2 TERMINOLOGÍA

La palabra *terminología* en una primera instancia se puede considerar como la materia de intersección que se ocupa de la designación de los conceptos de las lenguas de especialidad (Cabré, 1992). Un *lenguaje especializado* es un lenguaje que se usa en un campo del conocimiento y que se caracteriza por el uso de medios específicos de expresión lingüística (ISO 1087-1:2000, 2000). Por tanto, en otras palabras, la terminología, como disciplina, es una materia interdisciplinaria que se encarga de designar conceptos del lenguaje que se emplean en los campos del conocimiento y que tienen características específicas que las diferencian de la lengua general o cotidiana.

A lo largo de este tercer capítulo se abordará la terminología no solamente como disciplina sino también sus aplicaciones, su relación con el procesamiento de lenguaje y la aplicación de ambas materias en diversos sistemas y herramientas.

2.1 Terminología y terminografía

La *Terminología*, no sólo designa a una disciplina, sino también define el conjunto de unidades léxicas usadas con un valor preciso en los ámbitos de especialidad (Cabré, 1992). Es decir, todo el grupo de conceptos que la terminología, como disciplina, designa. Tomando en cuenta lo anterior, para Cabré (1992), existen cuatro puntos que muestran los distintos enfoques sobre el estudio y la práctica de la terminología:

- Para los lingüistas, la terminología es una parte del léxico delimitada por criterios temáticos y pragmáticos.
- Para los especialistas, la terminología es el reflejo formal de la organización conceptual de una especialidad, y un medio inevitable de expresión y de comunicación profesional.
- Para los usuarios (directos e intermediarios), la terminología es un conjunto de unidades de comunicación, útiles y prácticas, cuyo valor se mide en función de criterios de economía, de precisión y de adecuación.

- Para los planificadores lingüísticos, la terminología es un ámbito del lenguaje donde se debe intervenir para reafirmar la existencia, la utilidad y la pervivencia de una lengua, y para garantizar, mediante su modernización, su continuidad como medio de expresión.

2.1.1 Los términos

Una *unidad terminológica*, o *término*, es un símbolo convencional que representanta una noción definida en un cierto dominio del saber (Lérat, 1989). La unión de varios términos, forman la terminología del dominio de especialidad.

Existen distintos tipos de términos, estos se suelen clasificar de distinta manera, en torno a cuatro aspectos que son forma, función, significado y procedencia (Cabré, 1992).

El aspecto de forma es un conjunto de criterios que no son necesariamente excluyentes y que expresan la manera en que un término puede estar conformado. Estos criterios son los siguientes:

- **Número de morfemas[7]:** Dependiendo del número de morfemas un término puede ser simple o complejo. Ejemplo: cuadern-*o*, cuadern-*os*, *en*-cuadern-*ado.*
- **Tipos de morfemas:** Los distintos tipos de morfemas existentes en un término complejo determinan si es derivado o compuesto. Ejemplos de derivados son fruter-*ía*, libr-*ero*, verd-*oso*. En cambio, algunos ejemplos de términos compuestos son para-*brisas*, saca-*corchos*, balon-*cesto*.
- **Estructura:** Existen términos complejos que son la combinación de palabras que siguen una determinada estructura sintáctica. Algunos ejemplos de estructuras que se emplean en el español son sustantivo-preposición-sustantivo (método de Newton-Raphson), sustantivo-adjetivo (cristal líquido).

[7] Según el diccionario de la Real Academia Española un morfema es la unidad mínima analizable que posee sólo significado gramatical. En otras palabras es la parte variante de la palabra que otorga un significado y permite formar nuevas palabras. Ejemplo: niñ-*o*, niñ-*a*, niñ-*os*, niñ-*as*

- **Origen complejo:** En algunos casos los términos simples provienen de términos complejos; casos de este criterio son las abreviaturas (Del., av.), las siglas (SIDA, ONU), acrónimos (bit, sonar) o formas abreviadas (tele, cine).

El segundo aspecto existente es el de función, es decir, los términos siempre tienen una función determinada en las oraciones. Estas funciones pueden ser de nombres, adjetivos, verbos y adverbios. En el caso de las palabras funcionales, como las preposiciones, conjunciones, artículos, entre otros, Cabré (1992) indica que no tienen un carácter terminológico.

El aspecto siguiente es el de significado, el cual indica que un término denomina una determinada clase de conceptos. Para Cabré (1992) se pueden establecer cuatro grandes clases conceptuales que son las siguientes:

- Objetos o entidades: Nombres.
- Procesos, operaciones o acciones: Verbos, nominalizaciones de verbos[8].
- Propiedades, estados, cualidades: Adjetivos.
- Relaciones: Adjetivos, verbos.

El último aspecto que es mencionado por Cabré es el de procedencia lingüística, es decir, los términos pueden ser creados o construidos a partir de reglas del propio lenguaje o provenir de otras lenguas.

2.1.2 La terminografía

La *terminografía* es la rama aplicada de la terminología que se ocupa de la elaboración de diccionarios especializados o de glosarios terminológicos (Cabré, 1995). Esta tarea incluye además la compilación, la sistematización y la presentación de los términos de las áreas de especialización.

[8] Es el proceso de convertir un verbo en un sustantivo, por ejemplo gotear goteo.

Aunque la tarea de la terminografía es similar al de la lexicografía[9] (el de crear diccionarios y glosarios), estas dos tareas difieren en el método que emplean, la forma en que emplean los datos y la manera en que presentan los resultados.

Mientras que la lexicografía sigue un proceso semasiológico, es decir, a partir del término crea la definición; la terminografía parte de la definición o de una lista de conceptos para determinar su término (que corresponda a la forma en que se emplea en el área especializada), es decir, sigue un proceso onomasiológico.

De igual forma, dentro del proceso de la terminografía se lleva a cabo una normalización, esto quiere decir que se busca estandarizar los términos que se emplean dentro de un área especializada para conseguir una comunicación profesional precisa, moderna y unívoca (Cabré, 1995).

El proceso de la terminografía está conformado por seis fases que son las siguientes (Cabré, 1992):

- **Definición y delimitación del trabajo:** En esta primera fase se debe definir el tema a trabajar, cuál es el público al que va dirigido, cuál es la función que va a tener el trabajo y el alcance de la obra en función de las condiciones anteriores, pero también de las económicas, temporales, materiales, académicas, entre otras.
- **Preparación del trabajo:** Consiste en adquirir y reunir toda la información sobre el tema a trabajar, en la selección de asesores de trabajo, en la estructuración que se va emplear y en la propuesta del plan de trabajo.
- **Elaboración de la terminología:** En la tercera fase de la terminografía se localizan los términos en el corpus y se determina que pertenezcan al área analizada.
- **Presentación del trabajo:** En esta fase se crea la publicación que contendrá el trabajo realizado en las etapas anteriores.

[9] Es la rama aplicada de la lexicología. Según la RAE la lexicología es el estudio de las unidades léxicas de una lengua y de las relaciones sistemáticas que se establecen entre ellas.

- **Supervisión del trabajo:** Durante esta fase se juntan los expertos en terminología y los del área determinada para supervisar que el trabajo realizado no tenga problemas y sea el adecuado.
- **Tratamiento y resolución de los casos problemáticos:** Si existen casos problemáticos es necesario resolverlos; para ello se emplean diversos caminos dependiendo del caso, como consultar bibliografía complementaria, consultar a especialistas en la materia, lexicógrafos, especialistas multilingües o consultar a organismos oficiales de normalización.

2.1.3 Extracción de información terminológica

El desarrollo de nuevas materias de investigación y aplicación, como la informática o las ciencias computacionales, y su incursión dentro de diversas áreas, han hecho que muchas materias de investigación cambien su metodología, planteamiento o rendimiento. La terminología no es la excepción, ya que en la actualidad existe la terminótica. Para Cabré (1992) la *terminótica* es la materia que se ocupa, en general, de las relaciones entre la informática y la terminología; y, en particular, que trata de la aplicación de la informática al trabajo terminológico.

Esta incursión de la informática en el área de la terminología, de manera más específica en la terminografía ha adquirido cierto protagonismo en algunas de las tareas que se llevan a cabo en la metodología, como la documentación previa, la constitución del corpus, la verificación de la información, entre otras tareas. Pero también la extracción de términos ha sido una de las tareas donde la informática, específicamente el PLN, participa activamente por medio de la extracción de información, esto ha desarrollado la *extracción de información terminológica*, *extracción terminológica* o *terminology extraction (TE)*.

La extracción de información terminológica es el uso de métodos propios de la extracción de información con el objetivo de extraer los términos de un corpus apoyándose en el poder de procesamiento de las computadoras.

Cabe destacar que la extracción terminológica está altamente relacionada con la recuperación de información, no solamente porque la extracción de información está relacionada con esa tarea, sino por que frecuentemente los términos (empleando su sentido de la búsqueda de información) que indizan los documentos son los términos (en su sentido lingüístico) que conforman a un documento. La única diferencia es que la extracción terminológica busca obtener todas las unidades terminológicas y no sólo las más representativas de un documento. Por tanto, son constantemente empleadas técnicas que en un principio eran solamente de indización de documentos en sistemas de extracción de terminología.

2.2 Sistemas actuales de extracción terminológica

Según Cabré et al. (2001) desde el 2000 los lingüistas computacionales, los investigadores en lingüista aplicada, traductores, intérpretes, periodistas, científicos e ingenieros en computación han estado interesados en el aislamiento automático de la terminología de textos. La razón de ello es que la terminología no sólo sirve para crear diccionarios o glosarios, también es útil en la traducción automática, en el resumen automático, en bases de conocimiento, en sistemas expertos, entre otras tareas.

Por lo anterior se han desarrollado sistemas que extraigan de manera automática la terminología de grandes cantidades de texto, de una manera rápida. Sin embargo, con el paso del tiempo los desarrolladores de los sistemas de extracción terminológica han observado que existen diversas complicaciones la cuales, según Cabré et al. (2001), son las siguientes:

- Identificación de términos complejos, es decir, se necesita reconocer cuándo una unidad discursiva[10] constituye una frase terminológica y dónde comienza y termina ésta.
- Identificación de la naturaleza terminológica de una unidad léxica[11], esto es, conocer cuando dentro de un texto especializado una unidad léxica tiene una naturaleza terminológica o pertenece al lenguaje general.
- La propiedad y conveniencia de una unidad terminológica en un vocabulario dado.

---

[10] Una unidad discursiva es una estructura que puede ser identificable dentro de un texto (http://linguistics-ontology.org/gold/DiscourseUnit).

[11] Una unidad léxica es un elemento que es objeto de definición en un diccionario, vocabulario, glosario, etcétera (Luna Trail et al., 2005).

Los sistemas de extracción terminológica se basan en tres tipos de conocimientos que son los lingüísticos, los estadísticos y los híbridos. Cada uno de estos tipos de sistemas se explicará en los apartados siguientes, además de que se darán a conocer algunos sistemas de extracción terminológica.

2.2.1 Sistemas basados en conocimiento lingüístico

Como se indicó en el apartado anterior, los sistemas de extracción terminológica se basan en distintos tipos de conocimiento y uno de ellos es el lingüístico; su razón de uso es porque la terminología y los términos están ampliamente relacionados con la lingüística.

Para Pazienza et al. (2005) los sistemas con un acercamiento lingüístico tratan de identificar términos a través de sus propiedades sintácticas, esto se debe a que frecuentemente las unidades terminológicas tienen estructuras sintácticas definidas, como se vio en la sección 2.1.1. Estos sistemas se pueden basar en dos tipos de información (Cabré et al., 2001):

- **Término específico:** Este consiste en la detección de patrones recurrentes de unidades terminológicas complejas; en la Tabla 5 podemos ver algunas estructuras empleadas en el español que definen por lo general un término; en cambio en la Tabla 6 podemos observar algunas estructuras sintácticas que por lo general no forman un término. Los patrones que se buscan provienen de reglas que se obtienen de manera empírica a través del análisis de datos y se pueden programar a través de expresiones regulares o autómatas de estados finitos.
- **Lenguaje genérico:** Consiste en la detección de estructuras lingüísticas más básicas, como los sintagmas[12] nominales (por ejemplo: libro, campo de trigo), sintagmas preposicionales (de María, para cocinar), entre otros. Para ello se emplean herramientas de PLN complejas, como son los *analizadores sintácticos*, también conocidos como *parsers*, que son herramientas que analizan la estructura de un texto con base en una gramática.

[12] Un sintagma, según la Real Academia Española, es un conjunto de palabras. Por ejemplo: un sintagma nominal está construido en torno a un nombre o sustantivo. En cambio, uno preposicional, es el formado alrededor de una preposición.

Estructura sintáctica	Ejemplos
sustantivo	agua, planeta, protozoario, cimiento
sustantivo + adjetivo	plano inclinado, agua oxigenada
sustantivo + preposición + sustantivo	lámpara de halógeno, dióxido de carbono

Tabla 5. Ejemplos de estructuras sintácticas para términos en español

Estructura sintáctica	Ejemplos
artículo + sustantivo	la casa, el niño, los países
sustantivo + y/o + sustantivo	águila o sol, coseno y tangente

Tabla 6. Ejemplos de estructuras sintácticas que no forman por lo general términos en español

Los tipos de información explicados anteriormente se basan en el análisis morfológico.

Los sistemas terminológicos basados en conocimiento lingüístico tienen como ventaja que encuentran términos sin importar su frecuencia o importancia en el texto, pues se basan en su estructura. En cambio, su desventaja, es que son propensos al *ruido*, es decir, los sistemas son proclives a encontrar estructuras falsas debido a errores en la asignación de la categoría gramatical (análisis morfológico); de igual manera, los sistemas basados en conocimiento lingüístico son dependientes de la lengua, ya que las reglas generadas pueden no servir en otras lenguas.

2.2.1.1 LEXTER

El sistema de extracción de términos LEXTER (Bourigault, 1994) fue desarrollado para el francés basándose en conocimiento lingüístico; su objetivo principal era mejorar el sistema de indización de la compañía EDF (Electricité de France).

El principio básico de LEXTER es encontrar las fronteras de los sintagmas nominales, pero en lugar de hacerlo de manera "positiva", es decir, encontrando las estructuras que emplean los términos frecuentemente en francés, se realizó de manera "negativa", en otras palabras, era encontrar estructuras sintácticas que claramente no formaran un término.

La primera tarea que realiza LEXTER es un análisis morfológico y de desambiguación para cada uno de los textos que se va a analizar. Posteriormente, el sistema busca, dentro del texto preprocesado, patrones que no sean parte de un sintagma nominal y por tanto, de un término. Algunos casos según Bourigault et al. (1996) son verbos, pronombres, preposiciones unidos a artículos posesivos, entre otros. Este proceso deja secuencias de palabras que por lo general corresponden a sintagmas nominales y son candidatos a ser términos o partes de ellos son candidatos; a este conjunto de palabras le llamaron MLNP (Maximal-Length Noun Phrases).

La segunda tarea consiste en un analizador sintáctico que analiza los MLNP para dividir candidatos terminológicos complejos en partes más sencillas llamadas cabeza (head, H) y expansión (expansion, E). El módulo del analizador sintáctico se basa en reglas, que indican qué partes son la cabeza y qué partes son la expansión del MLNP; en caso de encontrar estructuras ambiguas, existe un algoritmo de desambiguación que ejecuta distintas formas de una regla si se hallan formaciones en la estructura ambigua que ya hubieran sido encontradas durante el análisis. A continuación, en la Tabla 7 se muestra una regla no ambigua, mientras que en la Tabla 8 se ejemplifica otra donde se presentan casos de ambigüedad.

Regla no ambigua

$sustantivo_1 + adj + prep + sustantivo_2$
$\rightarrow$
Cabeza: $sustanivo_1 + adj$
        Cabeza: $sustantivo_1$
        Extensión: adj
Extensión: $sustanivo_2$

Tabla 7. Ejemplo de una regla no ambigua empleada en LEXTER

Regla ambigua	
$sustantivo_1 + \ prep + sustantivo_2 + adj$	
Caso 1	Caso 2
$\rightarrow$	$\rightarrow$
Cabeza: $sustantivo_1$	Cabeza: $sustantivo_1 + prep + sustantivo_2$
Extensión: $sustantivo_2\ adj$	Cabeza: $sustantivo_1$
Cabeza: $sustantivo_2$	Extensión: : $sustantivo_2$
Extensión: adj	Extensión: adj

Tabla 8. Ejemplo de una regla ambigua empleada en LEXTER

La tercera parte del proceso es un módulo de estructuración que emplea la información dada por el paso anterior para crear una red terminológica. Este consiste en vincular las cabezas y extensiones de términos complejos con términos menos complejos, y estos, a su vez, vincularlos con términos todavía menos complejos hasta formar una red. En la Figura 6 se muestra un ejemplo[13] de la red terminológica generada por LEXTER.

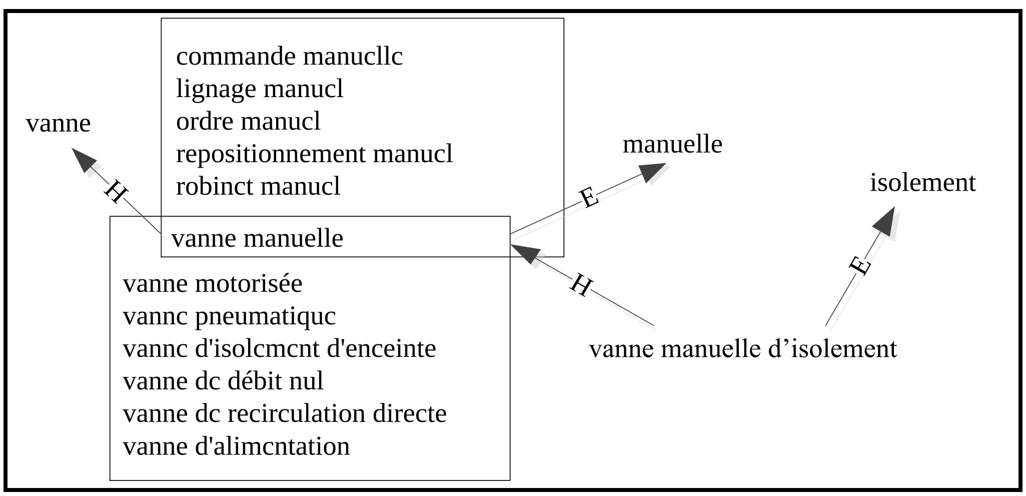

Figura 6. Ejemplo de una red terminológica creada por LEXTER

Al sistema extractor de términos LEXTER se le considera un sistema robusto, preciso e independiente del dominio desarrollado para el idioma francés. Sin embargo, LEXTER tiene algunos problemas de ruido por errores en el análisis morfológico, como ocurre en la mayoría de los sistemas basados en conocimiento lingüístico. Aun así, se le considera a este

[13] Ejemplo extraído de Bourigault et al. (1996)

extractor de términos un buen sistema por su habilidad de aprender conforme se van obteniendo unidades terminológicas.

2.2.1.2 HEID

HEID (Heid et al., 1996) es un sistema de extracción terminológica que se basa en conocimiento lingüístico para el idioma alemán. Su objetivo es aumentar la eficiencia del proceso de creación de glosarios en tareas relacionadas con la traducción de textos técnicos, en este caso de ingeniería automovilística.

El sistema de extracción está compuesto de dos partes, la primera de ellas es el análisis lingüístico y la anotación de los textos; la segunda es la extracción de términos por medio de consultas en el corpus.

El análisis lingüístico consiste en un tokenizador, un analizador morfosintáctico[14], un etiquetador POS[15] y un lematizador que se ejecutan al inicio del análisis. Posteriormente se extraen construcciones características de los sintagmas nominales, esto se debe a que no existía en el momento del desarrollo del extractor terminológico un analizador sintáctico de cobertura amplia para el alemán que pudiera extraer de manera total sintagmas nominales.

La extracción de términos está conformada por tres componentes principales:

- **Procesador de consultas de corpus general (CPQ):** Es un procesador que puede soportar expresiones complejas de consultas, como expresiones regulares, etiquetas POS, lemas, entre otras.
- **Macroprocesador para el lenguaje de consulta CPQ:** La extracción de términos en HEID se basa en listas de afijos y en la verificación de los contextos típicos de los candidatos a término (Heid et al., 1996); para llevar a cabo este proceso, dado un parámetro en consulta, ejecuta este en un gran número de palabras mientras mantiene los demás parámetros de la consulta iguales.

---

[14] Identifica las categorías gramaticales, morfosintácticas y características distribucionales (Heid et al., 1996)

[15] Es un etiquetador de partes de la oración, el cual según Heid et al. (1996) desambigua los casos identificados en el proceso morfosintáctico.

- **XKWIC:** Esta herramienta gráfica muestra los términos y sus concordancias[16]; también permite ordenar de manera automática el material extraído según las necesidades del usuario.

El extractor terminológico HEID fue evaluado empleando manuales de mantenimiento en alemán. Se buscó extraer principalmente términos monopalabra, que frecuentemente representan sintagmas nominales en alemán; en este tipo de casos se obtuvieron algunos problemas por ruido los cuales, según los desarrolladores, pueden ser eliminados con el uso de filtros (por frecuencia, por categoría gramatical, entre otros). Asimismo, HEID permite extraer colocaciones[17] combinando sustantivos y verbos, aunque, en este caso los resultados no son muy buenos.

2.2.2 Sistemas basados en conocimiento estadístico

Además de los sistemas basados en conocimiento lingüístico, existen aquellos que se basan en conocimiento estadístico, es decir, en el empleo de fórmulas matemáticas, modelos probabilísticos, modelos heurísticos, entre otros.

Estos sistemas, además de extraer términos, otorgan una calificación que permite clasificar los resultados en buenos o malos. Aunque lo anterior es algo ambiguo, lo que se busca es que los términos extraídos con una alta calificación expresen una mayor relevancia en el documento o corpus, mientras que uno con baja calificación indique lo opuesto.

Existen múltiples medidas estadísticas que se emplean en los extractores terminológicos, como el TF-IDF, el logaritmo de la verosimilitud (Log Likelihood), el T-score, entre otros.

La ventaja de estos sistemas de extracción es que son independientes de la lengua e indican una calificación para cada uno de los términos. El problema con este tipo de enfoque

[16] Las concordancias, según la Real Academia Española (RAE), es el índice de todas las palabras de un libro o del conjunto de la obra de un autor, con todas las citas de los lugares en que se hallan.

[17] Propiedad que tienen ciertos sustantivo y verbos, y algunos sustantivos y adjetivos de coincidir en estructuras sintagmáticas, gracias a su estructura semántica: *gato* y *ronronear*, *planta* y *marchita* (Luna Trail et al., 2005).

es que existen términos de baja frecuencia difíciles de manejar por los sistemas de extracción (Cabré et al., 2001), esto genera lo que se llama *silencio*.

2.2.2.1 ANA

El sistema ANA (Euguehard y Pantera, 1994), "Automatic Natural Acquisition", es un extractor terminológico basado en conocimiento estadístico. Se basó en la idea de que este sistema debía poder extraer los términos de cualquier texto, sin importar si estaba bien escrito o no, si eran textos escritos o transcripciones de conversaciones y sin la utilización de conocimiento lingüístico. El extractor estaba diseñado para funcionar con cualquier lengua europea que no fuera aglutinante; sus pruebas se basaron en el inglés y el francés.

El sistema está formado por dos módulos: el de familiarización y el de descubrimiento. El primero de estos determina tres listas que emplea como conocimiento de la lengua a analizar; este conocimiento es extraído de manera estadística sin el uso de diccionarios o gramáticas. Las listas empleadas como conocimiento son las siguientes:

- **Palabras funcionales:** Es un conjunto de palabras que aportan poco o ninguna información (Sección 1.1.6). En esta lista entran artículos, pronombres y algunos verbos recurrentes.
- **Palabras esquemáticas:** Son las palabras que establecen una relación semántica entre otras palabras. Por ejemplo, Euguehard y Pantera (1994) indican que en el fragmento "box of nails", la palabra "of" indica una cierta relación entre "box" y "nails", por lo tanto "of" es una palabra esquemática.
- **Palabras base (bootstrap):** Es el conjunto de términos base con el que se inicia el sistema, es decir, este grupo de unidades terminológicas es el núcleo del extractor terminológico ANA.

El segundo módulo que conforma ANA es el de descubrimiento y se basa en la adquisición de nuevos términos a través del descubrimiento, como lo hace una persona que aprende un idioma. Este proceso se apoya en la co-ocurrencia de las palabras, esto puede tener tres interpretaciones:

- **Expresiones:** Una expresión se genera y se agrega a la lista de términos (bootstrap) cuando dos términos co-ocurren frecuentemente, es decir, aparecen en estructuras similares. Por ejemplo, en las frases "the *diesel engine* is", "this *diesel engine* has", los términos "diesel" y "engine", que pertenecen al bootstrap, aparecen contiguos frecuentemente, por lo tanto es posible que "diesel engine" sea un término y se agrega a la lista de palabras base.
- **Candidato:** Cuando una palabra, llamémosla X, aparece seguidamente de una palabra esquemática y de términos pertenecientes al bootstrap, se le considera como un candidato a término y se agrega a la lista de palabras base. Ejemplo: en las frases "shade of wood", "shade of color", "shade of beech", donde "of" es una palabra esquemática y las palabras "wood", "color" y "beech" son términos, la palabra "shade" cumple con la interpretación de candidato.
- **Expansión:** Este caso es similar al anterior, la diferencia es que no existe ninguna palabra esquemática entre el término y la palabra X. Un ejemplo sería: "use any *soft woods* to", "this *soft woods* or", donde "wood"[18] es un término, por tanto la palabra "soft wood" se agregaría al conjunto de términos.

El proceso del módulo 2 se realiza de manera recursiva hasta que no se encuentre ningún término nuevo en el documento. Además, durante el proceso de descubrimiento se genera una red semántica, en el cual se muestran algunas relaciones morfológicas y las co-ocurrencias de los términos.

Con respecto a los resultados, el sistema ANA fue evaluado para el inglés y para el francés. En el caso del inglés se empleó un corpus de 25,000 palabras el cual no fue ejecutado en el módulo de familiarización, sino que se indicaron de manera manual cada una de las listas que se crean en este módulo por el pequeño tamaño que tenía el corpus; del uso de ANA en este corpus se obtuvieron 200 nuevos términos. Para el francés, en cambio, se

[18] Aunque la palabra en el ejemplo es "woods" y el término es "wood", ANA reconoce que son la misma palabra debido a que emplea una herramienta que llama Reconocimiento Flexible de Cadenas. Esta herramienta emplea la distancia de edición; por ejemplo, si se tiene "casa" y "casas" su distancia de edición es 1 (adición de una s), en cambio para "caza" y "casa" es de 2 (eliminación de z y adición de s); por tanto dos palabras se parecen si su distancia de edición es muy pequeña.

usó un corpus de 120,000 palabras el cual sí pasó por el módulo de familiarización; del proceso de extracción se obtuvieron más de 3,000 nuevos términos.

A pesar de los resultados obtenidos, los desarrolladores de ANA consideran que este sistema es un extractor terminológico especializado en corpus de gran tamaño pero que sean de mala calidad, ya que aprende sobre la lengua empleada.

2.2.2.2 Extractor de términos estadístico basado en corpus

Este extractor terminológico fue desarrollado por Pantel y Lin (2001) y se basa únicamente en conocimiento estadístico.

El extractor terminológico consta de dos partes; la primera consta de la extracción de candidatos de términos. Para ello primero se recuperan todas los bigramas que se encuentren en el texto y su frecuencia; esta información se almacena en una base de datos de proximidad[19]. Posteriormente, se eliminan los bigramas que no cumplen con una serie de valores que están relacionados con la frecuencia del bigrama, con el valor de información mutua entre bigramas adyacentes[20] y el valor del logaritmo de la verosimilitud entre las palabras que pertenecen a un mismo bigrama[21].

La segunda parte del extractor consiste en la extracción de términos multipalabra; en esta parte se realiza la extracción de todas las construcciones que puede tener un bigrama (extraído en el paso anterior) con sus palabras adyacentes, esto para obtener términos que sean más grandes que bigramas; de este proceso sólo se guardan las palabras adyacentes que aparecieron en una misma construcción con el bigrama en cuestión varias veces. En seguida, la base de datos de proximidad se actualiza con el bigrama formado por una palabra del término original y por la de la nueva palabra que se encontró en la construcción. Finalmente,

---

[19] Una base de datos de proximidad es una base de datos con dos tablas; en la primera se almacena el objeto o el registro, mientras que en la segunda se guardan vínculos; cada tabla además tiene algunos atributos, como el nombre o valor (http://c2.com/cgi/wiki/Wiki?ProximityDatabase; http://kdl.cs.umass.edu/software/about.html).

[20] Esto se lleva a cabo para eliminar bigramas que tengan una palabra que no esté altamente relacionada con un posible término.

[21] Esto se realiza para saber si las palabras dentro del bigrama están por casualidad o por una verdadera importancia.

el proceso se vuelve recursivo y se emplea la nueva información que se obtuvo en la base de datos de proximidad, para que se pueda extender un término y obtener sus variantes.

Este sistema de extracción terminológica se evaluó usando precisión y cobertura usando un corpus segmentado en el idioma chino, la razón de lo anterior es que dicen los desarrolladores del sistema que el detectar palabras en chino es similar a detectar frases en inglés. La precisión fue evaluada contra los valores que se obtuvieron del logaritmo de la verosimilitud, mientras que la cobertura contra la frecuencia mínima de las palabras. Este sistema de extracción terminológica obtuvo una precisión máxima de 74.4% y una cobertura del 62.3%

2.2.3 Sistemas basados en conocimiento híbrido

Los sistemas de extracción terminológica no sólo pueden estar basados en un tipo de conocimiento; pueden emplear tanto el lingüístico como el estadístico, de esta manera se forma un sistema con conocimiento híbrido. El objetivo de este tipo de extractores terminológicos es crear sistemas que aprovechen al máximo las ventajas tanto de la parte lingüística como de la estadística y disminuyan las desventajas que ambos tienen.

2.2.3.1 Termext

Termext (Barrón-Cedeño et al., 2009) es un extractor terminológico de tipo híbrido que se basa en una adaptación para el español del método de C-Value/NC-Value (Frantzi et al., 2000). Además el método fue modificado para que aceptara unigramas como términos.

Este extractor de términos está dividido a grandes rasgos en dos partes, la de C-Value, y la de NC-Value. La primera parte, a su vez, se divide en dos procesos, el lingüístico y el estadístico. El proceso lingüístico consiste en etiquetar con partes de la oración y lematizar cada uno de los textos a analizar por medio de la herramienta TreeTagger. Posteriormente, dentro de este mismo proceso, se aplica un filtro lingüístico que consiste en almacenar las estructuras que pueden formar un término en español; este filtro puede ser abierto o cerrado, si es abierto este es más flexible con los patrones de los términos, de lo contrario es estricto con los patrones encontrados. En el proceso estadístico se calcula cuál es la probabilidad de que una estructura extraída sea un término; es decir, el C-Value, y para tal fin se toma en cuenta la frecuencia de la estructura, la frecuencia de la estructura en estructuras más grandes, el número de ocurrencias de las estructuras más grandes anteriores y la longitud de la estructura.

La segunda parte que conforma a Termext es la del cálculo de NC-value. Este valor considera el contexto en el cual se encontraban los términos obtenidos en el proceso anterior, esto con base en que un término, por lo general, está rodeado de palabras que están altamente relacionadas y pueden ser un indicio que exprese qué tan representativo el término es o no. Para ello se obtienen las palabras que en el contexto del término tengan cierta relevancia y se les calcula un peso. Posteriormente, se calcula el NC-Value, usando estos pesos y el valor C-Value del término. Finalmente, los términos con valores más altos de NC-Value son los términos que son más importantes en el documento, mientras que los de menor valor, son términos no tan representativos.

El extractor Termext fue evaluado con precisión y cobertura cuatro veces, la primera de ella con un filtro abierto sin una lista de paro obtuvo 23% de precisión y 82.6% en cobertura. La segunda de evaluación fue con un filtro lingüístico abierto y con lista de paro, la cual tuvo una precisión de 26.5% y una cobertura de 79.4%. La tercera evaluación se llevó a cabo con un filtro cerrado sin lista de paro y la curta de ellas con un filtro cerrado y lista de paro, en precisión se obtuvo un 24% y 30.8% respectivamente mientras que en cobertura se alcanzó un 46.3% y 50.3% de manera respectiva. Además, para su uso, se indica que Termext obtiene los mejores resultados de precisión y cobertura cuando se emplea un corpus de carácter técnico o científico de alto nivel de especialización, de lo contrario se genera una gran cantidad de ruido.

2.2.3.2 YATE

YATE (Vivaldi, 2001) es un extractor terminológico que emplea conocimiento tanto estadístico como lingüístico. Permite extraer términos tanto en español como en catalán, en los dominios de medicina, economía y genética. Las principales características de YATE son dos: la primera es que emplea una combinación de varias técnicas de extracción de términos y la segunda, que usa EuroWordNet como recurso léxico principal; de este recurso se hablará más adelante en la sección 2.4.1.

Grosso modo, existen 3 procesos que conforman YATE, los cuales se explican a continuación:

- **Proceso lingüístico:** Este es el primer proceso del extractor YATE. En él se lleva a cabo la segmentación, un análisis morfológico y, finalmente, un etiquetado de partes de la oración. En este proceso se emplean recursos léxicos como diccionarios, EuroWordNet y un corpus de referencia.
- **Filtro lingüístico:** Este proceso filtra las construcciones sintácticas que tienden a generar términos ya sea en español o en catalán, dependiendo del texto analizado. De este proceso se obtienen los candidatos a término que serán utilizados en el siguiente proceso.
- **Analizador de candidatos a término:** Este es el último proceso que forma parte de YATE. En él se calculan las diversas métricas y los datos que emplea YATE para determinar si un candidato a término pertenece o no al dominio seleccionado. Algunos de sus módulos son los siguientes (Vivaldi et al., 2001):
    - **Sistema de combinación:** En este módulo se unen todos los resultados para crear la lista final de candidatos.
    - **Extractor de contenido semántico:** Este módulo emplea EuroWordNet para determinar cuándo una palabra dada pertenece al dominio analizado, empleando identificadores de dominio.
    - **Formas griegas y latinas:** En el vocabulario médico se emplean muchas palabras que contienen formas griegas y latinas; por lo tanto, el conocer los términos que contienen estas formas puede dar información útil.
    - **Análisis colocacional:** En este módulo se emplean algunas medidas estadísticas para clasificar los candidatos a término, como la información mutua y la información mutua cúbica (MI^3).

Para llevar a cabo la evaluación de YATE se empleó un corpus de 10,000 palabras que consistía en resúmenes de artículos médicos. Este sistema de extracción terminológica fue evaluado con las medidas de precisión y cobertura, donde obtuvo un 97.2% de exactitud para una cobertura del 30%.

2.3 Evaluación de los extractores terminológicos

Los sistemas de extracción terminológica, al igual que muchos otros sistemas realizados por el hombre, necesitan que se les evalúe, ya que se necesita ver que el sistema cumpla con los objetivos, funcione con los estándares adecuados y sea lo suficientemente bueno como para realizar la tarea de forma automática y no manual. Sin embargo, aun cuando la extracción y el reconocimiento automático de términos han sido trabajados por largo tiempo y desde diferentes perspectivas, ningún *gold standard*[22] de evaluación ha sido introducido para evaluar claramente y comparar distintos enfoques (Pazienza et al., 2005).

Aun así, se han desarrollado dos técnicas para la evaluación de los extractores terminológicos y se presentan a continuación.

2.3.1 Lista de referencia

Uno de los métodos utilizados para la evaluación de los sistemas de extracción terminológica es el empleo de una lista de referencia. En este caso, según Pazienza et al. (2005), una lista de referencia se toma como un gold standard; esta puede ser una lista de términos ya existente de un dominio o área específica, o puede ser construida por un experto analizando el corpus que se empleó para extraer los términos.

Con la lista de referencia, el extractor terminológico se evalúa mediante el empleo de las métricas de precisión y de cobertura que se vieron en el apartado 1.2.3.

Aunque la lista de referencia tiene sus ventajas, para Pazienza et al. (2005), en términos de eficiencia, la lista de referencia no es la mejor técnica para calcular la precisión. Esto se debe a que puede haber términos reales que no fueron colocados en la lista y, por tanto, se consideran como falsos, disminuyendo la precisión del sistema.

[22] Un gold standard o una prueba estándar es una prueba o punto de referencia que califica, en este sentido, un sistema; puede que esta prueba no sea la mejor, pero no existe alguna otra y cumple con los estándares más básicos (http://en.wikipedia.org/wiki/Gold_standard_%28test%29).

2.3.2 Validación

Otro de los métodos empleados para la evaluación de los extractores terminológicos es la validación. Este método es preferido cuando ningún gold standard está disponible o cuando algunas características particulares del proceso de extracción de términos tienen que ser explícitas (Pazienza et al., 2005).

Este método consiste en validar los términos que se encuentran en la lista creada por el sistema en evaluación. Para poder llevar esto a cabo, Pazienza et al. (2005) indican que es necesario que se cumplan dos cosas. La primera de ellas, es que la validación de la lista debe ser realizada por varios expertos, esto para tener una lista de términos mucho más confiable. El segundo parámetro a cumplir es que cada experto que va a participar en el análisis debe recibir una introducción a lo que es un término. De todas maneras, cabe aclarar que aun siguiendo estos dos parámetros, es posible que las listas resultantes sean diferentes, esto puede ser debido a los distintos conocimientos de los expertos, al juicio del experto o a la ambigüedad de lo que es una unidad terminológica; por tanto, es necesario que se llegue a un acuerdo entre los expertos para obtener una lista validada.

Con la lista de términos validada se emplean las métricas de precisión y de cobertura de la misma forma que ocurre en los sistemas de recuperación de información.

Al igual que la lista de referencia, este método de evaluación tiene sus desventajas, una de ellas es que no es el mejor método para calcular la cobertura del sistema. La razón de ello es que, al enfocarse en una lista extraída por el mismo sistema, se cierra la posibilidad de conocer si existen otros términos que se debieran haber obtenido.

2.4 Recursos electrónicos para la validación

Actualmente, existen algunos extractores terminológicos que validan cada uno de los términos encontrados en el documento antes de presentárselos al usuario; además algunos de ellos agregan información que podría ser de utilidad. Para ello emplean recursos semánticos, en su mayoría creados por expertos, que otorgan información sobre el dominio al que pertenecen, como sinónimos. Algunos extractores que emplean este tipo de validación, además de YATE, son MetaMap (Aronson y Lang, 2010) y TRUCKS (Maynard, 2000).

2.4.1 WordNet y EuroWordNet

WordNet es una base de datos léxica electrónica desarrollada por la Universidad de Princeton, la cual sirve como recurso para aplicaciones en PLN y recuperación de información (Fellbaum, 1998). Esta base de datos sólo maneja inglés y es de acceso libre por internet[23]. Su extensión a otros idiomas, como el español, se realizó por medio de *EuroWordNet* (EWN), que es de paga y actualmente está en crecimiento en algunas lenguas.

Dentro de WordNet y, por consiguiente, de EuroWordNet, existen tres estructuras que se encargan de las diversas categorías lingüísticas que maneja, es decir, hay una para sustantivos, otra para verbos y una para adjetivos y adverbios.

Esta base de datos se basa principalmente en conjuntos de sinónimos, llamados *synset*, que representan todo un concepto. Por ejemplo, en el caso del inglés, cuando se busca "elevator" también se muestra su variante británica que es "lift"; en el caso del español si buscamos "tepalcate" nos muestra que tiene como synset "tejoleta", "tiesto" y "casco".

La estructura de sustantivos, de WordNet y EWN, además de manejarse a través de los synset, se maneja por medio de relaciones de hiponimia e hiperonimia. La hiponimia es una relación que denota un subconjunto o subclase de una palabra; por ejemplo, en EWN la palabra "automóvil" tiene como hipónimos las palabras "limosina", "sedán", "jeep", entre otros. En cambio, la hiperonimia es una relación que expresa una superclase de una palabra; "vivienda", por ejemplo, es un hiperónimo de "casa", de "estudio" y de algunos otros más.

WordNet y EWN, además de contar con los synset, incluye definiciones tipo diccionario y ejemplos de uso.

[23] http://wordnetweb.princeton.edu/perl/webwn

2.4.2 Lexicón Specialist UMLS

Uno de los recursos léxicos electrónicos más importantes del área de la biomedicina es el lexicón *Specialist de UMLS*. Este lexicón es uno de los tres recursos que se generaron dentro del proyecto UMLS (Unified Medical Language System) creado por la Biblioteca Nacional de Medicina de los Estados Unidos de América (NLM).

Según Ananiadou y McNaught (2006), el lexicón Specialist es un diccionario general del inglés que contiene una gran cantidad de términos de biomedicina. Todos estos términos fueron extraídos de diversos recursos, como de los registros de MEDLINE/PubMed[24], del metatesauro UMLS[25] y de diccionarios médicos del inglés.

Cada una de las entradas del lexicón puede ser monopalabra o multipalabra; a su vez, estos términos tienen información como categoría gramatical, patrones complementarios permitidos, lema, variantes ortográficas y morfológicas.

2.4.3 Wikipedia

Otro de los recursos que se han estado empleando actualmente para la validación de extractores es Wikipedia[26]. La *Wikipedia* es una enciclopedia gratuita, multilenguaje, creada para la red y construida de manera colaborativa por voluntarios (Zesch et al., 2008).

Esta enciclopedia está formada por artículos que crean una red interconectada de conocimiento, adicionada con categorías y subcategorías (se podría decir que es un tipo de hiperonimia e hiponimia, aunque no cumplan forzosamente con las relaciones) que los voluntarios crean y organizan, y que permiten hasta cierto punto dividir los conocimientos en áreas o dominios. El uso de categorías y subcategorías forma lo que se conoce como una *taxonomía*, es decir una ordenación jerárquica y sistemática; aunque hay autores como Peters

---

[24] MEDLINE es una base de datos que almacena bibliografía médica que provienen desde 1950. Su motor de búsqueda es la herramienta de PubMed.

[25] Es otro de los recursos del proyecto de UMLS que incluye conceptos del área de biomedicina, nombres de conceptos, sinónimos, así como las relaciones entre los conceptos.

[26] http://www.wikipedia.org

(2009), que consideran esto realmente como una folksonomía[27], ya que es la gente quien desarrolla la jerarquización y sistematización de la Wikipedia.

Además Wikipedia contiene una gran cantidad de información semántica y léxica que se complementa con el conocimiento de entidades nombradas y términos de dominio específico o especializado que incluye el sitio. De igual forma, incluye un sistema de redireccionamiento, que podría ser considerado un diccionario de sinónimos en el cual se toman en cuenta variaciones ortográficas, morfológicas y de abreviaturas; por ejemplo, si se busca en la Wikipedia "ajolote", "axolote" o "axolotl" se redirecciona a "Ambystoma mexicanum", el nombre científico del ajolote. También el sistema de redireccionamiento funciona, en un menor grado, como un sistema que pasa de un tema específico a uno general, o de un verbo a un sustantivo.

Entre las ventajas con las que cuenta Wikipedia se puede mencionar que es un recurso libre, que se actualiza y crece rápidamente, que maneja una gran cantidad de dominios y que está en diversas lenguas, no solamente en las principales. Algunas de sus desventajas es que no existe un control editorial o por expertos, y que no se siguen lineamientos específicos para su construcción.

De este recurso electrónico se hablará más adelante, en la sección 3.4, donde se abordará la estructura interna y la manera en que fue empleada en el proyecto de tesis.

[27] Una folksonomía es un sistema de clasificación de contenidos desarrollado de manera colaborativa (Peters, 2009).