibir la función anómala de la conjunción disyuntiva.

20 *Vid*. 3.1.3.

A partir de los ejemplos analizados, que son representativos del tema en estudio, consideramos que debe implementarse un mecanismo que coadyuve a mejorar el alineamiento de definiciones que presentan relaciones de coordinación a diferentes niveles sintagmáticos. Proponemos que, en un alineamiento como el 9, se consideren tantas frases prepositivas en función de complemento adnominal como sustantivos contenga cada término de preposición. De esta manera, hablaríamos de tres frases desempeñando dicha función (*de materia*, *de energía* y *de espacio*, en las que la preposición de las últimas dos estaría elidida) para la primera definición; con lo anterior, estas estructuras podrían alinearse con las frases de la segunda, quedando un espacio vacío en donde aparecen las preposiciones que contiene esta última. También deberán alinearse con conjuntos vacíos los artículos contenidos en este tipo de complementos adnominales, pues su empleo no marca una diferencia semántica notable entre las construcciones que aparecen determinadas por ellos y las que no.

Los pares nulos formados por artículos o preposiciones y sus correspondientes elipsis deberán ser considerados semi nulos para el cómputo de LCC (evitando así que se corten las cadenas que asignan un valor a cada par correspondiente en relación con su entorno), siempre que se modifiquen los criterios para la asignación de este valor (2.3.2). De lo contrario, el exceso de pares semi nulos conllevará más dificultades que soluciones, pues hay que recordar que la aplicación de esta opción en el algoritmo flexibilizado incrementa considerablemente el número de pares vinculados, pero, en gran medida, se trata de malos pares.

Por otra parte, en un alineamiento, las palabras (o locuciones) enlazadas mediante una conjunción coordinante dentro de una misma definición deben poderse alinear por duplicado con una sola palabra en la definición con la que se compara.

Además de las conjunciones básicas, actúan como nexo gramatical construcciones perifrásticas denominadas *locuciones conjuntivas*. En las siguientes definiciones podemos ver estructuras que cumplen este oficio:

- **Dinámica:** Parte de la mecánica que estudia el movimiento **en relación con** las fuerzas que lo producen.[21]

- **Aceleración centrípeta:** Cuando un cuerpo se mueve en una trayectoria curva, siempre con la misma rapidez, su velocidad no es constante, **ya que** va cambiando de dirección. Entonces la velocidad cambia a cada instante **y por lo tanto** hay[22]…

La propuesta para este tipo de estructuras es no sólo tratarlas como unidad (3.1.1), sino que también, en la medida de lo posible, habrá que añadir una etiqueta adicional que explique cuál es su función y, así, se puedan alinear sólo con nexos.

21 Vox.
22 Física de emergencia.

Número de alineamiento: 9
Término: Física
Fuente de la definición 1: Física. Conceptos y aplicaciones
Fuente de la definición 2: DRAE 92

Def. 1	**F**	La	Física	puede	definirse	como		la	ciencia	que	investiga	los	conceptos	fundamentales	de	materia	energía	y	espacio	…		
Def. 2	**F**	Ciencia	que	estudia	las	propiedades	de	la	materia	**y**	de	la	energía	considerando	tan	sólo	los	atributos	capaces	…		
Tipo de par	I	C	C	C	C	C	N	I	C	C	C	C	C	C	C	C	C	C	C			
LCC	0	2	1	1	1	1	0	0	2	1	1	1	1	1	1	1	1	1	1			
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no			

Número de alineamiento: 10
Término: Primera ley de Newton
Fuente de la definición 1: Física universitaria Sears
Fuente de la definición 2: Física de emergencia

Def. 1	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en		su	estado	de	reposo	**o**	movimiento	rectilíneo	uniforme	a	menos	que	sea	impelido	…				
Def. 2	**P**	**l**	**d**	**N**	Todo	cuerpo	tiende	a	conservar	su	estado	de	reposo	**o**	**bien**	su	estado	de	movimiento	uniforme	rectilíneo		...				
Tipo de par	**I**	**I**	**I**	**I**	I	I	C	C	N	I	I	I	I	I	C	C	C	C	C	C	C	N					
LCC	**0**	**0**	**0**	**0**	0	0	7	1	0	0	0	0	0	0	6	1	1	1	1	1	1	0					
Par vinculado	**no**	**no**	**no**	**no**	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no					

Número de alineamiento: 11
Término: Ley de la conservación de la energía
Fuente de la definición 1: Enciclopedia Lexis 22
Fuente de la definición 2: Física de emergencia

Def. 1	**L**	**d**	**l**	**c**	**d**	**l**	**e**										La									energía	**no**	puede	crearse	**o**		destruirse	sólo					transformarse	…
Def. 2	**L**	**d**	**l**	**c**	**d**	**l**	**e**	Principio		fundamental		de		la	Física		que		afirma		que		la	energía	**no**	se	crea	**ni**	se	destruye		solamente		se	transforma	...
Tipo de par	I	I	I	I	I	I	I	N		N		N	I	N	N		N		N		N		I	I	C	I	C	N	I		I			N	I			
LCC	0	0	0	0	0	0	0	0		0		0	0	0	0		0		0		0		0	0	4	0	2	0	0		0			0	0			
Par vinculado	no	no	no	no	no	no	no	no		no		no	no	no	no		no		no		no		no	no	no	no	no	no	no		no			no	no			

3.2.2. Preposiciones

Las preposiciones son unidades invariables y "dependientes que incrementan a los sustantivos, adjetivos o adverbios como índices explícitos de las funciones que tales palabras cumplen bien en la oración, bien en el grupo unitario nominal."[23]

En este apartado, seleccionamos dos aspectos de las preposiciones que generan dificultades en los alineamientos y en los cuales vimos que era viable sistematizar procedimientos para su eliminación.

3.2.2.1. Verbos que rigen preposición

Al buscar pares semánticos manualmente, encontramos otro fenómeno que obstaculiza el correcto alineamiento de las definiciones. Se trata de los pares correspondientes que desencadena la unión de un verbo que rige preposición con otro verbo transitivo en un mismo par correspondiente.

Por su naturaleza, ciertos verbos —o algunas de sus acepciones— exigen el acompañamiento de un suplemento (frase compuesta por un sustantivo —o equivalente— precedido por una preposición), "pues su ausencia privaría de sentido cabal al enunciado."[24] En estos casos, "hay como una especie de concordancia semántica entre el significado de la raíz verbal y el de la preposición, con lo cual esta se convierte en un mero índice funcional obligatorio."[25]

Veamos algunos casos:

- **Fuerza conservativa:** Fuerza que cumple que cualquier trabajo que pueda realizar, al desplazar una partícula, no **depende** del cambio seguido sino únicamente de la posición inicial y final del trayecto.[26]

- **Física:** Ciencia que **se ocupa** de los componentes fundamentales del Universo.[27]

- **Éter:** Compuesto químico, orgánico que **resulta** de la combinación de un alcohol consigo mismo, con un ácido, o un alcohol.[28]

Los verbos resaltados son insuficientes para dar a conocer el sentido de las oraciones que los enmarcan y, por ello, aparece, inmediatamente después de cada

---

23 Emilio Alarcos, *op. cit*., p. 216.
24 *Ibid*., p. 286.
25 *Ibid*., p. 220.
26 Diccionario de Física Vox.
27 Encarta 2000.
28 Master. Diccionario enciclopédico.

uno, el suplemento que precisa su sentido. Como vemos, la función del suplemento es la de un modificador verbal, al igual que la del implemento. Tan estrecha es la relación entre ambos modificadores que algunos autores consideran que éstos desempeñan una misma función transitiva: “Con el término transitivo no se recoge aquí una estructura funcional específica sino más bien la relación de rección que se establece entre el predicado y un complemento.”[29] Éste es el criterio que nosotros seguiremos.

Uno de los argumentos con los que se sostiene la relación de identidad semántica entre los complementos oracionales mencionados es que en los suplementos hay una pérdida progresiva del significado propio de la preposición. Por tanto, si ésta carece de significado (salvo el de indicar relación), no queda nada que diferencie, en el plano del contenido, suplemento de implemento.

Si reconocemos que la preposición sirve sólo como el enlace que introduce en el discurso un complemento verbal, no hay por qué asignarle un espacio exclusivo dentro del alineamiento, que es lo que ocurre con el algoritmo básico y el algoritmo flexibilizado. Como consecuencia, tenemos que el nexo aparece formando un par correspondiente con la primera palabra del implemento y, por tanto, hay un desajuste de todo el alineamiento.

Considerando todo lo anterior, vemos que obtendríamos más pares semánticos acertados si se tomara la preposición como un incremento del verbo. Tendría, entonces, que incorporarse al sistema la información que permitiera a la preposición que sigue a verbos como *depender* y *ocuparse* aparecer como un elemento constitutivo suyo. Esto es factible en tanto son contados los verbos que rigen preposición y cada uno de ellos determina una específica. Además, los verbos que rigen preposición aparecen enlistados en gramáticas. Dichas listas podrían introducirse al motor del programa, a pesar de que no todos sus elementos están representados en el corpus, pues esto evitaría etiquetar manualmente cada ocurrencia.

En el alineamiento 12 aparece el par {trata, estudia} con un LCC=7, pero no es promovido a vinculado debido a que no cumple con la condición de frontera (1.2.1.2)**.** En consecuencia con nuestro análisis anterior, el par debería ser {trata de, estudia}.

Por otra parte, consideramos que, para nuestros propósitos, las partículas que son contracción de formas lingüísticas con diferente categoría gramatical deben separase en sus morfemas originales. De esta manera, en la etapa de preprocesamiento de las definiciones, se marcaría *del* como la suma de *de* + *el*, para que preposición y artículo puedan formar pares independientes en los alineamientos. Así que si *trata de* y *estudia* formaran un par, el siguiente par correspondiente sería un par igual {el, el}, por lo que se cumpliría la condición de frontera, y el par que manualmente hemos identificado como semántico sería promovido a vinculado con un LCC=9.

---

29 José María García-Miguel, *Transitividad y complementación preposicional en español*, p. 8.

El número 13 es un alineamiento en donde se comparan definiciones que contienen verbos de régimen preposicional, el segundo de ellos integrado a una perífrasis. Si apareciera como par correspondiente *se relaciona con* y *está referido a*, y se aplicaran otras modificaciones que proponemos en el capítulo, tendríamos además otro par semántico como correspondiente: {perpendiculares, ortogonales}.

La propuesta que manejamos en este apartado es sólo para efectos de Clustering, pues considerar la preposición como un elemento constitutivo del verbo nos pareció lo más atinado para reducir errores.

Número de alineamiento: 12
Término: Cinemática
Fuente de la definición 1: Vox
Fuente de la definición 2: Larousse

Def. 1	**C**	Parte	de	la	Mecánica	que	**trata**	del	movimiento	en	sus	condiciones			de	espacio	y	tiempo			
Def. 2	**C**	Parte	de	la	Mecánica	que	**estudia**	el	movimiento	de	los	cuerpos	con	independencia	de	las	fuerzas	que	lo	producen	
Tipo de par	I	I	I	I	I	I	C	C	I	C	C	C	N	N	I	C	C	C	N	N	
LCC	0	0	0	0	0	0	7	2	0	2	1	1	0	0	0	2	1	1	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 13
Término: Espacio
Fuente de la definición 1: Enciclopedia. Ciencia y tecnología
Fuente de la definición 2: Enciclopedia Britannia

Def. 1	**E**	Aquella	propiedad	del	universo	que	**se**	**relaciona**	**con**	la	extensión			de	tres	direcciones	mutuamente	perpendiculares				
Def. 2	**E**	**Está**	**referido**	**a**	una	terna	de	ejes	mutuamente	ortogonales	y	los	puntos	de	este	espacio	se	ponen	en	…		
Tipo de par	I	C	C	C	C	C	C	C	C	C	C	N	N	I	C	C	C	C	N			
LCC	0	2	1	1	1	1	1	1	1	1	1	0	0	0	2	1	1	1	0			
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no			

3.2.2.2. Locuciones prepositivas

Como en el caso de las conjunciones, pueden emplearse estructuras fijas superiores a la palabra para expresar las relaciones gramaticales propias de la preposición. Gili Gaya dice al respecto[30]:

> Además de las preposiciones que registran los diccionarios como tales, existen numerosas *frases prepositivas* en las cuales figuran ordinariamente un substantivo o un adjetivo: *alrededor de, encima de, dentro de, junto a, frente a, enfrente de, etc.*, y otras muchas que ocasionalmente pueden crearse para precisar así la relación, a veces poco definida de las preposiciones solas.

En las siguientes definiciones de *energía cinética* encontramos que la preposición *por* resaltada de la primera definición está manifestando que entre el adyacente circunstancial que introduce y su núcleo se establece una relación causativa, que es la misma función que cumplen las locuciones de las otras tres definiciones, las cuales perfectamente podrían ser sustituidas por *por* y entre sí, sin que esto significara ningún cambio en el significado global.

- La poseída por un cuerpo o una partícula material **por** su movimiento,…[31]
- La que posee un cuerpo **en virtud de** su movimiento.[32]
- La que posee un cuerpo **por razón de** su movimiento.[33]
- Es la energía que un objeto posee **debido a** su movimiento,…[34]

Nuestra propuesta, en este caso, es la misma que para las locuciones conjuntivas: que con las etiquetas correspondientes a la categoría gramatical se marquen en conjunto como preposiciones, para así delimitar las estructuras con las que pueden alinearse. Un ejemplo en este sentido lo tenemos en el alineamiento 1 del presente capítulo.

3.2.3. Verbos copulativos

Este tipo de verbos (*ser, estar, parecer*) tienen un significado tan amplio que requieren de un atributo que lo especifique. Una estructura de sujeto + atributo donde no aparece verbo presupone la existencia de uno copulativo que se ha elidido, y esto es bastante frecuente porque "en español moderno prescindimos muchas veces del verbo copulativo, especialmente cuando no interesa señalar el

30 Samuel Gili Gaya, *op. cit.*, p. 247.
31 María Moliner.
32 Vox.
33 DRAE 92.
34 Encarta 98.

tiempo."[35] Es frecuente que, cumpliendo la función de nexo, quede como rastro una coma.

Véase el siguiente alineamiento:

Def. 1[36]	**Móvil**	Cuerpo	que	**está**	en	movimiento
Def. 2[37]	**Móvil**	Cuerpo			en	movimiento
Tipo de par	I	I	N	N	I	I
LCC	0	0	0	0	0	0
Par vinculado	no	no	no	no	no	no

Proponemos que este tipo de pares, en los cuales un verbo copulativo (y, en este caso, el nexo que lo precede) encuentra un conjunto vacío en el alineamiento, se consideren pares semi nulos.

3.3. Adjetivación

Para abordar los problemas derivados de las diferentes formas de adjetivación en los alineamientos tenemos que empezar por comprender la noción de relaciones paradigmáticas, que son las que se establecen entre los miembros de una clase paradigmática.

Una clase paradigmática está formada por el conjunto de unidades que pueden ser de diferente clase sintagmática, pero que son susceptibles de realizar el mismo valor funcional en un contexto oracional.

Así, "la relación paradigmática es aquella entablada por una unidad lingüística perteneciente a una cadena y todas aquellas unidades que podrían desempeñar su misma función en un sintagma dado y que, por esta razón, su presencia, en principio, excluye de forma inmediata [sic.]."[38]

Por esto, encontramos en los alineamientos situaciones en donde una palabra aparece desempeñando cierta función y en la oración con la que se compara, esa misma función la cumple una frase o una oración transpuesta. Tal es el caso de los modificadores de la frase nominal que por formar una clase paradigmática pueden resultar recíprocamente sustituibles. En la siguiente definición de *período* encontramos, concatenados, estos posibles modificadores:

- Ciclo, porción de tiempo que comprende la duración total de una cosa.

En un primer nivel se observan, unidas por yuxtaposición, dos frases (*ciclo* y *porción de tiempo que comprende la duración total de una cosa)*. La segunda frase está funcionando como aposición explicativa de la primera. A su vez, al núcleo nominal

35 Samuel Gili Gaya, *op. cit.*, p. 58.
36 Vox.
37 Clave.
38 Guillermo Rojo y Tomás Jiménez, *Fundamentos del análisis sintáctico funcional*, p. 32.

de esta segunda frase lo modifica una frase preposicional (*de tiempo que comprende la duración total de una cosa*), cuyo término de preposición está formado por un sustantivo y una oración transpuesta adjetiva (*que comprende la duración total de una cosa*), dentro de la cual la función de implemento la cumplen un sustantivo (*duración*), el artículo que lo precede y un sintagma adjetivo especificativo, que a su vez se compone de núcleo (*total*) y un modificador representado por una frase preposicional (*de una cosa*).

En los siguientes alineamientos se puede observar la combinación de sintagmas adjetivos, sintagmas preposicionales y oraciones adjetivas ocupando la misma posición como modificadores de un sintagma nominal en estructuras oracionales semejantes. Así, en el alineamiento 14, el adjetivo *luminoso* cumple con respecto a *rayo* la misma función especificativa que la estructura *de luz.* El alineamiento 15 contiene un par semántico similar, formado por el adjetivo *rectilíneo* y la frase *en línea recta,* así como una oración *que actúen sobre él,* en donde el nexo *que* y el verbo conjugado se pueden sustituir por el participio en función adjetiva *impresas.* En el alineamiento 16, vemos que la oración transpuesta *que debe ser aplicada* desempeña una función adjetiva, la que cubre en la primera definición el participio *requerida*. En combinación con otras de las modificaciones que proponemos en el capítulo, si se considerara éste como par vinculado, no sólo obtendríamos un nuevo par semántico, sino otros dos que, con la ayuda de la especialista, identificamos manualmente como tales: {adentro, el centro} y {trayectoria, movimiento}[39].

La propuesta es considerar adjetivos, frases adjetivas, frases preposicionales en función de complemento adnominal y oraciones adjetivas como equivalentes en los alineamientos.

Para los alineamientos que contienen pares formados por un adjetivo y una de las frases mencionadas, o bien, por dos frases, esto puede realizarse —una vez etiquetadas las categorías gramaticales en las definiciones— con la aplicación de un etiquetador sintáctico parcial (denominado *chunking*), el cual es capaza de localizar frases y determinar con qué categoría gramatical se corresponden. Sin embargo, el procedimiento parece más complicado para los casos que involucran oraciones, por lo que sugerimos que este aspecto se considere a largo plazo.

39 *Vid*. Apéndice B.

Número de alineamiento: 14
Término: Difracción
Fuente de la definición 1: DRAE 92
Fuente de la definición 2: Larousse

Def. 1	**D**	Desviación	del		rayo	**luminoso**								al	rozar	el	borde	de	un	cuerpo	opaco					
Def. 2	**D**	Desviación	de	un	rayo	**de**	**luz**	o	de	otro	tipo	de	onda	al	rozar	el	borde	de	un	cuerpo	opaco	o	de	otro	obstáculo	
Tipo de par	I	I	C	N	I	C	N	N	N	N	N	N	N	I	I	I	I	I	I	I	I	N	N	N	N	
LCC	0	0	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 15
Término: Primera ley de Newton
Fuente de la definición 1: Física universitaria Sears
Fuente de la definición 2: El mundo de la Física 1

Def. 1	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o		movimiento	**rectilíneo**	uniforme				a	…				
Def. 2	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o	de	movimiento		uniforme	**en**	**línea**	**recta**	a	…				
Tipo de par	I	I	I	I	I	I	I	I	I	I	I	I	I	N	I	N	I	N	N	N	I					
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no					

Def. 1	…	menos	que	sea	impelido	a	cambiar	dicho	estado	por	fuerzas	**que**	**actúen**	sobre	él	
Def. 2	…	menos	que	sea	obligado	a	cambiar	este	estado	por	fuerzas	**impresas**		sobre	él	
Tipo de par		I	I	I	C	I	I	C	I	I	I	C	N	I	I	
LCC		0	0	0	7	0	0	6	0	0	0	4	0	0	0	
Par vinculado		no	no	no	sí	no	no	sí	no	no	no	no	no	no	no	

Número de alineamiento: 16
Término: Fuerza centrípeta
Fuente de la definición 1: Enciclopedia. Ciencia y tecnología
Fuente de la definición 2: Física Serway

Def. 1	**F**	**c**	La	fuerza		hacia	adentro	**requerida**					para	mantener	a	una	partícula	o	un	objeto	moviéndose	en	una		trayectoria	circular	
Def. 2	**F**	**c**	La	fuerza	dirigida	hacia	el	centro	**que**	**debe**	**ser**	**aplicada**	para	mantener	a				un	cuerpo		en	movimiento			circular	
Tipo de par	I	I	I	I	N	I	C	C	N	N	N	N	I	I	I	N	N	N	I	C	N	I	C	N	I		
LCC	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0		
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no		

3.4. Adverbios

Con respecto a los adverbios encontramos un fenómeno análogo al que se presenta con los diferentes modificadores nominales, visto en el apartado anterior (aunque en menor medida, pues las formas de adjetivar son más variadas y frecuentes). Aquí, la función de modificador verbal propia del adverbio puede estar representada por una palabra con esta categoría o bien por una frase.

De esta manera, en el alineamiento 17, el adyacente circunstancial expresado por el adverbio *libremente* que modifica al núcleo verbal, *puede oscilar*, de la primera definición —que se corresponde con una oración—, se manifiesta en la segunda definición mediante la frase *con libertad.* Considerándose esto junto con otras modificaciones que sugeriremos más adelante (3.7), el sistema asignaría un LCC suficiente para ser promovido a par vinculado al par correspondiente {punto, eje}, que en este contexto es semántico.

En el alineamiento 18 se observa una situación similar: al verbo *aumenta* de sendas definiciones lo modifica un aditamento introducido en la primera por el adverbio *proporcionalmente* y, en la segunda, por *de manera proporcional.*

No considerar situaciones de este tipo en los alineamientos supone que no se vinculen pares semánticos que realmente lo son; además, los pares nulos formados por elementos constitutivos de la frase adverbial y espacios vacíos desencadenan que el sistema asigne a otros pares semánticos un LCC insuficiente para considerarlos pares vinculados.

El procedimiento sugerido para detectar pares de adjetivos y frases análogas serviría también para los adverbios: Aplicar un *chunking* a las definiciones —después de haber llevado a cabo el etiquetado de las categorías gramaticales (POST)— permitiría al sistema reconocer de una sola vez frases adjetivas, adverbiales, sustantivas y preposicionales, para ser tratadas del mismo modo que las palabras con las que comparten la categoría gramatical que en conjunto desempeñan.

Número de alineamiento: 17
Término: Péndulo
Fuente de la definición 1: Clave
Fuente de la definición 2: Larousse

Def. 1	**P**	Cuerpo	que	suspendido	de	un	punto	que	está	por	encima	de	su	centro	de	gravedad	puede	oscilar	**libremente**		alrededor	de	dicho	punto	…
Def. 2	**P**	Cuerpo	rígido					que										oscila	**con**	**libertad**	alrededor	de	un	eje	…
Tipo de par	I	I	C	N	N	N	N	I	N	N	N	N	N	N	N	N	N	I	C	N	I	I	C	C	
LCC	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	3	1	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 18
Término: Movimiento uniformemente acelerado
Fuente de la definición 1: Desconocida
Fuente de la definición 2: Larousse

Def. 1	**M**	**u**	**a**	Aquél	en	que	la	velocidad	aumenta			**proporcionalmente**	al	tiempo	transcurrido														
Def. 2	**M**	**u**	**a**	Aquél	en	el	que	la	velocidad	aumenta	**de**	**manera**	**proporcional**		al	tiempo	transcurrido												
Tipo de par	I	I	I	I	I	N	I	I	I	I	N	N	I		I	I	I												
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0												
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no		no	no	no												

3.5. Negación

En este apartado abordaremos las dificultades que surgen cuando, para referirse a un concepto, el lexicógrafo o el autor de un libro optó por señalar las características que no le son propias en vez de mencionar las que sí posee. En las definiciones que componen el corpus, encontramos variadas formas de expresar ideas con una connotación negativa.

Para entender el fenómeno de la negación debe establecerse una diferencia entre negación gramatical (cuando ésta incide en las estructuras sintácticas) y negación léxica[40]; considerando que "negación no es estrictamente lo contrario de afirmación; porque la negación es una categoría semántica con repercusión sintáctica y en ocasiones morfosintáctica, como en el caso de los prefijos."[41]

A continuación enlistamos las formas lingüísticas encontradas en el corpus que implican negación:

- Los adverbios *no* y *nunca.*
- Los adjetivos *ningún* y *ninguna*.
- La preposición *sin,* que antepuesta a un infinitivo equivale a *no*.
- Formas que contienen el prefijo –in: *independiente, incapacidad*.
- Elementos coordinados en donde una de las conjunciones o ambas introducen oraciones con significado negativo: *no … pero sí…, no … sino…, ni… ni …, no … ni … , sin… ni …*
- Marcas que introducen restricciones: *si no, sin*; los adverbios *sola* y *únicamente*; las locuciones adverbiales *a menos que*, *a no ser que*, *aunque no, siempre que no* y algunos verbos con carga semántica negativa *excluir, prescindir, oponerse*, etc.

También observamos que las formas típicamente negativas no siempre expresan estructuras de este tipo:

- **Energía cinética:** Ésta **no sólo** depende de la velocidad, **sino** también de la masa.[42]

40 "…en elementos cuyo contenido o significado léxico es negativo, pero no podemos decir, evidentemente, que sea una negación gramatical, porque no afecta a la estructura, sino que el significado de esas unidades es negativo". Beatriz Sanz, *La negación en español*, p. 17.
41 *Ibid*., p. 16.
42 Atlas de Física.

En esta definición, la segunda parte amplía el contenido de la primera. No existe oposición o disyuntiva entre ambas, a pesar de contener un *no* antepuesto al núcleo oracional.

- **Onda luminosa:** Una onda luminosa de frecuencia determinada se propagará por los distintos medios con velocidades que son características de cada uno de ellos. Maxwell determinó que las ondas luminosas **no eran otra cosa que** ondas electromagnéticas.[43]

Aquí, la expresión negativa *no eran otra cosa que* equivale a decir *son* (o eran), es decir, marca una afirmación.

Atendiendo a este último aspecto, sería pertinente introducir a las definiciones donde aparecen este tipo estructuras y frases coloquiales con significado metafórico una etiqueta semántica que indique que el sentido de la expresión es positivo.

Observamos que si ambas ideas expresadas en las definiciones que se comparan son negativas, no generan dificultad para el análisis. El problema surge cuando una definición afirmativa se alinea con otra donde el significado del término se expresa negando ciertos atributos. Tal como sucede en los siguientes alineamientos.

En el alineamiento 19, consideramos una unidad sintáctica el verbo *ser* y su atributo, a la cual le corresponde alinearse con un verbo conjugado —en este caso unido a un adverbio de negación—. Aquí, nos encontramos ante formas de negación básicas y análogas: el significado original de la raíz léxica se ve modificado por la anteposición de un elemento lingüístico que expresa una negación, la diferencia es de carácter ortográfico: en el primer caso, la marca de negación se fusiona con el adjetivo original, dejando su rastro en un prefijo *in-*, mientras que el adverbio *no* debe permanecer separado en la escritura del verbo al que modifica.

En el alineamiento 20, para referirse a *estado de reposo* se opta por dos fórmulas distintas, una afirmación para la primera definición y una negación para la segunda. Sin embargo, hay que observar que en estas definiciones lo que se afirma o niega son procesos opuestos: En la segunda definición, el significado del verbo *cambiar* que funciona como núcleo del enunciado adquiere un carácter negativo debido al *no* antepuesto al verbo conjugado; mientras que el verbo de la primera oración es *permanecer* —antónimo de cambiar—. Como el significado de uno de los miembros de la pareja de antónimos se invierte, debido a la presencia del adverbio de negación, *permanece* y la estructura *no cambian* deben ser considerados como par semántico (la diferencia en el número se debe al criterio del lexicógrafo, pero se trata del mismo fenómeno).

Por otra parte, las dos definiciones que componen el alineamiento 21 comparten estructura sintáctica: Una oración principal en la que se inserta una transpuesta adjetiva, que en el primer caso es “que no modifican la estructura molecular de los cuerpos” y, para la segunda definición, “que ocurren en la materia excluyendo los

43 Atlas de Física.

que modifican la estructura molecular de los cuerpos"; ambas transpuestas adjetivan a *fenómenos*.

Las oraciones adjetivas, como ya hemos mencionado (3.3), presentan comportamiento y cualidades análogas a los de los adjetivos; así, tenemos tanto oraciones transpuestas explicativas como especificativas, tal es el caso de las de nuestro alineamiento. En la primera definición, "que no modifican la estructura molecular de los cuerpos" sirve para especificar de qué tipo de fenómenos se trata, por consiguiente, los fenómenos que sí modifican la estructura molecular de los cuerpos no son objeto de estudio de la Física, que es lo que con otras palabras señala la oración transpuesta de la segunda definición, "que ocurren en la materia **excluyendo** los que modifican…"; o lo que es lo mismo: cuáles fenómenos forman parte del campo de la Física, que son los que no entran en la restricción introducida por el gerundio *excluyendo.*

Como hemos visto, los recursos para expresar una negación en español (y también en nuestro corpus) son muy diferentes, por lo que no es posible ofrecer una fórmula que sea válida para identificar cuándo el contenido semántico de un enunciado es negativo de cuando no lo es. Sería óptimo que a cada estructura con connotación negativa se le incorporara una etiqueta semántica que lo hiciera explícito; hasta ahora, eso sólo es posible de manera manual, situación que hace inoperante la propuesta.

Lo que sí puede realizarse automáticamente es tratar el adverbio de negación como un elemento constitutivo del núcleo del enunciado, pues éste es el que aporta el contenido semántico básico a los diferentes tipos de sintagmas y, cuando ese contenido se ve modificado por la negación, se modifica el significado del enunciado en su conjunto. Con esto, prevemos que se eliminará el problema que surge cuando una definición negativa se compara en el alineamiento con una afirmativa, siendo que el significado global de ambas es equivalente.

Número de alineamiento: 19
Término: Fuerza conservativa
Fuente de la definición 1: Mecánica vectorial. Estática
Fuente de la definición 2: Diccionario de Física Vox

Def. 1	**F**	**C**	El	trabajo	realizado	por	una	fuerza	**es**	**independiente**	**de**	la	trayectoria	seguida	y	es	igual	a	menos	el	cambio	de	…
Def. 2	**F**	**C**	Fuerza	que	cumple	que	cualquier	trabajo	que	pueda	realizar	al	desplazar	una	partícula	**no**	**depende**	**de**l			cambio	seguido	…
Tipo de par	**I**	**I**	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	N	N	I	C	
LCC	**0**	**0**	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	2	
Par vinculado	**no**	**no**	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 20
Término: Estado de reposo
Fuente de la definición 1: Lexipedia
Fuente de la definición 2: Enciclopedia Lexis 22

Def. 1	**E**	**d**	**r**	Se	dice	que	un	cuerpo	está	en	reposo	cuando	**permanece**	en	el	mismo	lugar			
Def. 2	**E**	**d**	**r**	Estado	de	los		cuerpos		en	el	cual	**no**	**cambian**	de		lugar	o	de	posición
Tipo de par	I	I	I	C	C	C	N	I	N	I	C	C	C	C	C	N	I	N	N	N
LCC	0	0	0	4	1	1	0	0	0	0	2	1	1	1	1	0	0	0	0	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 21
Término: Física
Fuente de la definición 1: Vox
Fuente de la definición 2: María Moliner

Def. 1	**F**	Ciencia	que	estudia	la	materia			en	relación	con		los	fenómenos	que	**no**	modifican	la	estructura	molecular	de	los	cuerpos	
Def. 2	**F**	Ciencia	que	estudia	los	fenómenos	que	ocurren	en	la	materia	**excluyendo**	los		que		modifican	la	estructura	molecular	de	los	cuerpos	…
Tipo de par	I	I	I	I	C	C	N	N	I	C	C	N	I	N	I	N	I	I	I	I	I	I	I	
LCC	0	0	0	0	5	1	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

3.6. Sintagmas enfáticos

Con *sintagmas enfáticos* nos referiremos a todas aquellas estructuras que en determinados contextos actúan exclusivamente como recursos expresivos intensificadores, pues no aportan información semántica relevante a las definiciones en el proceso de alineamiento.

Para que exista una estructura oracional es necesario un núcleo verbal; todos los posibles sintagmas en torno a éste sirven para completar la información aportada por sus signos léxico y gramatical por lo que funcionalmente se definen como adyacentes. En este ámbito se establece la distinción fundamental que nos da Escandell "entre *argumentos* (aquellos constituyentes cuya presencia viene impuesta y exigida por el verbo, que restringe tanto su naturaleza categorial como sus propiedades semánticas) y *adjuntos*, o *complementos circunstanciales,* (aquellos cuya aparición es más libre)."[44] La autora señala que la distinción entre argumentos y adjuntos puede trasladarse al nivel sintagmático de la frase. En el caso de los adjetivos, éstos siempre exigen un antecedente; no ocurre lo mismo con los sustantivos, pues depende de la base léxica de cada uno.

Esta sección se refiere a los modificadores (independientemente de su categoría sintagmática) que presentan valor enfático o redundante en relación con su entorno. Por lo general, a lo largo del corpus, estos sintagmas se corresponden con la categoría gramatical de los adverbios.

Veamos los alineamientos. En el número 22 aparece integrada a una de las definiciones una oración transpuesta que no tiene par en la otra. Aunque el signo léxico del sustantivo al que se refiere —*capacidad*— presupone una estructura argumental que exigiría una propiedad y un propietario, éste quedará sobreentendido siempre que nos desenvolvamos en el dominio de la Física; por tanto, el hecho de que en la segunda definición no aparezca la acotación marcada por la oración *que tiene un sistema físico* no supone una carencia de información con respecto a la contenida en la definición con la que se compara.

En el alineamiento 23, el adjetivo *toda* que califica a *causa* es enfático, pues la frase adjetiva que tiene como núcleo *capaz* delimita ya a qué tipo de causas se refiere. En la misma definición encontramos un par nulo formado por un conjunto vacío y la preposición *de*. Ésta no debe ser tratada como redundante, sino considerarse implícita en la segunda definición, tal como explicamos más arriba (3.2.1).

Por otra parte, las definiciones que integran el alineamiento 24 presentan una estructura sintáctica semejante:

F. sust. = N + f. adj. ([adv.] + Núcleo(adj.) + f. prep. (enlace prep. + término + [f. prep.]))

44 Ma. Victoria Escandell, *Los complementos del nombre*, p. 19.

Entre corchetes incluimos los adyacentes que aparecen únicamente en la primera definición. Como puede observarse, ésta contiene dos adyacentes acompañando al núcleo *pequeña* en la frase adjetiva, pero entre ellos existe una diferencia notable, mientras que el pospuesto es necesario para restringir las posibilidades semánticas infinitas del adjetivo, el adverbio *muy* resulta prescindible, pues su función se reduce a intensificar una noción de por sí ambigua, ya que no se definen las dimensiones de *pequeña* y/*o muy pequeña.* El segundo adyacente que no tiene paralelo podría parecer también optativo, puesto que en la segunda definición no se contempló el rasgo *con entidad propia* para definir el concepto de *partícula*; sin embargo, no lo es: Esta frase hace que la primera definición sea más precisa, con lo que se reduce la ambigüedad presente en la segunda definición que podría remitirnos a otros conceptos físicos.

Consideramos que los sintagmas enfáticos que tuvieran como correspondiente un espacio vacío deberían tratarse como semi nulos. En este caso, el par {muy, ε} no cortaría la cadena de pares iguales dentro de la que aparece, con lo que (suponiendo que no se contabilizara para el cómputo de LCC, como es nuestra propuesta) el coeficiente de similitud del par {cantidad, parte}, se incrementaría de 2 a 5.

Sin embargo, el reconocimiento de este tipo de pares obedece más a criterios semánticos que a sintácticos; entonces, resulta muy importante la intervención humana para decidir cuándo un adyacente es innecesario. Por otra parte, en el recorrido de nuestro análisis hemos observado que la eliminación de problemas en algunos aspectos de los alineamientos se traduce en su reordenación general. Por tanto, dejamos esta línea abierta, esperando que las propuestas más viables, junto con los resultados de otras investigaciones que actualmente se desarrollan en el GIL, adelanten pasos en este sentido.

Número de alineamiento: 22
Término: Energía
Fuente de la definición 1: CIm1
Fuente de la definición 2: Física Weber

Def. 1	**E**	Capacidad	**que**	**tiene**	**un**	**sistema**	**físico**	de	realizar	trabajo
Def. 2	**E**	Es	la	capacidad				de	efectuar	trabajo
Tipo de par	I	C	C	C	N	N	N	I	C	I
LCC	0	2	1	1	0	0	0	0	3	0
Par vinculado	no	no	no	no	no	no	no	no	sí	no

Número de alineamiento: 23
Término: Fuerza
Fuente de la definición 1: DRAE 92
Fuente de la definición 2: Atlas de Física

Def. 1	**F**			Causa	capaz	de	modificar	el	estado	de	reposo	o	de	movimiento	de	un	cuerpo
Def. 2	**F**	Es	**toda**	causa	capaz	de	alterar	el	estado	de	reposo	o		movimiento	de	un	cuerpo
Tipo de par	I	N	N	I	I	I	C	I	I	I	I	I	N	I	I	I	I
LCC	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
Par vinculado	no	no	no	no	no	no	sí	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 24
Término: Partícula
Fuente de la definición 1: Vox
Fuente de la definición 2: DRAE 92

Def. 1	**P**	Cantidad	**muy**	pequeña	de	materia	**con**	**entidad**	**propia**
Def. 2	**P**	Parte		pequeña	de	materia			
Tipo de par	I	C	N	I	I	I	N	N	N
LCC	0	2	0	0	0	0	0	0	0
Par vinculado	no	no	no	no	no	no	no	no	no

3.7. Determinantes

En los alineamientos encontramos un fenómeno recurrente que puede ser atendido apelando a la noción de determinación, pues coincidimos con Manuel Leonetti en que éste es un "concepto muy útil que permite simplificar enormemente la descripción gramatical"[45].

La determinación es una propiedad que desempeñan clases de palabras con diferente categoría gramatical, la cual abarca las operaciones deícticas, anafóricas, referenciales y de cuantificación, "que ponen en relación las expresiones lingüísticas con las entidades representadas por ellas."[46] En este sentido, entendemos que son determinantes los **artículos** y los **demostrativos**, cuya función, ya sea adjetiva o pronominal, es deíctica; en consecuencia, son también determinantes los **pronombres** (los cuales, por otra parte, carecen de contenido conceptual), y por sus propiedades pronominales, los **posesivos**. También consideramos que pertenecen a esta clase paradigmática los **indefinidos**, tales como *todo, un, tan, tal, algún, cualquier,* etc.

Veamos lo que sucede en los alineamientos. En el número 25, el *la* que precede a recta y el *una* que hace lo propio con *flecha* coinciden en ser artículos con género femenino y número singular; el único rasgo que los diferencia es el de la *definitud*: El hablante —en este caso, el lexicógrafo— opta por el artículo definido cuando en el contexto de uso se cumple la *condición de unicidad* (es decir, que el referente sea el único objeto que satisfaga la descripción aportada), pero cuando la *flecha* no es la única posible se opta por el indefinido. Si *la* y *una* comparten la mayoría de sus propiedades esenciales, pertenecen a la misma categoría gramatical y desempeñan las mismas funciones sintácticas y semánticas, no hay razón para que Clustering las trate como palabras diferentes, sino que deben implementarse los mecanismos para que el programa las identifique como variantes de una sola expresión. El fenómeno se repite en el alineamiento: determinando a *vector,* encontramos en la primera definición, el artículo indefinido, *un,* mientras que en la segunda, esta función la cumple su correspondiente definido *el.*

En el mismo alineamiento aparece una muestra más de diferentes manifestaciones de determinantes que semánticamente cumplen una misma función. Se trata del posesivo *su* y el artículo *la*, que en la primera y segunda definición, respectivamente, afectan a *dirección.* En el primer caso se justifica el empleo del posesivo porque el verbo copulativo indica que la frase *su dirección* se refiere a *vector*; mientras que en la segunda definición esa relación de posesión se expresa con el genitivo *del* (de + el) *vector,* que afecta simultáneamente a dirección y sentido.

El alineamiento 26 también nos proporciona tres ejemplos de este tipo. Por un lado, *un* y *todo* que acompañan a *objeto* y a *cuerpo* respectivamente son determinantes pertenecientes a los indefinidos cuantitativos, semánticamente equivalentes. *Una* y *alguna,* modificando a *fuerza,* también pertenecen a la clase de los indefinidos. La marca de posesivo (*su*) que determina a *estado de reposo* en la segunda definición no tiene paralelo en la primera, sin que por ello se establezca una diferencia de sentido entre ambas.

45 Leonetti, Manuel, *Los determinantes*, p. 11.
46 Ma. Victoria Escandell, *op. cit*., p. 23.

Aparte de las que se muestran en estos ejemplos (artículo definido e indefinido, artículo definido y posesivo, posesivo y conjunto vacío, así como indefinido con indefinido), en el corpus se manifiestan las siguientes relaciones entre determinantes:

- Nombre / pronombre
- Nombre / indefinido
- Pronombre / indefinido
- Artículo definido / demostrativo
- Artículo definido / elemento vacío
- Demostrativo / demostrativo
- Posesivo / demostrativo
- Indefinido / elemento vacío

Como se puede apreciar, las relaciones son complejas. Nosotros proponemos que los pares correspondientes formados por palabras pertenecientes a estas categorías sean considerados como iguales por el sistema (y sólo para este programa), pues cumplen funciones idénticas. Si hay definiciones en singular y otras en plural que incorporan palabras con la misma base léxica pero diferente número y el programa las considera iguales, no existe razón para que *él* y *ellos*, por ejemplo, no sean tratados de la misma manera.

Intuitivamente, la versión flexibilizada del algoritmo ofrece una solución similar para el agrupamiento de determinantes, la consideración de pares semi iguales. Sin embargo, este criterio abarca a todos los determinantes y no establece distinciones de tipo funcional, por lo que no incorpora ninguna restricción al agrupamiento dentro de un par semi igual de palabras cuya función no es equiparable, sino que agrupa como tales a cualquier tipo de determinantes por considerarlos palabras funcionales, al igual que las preposiciones y las conjunciones. Consideramos que los pares formados por determinantes se consideren semi iguales sólo si los dos miembros del par pertenecen a la misma categoría gramatical.

En el alineamiento 27[47], obsérvese que el par correspondiente formado por los demostrativos *este* (elemento propiamente demostrativo) y *dicho* (que puede funcionar como tal), se identifica como par vinculado. Nosotros consideramos que entre estas dos palabras la relación semántica que se establece a partir de su función demostrativa es tan estrecha que deben considerarse más que par semántico, par semi igual (según hemos dicho, 2.3.1); lo cual tiene como implicación que el par que forman ayude a identificar pares semánticos en el mismo alineamiento. Si atendemos al par correspondiente {impelido, obligado}, veremos que éste es promovido a vinculado por poseer un LCC de 7; pero si se aplicaran sobre el alineamiento las modificaciones que se sugieren en la presente tesis, el coeficiente de similitud aumentaría a 26.

Finalmente, consideramos que la propuesta que hacemos a partir del análisis de los alineamientos es viable, puesto que los determinantes constituyen paradigmas cerrados cuya posición sintáctica es estable, por lo que pueden ser fácilmente identificables por el sistema de manera automática.

47 Este alineamiento ya fue materia de análisis en la sección 3.1.3

Número de alineamiento: 25
Término: Dirección de un vector
Fuente de la definición 1: Gran enciclopedia Larousse
Fuente de la definición 2: Teoría y problemas

Def. 1	**D**	**d**	**u**	**v**	**La**	recta	a	que	pertenece	la	longitud	del	segmento	de	**un**	vector			es	**su**	dirección		
Def. 2	**D**	**d**	**u**	**v**	Es	**una**	flecha	cuya	longitud	y	orientación	representa	el		módulo	y	**la**	dirección	y	sentido	**del**	vector	
Tipo de par	I	I	I	I	C	C	C	C	C	C	C	C	C		C	C	C	N	N	I	C	C	
LCC	0	0	0	0	5	1	1	1	1	1	1	1	1		1	1	1	0	0	0	2	1	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no		no	no	no	no	no	no	no	no	

Número de alineamiento: 26
Término: Primera ley de Newton
Fuente de la definición 1: Dictionary of Physics
Fuente de la definición 2: Física de emergencia

Def. 1	**P**	**l**	**d**	**N**	**Un**	objeto	continúa	en				estado	de	reposo	…								
Def. 2	**P**	**l**	**d**	**N**	**Todo**	cuerpo	tiende	a	conservar	**su**		estado	de	reposo	…								
Tipo de par	I	I	I	I	C	C	C	C	N	N	I	I	I										
LCC	0	0	0	0	5	1	1	1	0	0	0	0	0										
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no										

Def. 1	…	o	en	velocidad	constante	al	menos		que		actúe					sobre	él	**una**		fuerza	externa		
Def. 2	…	o	bien	**su**		estado	de	movimiento	uniforme		rectilíneo	a	no	ser	que	sobre	él	actúe	**alguna**	fuerza			
Tipo de par		I	C	C		C	C	C	C		C	N	N	N	N	I	I	C	N	I	N		
LCC		0	5	1		1	1	1	1		1	0	0	0	0	0	0	3	0	0	0		
Par vinculado		no	no	no		no	no	no	no		no	no	no	no	no	no	no	no	no	no	no		

Número de alineamiento: 27
Término: Primera ley de Newton
Fuente de la definición 1: Física universitaria Sears
Fuente de la definición 2: El mundo de la Física 1

Def. 1	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o		movimiento	rectilíneo	uniforme				a	menos	que	sea	impelido	a	cambiar	**dicho**	estado	…									
Def. 2	**P**	**l**	**d**	**N**	Todo	cuerpo	continúa	en	su	estado	de	reposo	o	de	movimiento			uniforme	en	línea	recta	a	menos	que	sea	obligado	a	cambiar	**este**	estado	…									
Tipo de par	I	I	I	I	I	I	I	I	I	I	I	I	I	N	I	N	I	N	N	N	I	I	I	I	C	I	I	C	I									
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	6	0									
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	sí	no	no	sí	no									

3.8. Relaciones léxicas de hiperonimia e hiponimia

Hay ocasiones en que las definiciones que se comparan se distinguen porque una describe su referente de manera general, mientras que la otra es más específica. Compárese la relación existente entre *objeto central* y *Sol* en la siguiente pareja de definiciones de *afelio*:

- Punto de la trayectoria elíptica de un planeta o satélite en el que se encuentra más lejos del objeto central que lo atrae.[48]

- Punto de la órbita de un planeta en que éste se aleja más del Sol.[49]

La segunda definición es correcta, pero está acotada al movimiento de los planetas alrededor del Sol; mientras que la primera abarca, además, la trayectoria de satélites que pueden moverse en torno a un planeta, e incluso fuera del sistema solar, ya que las órbitas elípticas no son exclusivas de éste.

Sol y *objeto central* forman un par semántico; si se consideran equivalentes de antemano, el usuario del diccionario onomasiológico tendrá mayores probabilidades de acceder con éxito al término *afelio* cuando haya olvidado esta palabra, independientemente de si está pensando en un significado restringido al sistema solar o más amplio.

En el alineamiento 28 sucede algo similar: un cronómetro es un instrumento, pero también un reloj, y todo reloj es un instrumento.

Las definiciones del alineamiento 29 buscan explicar la causa que produce la visión (la *luz*), pero también describen el lugar donde se produce. La primera definición es más específica, habla de *radiación* y *retina*, mientras que la segunda definición lo hace con conceptos más generales, *fenómeno físico* y *ojo*, respectivamente. Aquí es interesante observar que si las relaciones todo-parte se consideraran un criterio para la identificación de pares semánticos, esto podría mejorar los alineamientos, pues la información de que dispondría el sistema haría que se alinearan en pares, por un lado, *radiación* y *fenómeno físico*, y por otro, *retina* y *ojo*; una disposición del alineamiento a partir de estos pares permitiría entonces que se vinculara también el par semántico {incide, afecta}.

En el alineamiento 30 identificamos manualmente un par semántico que pertenece a esta clase de relaciones, pues *deslizamiento* es un tipo de *movimiento.* Considerarlo así podría permitir al sistema identificar otros pares semánticos en este contexto: *evita* y *retarda*, con *se opone*.

Actualmente, el algoritmo permite descubrir pares de este tipo, tal es caso de {telescopio, intrumento}, que se obtiene de un alineamiento que no pertenece a

48 Física de emergencia.
49 Larousse.

nuestro corpus. Sin embargo, la obtención de estos pares se debe a la posición que las palabras que los forman ocupan en el alineamiento. Lo que proponemos sobre este aspecto es que el motor del programa haga uso de redes semánticas existentes en el mercado[50] para identificar las relaciones de hiponimia e hiperonimia asociadas a las palabras contenidas en las definiciones, con la finalidad de establecer pares similares a los vistos en los ejemplos, los cuales servirán para que, a partir de ellos, se ordenen las estructuras que los integran dentro de los alineamientos y, así, ayudar al reconocimiento automático de otros pares semánticos.

50 Una importante red para el español hace parte de EuroWordNet, que es un sistema electrónico de referencia léxica-conceptual para varias lenguas europeas que sigue la estructura desarrollada con WordNet para el inglés. Ambas constituyen un sistema de redes compuesto por unidades léxicas y sus relaciones, las cuales pretenden reproducir, basándose en teorías psicolingüísticas, la organización de la información léxica en los hablantes.
EuroWordNet: http://www.illc.uva.nl/EuroWordNet/
WordNet: http://globalwordnet. com.mx

Número de alineamiento: 28
Término: Cronómetro
Fuente de la definición 1: Larousse de ciencias y técnicas
Fuente de la definición 2: Vox

Def. 1	**C**	**Instrumento**	de	precisión	destinado	a	conservar	la	hora	del	meridiano	de	origen	y	que	permite	fijar	la	posición	…
Def. 2	**C**	**Reloj**	de	alta	precisión	para	medir	fracciones				de	tiempo	muy	pequeñas					
Tipo de par	I	C	I	C	C	C	C	C	N	N	N	I	C	C	C	N	N	N	N	
LCC	0	3	0	2	1	1	1	1	0	0	0	0	2	1	1	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 29
Término: Luz
Fuente de la definición 1: Larousse
Fuente de la definición 2: Física de emergencia

Def. 1	**L**	**Radiación**	que	emitida	por	un	cuerpo	Incide	en	la	**retina**	provocando	la	sensación	de					visión
Def. 2	**L**	Podemos	definir	la	luz	como	aquel	**fenómeno**	**físico**	que	afecta	al	**ojo**	humano	produciendo	lo	que	llamamos	la	visión
Tipo de par	I	C	C	C	C	C	C	C	C	C	C	C	C	C	C	N	N	N	N	I
LCC	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 30
Término: Fuerza de fricción
Fuente de la definición 1: Ingeniería mecánica. Estática
Fuente de la definición 2: El mundo de la Física 1

Def. 1	**F**	**d**	**f**	Es	la							fuerza	de	resistencia	que	actúa	sobre	un	cuerpo	la	cual	evita	o	retarda	su	…
Def. 2	**F**	**d**	**f**		La	superficie	ejerce	sobre	el	objeto	una	fuerza	de	fricción	que	se	opone	siempre	al	**movimiento**	y	cuya	magnitud	es	propoprcional	…
Tipo de par	**I**	**I**	**I**	N	I	N	N	N	N	N	N	I	I	C	I	C	C	C	C	C	C	C	C	C	C	
LCC	**0**	**0**	**0**	0	0	0	0	0	0	0	0	0	0	4	0	2	1	1	1	1	1	1	1	1	1	
Par vinculado	**no**	**no**	**no**	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Def. 1	…	**deslizamiento**	en	relación	con	un	segundo	cuerpo	con	el	que	se	encuentra	en	contacto	Esta	fuerza	siempre	actúa	tangencial	a	la	superficie
Def. 2	…	a	la	fuerza	con	que	el	objeto	presiona	la	superficie	o	sea	a	la		fuerza	normal	de	contacto	n		
Tipo de par		C	C	C	I	C	C	C	C	C	C	C	C	C	C	N	I	C	C	C	C	N	N
LCC		1	1	2	0	2	1	1	1	1	1	1	1	1	1	0	0	2	1	1	1	0	0
Par vinculado		no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

3.9. Abreviaturas, siglas y símbolos

En los alineamientos se presentan situaciones en las que una expresión es descrita de manera extensa en un caso, y utilizando símbolos, abreviaturas o siglas en el otro. Esto es muy frecuente en Física porque muchos conceptos aparecen explicados con fórmulas donde se involucran símbolos. En el siguiente par de definiciones de *dimensión*, las magnitudes de longitud y tiempo, expresadas con símbolos en la primera definición, aparecen desarrolladas en la segunda.

- Es la combinación algebraica de [L], [T] y [M] a partir de las cuales se forma la cantidad.[51]

- Dimensión es el nombre que recibe cualquier cantidad física para poder expresar su medición como en magnitudes, longitud o tiempo.[52]

El alineamiento 31 aporta dos definiciones del término *energía potencial gravitacional.* En la segunda definición, éste se menciona de manera sintética, mientras que en el primero *ep* se utiliza para abreviar *energía potencial* y *gravitacional* no se abrevia. Esta situación genera que gravitacional se alinee con un conjunto vacío, cuando en el fondo sí tiene paralelo en el alineamiento. Si *ep gravitacional* se tomara como par semántico de *epg,* el LCC del par correspondiente {cuerpo, objeto} se incrementaría de 3 a 5, aunque con el inconveniente de que no cumpliría con la condición de frontera.

También se encuentran otras formas de abreviatura que no son símbolos. En las siguientes definiciones del *experimento de Michelson y Morley* encontramos que los nombres de los diseñadores del experimento aparecen completos en la primera definición y en la segunda se muestran los apellidos antecedidos por iniciales que representan al nombre de pila.

- Hasta 1887 no había aparecido ninguna grieta en la estructura de la Física clásica, que se estaba desarrollando con rapidez. Aquel año, el físico estadounidense Albert Michelson y el químico estadounidense Edward Williams Morley llevaron a cabo el llamado…[53]

- El experimento mas famoso diseñado para detectar pequeños cambios en la rapidez de la luz se realizó en 1887 por A.A. Michelson (1852-1931) y E.W. Morley (1838-1923). El experimento se diseñó para medir la velocidad de la tierra con respecto al éter…[54]

Sin embargo, en nuestro corpus no son representativos los casos en que el símbolo sustituye al término, ya que por tratarse de definiciones de obras de consulta o

51 Física para ciencias e Ingeniería.
52 Personal.
53 Encarta 2000.
54 Física Serway.

divulgación, éstos aparecen acompañando a las expresiones completas a las que se refieren. Tal como sucede en las siguientes definiciones:

- **Newton unidad:** Unidad del sistema internacional de unidades para medir la fuerza. Un newton (N) es la fuerza que aplicada a un kilogramo de masa le produce una aceleración de un metro sobre segundo cuadrado.[55]

- **Desplazamiento:** La forma del vector de la distancia, medida en metros (m) y que contempla la distancia y la magnitud.[56]

- **Momento angular:** Es lo mismo que momento de la cantidad de movimiento (m.c.m). Es una de las medidas del movimiento mecánico de un punto material o de un sistema.[57]

En lo que concierne a abreviaturas y símbolos, proponemos identificar éstos en el banco que nutre a Clustering y añadir a cada uno una etiqueta con la expresión completa a la que se refiere. Las formas sintéticas deberán considerarse iguales a las originales para efectos de los alineamientos, lo que permitirá tomar como par igual dos expresiones de un mismo término con apariencia diferente, pero también deberá servir para no separar las dos manifestaciones del término cuando aparecen en una misma definición. Así, *newton (N)* y *metros (m)* ocuparán un espacio en los alineamientos y no dos; asimismo, *momento de la cantidad de movimiento*, que es un término compuesto (3.1.2), junto con su expresión sintética *(m.c.m.)* deberán ocupar un mismo lugar.

Las fórmulas, que son recurrentes dentro del corpus, deben también considerarse una unidad, bajo el entendido de que son expresiones matemáticas que designan un solo concepto. Actualmente, el programa no las reconoce y, por tanto, no establece diferencia alguna con respecto al texto; marca separaciones donde, entre las mismas fórmulas, aparecen espacios en blanco, como puede observarse en el alineamiento 32.

Como ya adelantábamos (3.2.2.1), la presencia de las contracciones de preposición y artículo resultan problemáticas para el sistema. Proponemos que el mecanismo de Clustering las considere como la suma de los constituyentes que las originaron históricamente, con lo que *del* (*de*+*el*) y *al* (*a*+*el*) podrían formar cada una dos pares correspondientes con una preposición y un determinante, obstaculizando así que después de ellas aparezca sistemáticamente un par nulo o semi nulo, como sucede con las versiones del algoritmo básico y del flexibilizado respectivamente.

55 Física de emergencia.
56 Dictionary of Physics.
57 Diccionario enciclopédico de Física.

Número de alineamiento: 31
Término: Energía potencial gravitacional
Fuente de la definición 1: Física aplicada
Fuente de la definición 2: Física general Bueche-Schaum

Def. 1	**Energía**	**potencial**	**gravitacional**	La	**ep**	**gravitacional**	de	un	cuerpo										…
Def. 2	**Energía**	**potencial**	**gravitacional**	La	**Epg**		de	un	objeto	se	define	con	respecto	a	un	nivel	arbitrario	cero	…
Tipo de par	I	I	I	I	C	N	I	I	C	N	N	N	N	N	N	N	N	N	
LCC	0	0	0	0	5	0	0	0	3	0	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 32
Término: Energía cinética de rotación
Fuente de la definición 1: Enciclopedia Salvat
Fuente de la definición 2: Física Kane

Def. 1	**E**	**c**	**d**	**r**	Es	cuando	un	cuerpo	gira	alrededor	de	un	eje	fijo	con	una	velocidad	angular	de	magnitud	w	rad/s	está	dada	por	ek=1/2	iw2	…			
Def. 2	**E**	**c**	**d**	**r**		Cuando	un	objeto	se	mueve	en	un	camino	circular	V	=	Wr	donde	su	energía	cinética	es	como	1/2	m	r2	w2	…			
Tipo de par	I	I	I	I	N	I	I	C	C	C	C	I	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C				
LCC	0	0	0	0	0	0	0	3	1	1	2	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Par vinculado	no	no	no	no	no	no	no	No	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no				

3.10. Patrones definitorios

Existe una gran variedad de formas que se emplean para introducir un concepto —acompañando al término— en la redacción de un texto, tales como la cópula *es* o expresiones del tipo *significa*, *se define como, se refiere a,* y otras más complejas. En una investigación del GIL[58] encaminada a la identificación automática de contextos definitorios en textos especialidad se reúne una clasificación muy completa de estas fórmulas, que aquí denominamos patrones definitorios.

Este tipo de estructuras, mediante las que se relaciona un término con su acepción para un ámbito determinado, están presentes en revistas y libros especializados en la materia correspondiente. En el caso de los diccionarios, las definiciones ya fueron extraídas del contexto en que las reconoció el lexicógrafo. Nuestro corpus, como ya hemos mencionado, está formado por definiciones que pertenecen tanto a textos especializados en Física como a diccionarios de lengua general y sobre temas científicos. Entre estos dos tipos de fuentes se marca una diferencia con respecto a los mecanismos empleados para presentar una definición: En los libros de divulgación solemos encontrar las definiciones precedidas por estos patrones definitorios, en tanto en los diccionarios no, puesto que sus definiciones son resultado del trabajo lexicográfico mencionado.

Así por ejemplo, en los alineamientos 33 y 34, debido a la presencia de patrones definitorios en las definiciones de los textos de divulgación, se aprecia un desajuste importante en los alineamientos que impide que se vinculen posibles pares semánticos.

Considerando esta situación, proponemos que se incorpore a Clustering una lista con todos los patrones registrados en el trabajo mencionado, para que éstos se traten como palabras (o locuciones) pertenecientes a la misma categoría funcional, sólo vinculables con fórmulas de la misma clase. Cuando exista un patrón definitorio para cada definición del alineamiento, el par que resulte de la unión de ambos se considerará semi igual y cuando éste sólo aparezca en una de las definiciones, semi nulo.

58 Rodrigo Alarcón y Gerardo Sierra, *El rol de las predicaciones verbales en la extracción automática de conceptos*.

Número de alineamiento: 33
Término: Calor
Fuente de la definición 1: Clave
Fuente de la definición 2: Dictionary of Physics

Def. 1	**C**	Energía	que	al	ponerse	en	contacto	con	dos	cuerpos	pasa	del	cuerpo	con	mayor	temperatura	al	otro	cuya	temperatura	es	más	baja	…	
Def. 2	**C**	Nombre	dado	a	la	energía	que	es	transferida	de	regiones	de	alta			temperatura	a	aquellas	de	menor					
Tipo de par	I	C	C	C	C	C	C	C	C	C	C	C	C	N	N	I	C	C	C	C	N	N	N		
LCC	0	2	1	1	1	1	1	1	1	1	1	1	1	0	0	0	2	1	1	1	0	0	0		
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no		

Número de alineamiento: 34
Término: Difracción
Fuente de la definición 1: Larousse
Fuente de la definición 2: Física Weber

Def. 1	**D**	Desviación	de	un	rayo	de	luz	o	de	otro	tipo	de	onda	al	rozar	el	borde	de	un	cuerpo	opaco	o	de	otro	obstáculo			
Def. 2	**D**	Es	la	dispersión		de	la	luz	en	una	región	situada	tras						un						obstáculo			
Tipo de par	I	C	C	C	N	I	N	I	C	C	C	C	C	N	N	N	N	N	N	I	N	N	N	N	I			
LCC	0	2	1	1	0	0	0	0	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0			
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no			

3. 11. Consideraciones varias para mejorar el funcionamiento global de Clustering

Las propuestas de modificación que hemos agrupado en los ejes anteriormente descritos tienen en común que se fundamentan en una descripción sintáctica (y, en menor medida, morfológica) de los aspectos que resultan problemáticos dentro de los alineamientos, que es precisamente lo que nos hemos propuesto en esta tesis. No obstante, hemos querido mencionar otras observaciones que se han desprendido de nuestro análisis pero que no pueden explicarse sólo con bases sintácticas. Éstas son lo referente a *signos de puntuación*, *cálculo de LCC* y *reconsideración de pares semi iguales y semi nulos*, así como la *reconsideración de intercambios.*

3.11.1. Signos de puntuación

En posteriores versiones del algoritmo habría que considerar la función de los signos de puntuación. En el siguiente ejemplo lo que se señala con los paréntesis no forma parte de la definición sino que es otra manera de nombrar el término. Si no se considera (como sucede actualmente) que esas marcas ortográficas tienen una función, las palabras que, en conjunto, actúan como sinónimo[59] del término aparecen en los alineamientos como palabras independientes, paralelas a otras que forman parte de las definiciones con las que se compara.

- **Eje de rotación**: (Eje permanente de rotación). Recta alrededor de la cual un cuerpo libre, no solicitado por fuerza alguna, continuará indefinidamente girando si en un instante dado girase alrededor de ella.[60]

También son importantes las comas porque…

a) al igual que los paréntesis, se emplean para delimitar frases incidentales.

- **Cuerpo celeste:** Cuerpo que, para un observador terrestre, describe un movimiento periódico alrededor de la Tierra...[61]

b) sirven de marca ortotipográfica para señalar que un verbo copulativo se ha elidido. En el siguiente ejemplo la coma funciona como nexo, tal como lo haría un verbo copulativo.

- **Elongación:** Alargamiento, resultado de dar más longitud a una cosa.[62]

c) señalan relaciones de yuxtaposición equiparables a una conjunción. (ver 3.2.1)

---

59 Entendiendo sinónimo en un sentido amplio, tal como explica Cabré: “Dos unidades son sinónimas cuando designan un mismo concepto”. Ma. Teresa Cabré, *op. cit*., p. 216.
60 Gran enciclopedia Larousse.
61 Enciclopedia Salvat.
62 Larousse.

Por su parte, el punto señala cuando concluye una explicación e inicia otra.

- **Eje de rotación:** Rotación alrededor de un punto fijo. En este movimiento las partículas que forman el cuerpo rígido se mueven en planos paralelos a lo largo de círculos centrados sobre el mismo eje fijo.[63]

3.11.2. Cálculo de LCC

Clustering no distingue cuáles palabras corresponden a los términos y cuáles a las definiciones para establecer el grado de similitud de un par correspondiente. Como el programa alinea definiciones relativas a un mismo término, los pares que pertenecen a los términos son siempre iguales, y consecuentemente cada uno de ellos suma un punto al LCC del primer par correspondiente próximo al término (cuando no se interpone un par nulo).

Así por ejemplo, en el alineamiento 35 el término consta de cinco palabras (experimento de Michelson y Morley). Éstas, al sumar puntos al primer par correspondiente {lo, experimento}, le otorgan un LCC igual a 7, que lo promueve a par vinculado, lo que indudablemente es un error debido a los puntos que suma el término. Si éste no se considerara para el cómputo de LCC, su coeficiente de similitud tendría un valor de 2, que es el que realmente le corresponde.

Para ser verdaderamente rigurosos, debemos eliminar del cómputo de LCC los términos.

3.11.3. Reconsideración de pares semi nulos y semi iguales

En el capítulo anterior (2.3.2) propusimos que la categoría de pares semi nulos no se eliminara de las opciones del algoritmo flexibilizado (como sugiere Gabriel Castillo, debido a que ésta no produjo en sus pruebas el resultado esperado). Es importante que esta aplicación se mantenga, pues aunque los pares agrupados bajo ese criterio contienen palabras funcionales y, por lo mismo, no aportan contenido importante para fines de recuperación de información sirven para relacionar palabras de base léxica y determinan el tipo de relaciones que se establecen.

Para que dicha operación funcione mejor, retomamos nuestra propuesta de que los pares de este tipo no se consideren nulos (pues, de ser así, interrumpirían las cadenas discursivas de las que forman parte, como sucede en el algoritmo básico), pero tampoco sumen puntos a la asignación de valores de LCC de pares correspondientes próximos.

En el caso de los pares que involucran una preposición o un artículo con un espacio vacío, en donde hipotéticamente se encuentra elidido su equivalente, parece una

63 Mecánica vectorial. Dinámica.

incongruencia manejar que si un elemento se halla implícito, se considere inexistente en la oración en la que sí aparece. Bajo esta lógica, debería tratarse de pares semi iguales y no de semi nulos.

Entonces la división entre pares semi nulos y semi iguales no es muy consistente. Mientras los primeros vinculan palabras “irrelevantes” con conjuntos vacíos donde bien podría aparecer una de ellas, los segundos vinculan dos palabras pertenecientes a este tipo. Así que ambas categorías deberían aplicarse con el mismo criterio. Sin embargo, les daremos un trato diferente por una razón muy sencilla, los pares semi iguales involucran dos palabras y es posible, en un primer momento, analizar las relaciones entre ellas para, consecuentemente, imponer restricciones, lo que es impensable con los pares semi nulos.

El algoritmo flexibilizado permite que se agrupen en pares de este tipo palabras consideradas funcionales que previamente han sido introducidas a una lista, pero la selección no obedece a ningún criterio gramatical o semántico. En este sentido, nosotros proponemos que la lista de palabras “irrelevantes” se componga de las de base gramatical, pero por no tratarse de un conjunto homogéneo en lo que respecta a sus características funcionales, no debe permitirse que se vinculen sin tomar en cuenta otros criterios.

Debemos distinguir entre las palabras funcionales, cuatro subclases: determinantes, pronombres, conjunciones y preposiciones. Una palabra podrá agruparse sólo con las que pertenecen a su subclase; las conjunciones tendrán la restricción adicional de vincularse sólo entre las que se comparta el tipo de relación (copulativa, disyuntiva, adversativa,…). Por la complejidad en las relaciones semánticas que establecen las preposiciones no nos atrevemos a hacer ninguna propuesta en torno a ellas; sabemos de antemano que seguirán siendo problemáticas para los alineamientos. A mediano plazo, habrá que buscar los mecanismos para que un pronombre pueda alinearse con la palabra a la que representa en un contexto determinado.

El apartado de determinantes (3.7) refuerza la idea de que es necesario agrupar de algún modo las palabras funcionales, pero lo que planteamos ahí abre una pregunta: ¿Qué criterio existe para considerar como par igual el formado por {el, el} o {un, un} y como semi igual {el, un}? Ninguno.

Como se puede observar en el alineamiento 36, la consideración de pares semi iguales y semi nulos incrementa el valor de LCC de un par que no es semántico {especificación, conexión} de 2 a 5, promoviéndolo a vinculado; situación que no es una excepción, sino una consecuencia de que en el algoritmo flexibilizado ese tipo de pares se tomen en cuenta para determinar el coeficiente de similitud de pares correspondientes. Aunque en menor medida, en el algoritmo básico también se dan casos en los que el sistema vincula pares debido a la presencia de pares formados por palabras funcionales; así, en el alineamiento 37 el par {interacción, observación} adquiere un LCC de 5 debido a que colinda a la izquierda con dos pares iguales integrados por preposiciones y artículos respectivamente, lo que se repite a la derecha del par correspondiente.

Por tanto, un par semi igual debe componerse por dos palabras de base gramatical, sean diferentes o la misma. La consecuencia tangible de esto, aplicándose la modificación correspondiente a la asignación de valores para el coeficiente de similitud, será que ningún par formado por palabras funcionales sume puntos al LCC de pares en su entorno inmediato.

Por otra parte, sugerimos ampliar el criterio de pares semi iguales para los pares semánticos que previamente hayan sido identificados por el sistema en otros alineamientos, con lo que contribuirían a incrementar el valor de LCC de nuevos pares correspondientes.

Asimismo, en este capítulo hemos propuesto tratar como semi iguales a los pares formados por dos patrones definitorios y como semi nulos a los que asocian un patrón con un conjunto vacío. En esta última categoría entrarían también los pares nulos que contienen palabras de base léxica catalogadas como sintagmas enfáticos.

Número de alineamiento: 35
Término: Experimento de Michelson y Morley
Fuente de la definición 1: Encarta 2000
Fuente de la definición 2: Diccionario Mc. Graw Hill de Física

Def. 1	**E**	**d**	**M**	**y**	**M**	**Lo**	que	pretendía	detectar	el		experimento	de	Michelson-Morley	era		una	diferencia	en	…
Def. 2	**E**	**d**	**M**	**y**	**M**	**Experimento**	que	utiliza	un	interferómetro	de		Michelson	para		determinar	la	diferencia	de	…
Tipo de par	I	I	I	I	I	**C**	I	C	C	C	C		C	C		C	C	I	C	
LCC	0	0	0	0	0	**7**	0	2	1	1	1		1	1		1	2	0	2	
Par vinculado	no	no	no	no	no	**sí**	no	no	no	no	no		no	no		no	no	no	no	

Número de alineamiento: 36
Término: Causalidad
Fuente de la definición 1: Diccionario de términos científicos
Fuente de la definición 2: Personal

Def. 1	**C**	Principio		en	que	la	**especificación**	de	las	variables	dinámicas	de	un		sistema	en	un	tiempo	dado	…	
Def. 2	**C**	Relación	existente	en		la	**conexión**		que	hay		de	causa-efecto		o	bien				…	
Tipo de par	I	C	N	I	SN	I	**C**		SI	C	N	N	I	C	C	C	SN	N	N			
LCC	0	2	0	0	0	0	**5**		0	2	0	0	0	2	1	2	0	0	0			
Par vinculado	no	no	no	no	no	no	**sí**		no	no	no	no	no	no	no	no	no	no	no			

Número de alineamiento: 37
Término: Física
Fuente de la definición 1: Encarta 2000
Fuente de la definición 2: Física. Fundamentos y aplicaciones

Def. 1	**F**	La	Física	está	estrechamente	relacionada	con	las	demás	ciencias	naturales	y	en	cierto	modo	las	engloba	a	todas	…				
Def. 2	**F**	La	Física	es	una	ciencia	empírica	Todo	lo	que	sabemos	del	mundo	físico	y	de	los	principios	que	…				
Tipo de par	I	**I**	**I**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**					
LCC	0	0	0	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no					

Def. 1	…	la	química	por		ejemplo	se	ocupa			de	la	**interacción**	de	los	átomos		…						
Def. 2	…	rigen	su	comportamiento		ha	sido	aprendido	a	través	de	la	**observación**	de	los	fenómenos								
Tipo de par		C	C	C		C	C	C	N	N	I	I	**C**	I	I	C								
LCC		1	1	1		1	1	1	0	0	0	0	**5**	0	0	3								
Par vinculado		no	no	no		no	no	no	no	no	no	no	**sí**	no	no	no								

3.11.4 Reconsideración para la aplicación de intercambios

En el capítulo 2 dejamos dicho que era necesario averiguar por qué las opciones de intercambio no han incidido en la obtención de pares semánticos como se esperaba (2.3.4); pues bien, basándonos en nuestro análisis podemos decir que esto obedece, una vez más, a que el sistema no considera la gramática de nuestro idioma. Lo explicaremos con ejemplos:

En el alineamiento 38 Clustering ubica como equivalentes a *estudia* y *estudio* cuando se trata de lexemas distintos cuyo origen etimológico es común, pero no así su categoría gramatical; este intercambio (marcado con *swap*) mal aplicado desencadena un mal par vinculado. Si tuvieramos el etiquetado gramatical (POST) para las bases del banco terminológico con que trabaja el programa, esto no sucedería. Las etiquetas, además de ayudar a impedir que se vincularan pares de palabras que no comparten tal categoría, podrían hacer operante cierta información sobre las clases de palabras susceptibles de intercambio según la lógica del español; así, siguiendo el mismo ejemplo, el sistema estaría programado para no realizar intercambios de los que resulten sustantivos antepuestos a su determinante.

En el alineamiento 39 sucede algo muy parecido que en el caso anterior. El error que conduce al intercambio no conjuntivo es que, según esta operación, el sustantivo al que califica el complemento adnominal podría aparecer entre el enlace preposicional y el término del mismo complemento; situación que no se presenta en la lengua. Por tanto, debería existir una restricción que impidiera el intercambio de preposición y sustantivo y facilitara el de secuencias como adjetivo y sustantivo, aún cuando entre ellos apareciera un sintagma enfático (alineamiento 40). Para su instrumentación se puede retomar el mecanismo diseñado para el intercambio conjuntivo.

Sobre todo, el intercambio conjuntivo se vería favorecido con la incorporación de información gramatical pues las situaciones en las que puede aprovecharse son más recurrentes en las definiciones. Las conjunciones coordinantes enlazan elementos al mismo nivel sintagmático[64] y no sólo palabras; las construcciones que coordinan muchas veces comienzan por preposiciones o determinantes. El algoritmo flexibilizado no contempla esto y por el mismo motivo sólo identificamos un alineamiento en donde su aplicación redunda en un par vinculado (2.3.4.1). De efectuarse un intercambio conjuntivo dentro de un par de definiciones como el que se muestra en el alineamiento 41, se haría patente que la conjunción está sirviendo para enlazar los mismos elementos, aunque en orden inverso, y por tanto, los pares son iguales y no correspondientes.

En síntesis, para que el intercambio se aplique con mejores resultados es fundamental el etiquetado de las categorías gramaticales (POST). En esta aplicación, las palabras intercambiadas sí deben contabilizarse para medir la similitud de un par correspondiente cuando se trate de palabras de base léxica.

---

[64] *Vid.* 3.2.1

Número de alineamiento: 38
Término: Mecánica
Fuente de la definición 1: Física Resnick
Fuente de la definición 2: Gran diccionario enciclopédico

Def. 1	**M**	La	más	antigua	de	las	ciencias	físicas	**Es**	**el**	**estudio**	del	movimiento	de	los	cuerpos			
Def. 2	**M**	Parte			de	la		Física	**que**	**estudia**	**el**		movimiento	y	las	fuerzas	capaces	de	producirlo
Tipo de par	I	C	N	N	I	SI	N	I	**C**	**swap**	**swap**	SN	I	SI	SI	C	N	N	N
LCC	0	2	0	0	0	0	0	0	**8**	**0**	**0**	0	0	0	0	6	0	0	0
Par vinculado	no	no	no	no	no	no	no	no	**sí**	**no**	**no**	no	no	no	no	no	no	no	no

Número de alineamiento: 39
Término: Frecuencia
Fuente de la definición 1: CIm1
Fuente de la definición 2: Encarta 2000

Def. 1	**F**			Las	unidades							**de**	**frecuencia**	se	definen	como	revoluciones/segundo						
Def. 2	**F**	En	todas	las	clases	de	movimiento	ondulatorio	la	**frecuencia**	**de**	la	onda	suele	darse			indicando	el	número	…		
Tipo de par	I	N	N	I	C	N	N	N	N	**swap**	**swap**	C	C	C	C			N	N	N			
LCC	0	2	2	O	2	0	0	0	0	**0**	**0**	3	1	1	1			0	0	0			
Par vinculado	no	no	no	no	no	no	no	no	no	**no**	**no**	no	no	no	no			no	no	no			

Número de alineamiento: 40
Término: Choque
Fuente de la definición 1: Cinemática y dinámica básica
Fuente de la definición 2: Física Resnick

Def. 1	**Ch**	Es	un	proceso	dinámico	que	se	caracteriza	por	la	presencia	de	**grandes**	**fuerzas**	durante	un	intervalo	de	tiempo	…
Def. 2	**Ch**	Es	el	impacto	entre	dos	cuerpos	o	partículas	donde	obra	una		**fuerza**	relativamente	**grande**	sobre	cada	una	…
Tipo de par	I	I	C	C	C	C	C	C	C	C	C	C	N	I	C	C	C	C	C		
LCC	0	0	1	1	1	1	1	1	1	1	1	1	0	0	2	1	1	1	1		
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no		

Número de alineamiento: 41
Término: Mecánica
Fuente de la definición 1: DRAE 92
Fuente de la definición 2: María Moliner

Def. 1	**M**	Parte	de	la	física	que	trata	del	**equilibrio**	y	del	**movimiento**		de	los	cuerpos	sometidos	a	cualesquiera	fuerzas	
Def. 2	**M**	Parte	de	la	física	que	trata	del	**movimiento**	y	el	**equilibrio**	y	de	las	fuerzas	que		los	producen	
Tipo de par	I	I	I	I	I	I	I	I	C	I	C	C	N	I	C	C	C	C	C	N	
LCC	0	0	0	0	0	0	0	0	10	0	2	1	0	0	2	1	1	1	1	0	
Par vinculado	no	no	no	no	no	no	no	no	sí	no	no	no	no	no	no	no	no	no	no	no	

3.12. Conclusiones

La primera conclusión que se desprende del estudio que registramos en este capítulo es que para analizar definiciones lingüísticamente debe entenderse que se está trabajando con una porción de lengua que en sí comprende un lenguaje con códigos específicos. Por tanto, es necesario conocer las características que le son propias: ¿Cómo se componen los términos, conceptos y definiciones?

En relación con la anterior, los sintagmas que forman las definiciones no se corresponden con las secuencias sintácticas identificables en la oración. Según hemos visto, una definición puede constituirse por una o varias oraciones, pero también puede tratarse de una sencilla frase. Esto nos ha obligado a rebasar los límites de la sintaxis y a aventurar teorías fuera del ámbito oracional y fuera también de los alcances de esta tesis.

Con base en los exámenes aplicados a los alineamientos, consideramos que la agrupación separada de las palabras funcionales en conjunciones, preposiciones y determinantes, así como el tratamiento propuesto para las siglas y la unión de construcciones perifrásticas (incluyendo los patrones definitorios) son fácilmente sistematizables. Esto porque las palabras de base gramatical pertenecen a un paradigma cerrado. Con las siglas se observa algo similar, pues aunque la lista es mucho más extensa, caben pocas situaciones de ambigüedad. Por su parte, las perífrasis verbales se disponen en torno a fórmulas estables y sistemáticas que ya han sido desentrañadas[65]; las locuciones de uso frecuente aparecen registradas en gramáticas y los términos compuestos se encuentran igualmente documentados. Consideramos, por tanto, que éstas son las primeras modificaciones que deben integrarse al programa.

Dejamos para un segundo momento la modificación que vincula las preposiciones que cumplen la función de término de un suplemento con el verbo que rige a cada una, pues aunque estas construcciones son fácilmente reconocibles, para su localización automática, el sistema requiere de información adicional más elaborada que para el caso de las líneas anteriores. Por otra parte, es posible acelerar la materialización de esta propuesta ya que los verbos que rigen preposición cuentan con la ventaja de formar un inventario reducido con cuyas ocurrencias más frecuentes bien puede incorporarse una lista al sistema (aunque esta opción no es la ideal, pues se trata de una enumeración de las estructuras fuera de contexto, en donde se elimina la información asociada a su funcionamiento).

Posteriormente, debe entrar el intercambio de adjetivos por frases prepositivas y oraciones adjetivas porque, según lo visto, dicho proceso no parece oponer muchas dificultades para la sistematización, además de que por la alta incidencia del fenómeno sería positivo para el programa intentar un esfuerzo en este sentido. Al desarrollarse los mecanismos para la identificación de

65 *Vid*. 3.1.3

estructuras equivalentes a un adjetivo se tendrían las bases para hacer lo propio con los adverbios.

Para un plazo largo dejamos los sintagmas enfáticos, las conjunciones y las diferentes formas de negación, pues su integración a un sistema recuperador de información con los criterios que aquí se sugieren requiere de un estudio exhaustivo en cuanto a sus relaciones semánticas.

3.13. Recapitulación

Este capítulo es resultado, principalmente, de la observación cuidadosa de los alineamientos que componen nuestro corpus con la debida reflexión sobre la lengua a partir de conocimientos gramaticales. Esta situación nos llevó a definir algunos aspectos que consideramos interesantes para abordarse desde una perspectiva sintáctica. Con este interés nos acercamos a textos de apoyo y encontramos información que nos fue muy útil para delimitar los temas que son los apartados de este capítulo: *Perífrasis gramaticales, nexos, adjetivación, adverbios, negación, sintagmas enfáticos, determinantes, relaciones léxicas, abreviaturas y símbolos*, y *patrones definitorios*; en algunos de estos temas se engloban varios subapartados. En cada inciso se describe lo que significa la línea de investigación, se proporcionan algunos ejemplos que la ilustran y se hace una propuesta para su tratamiento en posteriores versiones del programa. En el Apéndice C se proporciona una síntesis de las propuestas vertidas en este capítulo.

Finalmente, resta señalar que la mejor vía para evaluar la operatividad de todas estas propuestas es su traducción en algoritmos y su aplicación al sistema, más allá de de realizar especulaciones sin sustento práctico.

El siguiente capítulo mostrará los resultados de un pequeño experimento. Consiste en correr Clustering con las definiciones de nuestro corpus, que han sido modificadas manualmente atendiendo a algunos criterios enunciados en el presente capítulo, lo cual servirá como una prueba piloto para confirmar o rechazar nuestras hipótesis en cuanto a sus ventajas.

4. EJERCICIO

Como describimos en el capítulo 2, con la revisión humana de nuestro corpus fueron identificados pares semánticos que Clustering no considera vinculados. En los alineamientos que contienen dichos pares semánticos, además de los pares, encontramos patrones sintácticos que se repiten, por lo que consideramos que al incorporar al sistema la información que facilite su correcto tratamiento, se eliminarán errores y omisiones en la obtención de pares semánticos.

El presente capítulo describirá el procedimiento seguido para la aplicación de una serie de pruebas al sistema que incorporaron las propuestas de modificaciones hechas en el capítulo anterior; asimismo, registrará los resultados obtenidos.

4.1. Descripción del experimento

Con base en las conclusiones del Capítulo 3, se definieron las modificaciones a aplicarse en esta prueba piloto, cuya incorporación al sistema, según nuestro análisis, resulta viable en un corto plazo. Los ejes que contienen estas modificaciones son: *separación de contracciones, tratamiento independiente de las palabras funcionales*, *tratamiento de verbos* y *tratamiento de términos compuestos, siglas y símbolos.*

Se realizaron cinco pruebas, una por cada línea mencionada y otra que combina todas las modificaciones. El mecanismo seguido para este experimento fue realizar los cambios pertinentes a las definiciones que componen el corpus que se ha venido trabajando en esta investigación, y una vez con las definiciones marcadas, se procedió a correr el algoritmo básico, tal como se hizo con cada una de las 32 pruebas mencionadas en el capítulo 2.

Sin embargo, hay que destacar que estos cambios se realizaron en forma semiautomática y que no se aplicó ninguna modificación al sistema. Ahora describiremos como se llevó a cabo cada operación.

4.1.1. Separación de contracciones

La posibilidad de realizar la separación de palabras que son contracción de otras dos con diferente categoría gramatical, es decir *del* y *al*, fue prevista con la finalidad de que preposición y determinante pudieran formar en los alineamientos pares independientes. Su incidencia en el corpus es muy alta por

ser contracciones de uso obligado en español; la separación puede realizarse automáticamente con el empleo de las herramientas **buscar** y **reemplazar** en un archivo de texto que contenga el corpus.

Para poder determinar mejor si esta sencilla operación tiene repercusiones en la obtención de pares semánticos, decidimos manejar esta modificación como una prueba independiente de la relativa a las abreviaturas.

4.1.2. Tratamiento de las palabras funcionales

Actualmente, la opción de pares semi iguales se sustenta en el concepto de palabras irrelevantes, lo que implica homologar arbitrariamente determinantes, conjunciones y preposiciones (sea cual fuere su función en una estructura oracional). En este experimento consideraremos pares iguales aquellos formados por dos palabras de estas categorías, siempre y cuando se tratara de la misma clase y desempeñaran una función equivalente.

Identificamos las preposiciones simplemente como tales, con la finalidad de que cada una de ellas pudiera vincularse con cualquier otra preposición. Para esto, se sustituyeron las preposiciones en sentido estricto, así como las locuciones que desempeñan la misma función conectora, por una etiqueta que las identifica como elementos de una misma clase.

En lo que se refiere a las conjunciones, tal como se mencionó en el apartado correspondiente (3.2.1), nuestra propuesta es identificar así —por el momento— sólo a las copulativas y disyuntivas. Las conjunciones copulativas fueron reemplazadas por una etiqueta y las disyuntivas por otra. Las locuciones conjuntivas que no se correspondieron con estas subclases, simplemente quedaron unidas mediante guiones por ser una unidad gramatical, pero no se identificaron como conjunciones.

Además, etiquetamos por separado artículos y pronombres. Nuestro corpus quedó marcado de la siguiente forma:

- Con la etiqueta PREP:
  - a
  - bajo
  - con
  - contra
  - de
  - desde
  - en
  - entre
  - hacia
  - hasta
  - para
  - por

- según
- sin
- sobre
- alrededor de
- además de
- de parte de
- debido a
- dentro de
- gracias a
- por encima de
- por razón de
- sobre todo

- Con la etiqueta CONJC: y[1], e, ni.

- Con la etiqueta CONJD: o, u, o bien, ya sea, ya.

- Locuciones conjuntivas que se unieron simplemente mediante guiones por no ser ni copulativas ni disyuntivas:

  - en relación con
  - o sea
  - por lo tanto
  - tal que
  - ya que

- Con la etiqueta ART:

  - Los artículos definidos: el, la, lo, los, las.
  - Los artículos indefinidos un, una, unas.
  - Los adjetivos demostrativos este, esta, estos, estas, ese, esa, esos, aquel, aquella, aquellos, dicho y tal.
  - Los posesivos mi, nuestro, nuestros y nuestras, su, sus.

- Con la etiqueta PRON:

  - aquél
  - aquélla
  - aquéllas
  - aquello
  - él
  - ella
  - ellas
  - ellos
  - ésta
  - éstas

1 Este símbolo puede significar una conjunción o una variable dependiente.

- o éste
- o esto
- o suyo

4.1.3. Tratamiento de verbos compuestos

En una tercera prueba decidimos concentrar las modificaciones que tienen que ver con el tratamiento de verbos. Por una parte, cuando éstos se expresan con más de una palabra ortográfica: los tiempos compuestos de la conjugación, los verbos pronominales en donde el pronombre *se* aparece en posición proclítica y las perífrasis verbales; por otra, los verbos que rigen suplemento.

En este sentido, en el capítulo anterior propusimos incorporar a Clustering los mecanismos para reconocer los tiempos compuestos y los verbos pronominales automáticamente; para nuestra prueba los identificamos nosotros mismos y los unimos mediante guiones; de esta manera el sistema los considera una sola palabra. El *se* que es marca de una oración pasiva refleja se unió de la misma forma al núcleo verbal. A manera de ejemplo incluimos las siguientes definiciones marcadas:

- **Dimensión:** Es el nombre que se_da a una cantidad física que ha_sido_seleccionada para formar la estructura fundamental de un sistema de unidades.[2]

- **Ley de la conservación de la energía:** Principio fundamental de la física que afirma que la energía no se_crea ni se_destruye, solamente se_transforma de unas formas en otras. Esto quiere decir que la energía total de un sistema físico cerrado es siempre la misma.[3]

Sobre las perífrasis, habíamos adelantado que su identificación sería más complicada. Como no existe ninguna diferencia formal que indique cuándo la unión de un verbo conjugado y una forma no personal implica una perífrasis y cuándo no la forma (3.1.3) cualquier sistematización en este sentido se dificulta. Por ello, la localización se realizó caso por caso, tomando por perífrasis las secuencias verbales que se refieren a un solo proceso o estado. Incluimos un par de definiciones tal como fueron modificadas:

- **Fuerza:** Acción entre dos cuerpos que cambia o tiende_a_cambiar cualquier relación física entre ambos.[4]

- **Aceleración centrípeta:** Es lo mismo que aceleración normal, pero por lo general se usa cuando un punto se_mueve_describiendo una circunferencia y por lo tanto su aceleración está dirigida hacia el centro de la misma.[5]

---

2 Antecedentes de Física.
3 Física de emergencia.
4 Vox.
5 Diccionario enciclopédico de Física.

Al momento del etiquetado nos encontramos con la dificultad para marcar algunas perífrasis en donde aparece un adverbio entre las palabras que las forman; tal situación ya había sido advertida durante el análisis sintáctico de los alineamientos. Para esta prueba no realizamos ningún tipo de marcación y las perífrasis en dicha condición no fueron consideradas como tales.

Por otra parte, localizamos manualmente los verbos que rigen preposición y, por ser obligada su asociación (3.2.2.1), los unimos mediante guiones. Las estructuras enlazadas fueron:

- depender de
- ocuparse de (se_ocupa_de)
- referirse a (se_refiere_a)
- relacionarse con (se_relaciona_con)
- resultar de
- tratar de

En este punto, al hacer un experimento nos encontramos con una dificultad para marcar correctamente muchas de las ocurrencias de verbos que rigen preposición cuando ésta es *de* y forma parte de una contracción. Optamos, entonces, por repetir la operación de la prueba 1 y dividimos las contracciones.

4.1.4. Términos compuestos, siglas y símbolos

Para esta prueba se realizaron las modificaciones propuestas en torno a dos líneas del capítulo anterior: *términos compuestos* (3.1.2) y *abreviaturas, siglas y símbolos* (3.9).

Con relación al corpus que hemos analizado, resultaba particularmente necesario aportar iniciativas concernientes al tratamiento de símbolos y siglas, ya que las definiciones en Física muchas veces los contienen. En este sentido, el tipo de prueba escigida nos permitió aplicar la propuesta de considerar como una unidad las formas sintéticas y las desarrolladas cuando aparecen juntas en una definición, lo cual se hizo apoyándonos en los guiones.

Por otra parte, instrumentamos la propuesta de tratar las fórmulas en su conjunto como una sola expresión.

Los términos compuestos que habían sido identificados, con la ayuda de la especialista, durante el análisis manual del corpus se unieron, también con guiones, dentro de los alineamientos.

Como ejemplo de las marcaciones realizadas para esta prueba incluimos una definición para el término *energía potencial gravitacional*:

> La Ep_gravitacional de un cuerpo de masa_m y que está a una altura_h por encima de un determinado nivel de referencia es Epg_=_Ep_=mgh

donde g es la aceleración de la gravedad. En términos del peso_w del cuerpo Ep=wh[6]

Con relación a los términos compuestos nos encontramos con una dificultad que obstaculizó su vinculación. ¿Qué hacer cuándo dos términos se combinan en una estructura oracional?

Así, en la definición asociada a Galileo tenemos el enunciado “descubrió las leyes de la caída de los cuerpos y del movimiento de los proyectiles”[7] donde al sustantivo **leyes** lo determinan, por separado, dos complementos adnominales que contienen parte de un término que se inicia con *ley*.

La situación fue muy recurrente al momento de marcar el corpus para las pruebas y evitó la identificación automática de pares semánticos en los alineamientos resultantes, así que debe dársele prioridad en posteriores modificaciones al sistema. Su tratamiento puede generalizarse: ¿Qué hacer cuando en un sintagma existe más de un modificador por cada núcleo? El problema quedó esbozado en el apartado de conjunciones del capítulo 3.

4.2. Resultados

El algoritmo se corrió cinco veces, una por cada una de las bases de texto resultantes de las modificaciones descritas en el apartado anterior; la última prueba consistió en la aplicación simultánea de todas las modificaciones. Revisamos los alineamientos y los pares vinculados que se generaron con cada prueba y confrontamos las corridas con el desempeño del algoritmo original. Comparando los pares generados con los que habíamos identificado manualmente al principio de la investigación, establecimos los valores de *precision* y *recall* para cada prueba. En ambos casos hicimos un cálculo que contempla sólo los pares simples ($precision_1$ y $recall_1$) y otro que abarca el total de los pares semánticos ($precision_2$ y $recall_2$).

Los resultados numéricos de las primeras cuatro pruebas de nuestro experimento, así como los de la que corre el algoritmo original sin modificaciones, con los consecuentes valores de *precision* y *recall* para cada caso, se registran en la siguiente tabla:

6 Física Aplicada.
7 Encarta 98.

Corrida	Algoritmo básico	Prueba 1	Prueba 2	Prueba 3	Prueba 4
Pares generados	53	55	205	63	56
Pares simples generados	53	55	205	52	52
Pares semánticos acertados	30	30	42	38	34
Pares semánticos Simples acertados	30	30	42	29	30
Pares semánticos identificados manualmente	364	364	364	364	364
Pares semánticos simples identificados manualmente	167	167	167	167	167
***Precision*$_1$**	0.5660	0.5455	0.2049	0.5577	0.5769
***Precision*$_2$**	0.5660	0.5455	0.2049	0.6032	0.6071
***Recall*$_1$**	0.1796	0.1796	0.2515	0.1737	0.1796
***Recall*$_2$**	0.0824	0.0824	0.1154	0.1044	0.0934

Tabla 7: Resultados cuantitativos de las pruebas y valores de *precision* y *recall* provenientes de correr el algoritmo básico con y sin modificaciones a las definiciones.

Con la finalidad de ilustrar mejor la relación entre los valores de *precision* y *recall* para los pares semánticos y los pares semánticos simples incluimos la gráfica siguiente:

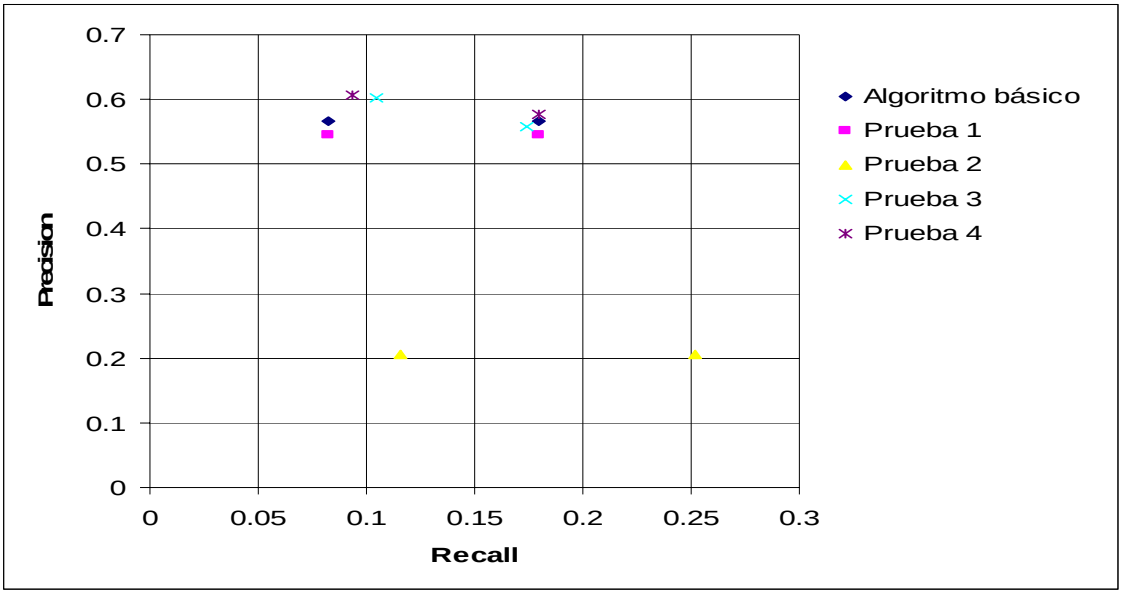

Figura 3: Índices de *recall* y *precision*, correspondientes a los pares semánticos simples y al total los pares semánticos, para el algoritmo básico y las pruebas de aplicación de nuestras propuestas.

Como puede observarse, con la prueba 2 se incrementa el número de pares semánticos identificados, pero también se recupera mucha basura. Esto se explica por una deficiencia en la aplicación de dicha prueba, debido a que no hicimos modificaciones al sistema y, por tanto, las transformaciones que llevamos a cabo en el corpus fueron insuficientes para reproducir la propuesta vertida en el capítulo anterior (3.7). En la descripción de la prueba (4.1.2) decíamos que marcamos las palabras de base gramatical con cinco tipos de etiquetas con la finalidad de que, en un alineamiento, sólo pudieran formar par las pertenecientes a una misma clase; sin embargo, no consideramos que había que hacer cambios en la configuración del programa para que la restricción se hiciera efectiva.

Si bien, como esperábamos, pares formados por preposiciones diferentes se unieron en un par igual y, por otro lado, dos artículos diferentes formaron un mismo par igual, gracias a las etiquetas PREP y ART, también aparecieron pares correspondientes del tipo {PRON, experimento}, {PREP, en_relación_con}, y otros más constantes como {PREP, ART}, {CONJC, CONJD}, {ART, CONJC}, los cuales alteraron todos los alineamientos. Además, al registrarse como iguales, los pares formados por palabras funcionales, sumaron puntos al coeficiente de similitud de pares correspondientes. Ambas situaciones impidieron que mejorararan los resultados obtenidos con la opción de semi iguales del algoritmo flexibilizado y probar realmente nuestra propuesta.

El motivo por el que en la tabla no aparecen registrados los resultados obtenidos con la prueba 5, que corresponde a la aplicación simultánea de las modificaciones contenidas en las otras cuatro propuestas, es que no sirve como indicador para determinar la eficacia combinada de nuestras iniciativas ya que reproduce los errores de la prueba 2.

En el siguiente apartado estudiaremos los resultados obtenidos con las tres pruebas que arrojaron datos analizables.

4.3. Análisis

Aquí explicaremos cómo funcionaron en nuestro corpus las modificaciones contenidas en las pruebas 1, 3 y 4.

4.3.1. Prueba 1

Esta prueba consistió en la separación de contracciones. Como lo muestra la tabla, con esta modificación se obtuvieron dos pares más que con el algoritmo original, {combinación, substitución} y {valor, cuadrado}; en ambos casos se debe a la separación de *del* en *de* y *el.* Lamentablemente, los nuevos pares no son semánticos, por lo que disminuyó el índice de *precision.* Veamos dichos pares en el contexto de los alineamientos que se desprenden del algoritmo

original y en el de los alineamientos que produce la operación de separar las contracciones al interior de las definiciones:

Número de alineamiento: 42
Término: Éter
Fuente de la definición 1: Master. Diccionario enciclopédico
Fuente de la definición 2: Enciclopedia internacional

Def. 1	**É**				Compuesto	químico	orgánico				que	resulta	de	la	**combinación**					de	un			…
Def. 2	**É**	Cualquiera	de	los	compuestos	químicos	gases	líquidos	o	sólidos	que	resultan	de	la	**substitución**	del	átomo	de	hidrógeno	de	un	hidroxilo	por	…
Tipo de par	I	N	N	N	I	I	C	N	N	N	I	I	I	I	**C**	N	N	N	N	I	I	N	N	
LCC	0	0	0	0	0	0	3	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	**no**	no	no	no	no	no	no	no	no	

Número de alineamiento: 43
Término: Éter
Fuente de la definición 1: Master. Diccionario enciclopédico
Fuente de la definición 2: Enciclopedia internacional

Def. 1	**É**				Compuesto	químico	orgánico				que	resulta	de	la	**combinación**	de						un			…
Def. 2	**É**	Cualquiera	de	los	compuestos	químicos	gases	líquidos	o	sólidos	que	resultan	de	la	**substitución**	de	el	átomo	de	hidrógeno	de	un	hidroxilo	por	…
Tipo de par	I	N	N	N	I	I	C	N	N	N	I	I	I	I	**C**	I	N	N	N	N	N	I	N	N	
LCC	0	0	0	0	0	0	3	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	**sí**	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 44
Término: Órbita elíptica
Fuente de la definición 1: Ingeniería mecánica. Dinámica
Fuente de la definición 2: Science Study Comitte

Def. 1	**Ó**	**e**	Tipo	de	trayectoria	que	toma	un	satélite	se	determina	a	partir							del	**valor**	de	la	excentricidad	de	…
Def. 2	**Ó**	**e**	Siempre	que	una	masa	se	mueve	bajo	la	influencia	de	una	fuerza	que	varía	con	la	inversa	del	**cuadrado**	de	la	separación	de	un	…
Tipo de par	I	I	C	C	C	C	C	C	C	C	C	C	C	N	N	N	N	N	N	I	**C**	I	I	C	I	N					
LCC	0	0	3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	**4**	0	0	4	0	0					
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	**no**	no	no	no	no	no	no				

Número de alineamiento: 45
Término: Órbita elíptica
Fuente de la definición 1: Ingeniería mecánica. Dinámica
Fuente de la definición 2: Science Study Comitte

Def. 1	**Ó**	**e**	Tipo	de	trayectoria	que	toma	un	satélite	se	determina	a	partir							de	el	**valor**	de	la	excentricidad	de	…
Def. 2	**Ó**	**e**	Siempre	que	una	masa	se	mueve	bajo	la	influencia	de	una	fuerza	que	varía	con	la	inversa	de	el	**cuadrado**	de	la	separación	de	un	…
Tipo de par	I	I	C	C	C	C	C	C	C	C	C	C	C	N	N	N	N	N	N	I	I	**C**	I	I	C	I	N				
LCC	0	0	3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	**5**	0	0	4	0	0				
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	**sí**	no	no	no	no	no	no			

Los alineamientos 42 y 43 comparan las mismas dos definiciones; el primero se obtuvo con el algoritmo básico, mientras que el segundo es resultado de correr Clustering con el mismo algoritmo pero después de haber aplicado al corpus las modificaciones correspondientes a la prueba 1. Como puede observarse, la manipulación de las definiciones incidió en la obtención de un par vinculado {combinación, substitución}, el cual no es considerado así por la versión original del programa, a pesar de tener un LCC = 5, ya que no cumple con la condición de frontera que indica que todo par correspondiente debe tener al menos un par igual a la izquierda y uno a la derecha para ser promovido a vinculado.

La separación de *del* en *de* y *el* (alineamiento 43) permitió que se alinearan estas dos palabras de manera independiente, así *de* pudo formar parte de un par igual, lo que permitió que el par en estudio alcanzara la condición de frontera (además de sumar un punto a su coeficiente de similitud) para ser promovido a vinculado.

La relación entre el alineamiento 44 y 45 es equivalente a la que describimos para el 42 y 43. La diferencia que hizo que el alineamiento 45 el par {valor, cuadrado} se promoviera a vinculado responde a la misma operación. En este caso la separación de *del* dio como resultado que lo que era un par igual se dividiera en dos, aportando de esta manera la unidad para que el par correspondiente alcanzara el LCC de 5 que lo convierte en vinculado.

Aquí hay que destacar que con la aplicación del conjunto de las modificaciones descritas en el capítulo anterior estos falsos pares semánticos no habrían alcanzado la categoría de vinculados, pues la propuesta contempla que las palabras funcionales no deben ser consideradas para el cómputo de LCC de pares correspondientes (3.11.3). Con un cambio al mecanismo del sistema el coeficiente de similitud de {combinación, substitución} desendería a 2 y el de {valor, cuadrado} a 1.

A diferencia de lo que pensábamos, esta operación no tiene repercusiones en la obtención de pares semánticos.

4.3.2. Prueba 3

En lo que se refiere al tratamiento de perífrasis verbales, el balance es positivo. Aunque los índices de *precision* y *recall* para los pares simples disminuyeron en comparación con los obtenidos con el algoritmo original, la operación generó nueve pares semánticos compuestos.

El reordenamiento de los alineamientos que derivó de la consideración de las formas perifrásticas no produjo nuevos pares semánticos simples, pero sí eliminó dos que no lo son y que el resto de las pruebas recuperan como vinculados, {interacción, observación} y {producen, cambia}.

Abajo, reproducimos el alinemaiento que contiene este último par para las versiones del algoritmo básico y flexibilizado (alineamiento 46) y el que se obtiene con esta prueba (alineamiento 47):

Número de alineamiento: 46
Término: Fuerza
Fuente de la definición 1: Gran enciclopedia del mundo
Fuente de la definición 2: Vox

Def. 1	**F**	Nombre	genérico	que	se	aplica	a	todos	aquellos	agentes	que	**producen**	o	tienden	a	producir	una	variación	de	la	forma	o	las	dimensiones	…
Def. 2	**F**	Acción	entre	dos	cuerpos						que	**cambia**	o	tiende	a	cambiar	cualquier	relación	física	entre	ambos				
Tipo de par	I	C	C	C	C	N	N	N	N	N	I	**C**	I	I	I	C	C	C	C	C	C	N	N	N	
LCC	0	2	1	1	1	0	0	0	0	0	0	**5**	0	0	0	4	1	1	1	1	1	0	0	0	
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	**sí**	no	no	no	no	no	no	no	no	no	no	no	no	

Número de alineamiento: 47
Término: Fuerza
Fuente de la definición 1: Gran enciclopedia del mundo
Fuente de la definición 2: Vox

Def. 1	**F**	Nombre	genérico	que	se	aplica	a	todos	aquellos	agentes	que	**producen**	o	tienden_a_producir	una	variación	de	la	forma	o	las	dimensiones	…		
Def. 2	**F**	Acción	entre	dos	cuerpos						que	**cambia**	o	tiende_a_cambiar	cualquier	relación	física	entre	ambos						
Tipo de par	I	C	C	C	C	N	N	N	N	N	I	**C**	I	C	C	C	C	C	C	N	N	N			
LCC	0	2	1	1	1	0	0	0	0	0	0	**3**	0	2	1	1	1	1	1	0	0	0			
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	**no**	no	no	no	no	no	no	no	no	no	no			

Como se observa, la disminución del LCC del par {producen, cambia} de 5 a 3 depende plenamente del hecho de considerar perífrasis como verbos. Actualmente Clustering contabiliza como palabra cada segmento de una perífrasis (tal como sucede con “tienden / a / producir” y “tiende / a / cambiar”), lo que incrementa exageradamente el coeficiente de similitud de pares correspondientes cercanos. La operación que proponemos corrige esta situación, pues considerar expresiones perifrásticas como unidades léxicas no sólo sirve para obtener automáticamente pares semánticos formados por expresiones de este tipo, sino que también ubica a dicha clase de pares en la posición que les corresponde para el cómputo de LCC en el contexto de los alineamientos.

Por otra parte, el error que determinó la disminución de *recall* para los pares simples en esta prueba se debe a que el sistema está dejando de reconocer un par semántico incluido en una perífrasis, {impelido, obligado}. El problema ya había sido advertido cuando abordamos el tema en el capítulo 3 (alineamiento 8). Para subsanar esta deficiencia consideramos entonces que deben implementarse los mecanismos para que una vez identificados los pares semánticos el programa repita la búsqueda, pero esta vez realizando alineamientos al interior de las perífrasis.

Como en ésta se incorporaron las modificaciones llevadas a cabo en la primera prueba, *precision* disminuyó (para los pares simples) debido a los dos errores de dicha prueba; sin embargo, aquí cabe hacer una anotación importante: gracias a la separación de contracciones se identificaron los pares {trata de, estudia}, {estudia, se ocupa de}, {trata de, se ocupa de} que ninguna de las otras pruebas ni los algoritmos básico y flexibilizado reconocen, así sea como {trata, estudia}, ….

Los siguientes ejemplos muestran dos alineamientos para un mismo par de definiciones que involucran uno de estos casos. En el primero de ellos (alineamiento 48) el par {trata, estudia}, con un LCC de 7 y que a simple vista es semántico, no es promovido a par vinculado por no cumplir con la condición de frontera de terner un par igual inmediatamente después del par correspondiente (además de uno anterior). Esto mismo sucede con todas las ocurrencias del par en cualquier versión del programa porque el *de* que sigue a *trata* no puede verse correspondido con la misma preposición cuando la forma con que se alinea el verbo es *estudia.* Al tomar *de* como un elemento regido (y formante) del verbo, el par correspondiente es entonces {trata_de, estudia} y los pares que le siguen son iguales, con lo que se alcanza la condición de frontera y el par se promueve a vinculado (alineamiento 49).

La aplicación de está prueba, por sí misma y especialmente en combinación con la anterior, repercutió favorablemente a la localización de nuevos pares. Entonces, la operación de dividir contracciones no puede ser descartada para los propósitos de Clustering. Sobre los pares erróneos que se generaron con la prueba 1 ya hemos dicho que quedarán eliminados una vez que se apliquen las modificaciones referentes al cómputo de LCC.

Número de alineamiento: 48
Término: Mecánica
Fuente de la definición 1: María Moliner
Fuente de la definición 2: Gran diccionario enciclopédico

Def. 1	**M**	Parte	de	la	física	que	**trata**	**del**	movimiento	y	el	equilibrio	y	de	las	fuerzas	que	los	producen
Def. 2	**M**	Parte	de	la	física	que	**estudia**	el	movimiento	y					las	fuerzas	capaces	de	producirlo
Tipo de par	I	I	I	I	I	I	C	C	I	I	N	N	N	N	I	I	C	C	I
LCC	0	0	0	0	0	0	7	3	0	0	0	0	0	0	0	0	3	2	0
Par vinculado	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Número de alineamiento: 49
Término: Mecánica
Fuente de la definición 1: María Moliner
Fuente de la definición 2: Gran diccionario enciclopédico

Def. 1	**M**	Parte	de	la	física	que	**trata_de**	el	movimiento	y	el	equilibrio	y	de	las	fuerzas	que	los	producen
Def. 2	**M**	Parte	de	la	física	que	**estudia**	el	movimiento	y					las	fuerzas	capaces	de	producirlo
Tipo de par	I	I	I	I	I	I	C	I	I	I	N	N	N	N	I	I	C	C	I
LCC	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	3	2	0
Par vinculado	no	no	no	no	no	no	sí	no	no	no	no	no	no	no	no	no	no	no	no

4.3.3. Prueba 4

Esta prueba se enfocó principalmente en la unión de términos compuestos. Con ella se obtuvieron los mismos 30 pares vinculados que con el algoritmo básico y cuatro más, todos compuestos. Aunque modestamente, esta aplicación ha repercutido en la obtención de nuevos pares.

La operación de vincular —con guiones— los símbolos al término al que hacen referencia no mostró ninguna incidencia favorable a nuestro propósito. Probablemente esto se deba a que la propuesta relacionada con siglas y símbolos no se aplicó por completo debido a las limitaciones del tipo de etiquetado que llevamos a cabo.

Por otra parte, para todas las corridas de Clustering analizadas se obtuvieron dos pares de fórmulas. En esta prueba las expresiones generadas fueron diferentes a las que se obtienen al correr el algoritmo básico y las opciones del flexibilizado; éstas son efectivamente fórmulas completas —puesto que fueron etiquetadas para recibir un tratamiento unitario— y sí se corresponden con equivalencias; no obstante, las registramos como error puesto que el interés del programa no es obtener pares de expresiones matemáticas, sino de palabras. En consecuencia, sugerimos que se establezca un filtro que evite que se reporten pares vinculados de este tipo.

4.4. Evaluación

De la observación de las tres pruebas que arrojaron resultados analizables podemos concluir que la opción que redunda en un incremento importante en la identificación de pares semánticos es la que corresponde al tratamiento de verbos compuestos como unidades léxicas; mientras que la separación de las contracciones en sus morfemas originales no aporta por sí misma resultados satisfactorios, sino en combinación con otras modificaciones a las definiciones, al menos con la anterior, por lo que no se puede catalogar como superflua. En lo que se refiere a términos compuestos se registran mejoras, pero insuficientes.

Por otra parte, la prueba relacionada con las palabras funcionales falló y en consecuencia la prueba final, que involucraba todas las modificaciones, por lo que no tenemos un referente de lo que podría ocurrir con la aplicación simultánea de las modificaciones exitosas.

Enseguida, mostramos las gráficas que contrastan los parámetros de evaluación referidos a la productividad de Clustering para el algoritmo básico de agrupamiento semántico, para las pruebas al algoritmo flexibilizado que se registran en el capítulo 2 y para las pruebas descritas en el presente capítulo:

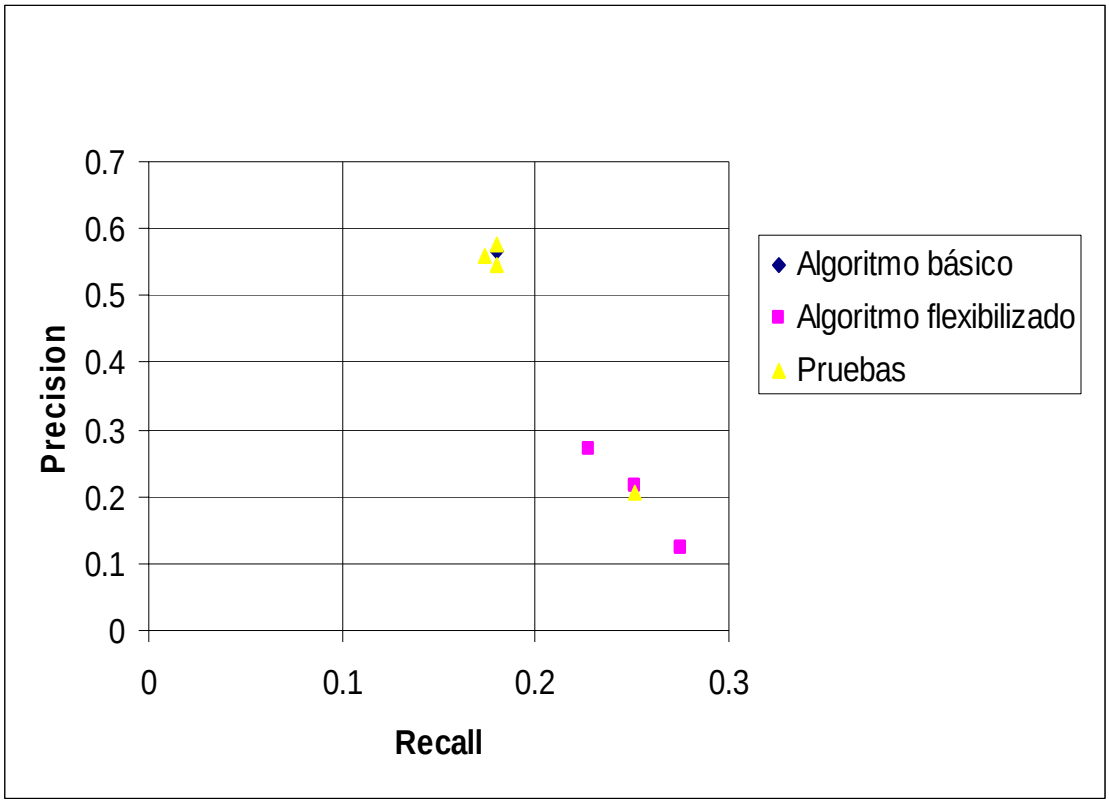

Figura 4: R*ecall* y *precision* correspondientes a las dos versiones del algoritmo y a las pruebas que aplican las propuestas de esta tesis.

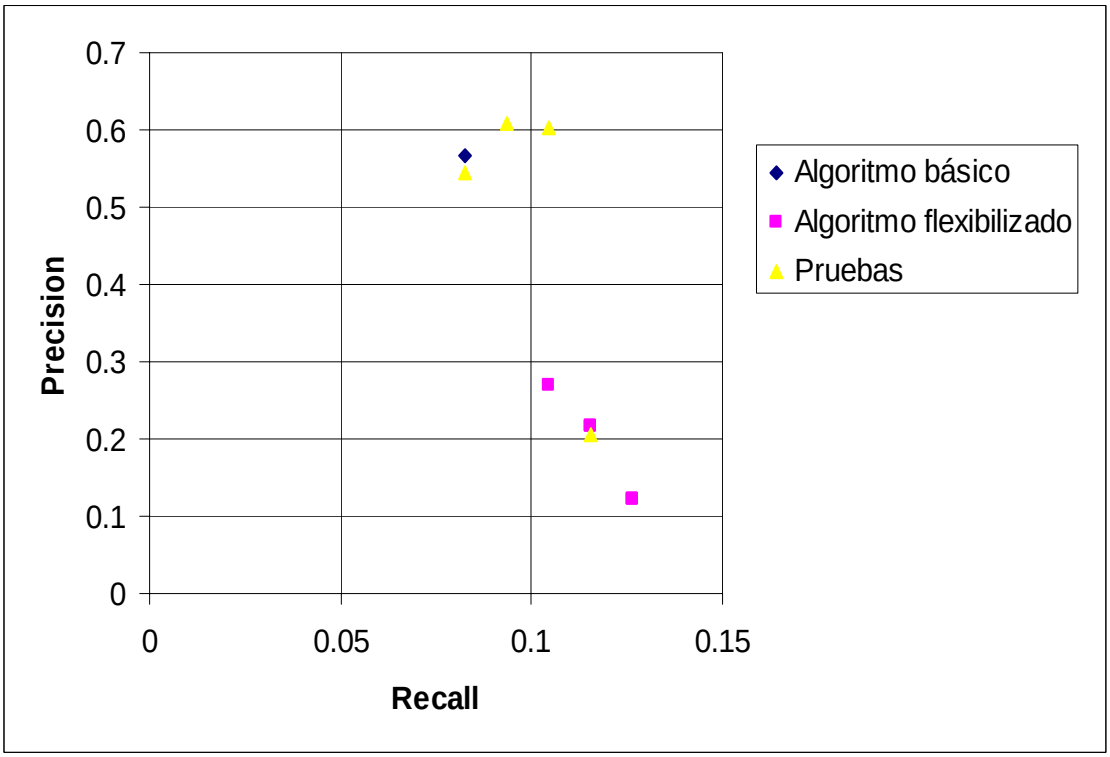

***Figura 5: Recall* y *precision* correspondientes a los pares semánticos (simples y compuestos) para las dos versiones del algoritmo y a las pruebas que aplican las propuestas de esta tesis.**

Las pruebas 1, 3 y 4 mejoran los resultados obtenidos con las pruebas que combinan los parámetros del algoritmo flexibilizado y se localizan en torno a los valores de recuperación del algoritmo básico. El punto que se observa entre los índices de las pruebas al algoritmo flexibilizado corresponde a nuestra prueba 2; en este caso la recuperación de pares semánticos disminuyó drásticamente porque no fue aplicada correctamente, conforme hemos explicado.

La diferencia positiva más marcada de *precision* y *recall* de las pruebas correspondientes a este capítulo con respecto a las otras corridas de Clustering es la que corresponde a la obtención absoluta de pares semánticos (simples o compuestos), debido a que sólo nuestras modificaciones consideran la recuperación de pares semánticos compuestos.

Finalmente, el incremento numérico en la obtención de pares semánticos con las pruebas realizadas no se dio en las dimensiones previstas; sin embargo, a partir de las dificultades que encontramos para reproducir cada propuesta, consideramos que este hecho no obedece a un mal planteamiento de las modificaciones sino al método que seguimos para aplicar el experimento. Éste fue muy rudimentario porque quisimos simplificar el proceso y por no saber programar, lo que se tradujo en resultados deficientes.

CONCLUSIONES

O CONSIDERACIONES FINALES

Después de haber concluido la investigación podemos decir, en primer lugar, que se cubrió por completo el planteamiento original de esta tesis. Conforme a lo previsto, se realizaron pruebas al sistema que permitieron hacernos una idea certera del desempeño del programa; para ello tuvimos que identificar manualmente los pares semánticos contenidos en el corpus, lo que, al mismo tiempo, nos fue útil para conocerlo y perfilar las líneas de trabajo que se definieron con un análisis exhaustivo de los alineamientos. Cada eje de análisis desembocó en una propuesta —sustentada en las necesidades específicas del programa y apoyada en un aparato crítico— para replantear el proceso de alineamiento de las definiciones con miras a lograr un mejor rendimiento en la obtención de pares semánticos. Finalmente, el ejercicio para aplicar algunas de las propuestas se llevó a término.

Por otra parte, nuestra investigación respondió a la necesidad de aplicar un análisis de corte lingüístico a las definiciones del banco terminológico que muy claramente plantea Gabriel Castillo en el apartado *trabajos futuros* (5.6) de su tesis con el propósito de mejorar los resultados obtenidos con su algoritmo.

En cuanto a los objetivos trazados, se han delimitado los aspectos sintácticos que deben tomarse en consideración para las posteriores modificaciones al programa de agrupamiento semántico. Incluso hemos propuesto, en muchos de los casos, alternativas específicas sobre cómo podría aprovecharse el análisis que hacemos sin realizar cambios profundos al sistema. También sugerimos el orden en que pueden llevarse a cabo estas modificaciones, priorizando lo que, a partir de nuestro estudio, parece más sencillo de implementar.

El experimento que realizamos no cubrió el objetivo de permitirnos reafirmar y/o descartar líneas de investigación; no obstante, confirmó nuestra hipótesis de que programar Clustering para reconocer rasgos fundamentales de la gramática española mejora el proceso de recuperación de pares semánticos, como ejemplo sirven los pares generados a partir de la identificación de verbos compuestos. Además, el ejercicio hizo patente el riesgo de que las iniciativas que puntualizamos no puedan ser fácilmente sistematizadas. Hacer un análisis más profundo en este sentido corresponde a un computólogo.

Por otra parte, todos los análisis refuerzan la idea de que es importantísimo realizar el etiquetado de categorías gramaticales y aplicarlo simultáneamente

con las modificaciones que deriven de las propuestas que aquí se vertieron. Para instrumentar esto último resulta necesario, en un segundo momento, etiquetar las definiciones con un *chunking* (análisis sintáctico parcial), a fin de que el sistema sea capaz de reconocer frases que puedan alinearse con palabras. Y no sólo eso, con la finalidad de extender las búsquedas más allá de lo que literalmente expresan las definiciones que se comparan en los alineamientos, deben crearse etiquetas que contengan información de carácter pragmático como la referencia a expresiones implícitas en las definiciones que se manifiestan mediante pronombres o a través de la negación de opuestos. Asimismo, deben incorporarse al mecanismo del sistema redes semánticas que den cuenta de las relaciones léxicas de hiponimia e hiperonimia.

Una consideración que se desprende del conjunto de los análisis realizados es que de aplicarse las modificaciones en cuanto a la extensión de semi iguales a todo los pares formados por palabras funcionales de la misma clase y a que éstos no intervengan en el cálculo de coeficientes de similitud, disminuiría drásticamente el número de pares vinculados, pero estos serían mejores pares. Tal vez, incluso, fuera necesario reducir el valor de LCC que debe considerarse para promover a vinculados pares correspondientes.

Debemos señalar que el trabajo realizado consistió en un estudio sintáctico acotado a la lexicografía en el dominio de Física, la cual incorpora características que le son exclusivas; pero la mayoría de las iniciativas aportadas para una mejor recuperación de pares semánticos[1] puede extenderse a solucionar problemas en la extracción de información relevante en otras bases de datos léxicos. Siempre que un proceso sea automatizable, se cuenta con la ventaja adicional de que no es sólo útil para la colección a la que aplicamos el estudio, sino para cualquier texto que cumpla con las características requeridas. Por lo anterior, la presente tesis hace un aporte a los sistemas de recuperación de información.

De este trabajo no se desprenden fórmulas para la resolución de problemas en la obtención de pares semánticos detectados en el transcurso del mismo y en otros anteriores. Se ha abocado simplemente a elaborar propuestas que pueden ser formalizadas en algoritmos. Sólo con el diseño de herramientas computacionales que contemplen lo aquí trabajado para coadyuvar en el mejoramiento del funcionamiento de Clustering cobrará sentido esta investigación, de la misma manera en que ella es continuación y complemento del esfuerzo de otros compañeros dentro del GIL. Una vez integradas estas propuestas al sistema se podrá hacer un análisis a detalle de su pertinencia para el proceso de identificación de pares semánticos.

Finalmente, en la misma medida en que esta tesis habla de problemas concernientes a la ingeniería lingüística, para su elaboración nos hemos valido de los recursos aportados por esta disciplina, lo cual facilitó enormemente su realización. Debido a lo anterior podemos dar testimonio de la utilidad de las tecnologías del lenguaje para quienes emplean la computadora como herramienta de apoyo a la investigación.

---

1 La única iniciativa que quedaría fuera de esta generalización es la referida a los patrones definitorios.

TRABAJOS FUTUROS

Como ya adelantábamos en las conclusiones, el siguiente trabajo enfocado a Clustering que deberá realizarse es la instrumentación de las propuestas reunidas en esta tesis.

Para posteriores trabajos lingüísticos, además de profundizar en la investigación que conduzca a reconocer y abordar correctamente las diferentes formas de negación; a identificar los sintagmas enfáticos, y a incorporar al sistema la información gramtical contenida en las conjunciones y en los signos de puntuación, queremos plantear tres posibilidades:

La primera obedece a un problema en cuanto a la obtención de pares semánticos que no contempla nuestra propuesta. Con las modificaciones sintácticas esbozadas se recuperarán más pares semánticos, pero dicho incremento llegará a un límite que no podrá rebasar; lo cual responde al hecho de que muchos de los pares identificados manualmente no ocupan posiciones paralelas en los alineamientos. Si bien es cierto que cuando —por ejemplo— una frase adjetiva se pueda alinear con un adjetivo[2] se producirá un reordenamiento general que, sin lugar a dudas, será positivo para la vinculación de pares cercanos; esto no soluciona el problema en alineamientos donde la primera definición comienza por un sujeto y la segunda por un adjunto.

En relación con lo anterior, se podría hacer un etiquetado de las definiciones a nivel de frase y reordenarlas todas de acuerdo con una forma canónica. Sin embargo, debe considerarse que esto tal vez desembocaría en otras dificultades pues, por ejemplo, podemos pasar por alto pares relevantes formados por dos adjetivos dentro de frases que actuarían como par pero en conjunto con los sustantivos a los que determinan (que fue lo que ocurrió con algunas formas verbales al interior de las perífrasis en el experimento del capítulo 4). En este sentido, en 3.1.3 propusimos que el programa, una vez que haya identificado los pares a nivel de frase, corra el algoritmo recursivamente para localizar posibles nuevos pares al interior de las frases.

Este reordenamiento también podría considerar atender los problemas para el etiquetado que se derivan de la existencia de más de un modificador por cada núcleo, tales como los detectados en 3.2.1 y 4.1.4.

En segundo lugar, queremos destacar otro aspecto que genera dificultades para los alineamientos y que no fue atendido en esta tesis. Las definiciones con las que trabaja Clustering no siguen un mismo esquema puesto que su origen se fundamenta en conceptualizaciones distintas. Las cuales, según hemos

2 *Vid.* 3.3

visto, pueden agruparse en torno a unas líneas generales. Así tenemos definiciones teóricas (la mayoría), históricas (que ponen énfasis en datos como quiénes son los formuladores de tal teoría o la fecha de formulación), las etimológicas, las lógicas (a través de fórmulas), o las que describen experimentos.

Su tratamiento parece más sencillo, pues cabe la posibilidad de determinar con mayor exactitud los tipos de definiciones existentes en cuanto al modo de conceptualizar y añadir a cada definición una etiqueta con la información que le corresponda, a fin de que se impida alinear, por ejemplo, una definición histórica y una lógica. Asimismo, cada definición podría recibir un orden considerando que existe esta estructura.

Por último y recordando que la utilidad final de Clustering es el diccionario onomasiológico, sugerimos adelantarnos a problemas que puedan aparecer cuando éste se encuentre en funcionamiento. Para que el DEBO sea funcional debe estar equipado para "comprender" lo que un usuario expresa en lenguaje natural, lo cual requiere de un conocimiento profundo del uso que el hablante hace de las estructuras de la lengua.

El canal de comunicación para que el usuario introduzca sus definiciones es la escritura, porque así se ha diseñado el sistema, pero no debe perderse de vista que puede ocurrir que sus conceptualizaciones se realicen con expresiones más cercanas a la oralidad de un habla coloquial (tal como sucede con el *chat*) que a la jerga terminológica.

En este sentido, parece pertinente recoger más definiciones **personales** y hacer nuevas pruebas al sistema considerando sólo a éstas.

BIBLIOGRAFÍA DIRECTA

Alarcón Martínez, Rodrigo, *Análisis lingüístico de contextos definitorios en textos de especialidad.* Tesis de licenciatura. México: UNAM, Facultad de Filosofía y Letras, 2003.

Alarcón, Rodrigo y Gerardo Sierra, "El rol de las predicaciones verbales en la extracción automática de conceptos" en *Estudios de lingüística aplicada.* Eds. Natalia Ignatieva y Gerardo Sierra. CELE, UNAM, año 22, núm. 38, (México, diciembre de 2003).

Alarcos Llorach, Emilio, *Gramática de la lengua española*. Madrid: Espasa-Calpe, 1994.

Barrón C., Alberto, Gerardo Sierra y Elio Villaseñor, "C-*value* aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español" [artículo aceptado para presentarse en el Encuentro Nacional de Ciencias de la Computación. San Luis Potosí, septiembre 2006]
http://ccc.inaoep.mx/~tec_lenguaje06/trabajosAceptados.html

Bosque, Ignacio, "Artículo y pronombre: relaciones y diferencias" en *Las categorías gramaticales: relaciones y diferencias.* Madrid: Síntesis, 1990. (Lingüística, 11).

Cabré, María Teresa, *La terminología: teoría, metodología, aplicaciones.* Barcelona: Antártida/Empúries, 1993.

Casares, Julio, *Introducción a la lexicografía moderna,* Madrid: CSIC, 1950.

Castillo Hernández, Gabriel, *Algoritmo revisado para la extracción automática de agrupamientos semánticos*. México: UNAM, 2002.

Castillo H., Gabriel y Gerardo Sierra, "Algoritmo flexibilizado de agrupamiento semántico" en *Estudios de lingüística aplicada.* Eds. Natalia Ignatieva y Gerardo Sierra. CELE, UNAM, año 22, núm. 38, (México, diciembre de 2003).

Escandell Vidal, Ma. Victoria, *Los complementos del nombre.* Madrid: Arco/Libros, 1995. (Cuadernos de lengua española, Y).

Frakes, William B., "Introduction to Information Storage and Retrieval Systems" en *Information Retrieval*. *Data Structures and Algorithms*. Eds. William B. Frakes y Ricardo Baeza-Yates. Nueva Jersey: Prentice Hall, 1992.

García-Miguel, José María, *Transitividad y complementación preposicional en español.* Santiago de Compostela: Universidad de, 1995. (Verba, anexo nº 40).

Gili Gaya, Samuel, *Curso superior de sintaxis española.* Barcelona: Biblograf, 1969.

Leonetti, Manuel, *Los determinantes.* Madrid: Arco/Libros, 1999. (Cuadernos de lengua española, M).

Morales, Eduardo F. *Descubrimiento de conocimiento en bases de datos.* Capítulo 2 [en línea]
http://dns1.mor.itesm.mx/~emorales/Cursos/KDD03/node45.html
[Consulta: 22/03/2005].

Pamies, Antonio, Martina Bálmacz y Eva Ma. Inesta, "Criterios para una fraseografía onomasiológica automatizable" en *Léxico y fraseología.* Eds. Juan de Dios Luque y Antonio Pamies. Granada: Método, 1998.
Tomado de la versión electrónica:
http://www.ashda.ugr.es/laboratorio/liptovesp.pdf [Consulta: 24/05/2005].

Porter, M.F. "An algorithm for suffix stripping" en *Program*, vol. 14, núm. 3, julio de 1980.

Reyes Pérez, Antonio, *Extracción automática de terminología en el léxico de Física.* Tesis de licenciatura. México: UNAM, Facultad de Filosofía y Letras, 2003.

Rojo, Guillermo y Tomás Jiménez Juliá, *Fundamentos del análisis sintáctico funcional.* Santiago de Compostela: Universidad de, 1989.

Sanz Alonso, Beatriz, *La negación en español.* Madrid: Ediciones Colegio de España, 1996. (Español, lengua extranjera, 9).

Sidorov, Gregory, "Problemas actuales de lingüística computacional" en *Revista digital universitaria* [en línea]. Vol. 2, No. 1, 31/03/2001.
http://www.revista.unam.mx/vol.2/num1/art1/ [Consulta: 27/02/2005].

Sierra Martínez, Gerardo, "Avances en el desarrollo del Diccionario Electrónico de Búsqueda Onomasiológica" en *Actas del V Simposio iberoamericano de terminología RITerm.* Ciudad de México, del 3 al 8 de noviembre de 1996.

Sierra Martínez, Gerardo, *Design of a concept-oriented tool for terminology.* Tesis doctoral. Manchester: Institute of Science and Technology, 1999.

Sierra Martínez, Gerardo y John y McNaught, "Serendipitous Wording of POS Tags to Extract Semantic Pairs of Words from Dictionary Definitions" en *Computacional Intelligence* (actas de la Conferencia Internacional en Inteligencia Computacional 2004). Okatan, Ali, ed. Estambul, diciembre de 2004.

Wagner, R.A. y M.J. Fisher "The String To String Correction Problem" en *Journal Association for Computing Machinery*, vol. 21, núm.1, (1974).

Zampolli, Antonio, “Los bancos de datos léxicos: bases multifuncionales de datos léxicos” (trad. E. Lavín) en *Las industrias de la lengua*. Dir. José Vidal Beneyto. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

BIBLIOGRAFÍA DE CONSULTA

Alcaraz Varó, Enrique y Ma. Antonia Martínez Linares. *Diccionario de lingüística moderna*. Barcelona: Ariel, 1997.

García-Pelayo y Gross, Ramón (dir.), *Gran Diccionario Larousse Español – Inglés.* México: Ediciones Larousse, 1984.

Lázaro Carreter, Fernando. *Diccionario de términos filológicos,* 3ª. ed. Madrid: Gredos, 1968. (Biblioteca Románica Hispánica: III. Manuales, 6)

Seco, Manuel, Olimpia Andrés y Gabino Ramos, *Diccionario del español actual*. Madrid: Aguilar, 1999.

Real Academia Española. *Diccionario de la lengua española (RAE Usual)*, 21 ed. Madrid: Espasa-Calpe, 1992.

Webster's Universal College Dictionary. New York: Gramercy Books, 1997.

BIBLIOGRAFÍA INDIRECTA

(CONSULTADA A TRAVÉS DEL BANCO TERMINOLÓGICO)

Código de fuente	Referencia
Antecedentes de Física	Jaramillo Morañes, Gabriel A., *Antecedentes de Física.* México: Trillas, 1990.
Atlas de Física	Fernández, Ferrer, J., *Atlas de física.* Barcelona: Jover, 1974
Britanica.com	Enciclopaedia Britannia. [en línea]
CIm1	*Archivo de mecánica*. México: Centro de Instrumentos – UNAM, 1999.
Cinemática y dinámica básica	Solar González, Jorge, *Cinemática y dinámica básica para ingenieros.* México: Trillas, 1991(1a. reimp.)
Clave	*Diccionario de uso del español actual (CLAVE)*. Madrid: Diccionarios SM, 1997.
Curso de Física moderna	Acosta, Virgilio y Gram Cowan, *Curso de Física Moderna.* México: Harla, 1975.
Diccionario de electrónica	Amos, S. W., *Diccionario de electrónica español-inglés/inglés-español.* Madrid: Paraninfo, 1998.
Diccionario de energía	Hunt, V. Daniel, *Diccionario de energía.* México: Marcombo, 1984.
Diccionario Oxford-Complutense	*Diccionario Oxford-Complutense de física.* Madrid: Complutense, 1998.
Diccionario de ciencia y tecnología	*Diccionario de ciencia y tecnología* [en línea] http://harcourt.com

Diccionario de Física Vox	*Diccionario de física Vox.* Barcelona: Biblograf, 1990.
Diccionario enciclopédico de Física	*Diccionario Enciclopédico de física.* Moscú-Madrid: Mir, 1995.
Diccionario enciclopédico Planeta	*Diccionario enciclopédico Planeta.* Barcelona: Planeta, 1984.
Diccionario Mc. Graw Hill de Física	Parker, Sybil P., *Diccionario McGraw-Hill de Física.* México: McGraw-Hill, 1991.
Dictionary.msn.com	Dictionary.msn.com
Dictionary of Physics	Daintith, John, *Dictionary of physics.* Nueva York: Facts on File, 1981.
Dinámica (Interamericana)	Ginsberg, Jerry H. y Joseph Genin, *Dinámica.*México: Nueva editorial Interamericana, 1980.
DRAE 92	Real Academia Española, *Diccionario de la lengua española (RAE Usual),* 21 ed. Madrid: Espasa-Calpe, 1992.
El mundo de la Física 1	Cetto K., Ana María y otros, *El mundo de la física 1.* México: Trillas, 1991.
Electrónica y técnica nuclear	Markus, John, *Diccionario de electrónica y técnica nuclear.* Barcelona: Marcombo, 1972
Encarta 98	*Enciclopedia Encarta* [electrónica]. Microsoft, 1998.
Encarta 2000	*Enciclopedia Encarta* [electrónica]. Microsoft, 2000.
Enciclopedia Britannia	*Enciclopedia Britannia.* México: Océano, 1995.
Enciclopedia Ciencia y tecnología	*Enciclopedia ciencia y tecnología.* México: McGraw-Hill, 1993.
Enciclopedia de las ciencias	*Enciclopedia de las Ciencias. México: Editorial Cumbre, 1987.*

Enciclopedia Lexis 22	*Diccionario enciclopedico Lexis*, Barcelona: Círculo de Lectores,1978.
Enciclopedia Salvat	*Enciclopedia Salvat*. Barcelona: Salvat, 1999.
Física 2. Electricidad y magnetismo	Gartenhaus, Solomon, *Física 2. Electricidad y magnetismo.* México: Nueva Editorial Interamericana, 1979.
Física aplicada	Beiser, Arthur, *Física aplicada,* 2ed. México, Mc Graw-Hill Interamericana, 1995.
Física con aplicaciones	Wilson, Jerry D., *Física con aplicaciones.* México: Nueva Editorial Interamericana, 1984.
Física. Conceptos y aplicaciones	Tippens, Paul E., *Física conceptos y aplicaciones.* México: Mc Graw Hill Interamericana, 1992.
Física conceptual	Hewitt, Paul G., *Física conceptual.* México: Pearson, 1999.
Física de emergencia	Noreña, Francisco, *Física de emergencia*. *Diccionario enciclopédico de física para jóvenes*. México: Pangea Editores, 1995. (Serie Acordeones).
Física. Fundamentos y aplicaciones	Eisberg, Robert M. y Lawrence S. Lerner, *Física*. *Fundamentos y aplicaciones*. México: McGraw-Hill, 1990.
Física general Alvarenga	Alvarenga Alvarez, Beatriz y Antonio Máximo Ribeiro Da Luz, *Física general: con experimentos sencillos.* México: Harla, 1983.
Física general Benson	Benson, Harris, *Física general*. México: CECSA, 1997.
Física general Bueche-Schaum	Bueche, Frederic, *Física general*, 9ª ed. México, , McGraw-Hill, 1991. (serie Schaum).
Física general van der Merwe	Van der Merwe, Carel W. *Física general*. México: McGraw-Hill, 1998.
Física I. Guía escolar Vox	*Física I. Guía escolar VOX.* México: Editorial Patria, 1993.

Física moderna	White, Harvey E., *Física moderna*, Barcelona: UTHEA, 1965.
Física para ciencias e Ingeniería	Fishbane, Paul, Gasiorowicz y Thornton, *Física para ciencias e ingeniería.*México: Prentice-Hall hispanoamericana, 1994,
Física para Ingeniería y ciencias	Wells, Dare A. y Harold S. Slusher, *Física para ingeniería y ciencias*. Traducción, Antonio Ortiz Herrera; revisión técnica, Miguel Irán Alcérreca Sánchez. México: Mc-Graw Hill, 1985.
Física Resnick	Resnick, Robert, David Halliday y Keneth Krane, Física, México, CECSA, 1996
Física Serway	Serway, Raymond, *Física general*, México: McGraw-Hill Interamericana, 1997.
Física universitaria Sears	Sears, Francis W, Mark W. Zemansky y Hugh D. Young, *Física Universitaria*, Wilmington: Addison Wesley Iberoamericana, 1988.
Física Weber	Weber, Robert L., Kenneth V. Manning y Marsh W. White, Física, Barcelona, Reverté, 1970.
Fundamentos de Física	Blatt Frank J. *Fundamentos de Física*. México: Ed. Prentice Hall, 1991.
Gran diccionario enciclopédico	*Gran Diccionario Enciclopédico Visual. Barcelona:*Océano, 1997.
Gran enciclopedia del mundo	Menendez Pidal, Ramón, *Gran enciclopedia del mundo.* Bilbao: Durvan, 1961.
Gran enciclopedia Larousse	*Gran enciclopedia Larousse.* Barcelona: Planeta, 1991.
Ingeniería mecánica. Dinámica	Hibbeler, Russell C., *Ingeniería mecánica dinámica*, 7ª ed. México: Prentice Hall, 1996.
Ingeniería mecánica. Estática	Hibbeler, Russell C., *Ingeniería mecánica estática.* México: Prentice Hall, 1995.

Introducción al estudio de la mecánica	Ingard, U. y W. L. KRAUSHAAR, *Introducción al estudio de la mecánica materia y ondas.* Barcelona: Reverté, 1960.
Larousse	Gran diccionario de la lengua española. Barcelona: Larousse-Planeta, 1996.
Larousse de ciencias y técnicas	*Pequeño Larousse de ciencias y técnicas.* París: Larousse, 1967.
Lexipedia	*Lexipedia: Diccionario enciclopédico.* México : Enciclopedia Británica Publisher, 1999.
María Moliner	Moliner María, *Diccionario de uso del español.* Madrid: Gredos, 1992.
Master. Diccionario enciclopédico	*Master. Diccionario enciclopédico.*México: Ediciones culturales internacionales, 1998.
Mecánica I. Estática	Ocampo Canaval, Fernando, *Mecánica I. Estática*. México: Limusa, [s.f.].
Mecánica para ingenieros Huang	Huang, T. C., *Mecánica para Ingenieros*. *Estática.* México: Alfaomega, 1993.
Mecánica para ingenieros Singer	Singer, Ferdinand, *Mecánica para ingenieros. Dinámica*, México: Harla, 1982.
Mecánica vectorial. Dinámica	Beer, Ferdinand P., *Mecánica vectorial para ingenieros. Dinámica.* México: Mc Graw-Hill, 1990.
Mecánica vectorial. Estática	Beer, Ferdinand P., *Mecánica vectorial para ingenieros*. *Estática*, México: Mc Graw-Hill, 1990.
Océano Uno Color	*Diccionario enciclopédico a color Océano*. *Uno.* México: Océano, 1998.
Química Chang	Chang, Raymond, *Química.* México: McGraw-Hill, 1992.
Salvat Multimedia	Enciclopedia Salvat Multimedia. 2001
Science Study Comitte	*Física.* España: Physical Science Study Comité, 1966.
Teoría y problemas	Beiser, Arthur, *Teoría y problemas en ciencias físicas.* México: McGraw-Hill, 1976.

Vox *Vox: Diccionario general de la lengua española* . Barcelona: Biblograf, 1997.

APÉNDICE A

ESQUEMA DEL CORPUS TRABAJADO

En este apéndice se puede reconocer a qué términos de la base de Física del banco terminológico corresponden las definiciones que integran el corpus trabajado en esta tesis. Se proporciona el código numérico que el programa asigna a cada término, así como el número de definiciones y alineamientos que le corresponden.

Cualquier consulta sobre el banco terminológico se puede realizar en la dirección:
http://iling.torreingenieria.unam.mx/diccionarios

Si se desea observar los alineamientos con cualquiera de las versiones del programa, consúltese:
http://tabasco.torreingenieria.unam.mx/scripts/clusters.exe/Alineamiento

Esquema del corpus trabajado en esta tesis

Nº de término que el programa asigna	Término	Nº de definiciones	Nº de alineamientos
1	Mecánica	15	104
2	Cinemática	17	134
3	Dinámica	17	133
5	Partícula	7	21
8	Velocidad angular	7	19
9	Período	13	74
13	Energía	18	150
14	Energía potencial gravitacional	9	35
15	Movimiento rectilíneo y uniforme	4	6
16	Movimiento uniformemente acelerado	5	10
17	Primera ley de Newton	10	44
18	Segunda ley de Newton	3	3
19	Tercera ley de Newton	3	3
20	Velocidad virtual	2	1
21	Principio de Galileo	3	3
23	Energía cinética	14	90
24	Movimiento uniformemente retardado	2	1
25	Ley de la conservación de la energía	8	27
26	Conservación	2	1
27	Fuerzas disipativas	3	3
32	Péndulo	8	28
35	Movimiento de traslación	5	10
38	Tiro parabólico	2	1
39	Acción	3	3
40	Aceleración centrípeta	4	6
41	Afelio	5	10
45	Calor	14	90
46	Causalidad	6	5
47	Centro de equilibrio	5	10
48	Centro de masa	4	6
49	Choque	6	15
54	Coeficiente de fricción estática	5	10
57	Cronómetro	11	54
58	Cuerpo	8	27
59	Cuerpo celeste	2	1
61	Desplazamiento	7	21
62	Día solar	5	10

65	Difracción	10	45
66	Dilatación del espacio	6	14
67	Dilatación del tiempo	2	1
68	Dimensión	13	74
69	Dirección de un vector	5	10
70	Eje de giro	5	9
71	Eje de rotación	5	9
72	Elongación	8	26
73	Energía cinética de rotación	5	10
74	Energía eléctrica	8	27
75	Energía mecánica	8	28
76	Energía química	9	36
77	Energía térmica	6	15
78	Espacio	9	36
79	Estado de movimiento	5	10
80	Estado de reposo	6	15
81	Éter	13	76
82	Eventos simultáneos	6	14
83	Experimento de Michelson y Morley	8	25
84	Experimento de Young	6	15
85	Física	13	77
86	Frecuencia de resonancia	4	5
87	Frecuencia natural	5	10
88	Fricción	8	28
89	Fuerza	12	66
90	Fuerza centrípeta	8	28
91	Fuerza conservativa	8	28
92	Fuerza de atracción gravitacional	5	10
93	Fuerza de fricción	6	15
94	Fuerza de fricción cinética	5	10
95	Fuerza disipativa	5	10
96	Fuerza ficticia o de inercia	6	15
97	Fuerza resultante	4	6
98	Giro	2	1
99	Gravedad	6	15
100	Hertz (unidad)	5	10
101	Inercia	9	36
102	Interferencia	6	15
103	Ley de Hook	4	6
104	Ley de la acción	5	10
105	Ley de la conservación del ímpetu	5	9
106	Ley de la conservación del momento angular	2	1

107	Ley de la gravitación universal	3	3
108	Ley de las áreas	3	2
109	Leyes de Kepler	2	1
110	Leyes de Newton	2	1
111	Línea de acción de una fuerza	3	3
112	Luz	13	78
113	Magnitud de un vector	6	14
114	Marco de referencia	5	10
115	Marco de referencia inercial	5	10
116	Marco estacionario	5	10
117	Masa	15	104
118	Masa puntual	5	10
119	Masa total de un sistema	5	10
120	Momento angular	4	6
121	Momento de inercia	4	6
122	Móvil	4	6
123	Movimiento armónico	5	10
124	Movimiento de caída libre	2	1
125	Movimiento lineal	2	1
126	Movimiento rotacional	2	1
127	Newton unidad	6	15
128	Objeto	3	3
129	Observador	2	1
130	Onda	7	20
131	Onda luminosa	3	3
132	Onda senoidal pura	2	1
133	Órbita	9	33
134	Órbita circular	5	10
135	Órbita elíptica	5	10
160	Segundo (unidad de tiempo)	5	10
194	Einstein, Alberto (1879-1955)	4	6
196	Galilei, Galileo (1564-1642)	4	6

APÉNDICE B

PARES SEMÁNTICOS IDENTIFICADOS MANUALMENTE

En este apéndice se pueden consultar los pares semánticos que fueron detectados durante el análisis manual de los alineamientos. La primera tabla corresponde a los pares semánticos simples y la segunda, a los compuestos. Para cada par proporcionamos el alineamiento en que fue hallado (nombre del término, código de fuente de la definición 1 y código de fuente de la definición 2). La mayoría de los pares presentan repeticiones a lo largo del corpus, pero proporcionamos sólo la ubicación de la primera ocurrencia.

Pares semánticos simples identificados manualmente

PAR SEMÁNTICO	TÉRMINO	DEFINICIÓN 1	DEFINICIÓN 2
{parte, rama}	Mecánica	DRAE 92	Física general Alvarenga
{parte, disciplina}	Mecánica	María Moliner	Larousse
{rama, disciplina}	Mecánica	Física general Alvarenga	Larousse
{describir, estudiar}	Cinemática	CIm1	DRAE 92
{movimiento, trayectoria}	Cinemática	CIm1	Definición en la web
{razón, causa}	Cinemática	Desconcida	Fundamentos de Física
{considerar, importar}	Cinemática	Desconcida	Física de emergencia
{descripción, estudio}	Cinemática	Física. Fundamentos y aplicaciones	Cinemática y dinámica básica
{fuerza, causa}	Cinemática	DRAE 92	Desconocida
{fuerza, razón}	Cinemática	DRAE 92	Fundamentos de Física
{originar, producir}	Dinámica	Física universitaria Sears	María Moliner
{cantidad, parte}	Partícula	Vox	DRAE 92
{poseer, tener}	Partícula	Ingeniería mecánica. Estática	Mecánica para ingenieros Huang
{sólo, solamente}	Ley de la conservación de la energía	Enciclopedia Lexis 22	Personal
{vibración, oscilación}	Período	Vox	Larousse
{móvil, objeto}	Período	CIm1	Física de emergencia
{móvil, péndulo}	Período	CIm1	DRAE 92
{capacidad, aptitud}	Energía	CIm1	María Moliner
{realizar, efectuar}	Energía	CIm1	Física Weber
{capacidad, propiedad}	Energía	CIm1	Física de emergencia
{producir, hacer}	Energía	María Moliner	Dictionary of Physics
{producir, efectuar}	Energía	María Moliner	Física Weber
{materia, sistema}	Energía	María Moliner	Física de emergencia
{producir, realizar}	Energía	Océano Uno Color	Física de emergencia
{hacer, realizar}	Energía	Dictionary of Physics	Física de emergencia
{hacer, efectuar}	Energía	Dictionary of Physics	Física Weber
{trabajo, efecto}	Energía	Desconocida	Diccionario enciclopédico Planeta
{aumentar, cambiar}	Movimiento uniformemente acelerado	Desconocida	Física de emergencia
{impelido, obligado}	Primera ley de Newton	Física universitaria Sears	El mundo de la Física 1
{preservar, continuar}	Primera ley de Newton	Física moderna	Enciclopedia de las ciencias

{exterior, externo}	Primera ley de Newton	Física moderna	Dictionary of Physics
{ejercer, actuar}	Primera ley de Newton	Personal	Física de emergencia
{pero, mas}	Tercera ley de Newton	CIm1	Física de emergencia
{misma, igual}	Tercera ley de Newton	CIm1	Física de emergencia
{infinitamente, muy}	Velocidad virtual	DRAE 92	Larousse
{forma, clase}	Ley de la conservación de la energía	CIm1	Física Resnick
{acción, influencia}	Péndulo	Vox	Larousse
{punto, soporte}	Péndulo	DRAE 92	Física de emergencia
{cuerpo, astro}	Movimiento de traslación	DRAE 92	María Moliner
{duración, tiempo}	Acción	Vox	Dictionary of Physics
{energía, fuerza}	Acción	Dictionary of Physics	Física Weber
{alejado, distante}	Afelio	Vox	María Moliner
{alejado, lejos}	Afelio	Vox	Física de emergencia
{pasar, transmitirse}	Calor	DRAE 92	Física de emergencia
{movimiento, agitación}	Calor	Vox	Física Weber
{corto, breve}	Choque	Cinemática y dinámica básica	Física general Alvarenga
{grande, intenso}	Choque	Cinemática y dinámica básica	Física general Alvarenga
{cambio, variación}	Choque	Cinemática y dinámica básica	Personal
{intervalo, espacio}	Choque	Cinemática y dinámica básica	Personal
{colisión, impacto}	Choque	Física general Alvarenga	Física Resnick
{intervalo, espacio}	Choque	Física general Alvarenga	Personal
{cociente, relación}	Coeficiente de fricción estática	Física Preuniversitaria	Física general van der Merwe
{instrumento, aparato}	Cronómetro	Larousse de ciencias y técnicas	Enciclopedia de la técnica y la mecánica
{instrumento, reloj}	Cronómetro	Larousse de ciencias y técnicas	Vox
{aparato, reloj}	Cronómetro	Enciclopedia de la técnica y la mecánica	Vox
{alta, gran}	Cronómetro	Vox	DRAE 92
{gran, especial}	Cronómetro	DRAE 92	María Moliner
{precisión, exactitud}	Cronómetro	DRAE 92	María Moliner
{objeto, sustancia}	Cuerpo	Definición en la WEB	Larousse de ciencias y técnicas
{sustancia, materia}	Cuerpo	Larousse de ciencias y técnicas	Larousse

{lugar, posición}	Desplazamiento	Física de emergencia	Física Weber
{pasos, transiciones}	Día solar	Encarta 2000	Dictionary of Physics
{desviación, dispersión}	Difracción	Vox	María Moliner
{bordear, encontrar}	Difracción	María Moliner	Física de emergencia
{parte, región}	Difracción	María Moliner	Física Weber
{cuerpo, obstáculo}	Difracción	Vox	Atlas de Física
{radiación, onda}	Difracción	Dictionary of Physics	Física de emergencia
{definir, determinar}	Dimensión	DRAE 92	María Moliner
{cosa,objeto}	Dimensión	María Moliner	Clave
{cosas, algo}	Dimensión	María Moliner	Clave
{cantidad, magnitud}	Dimensión	Antecedentes de Física	Clave
{expresar, definir}	Dimensión	Personal	Clave
{varilla, espiga}	Eje de giro	Océano Uno Color	Encarta 98
{varilla, recta}	Eje de giro	Diccionario enciclopédico Planeta	Encarta 98
{eje, varilla}	Eje de rotación	Diccionario enciclopédico Planeta	Personal
{girar, rodar}	Eje de rotación	Diccionario enciclopédico Planeta	Personal
{moverse, girar}	Eje de rotación	Mecánica vectorial. Dinámica	Gran enciclopedia Larousse
{momento, instante}	Elongación	Océano Uno Color	Atlas de Física
{objeto, pieza}	Elongación	Personal	DRAE 92
{pieza, cosa}	Elongación	DRAE 92	Larousse
{cuerpo, masa}	Energía cinética de rotación	Física aplicada	Enciclopedia de la Física
{derivar, proceder}	Energía química	Gran enciclopedia Larousse	Personal
{desprendido, liberado}	Energía química	Enciclopedia Espasa - Calpe	Personal
{extensión, superficie}	Espacio	Personal	Larousse
{medio, extensión}	Espacio	Vox	Larousse
{medio, superficie}	Espacio	Vox	Larousse
{ortogonales, perpendiculares}	Espacio	Enciclopedia Ciencia y tecnología	Enciclopedia Britannia
{lugar, sitio}	Estado de reposo	Enciclopedia Lexis 22	Física Cutnell-Limusa
{combinación, unión}	Éter	Master. Diccionario enciclopédico	Enciclopedia Ciencia y tecnología
{compuesto, sustancia}	Éter	Master. Diccionario enciclopédico	Física de emergencia
{compuesto, fluido}	Éter	Enciclopedia Ciencia y tecnología	DRAE 92
{invisible, hipotético}	Éter	Larousse	Dictionary of Physics

{creer, pensar}	Éter	Larousse	Física de emergencia
{creer, suponer}	Éter	Larousse	Vox
{pensar, suponer}	Éter	Física de emergencia	Vox
{tren, furgón}	Eventos simultáneos	Física universitaria Sears	Física Serway
{rayo, relámpago}	eventos simultáneos	Física universitaria Sears	Física Serway
{caer, golpear}	eventos simultáneos	Física universitaria Sears	Física Serway
{medir, determinar}	Experimento de Michelson y Morley	Física 2. Electricidad y magnetismo	Diccionario Mc. Graw Hill de Física
{cambio, diferencia}	Experimento de Michelson y Morley	Física Serway	Encarta 2000
{detectar, obtener}	Experimento de Michelson y Morley	Física Serway	Personal
{detectar, determinar}	Experimento de Michelson y Morley	Física Serway	Diccionario Mc. Graw Hill de Física
{espacio, éter}	Experimento de Michelson y Morley	Encarta 2000	Personal
{demostración, evidencia}	Experimento de Young	Física. Fundamentos y aplicaciones	Física. Conceptos y aplicaciones
{demostración, experimento}	Experimento de Young	Física. Fundamentos y aplicaciones	Personal
{demostrar, mostrar}	Experimento de Young	Personal	Física de emergencia
{ranura, rendija}	Experimento de Young	Física Serway	Física. Conceptos y aplicaciones
{investigar, estudiar}	Física	Física. Conceptos y aplicaciones	Personal
{concepto, propiedad}	Física	Física. Conceptos y aplicaciones	DRAE 92
{vibrante, oscilante}	Frecuencia de resonancia	Diccionario de electrónica	Física Resnick
{oscilar, vibrar}	Frecuencia natural	Diccionario de electrónica	Ingeniería mecánica. Dinámica
{sistema, cuerpo}	Frecuencia natural	Diccionario de electrónica	Electrónica y técnica nuclear
{movimiento, deslizamiento}	Fricción	Física general Alvarenga	Enciclopedia Ciencia y tecnología
{frotamiento, rozamiento}	Fricción	Clave	Larousse
{agente, causa}	Fuerza	Gran enciclopedia del mundo	Enciclopedia Ciencia y tecnología
{cambiar, alterar}	Fuerza	Dictionary of Physics	Atlas de Física

{modificar, alterar}	Fuerza	DRAE 92	Atlas de Física
{modificar, cambiar}	Fuerza	DRAE 92	Dictionary of Physics
{cuerpo, móvil}	Fuerza centrípeta	Física Serway	Vox
{estado, posición}	Fuerza conservativa	Física Resnick	Diccionario de Física Vox
{cuerpo, partícula}	Fuerza de atracción gravitacional	Mecánica I. Estática	Mecánica para ingenieros Huang
{material, cuerpo}	Fuerza de fricción	Física conceptual	Personal
{superficie, material}	Fuerza de fricción	Mecánica I. Estática	Física conceptual
{deslizamiento, desplazamiento}	Fuerza de fricción cinética	Ingeniería mecánica. Estática	Personal
{existe, ocurre}	Fuerza de fricción cinética	Física conceptual	Personal
{suma, conjunto}	Fuerza resultante	Estática para ingenieros	Enciclopedia Salvat
{actuar, interactuar}	Fuerza resultante	Física general Alvarenga	Personal
{fuerza, atracción}	Gravedad	María Moliner	Clave
{preservar, mantener}	Inercia	Física de emergencia	El mundo de la Física 1
{incapacidad, resistencia}	Inercia	DRAE 92	Clave
{causa, ayuda}	Inercia	María Moliner	Larousse
{variar, modificar}	Inercia	Clave	Larousse
{continuar, mantener}	Inercia	Dictionary of Physics	El mundo de la Física 1
{materia, cuerpo}	Inercia	El mundo de la Física 1	Física Weber
{región, lugar}	Interferencia	Dictionary of Physics	Física de emergencia
{punto, lugar}	Interferencia	Vox	Física de emergencia
{sentido, signo}	Ley de la acción	Física general Alvarenga	Enciclopedia Salvat
{acción, fuerza }	Ley de la acción	Física. Conceptos y aplicaciones	Enciclopedia Salvat
{permanecer, conservarse}	Ley de la conservación del momento angular	Encarta 98	Fundamentos de Física
{deducir, establecer}	Leyes de Kepler	Dictionary of Physics	Física de emergencia
{agente, fenómeno}	Luz	DRAE 92	Física de emergencia
{impresionar, afectar}	Luz	Física general aplicada	Física de emergencia
{intervalo, rango}	Luz	Física. Fundamentos y aplicaciones	Dictionary of Physics
{actuar, incidir}	Luz	Vox	Larousse
{ojo, retina}	Luz	Vox	Larousse
{trasladarse, desplazarse}	Marco de referencia inercial	Dinámica (Interamericana)	Ingeniería mecánica. Dinámica
{grupo, sistema}	Marco de referencia	Física moderna	Personal
{trasladarse, moverse}	Marco de referencia inercial	Dinámica (Interamericana)	Personal
{rapidez, velocidad}	Marco de referencia inercial	Mecánica para ingenieros Singer	Física para ciencias e Ingeniería

{coordenadas, referencia}	Marco de referencia inercial	Ingeniería mecánica. Dinámica	Personal
{coordenadas, ejes}	Marco estacionario	Ingeniería mecánica. Dinámica	Mecánica vectorial. Dinámica
{cuerpo, objeto}	Masa	Ingeniería mecánica. Dinámica	Dictionary of Physics
{inercia, resistencia}	Masa	Física, fundamentos y fronteras	Britanica.com
{cuantitativa, numérica}	Masa	Física, fundamentos y fronteras	Física Weber
{imprimir, aplicar}	Masa	María Moliner	Personal
{contener, poseer}	Masa	Vox	Clave
{básica, fundamental}	Masa puntual	Dictionary.msn.com	Harcourt.com
{básica, elemental}	Masa puntual	Dictionary.msn.com	Personal
{desplazamiento, movimiento}	Movimiento armónico	Ingeniería mecánica. Dinámica	Física. Conceptos y aplicaciones
{oscilatorio, periódico}	Movimiento armónico	Ingeniería mecánica. Dinámica	Mecánica vectorial. Dinámica
{vacío, espacio}	Onda	Larousse	Física de emergencia
{vibración, perturbación}	Onda	Clave	Física de emergencia
{amplitud, enlogación}	Onda senoidal pura	Diccionario Mc. Graw Hill de Física	Diccionario de Física Gran Vox
{masa, objeto}	Órbita elíptica	Science Study Comitte	Personal
{importante, grande}	Einstein, Alberto (1879-1955)	Física de emergencia	Diccionario de Física Gran Vox
{científicos, físico}	Einstein, Alberto (1879-1955)	Física de emergencia	Diccionario de Física Gran Vox
{aportaciones, contribuciones}	Galilei, Galileo (1564, 1642)	Enciclopedia Salvat	Encarta 98

Pares semánticos compuestos identificados manualmente

PAR SEMÁNTICO	TÉRMINO	DEFINICIÓN 1	DEFINICIÓN 2
{capaz de producirlo, que lo produce}	Mecánica	Gran diccionario enciclopédico	Vox
{tratado, estudio}	Mecánica	Tratado popular de Física	Dictionary of Physics
{tratado, disciplina que estudia	Mecánica	Tratado popular de Física	Larousse
{tratado, parte … que estudia}	Mecánica	Tratado popular de Física	Física de emergencia
{disciplina que estudia, estudio}	Mecánica	Larousse	Dictionary of Physics
{estudio, parte … que estudia}	Mecánica	Dictionary of Physics	Física de emergencia
{con independencia, independiente}	Cinemática	Larousse	Atlas de Física
{describir, ocuparse de}	Cinemática	CIm1	Cinemática y dinámica básica
{sin considerar, sin relacionarlo con}	Cinemática	Desconcida	Clave
{considerar, tener en cuenta}	Cinemática	Desconcida	Física. Fundamentos y aplicaciones
{considerar, atenerse a}	Cinemática	Desconcida	Definición en la WEB
{considerar, atender a}	Cinemática	Desconcida	Cinemática y dinámica básica
{atenerse a, atender a}	Cinemática	Definición en la WEB	Cinemática y dinámica básica
{sin tener en cuenta, sin importar}	Cinemática	Física. Fundamentos y aplicaciones	Física de emergencia
{tener en cuenta, atender a}	Cinemática	Física. Fundamentos y aplicaciones	Cinemática y dinámica básica
{tener en cuenta, atenerse a}	Cinemática	Física. Fundamentos y aplicaciones	Definición en la WEB
{atenerse a, atender a}	Cinemática	Definición en la WEB	Cinemática y dinámica básica
{atenerse a, relacionar(lo) con}	Cinemática	Definición en la WEB	Clave
{atenerse a, importar}	Cinemática	Definición en la WEB	Física de emergencia
{atender a, prescindir de}	Cinemática	Cinemática y dinámica básica	DRAE 92
{atender a, relacionar(lo) con}	Cinemática	Cinemática y dinámica básica	Clave
{atender a, importar}	Cinemática	Cinemática y dinámica básica	Física de emergencia
{prescindir de, sin importar}	Cinemática	DRAE 92	Física de emergencia
{prescindir de, sin considerar}	Cinemática	DRAE 92	Desconocida
{prescindir de, sin tener en cuenta}	Cinemática	DRAE 92	Física. Fundamentos y aplicaciones
{prescindir de, sin atenerse a}	Cinemática	DRAE 92	Definición en la WEB

{prescindir de, sin relacionarlo con}	Cinemática	DRAE 92	Clave
{sin relacionarlo con, sin importar}	Cinemática	Clave	Física de emergencia
{tratar de, ocuparse del estudio}	Cinemática	Vox	María Moliner
{relacionar, estudiar conjuntamente}	Dinámica	CIm1	Física universitaria Sears
{relacionar, estudiar la relación}	Dinámica	CIm1	Atlas de Física
{estudiar, tratar de}	Dinámica	Física Cutnell-Limusa	DRAE 92
{estudiar, ocuparse de}	Dinámica	DRAE 92	María Moliner
{estudia, ayuda a estudiar}	Dinámica	DRAE 92	Personal
{tratar de, ocuparse de}	Dinámica	DRAE 92	María Moliner
{en torno, alrededor}	Velocidad angular	DRAE 92	Larousse
{móvil, partícula vibrante}	Período	CIm1	María Moliner
{vuelta, vibración completa}	Período	CIm1	María Moliner
{revolución, vibración completa}	Período	CIm1	María Moliner
{vibración, oscilación}	Período	Vox	Larousse
{oscilación, movimiento periódico}	Período	Dictionary of Physics	Física de emergencia
{que tiene, de}	Energía	CIm1	Encarta 98
{sistema físico, materia}	Energía	CIm1	María Moliner
{sistema físico, sistema de cuerpos}	Energía	CIm1	Atlas de Física
{sistema físico, sistema material}	Energía	CIm1	Diccionario enciclopédico Planeta
{capacidad, causa capaz}	Energía	CIm1	Clave
{causa capaz, aptitud}	Energía	DRAE 92	María Moliner
{materia, sistema material}	Energía	María Moliner	Diccionario enciclopédico Planeta
{fenómenos físicos o químicos, trabajo}	Energía	María Moliner	Física Weber
{materia, cuerpo}	Energía	María Moliner	Física Weber
{materia, sistema físico de cuerpos}	Energía	María Moliner	Física Weber
{etc, otra transformación}	Energía	María Moliner	Larousse
{sistema material, sistema de cuerpos}	Energía	Diccionario enciclopédico Planeta	Atlas de Física
{para, en virtud de la cual se puede}	Energía	Diccionario enciclopédico Planeta	Física Weber
{cuerpos, sistema material}	Energía	Diccionario enciclopédico Planeta (2)	Encarta 98
{cuerpos, sistema físico}	Energía	Desconocida	Larousse

{de, que posee}	Energía potencial gravitacional	Enciclopedia Salvat	Personal
{por encima, sobre}	Energía potencial gravitacional	Física aplicada	Física General Bueche-Schaum
{debido a, por}	Energía potencial gravitacional	Desconocida	María Moliner
{se mantiene, es}	Movimiento rectilineo y uniforme	CIm1	Vox
{movimiento rectilíneo uniforme, velocidad constante}	Primera ley de Newton	Física universitaria Sears	Dictionary of Physics
{rectilíneo, en línea recta}	Primera ley de Newton	Física universitaria Sears	El mundo de la Física 1
{continuar, tender a conservar}	Primera ley de Newton	Física universitaria Sears	Física de emergencia
{no cambia, continúa}	Primera ley de Newton	CIm1	Física universitaria Sears
{no cambia, preserva}	Primera ley de Newton	CIm1	Física moderna
{a menos que, a no ser que}	Primera ley de Newton	Física moderna	Física de emergencia
{preservar, tender a conservar}	Primera ley de Newton	Física moderna	Física de emergencia
{movimiento uniforme, velocidad constante}	Primera ley de Newton	Física moderna	Dictionary of Physics
{aplicarse, actuar sobre}	Primera ley de Newton	Física moderna	Dictionary of Physics
{fuerza externa, fuerza neta}	Primera ley de Newton	Dictionary of Physics	El mundo de la Física 1
{acción, fuerza de acción}	Tercera ley de Newton	CIm1	Física de emergencia
{reacción, fuerza de reacción}	Tercera ley de Newton	CIm1	Física de emergencia
{mismas velocidades, velocidad constante}	Principio de Galileo	CIm1	Introducción al estudio de la mecánica, materia y ondas
{que posee, poseída por}	Energía cinética	Desconocida	María Moliner
{debido a, ocasionado por}	Energía cinética	Desconocida	Personal
{debido a, en virtud de}	Energía cinética	Desconocida	Vox
{debido a, ocasionado por}	Energía cinética	Desconocida	Personal
{debido a, por razón de}	Energía cinética	Encarta 98	DRAE 92
{en virtud de, ocasionado por}	Energía cinética	Diccionario enciclopédico Planeta	Personal
{en virtud de, por razón de}	Energía cinética	Diccionario enciclopédico Planeta	DRAE 92
{por, en virtud de}	Energía cinética	María Moliner	Larousse
{poseída por, que tiene}	Energía cinética	María Moliner	Física de emergencia
{proporcionalmente, de manera proporcional}	Movimiento uniformemente retardado	Desconocida	Larousse

{poderse transformar, poder ser transformado}	Ley de la conservación de la energía	CIm1	Física Resnick
{ser transformado, transformarse}	Ley de la conservación de la energía	Física Resnick	Enciclopedia Lexis 22
{ser creado, crearse}	Ley de la conservación de la energía	Física Resnick	Enciclopedia Lexis 22
{destruirse, se destruye}	Ley de la conservación de la energía	Enciclopedia Lexis 22	Personal
{transformarse, se transforma}	Ley de la conservación de la energía	Enciclopedia Lexis 22	Personal
{transformarse, poderse transformar}	Ley de la conservación de la energía	Enciclopedia Lexis 22	Física general Alvarenga
{poder crearse, crearse}	Ley de la conservación de la energía	Enciclopedia Lexis 22	Personal
{se puede transformar, se transforma}	Ley de la conservación de la energía	Física general Alvarenga	Física de emergencia
{puede ser creada, se crea}	Ley de la conservación de la energía	Física general Alvarenga	Física de emergencia
{puede ser destruida, se destruye}	Ley de la conservación de la energía	Física general Alvarenga	Física de emergencia
{cuerpo grave, cuerpo pesado}	Péndulo	DRAE 92	María Moliner
{cuerpo grave, cuerpo rígido}	Péndulo	DRAE 92	Larousse
{cuerpo pesado, cuerpo rígido}	Péndulo	María Moliner	Física de emergencia
{suspendido, que cuelga}	Péndulo	María Moliner	Física de emergencia
{acción de la gravedad, acción de su peso}	Péndulo	María Moliner	Física de emergencia
{acción de la gravedad, acción de su peso}	Péndulo	María Moliner	Física de emergencia
{libremente, con libertad}	Péndulo	Clave	Larousse
{trayectoria curva, circunferencia}	Aceleración centrípeta	Física de emergencia	Diccionario enciclopédico de Física
{estar dirigido hacia, buscar}	Aceleración centrípeta	Diccionario enciclopédico de Física	El mundo de la Física 1
{que dista, distante}	Afelio	DRAE 92	María Moliner
{dista, se aleja}	Afelio	DRAE 92	Larousse
{dista, se encuentra lejos}	Afelio	DRAE 92	Física de emergencia
{se aleja, se encuentra lejos}	Afelio	Larousse	Física de emergencia
{sol, objeto central}	Afelio	Larousse	Física de emergencia

{órbita, trayectoria elíptica}	Afelio	Vox	Física de emergencia
{alejado, que se aleja}	Afelio	Vox	Larousse
{alejado, que dista}	Afelio	Vox	DRAE 92
{debido a, originado … por}	Calor	Salvat Multimedia	Vox
{manifestación de, forma de}	Calor	Oceano Uno Color	María Moliner
{es transferida, cede}	Calor	Dictionary of Physics	Atlas de Física
{pasa de, es transferida de}	Calor	DRAE 92	Dictionary of Physics
{atómicomolecular, de las moléculas}	Calor	Vox	Física Weber
{originada probablemente, asociada}	Calor	Vox	Física Weber
{es lo mismo que, [es] equivlaente a}	Centro de masa	Diccionario Oxford-Complutense	Diccionario de Física Vox
{colisión, encuentro violento}	Choque	Física general Alvarenga	DRAE 92
{pasar por los bordes, bordear}	Difracción	Vox	María Moliner
{de, que sufre}	Difracción	Vox	María Moliner
{rayos luminosos, luz}	Difracción	Vox	Física Weber
{pasar por los bordes, rozar}	Difracción	Vox	DRAE 92
{pasar por los bordes, rozar el borde}	Difracción	Vox	Larousse
{luminosos, de luz}	Difracción	Vox	Larousse
{rayo luminoso, onda}	Difracción	Vox	Física de emergencia
{rayo luminoso, radiación}	Difracción	DRAE 92	Dictionary of Physics
{rozar el borde, bordear}	Difracción	DRAE 92	María Moliner
{rayo de luz, luz}	Difracción	Larousse	Física Weber
{sirven para, se consideran para}	Dimensión	DRAE 92	María Moliner
{se da a, recibe}	Dimensión	Antecedentes de Física	Personal
{formar, poder expresar}	Dimensión	Antecedentes de Física	Personal
{eje horizontal, eje x}	Dirección de un vector	Física Universitaria Benson	Física universitaria Sears
{varilla, pieza cilíndrica}	Eje de giro	Océano Uno Color	Encarta 98
{moverse en círculos, rodar}	Eje de rotación	Mecánica vectorial. Dinámica	Personal
{producida por, debida a}	Energía eléctrica	Larousse	Física de emergencia
{asociada con, producida por}	Energía eléctrica	Diccionario de energía	Larousse
{cargas eléctricas, campo eléctrico}	Energía eléctrica	Larousse	Física de emergencia
{asociada con, debida a}	Energía eléctrica	Diccionario de energía	Física de emergencia
{energía cinética, movimiento}	Energía mecánica	Personal	María Moliner

{que deriva, producida}	Energía química	Gran enciclopedia Larousse	Larousse
{que prodede, producida}	Energía química	Personal	Larousse
{asociada a, que procede de}	Energía térmica	Física con aplicaciones	María Moliner
{sustancia, compuesto químico}	Energía química	Química Chang	Dictionary of Physics
{tres, terna de}	Espacio	Enciclopedia Ciencia y tecnología	Personal
{cambia, no permanece}	Estado de movimiento	Física I. Guía escolar Vox	Física Cutnell-Limusa
{experimenta un cambio, no permanece}	Estado de movimiento	Física con aplicaciones	Física Cutnell-Limusa
{no cambia, permanece}	Estado de reposo	Enciclopedia Lexis 22	Lexipedia
{resulta de, obtenido por}	Éter	Master. Diccionario enciclopédico	Lexipedia
{obtenido, que se obtiene}	Éter	Lexipedia	Personal
{se supone, se cree}	Éter	Vox	Larousse
{puede definirse como, se define como}	Física	Física. Conceptos y aplicaciones	Personal
{se ocupa de, investiga}	Física	Encarta 2000	Física. Conceptos y aplicaciones
{se ocupa de, se dedica a estudiar}	Física	Encarta 2000	El mundo de la Física 1
{materia, propiedades de la materia}	Física	Vox	DRAE 92
{sistema vibrante, circuito resonante}	Frecuencia de resonancia	Diccionario de electrónica	Electrónica y técnica nuclear
{fuerza de contacto, interacción de contacto}	Fricción	Física general Benson	Física Resnick
{producir una variación, cambiar}	Fuerza	Gran enciclopedia del mundo	Vox
{modificar la forma de, deformar}	Fuerza	Clave	Larousse
{deformar, modificar la forma}	Fuerza	Larousse	Clave
{debe ser aplicada, es preciso aplicar}	Fuerza centrípeta	Física Serway	DRAE 92
{adentro, el centro}	Fuerza centrípeta	Física Resnick	Física para Ingeniería y ciencias
{depende sólo, es independiente}	Fuerza conservativa	Mecánica vectorial. Estática	Física Resnick
{es independiente, no depende}	Fuerza conservativa	Física Resnick	Diccionario de Física Vox
{evitar, oponerse}	Fuerza de fricción	Ingeniería mecánica. Estática	El mundo de la Física 1
{retardar, oponerse}	Fuerza de fricción	Ingeniería mecánica. Estática	El mundo de la Física 1
{están en contacto, se tocan}	Fuerza de fricción	Mecánica para ingenieros Huang	Física conceptual
{encontrarse en contacto, tocarse}	Fuerza de fricción	Ingeniería mecánica. Estática	Física conceptual

{trayectoria seguida, camino recorrido}	Fuerza disipativa	Física Serway	Física general Alvarenga
{tener, manifestarse en}	Inercia	Vox	El mundo de la Física 1
{incapacidad … de modificar, tendencia … a preservar}	Inercia	Vox	Física de emergencia
{incapacidad … de modificar, tendencia … a mantener}	Inercia	Vox	El mundo de la Física 1
{movimiento, velocidad constante}	Inercia	Vox	Dictionary of Physics
{corresponde, se opone}	Ley de la acción	Física. Conceptos y aplicaciones	Enciclopedia Salvat
{puntos de masa, cuerpos}	Ley de la gravitación universal	Dictionary of Physics	Atlas de Física
{onda electromagnética, radiación electromagnética}	Luz	Física con aplicaciones	Física. Fundamentos y aplicaciones
{energía radiante, radiación}	Luz	Física general aplicada	Larousse
{energía radiante, forma de movimiento}	Luz	Física general aplicada	El mundo de la Física 1
{radiación electromagnética, forma de energía}	Luz	Física. Fundamentos y aplicaciones	Vox
{radiación electromagnética, agente físico}	Luz	Física. Fundamentos y aplicaciones	DRAE 92
{forma de energía, radiación}	Luz	María Moliner	Larousse
{actuar sobre, afectar}	Luz	María Moliner	Física de emergencia
{forma de energía, agente físico}	Luz	Vox	DRAE 92
{forma de energía, fenómeno físico}	Luz	Clave	Física de emergencia
{moverse, describirse los movimientos}	Marco de referencia	Curso de Física moderna	Salvat Multimedia
{ocurrir, se describirse}	Marco de referencia	Dinámica (Interamericana)	Salvat Multimedia
{describir, permitir conocer}	Marco de referencia	Dinámica (Interamericana)	Personal
{marco de referencia, sistema}	Marco de referencia inercial	Dinámica (Interamericana)	Ingeniería mecánica. Dinámica
{ser capaz de, poder}	Masa	Tratado popular de Física	Personal
{masa, cantidad de materia}	Masa total de un sistema	Personal	Gran diccionario enciclopédico
{estar en movimiento, moverse}	Móvil	Vox	Física de emergencia
{(fuerza) que comunica, (fuerza) necesaria para comunicar}	Newton (unidad)	Vox	Clave
{equivale a, es}	Newton (unidad)	Larousse	Física de emergencia
{movimiento vibratorio, perturbación periódica}	Onda	Larousse	Física de emergencia
{movimiento vibratorio, vibración periódica}	Onda	Larousse	Clave

{vibración, movimiento vibratorio}	Onda	Clave	Larousse
{que recorren, descrita por}	Órbita	DRAE 92	Clave
{de, que recorre}	Órbita	Vox	DRAE 92

APÉNDICE C

RESUMEN DE LAS PROPUESTAS VERTIDAS EN EL CAPÍTULO 3

En este apéndice se encontrará un resumen de las propuestas concernientes a los ejes trabajados en el capítulo 3: Perífrasis gramaticales, nexos, adjetivación, adverbios, negación, sintagmas enfáticos, determinantes, relaciones léxicas de hiperonimia e hiponimia, abreviaturas, siglas y símbolos, patrones definitorios y consideraciones para mejorar el funcionamiento global de Clustering.

En cada caso se proporciona el nombre del eje, su descripción y el comportamiento que manifiesta aquello a lo que hace referencia dentro de las actuales versiones del programa de agrupamiento semántico.

Resumen de las propuestas vertidas en el capítulo 3

Nombre del eje	Descripción	Comportamiento actual	Propuesta
Perífrasis gramatical (3.1)	Se refiere a aquellas construcciones gramaticales formadas por más de una palabra ortográfica que se comportan como unidad indivisible para efectos semánticos y sintácticos. Consideramos tres tipos de perífrasis:	Las palabras que hacen parte de una perífrasis se agrupan por separado. En consecuencia, su presencia desencadena desajustes dentro de los alineamientos.	
	Locuciones (3.1.1)		- Que las construcciones que han mostrado inestabilidad en la escritura como *a cabo* y *a partir* sean tratadas como unidades.
	Términos compuestos (3.1.2)		- Localizar los términos compuestos insertos en las definiciones. (Esta identificación podrá realizarse automáticamente en un corto plazo). - Que provisionalmente se anexe, para cada base de definiciones, una lista con los términos compuestos que sean entrada de diccionarios especializados en la materia respectiva.
	Perífrasis verbales (3.1.3)		- Generar los mecanismos para la identificación automática de las perífrasis verbales, considerando como tales a toda reunión de formas verbales que en conjunto refieran a un solo proceso. - Identificar, además, los tiempos compuestos y los verbos pronominales a partir de información gramatical.

Nombre del eje	Descripción	Comportamiento actual	Propuesta
Nexos (3.2)	En esta categoría funcional se reúnen los elementos gramaticales que sirven para enlazar a otros. Las clases de palabras que funcionalmente actúan como nexos son:		
	Conjunciones (3.2.1)	No se vinculan palabras funcionalmente equivalentes porque el programa no distingue si las conjunciones afectan a palabras, frases u oraciones, o a estructuras de diferente tipo.	- Implementar los mecanismos para identificar a qué tipo de sintagmas enlaza una conjunción coordinante. - Las palabras unidas mediante una conjunción coordinante dentro de una definición deben poderse alinear por duplicado con una sola palabra en la definición con la que se compara. - Marcar las locuciones conjuntivas con una etiqueta que especifique su función.
	Preposiciones (3.2.2)	No reciben ningún tratamiento especial las preposiciones cuando obligatoriamente acompañan a un verbo.	- Incorporar al motor del sistema una lista con verbos que rigen preposición para que éstas sean tratadas como un incremento verbal y se ubiquen en el mismo espacio dentro de los alineamientos. - Marcar las locuciones prepositivas con una etiqueta que especifique su función.
	Verbos copulativos (3.2.3)	Cuando el verbo copulativo de una definición se alinea con un conjunto vacío, el sistema registra un par nulo.	- Que los verbos copulativos (y, en su caso, el nexo que los precede) que aparecen alineados con conjuntos vacíos, se consideren pares semi nulos.

Nombre del eje	**Descripción**	**Comportamiento actual**	**Propuesta**
Adjetivación (3.3)	Se considera adjetivo a toda unidad gramatical, independientemente de su clase sintagmática, que desempeña la función de modificador nominal.	El sistema no realiza ninguna identificación en este sentido.	- Etiquetar las definiciones con un *chunking*, para así localizar frases y determinar con qué categoría gramatical se corresponden. - Considerar, en un primer momento, adjetivos, frases adjetivas y frases preposicionales en función de complemento adnominal como equivalentes en los alineamientos.
Adverbios (3.4)	Se considera adverbio a toda unidad gramatical, independientemente de su clase sintagmática, que desempeña la función de adyacente circunstancial.	No se hace ningún reconocimiento de este tipo.	- Etiquetar las definiciones con un *chunking* con el fin de localizar frases y determinar con qué categoría gramatical se corresponden. - Habilitar el sistema para alinear frases adverbiales con adverbios.
Negación (3.5)	Trata de las dificultades que se presentan en los alineamientos cuando en las definiciones se aprovecha alguna forma de negación para apoyar una descripción.	El hecho de que una definición afirmativa se alinea con otra donde el significado del término se expresa negando ciertos atributos redunda en malos pares.	- Tratar el adverbio de negación como un elemento constitutivo del núcleo del enunciado. - Introducir a las definiciones donde aparecen estructuras con palabras típicamente negativas cuyo significado es positivo una etiqueta semántica que indique cuál es el sentido de la expresión. Entonces, estas negaciones quedarían exentas del tratamiento anteriormente mencionado.
Sintagmas enfáticos (3.6)	Se refiere a todas aquellas estructuras cuya función es puramente discursiva en un contexto dado.	El programa los alinea con palabras que cumplen otras funciones.	- Identificar los sintagmas enfáticos. - Priorizar el alineamiento de sintagmas enfáticos con espacios vacíos. Tratar los pares resultantes como semi nulos.

Nombre del eje	**Descripción**	**Comportamiento actual**	**Propuesta**
Determinantes (3.7)	La determinación es una categoría funcional que abarca las operaciones deícticas, anafóricas, referenciales y de cuantificación. Las categorías gramaticales que cumplen estas funciones son artículos, demostrativos, pronombres, posesivos e indefinidos.	El algoritmo alinea palabras pertenecientes a esta categoría con palabras de cualquier otra categoría. Cuando dos palabras de este tipo se unen en un par, el algoritmo básico lo puede considerar vinculado; el flexibilizado, le da el tratamiento de par semi igual pero sin hacer ninguna clasificación entre tipos de determinantes.	- Que los pares correspondientes formados por palabras pertenecientes a estas categorías sean considerados semi iguales por el sistema siempre y cuando los dos miembros del par se identifiquen con la misma categoría gramatical.
Relaciones léxicas de hiperonimia e hiponimia (3.8)	Se trata de aquellas palabras que describen un solo concepto refiriéndose al todo por la parte y viceversa.	No se hace ningún reconocimiento en este sentido. Cuando se encuentran en un par es por la posición que ocupan en el alineamiento.	- Localizar, mediante redes semánticas, los pares de palabras que establecen relaciones de este tipo. - Contribuir, a partir de ellos, al acomodo del alineamiento.

Nombre del eje	Descripción	Comportamiento actual	Propuesta
Abreviaturas, siglas y símbolos (3.9)	Contempla las situaciones en que una definición aprovecha abreviaturas (que pueden ser siglas o símbolos) y la otra presenta las expresiones destrabadas; o bien, se muestran concatenadas ambas formas en una definición.	No identifica como equivalentes expresiones sintéticas y extensas.	- Etiquetar las abreviaturas con el nombre del término que les corresponde. - Considerar pares iguales los formados por diferentes manifestaciones del mismo término. - Que las abreviaturas precedidas por la expresión amplia ocupen un solo espacio en los alineamientos. - Que las fórmulas se traten unitariamente. - Separar las contracciones en sus morfemas originales.
Patrones definitorios (3.10)	Se refiere a las palabras y locuciones que, en la redacción de un texto, sirven para introducir un concepto.	Las trata como parte de la definición.	- Introducir al sistema una lista amplia de patrones. - Tratar como semi iguales y semi nulos a los pares formados con patrones definitorios.

Nombre del eje	Descripción	Comportamiento actual	Propuesta
Consideraciones varias para mejorar el funcionamiento global de Clustering (3.11)	Contempla cuatro observaciones sobre el desempeño del algoritmo que en principio no son sintácticas:		
	Signos de puntuación (3.11.1)	Elimina los signos de puntuación.	- Mantener la información sobre la función gramatical que imprimen los signos de puntuación.
	Cálculo de LCC (3.11.2)	Contabiliza las palabras correspondientes al término para la asignación de valores de LCC.	- Eliminar del cómputo de LCC a los términos.
	Reconsideración de pares semi iguales y semi nulos (3.11.3)	El algoritmo básico no los considera. El algoritmo flexibilizado agrupa las palabras funcionales pertenecientes a una lista sin ninguna restricción.	- Que las palabras funcionales se agrupen sólo si pertenecen a la misma categoría: determinantes, pronombres, conjunciones y preposiciones. - Que los pares semi iguales se formen con dos palabras de base gramatical, aunque se trate del mismo lexema. - Que ni los pares semi iguales ni los pares semi nulos sumen puntos para la asignación de valores de LCC.
	Reconsideración para la aplicación de intercambios (3.11.4)	Realiza intercambios entre cualquier secuencia de dos palabras con apariencia semejante o cuando aparece la conjunción “y” entre ellas.	- Que el sistema aproveche las etiquetas que hacen patente la categoría gramatical de las palabras para aplicar a esta operación ciertas restricciones que la misma lengua impone a los intercambios. - Extender el intercambio conjuntivo al nivel sintagmático de la frase donde los núcleos de dos de ellas relacionadas coordinantemente son susceptibles de intercambio.