Universidad
Carlos III de Madrid

EXTRACCIÓN DE PATRONES SINTÁCTICO-SEMÁNTICOS DE DOCUMENTOS DE PATENTES

PROYECTO FINAL DE CARRERA

Ingeniería Técnica en Informática de Gestión

ESCUELA POLITÉCNICA SUPERIOR

Autor: Leticia Arroyo Minguela

Tutores: Prof. Dr. Anabel Fraga y Prof. Dr. Valentín Moreno

Leganés, Octubre 2015

Agradecimientos

No quería comenzar este documento sin antes agradecer a mi marido, mi familia, mis amigos y mis compañeros de trabajo las fuerzas y el apoyo emocional recibido para realizar el proyecto. Ha sido un año complicado en mi vida, con muchos cambios que por suerte van saliendo todos bien gracias a ellos. Muchas gracias por la paciencia demostrada y la confianza depositada en mí.

También mostrar mi mayor agradecimiento a mis tutores Anabel Fraga y Valentín Moreno de la Universidad Carlos III por haber hecho posible este estudio. Por orientarme, guiarme y ayudarme, sin olvidar la cooperación de Eugenio Parra que me ha dado soporte y me ha prestado su herramienta BoilerPlates.

Con este proyecto puedo dar por finalizada una etapa muy importante de mi vida. Es una espinita menos para seguir creciendo como profesional y como persona.

Muchas gracias a todos de corazón.

State of the art

The objective of this project is the analysis of patent documents, it is to analyze the contents of the documents and to do natural language processing is required.

The natural language processing (NLP) is the science studied by computational linguistics, in order that a computer can understand human language.

We find the issue with the ambiguity, the same expression can have more than one meaning. All depends on the context in which the term is found, the same word may be different semantics.

For the analysis in this project, we will see the difference when analyzing the text as simple words, compared to the analysis of the text semantic distinction that each one of them.

In the methodology used to generate patterns, three types of analysis are performed:

a. **Lexical analysis:** Lexical analysis is the first step in most applications for word processing, where the process is to convert a flow of input characters into a flow of words or token. Tokens are identified because words are segmented by spaces, commas, periods, among others. These tokens are used by the syntactic analysis.

b. **Syntactic analysis or parsing:** text morphology is analyzed, allowing search and word count. The parser identifies the grammatical structure of the sentence. By performing from the decomposition of their sentences in their nominal syntagma and verb syntagma until reaching identify the appropriate grammatical category for each word.

c. **Semantic analysis**: With it search the sense to the words, it refers to the aspects of meaning, sense or interpretation of the meaning of a particular element, symbol, word, expression or formal representation. The semantic analysis is performed after syntactic analysis, and is more complicated by perform, because it's necessary to interpret ambiguities.

Applying the analysis we have just explained, we obtain the basic patterns or token through boilerplates tool. Each basic pattern is formed by a single token and come identified by its grammatical and semantic category if any.

Pattern define it as a minimum group of two words that are consecutive in the same text. Four types of patterns are set: word + word; word + pattern; pattern + word and pattern + pattern.

Within three types of pattern is again include the expression pattern, these would be the subpatterns. Subpattern will be the one that is part of another pattern.

Ontology. According to Thomas Gruber: "To support the sharing and reuse of formally represented knowledge among AI systems, it is useful to define the common vocabulary in which shared knowledge is represented. A specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects — is called an ontology."

WordNet it is a lexical database of English. It groups English words into sets of synonyms called synsets, providing short and general definitions, and stores the semantic relationships between sets of synonyms. Its purpose is double: to produce a combination of dictionary and thesaurus whose use is more intuitive, and support automatic text analysis and artificial intelligence applications.

Methodology and development

The objective of this project is to perform the extraction of syntactic-semantic patterns found within documents on patents.

Patent documents are writing by experts, therefore we are saying that we will have very well written documents and high quality grammatical.

When the investigation is complete, we have a list sorted by frequency patterns. We will know the syntactic-semantic patterns that are most used when writing a patent.

Addition to patterns are also know what the most recurrent words are, we will know the most common words in the writing of patterns.

The phases defined here are needed to meet the objectives:

PHASE 1: Search for patent sources where they can download patents documents public and registered in PDF format. The documents must be converted to TXT format using pdf2txt program.

PHASE 2: Download at least about 500 documents.

PHASE 3: Convert the PDF documents to TXT using the pdf2txt program.

PHASE 4: Get WordNet dictionary to form the ontology. This phase can be performed in parallel to steps 1, 2 and 3.

PHASE 5: Manage the ontology with KnowledgeMANAGER. Adding vocabulary obtained in phase 4.

PHASE 6: Add the new ontology in BoilerPlates tool.

PHASE 7: Define study scenarios and using ontology created, generating patterns with the BoilerPlates tool.

PHASE 8: TXT documents will be included one by one on the BoilerPlates tool, with this first step in the tool will generate the basic patterns.

PHASE 9: Representing one to one each scenario in BoilerPlates tool and start pattern generation.

PHASE 10: Analyze the results obtained by scenario.

PHASE 11: Analyze and compare the results of all scenarios.

In this project a syntactic-semantic analysis is performed, of a sample of registered patents and made public, through an ontology based on natural language words.

To get a larger sample of patent documents to analyze, it has decided to use English as the language of analysis. Therefore all patents that are used in this investigation will be written in the English language.

All patents are search in Internet and document must be PDF formats.

It does not establish any particular subject, and not any particular area of investigation, the investigation developed here is valid for all subjects.

We have two samples of patents, on one hand analyze documents of the United States Patent and Trademark Office, we have 359 documents, and secondly analyze documents of the European Patent Office, we have 379 documents Europeans different.

The study will be made with over 700 patent documents, all documents be analyzed with the BoilerPlates tool.

The ontology that includes into the boilerplates tool, will be managed with the KnowledgeMANAGER tool of REUSE Company. The vocabulary will form the ontology is providing by WordNet.

WordNet used as a basis for the ontology of data recovery, we will have a language general controlled (not specialized by subject) and to language English. Into the WordNet we obtain nouns, verbs, adjectives and adverbs.

The investigation done here is interesting because we discover how the pattern of professional experts document their investigations, findings and studies.

Here art to documentation is analyzed, so important it is to have an idea as important is knowing it registered.

The patterns that are obtained in the investigation may be useful in the future to guide the new professionals in the time of writing.

Analysis of result

After the analysis of the US patents documents and European patents documents we can conclude the following:

The basic patterns obtained are independent of the frequency and the selection of grammatical categories in the boilerplates tool. All basic patterns are common within the same sample.

In the boilerplates tool, the higher the minimum frequency used, is less the number of patterns obtained and is shorter the time necessary to obtain them.

Differentiation has been made by their semantic patterns in the minimum frequencies of 1, 20 and 100 to US samples, and 20 and 100 for European samples. For frequency 1 it has not been possible to obtain results due to the high volume of information that we have handled. More than 25 days after running the tool, it has had to reject frequency 1 for the study. About the other two frequencies, we can say that the higher the frequency the number of patterns obtained is less.

Patterns are calculated without differentiation of semantics for the minimum frequencies of 1 and 100 with US sample. It is also calculated with the European sample for the minimum frequency of 100, without differentiation patterns by their semantics. It can be concluded that the same patterns are obtained with different semantics.

With increasing frequency we lose patterns that have longer decomposition. Because the number of repetitions is less. The longest pattern shown in the graph 17 of section 3.

After using different frequencies to generate patterns in boilerplates, we can say that the intermediate frequency is what has given us the best results.

In both samples the unclassified words are very present.

The patterns obtained in all scenarios can assist the writing for any user who need to write a patent.

After the investigation, with the knowledge obtained now, we can give some recommendations to people who will do a similar study in the future.

The ontology can be improved, the ontology has 73 grammatical categories to define their vocabulary. For this project has not been completed because all the most important words are covered. The pending grammar to define are the type of punctuation, dates, email, arithmetic symbols, acronyms, etc. The undefined categories are shown in Table 8.

For future projects, scenarios of using a minimum frequency of 100 can be applied to search which is the minimum frequency that will create zero patterns.

It is possible create a new analysis with minimum frequency greater than 100, because we obtained patterns where their repetition frequency is greater than 100. But before begin studies with a higher minimum frequency, we recommend you should not consider words that do not correspond to a grammar of the ontology.

Índice de contenido

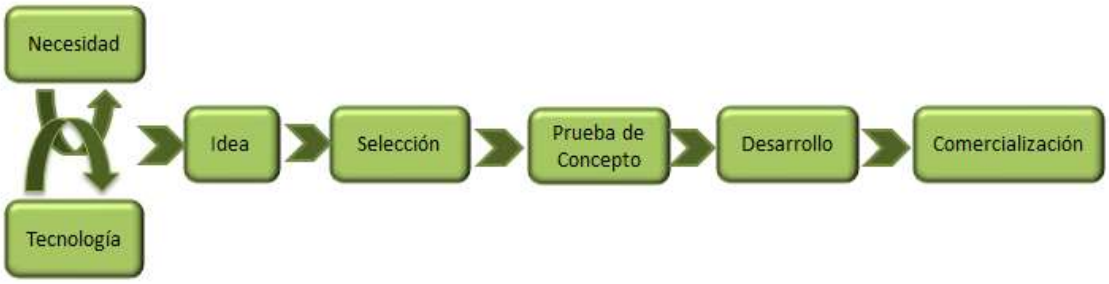
Índice de ilustraciones

1. Introducción

Los análisis de patentes no son innovación con esta investigación, ya existen muchos análisis y mucha información la que se extrae de las patentes. Nos ha gustado como han expresado los antecedentes históricos sobre la investigación y hemos querido citarlo en esta pequeña introducción:

La innovación ocurre cuando las necesidades se juntan a las tecnologías que tratan esas necesidades y el proceso nuevo resultante, el producto y las ideas del servicio se desarrollan de una manera responsable que balancea los riesgos y las recompensas de hacer algo nuevo.

El concepto básico es simple: la innovación ocurre cuando hay un empate de una necesidad nueva o emergente con una tecnología existente o emergente, y las empresas industriales seleccionan y desarrollan las mejores ideas usando un proceso dirigido que balancee los riesgos y las variables desconocidas. En la Figura 1 se adapta del modelo de la innovación de Marquis y Myers.



Gráfica 1. Modelo de innovación (adaptado de Marquis y Myers)

Se trata del documento "Análisis Morfológico de Patentes para Desarrollar un Producto de Seguridad Vehicular" y la referencia es el número [1] de la biografía incluida en este documento.

Los objetivos de nuestra investigación son diferentes a la investigación citada. En el siguiente punto se definen.

1.1 Objetivos

El objetivo de este proyecto consiste en realizar la extracción de patrones sintáctico-semánticos que se encuentran dentro de los documentos de patentes publicadas y públicas.

Los documentos de patentes son redactados por profesionales expertos, por ello estamos hablando de que contaremos con documentos muy bien redactados y de gran calidad gramatical.

Al finalizar la investigación contaremos con un listado de patrones ordenados por frecuencia. Conoceremos los patrones sintáctico-semánticos que son más utilizados a la hora de redactar una patente.

Además de patrones, también conoceremos cuáles son las palabras que más se repiten, sabremos las palabras más comunes en la redacción de patrones.

1.2 Metodología

La metodología empleada se describe en las fases que aquí se definen, son los pasos necesarios para cumplir con los objetivos:

FASE 1: Búsqueda de fuentes de patentes dónde se puedan descargar documentos en PDF de patentes registradas y públicas. Los documentos deben ser convertibles a formato TXT.

FASE 2: Descargar al menos unos 500 documentos.

FASE 3: Convertir los documentos PDF en TXT utilizando el programa pdf2txt.

FASE 4: Obtener diccionario WordNet para formar la ontología. Esta fase puede realizarse en paralelo a las fases 1, 2 y 3.

FASE 5: Gestionar la ontología con KnowledgeMANAGER. Añadiendo el vocabulario obtenido en la fase 3.

FASE 6: Añadir la nueva ontología a la herramienta BoilerPlates

FASE 7: Definir escenarios de estudio y haciendo uso de la ontología creada, generar patrones con la herramienta BoilerPlates.

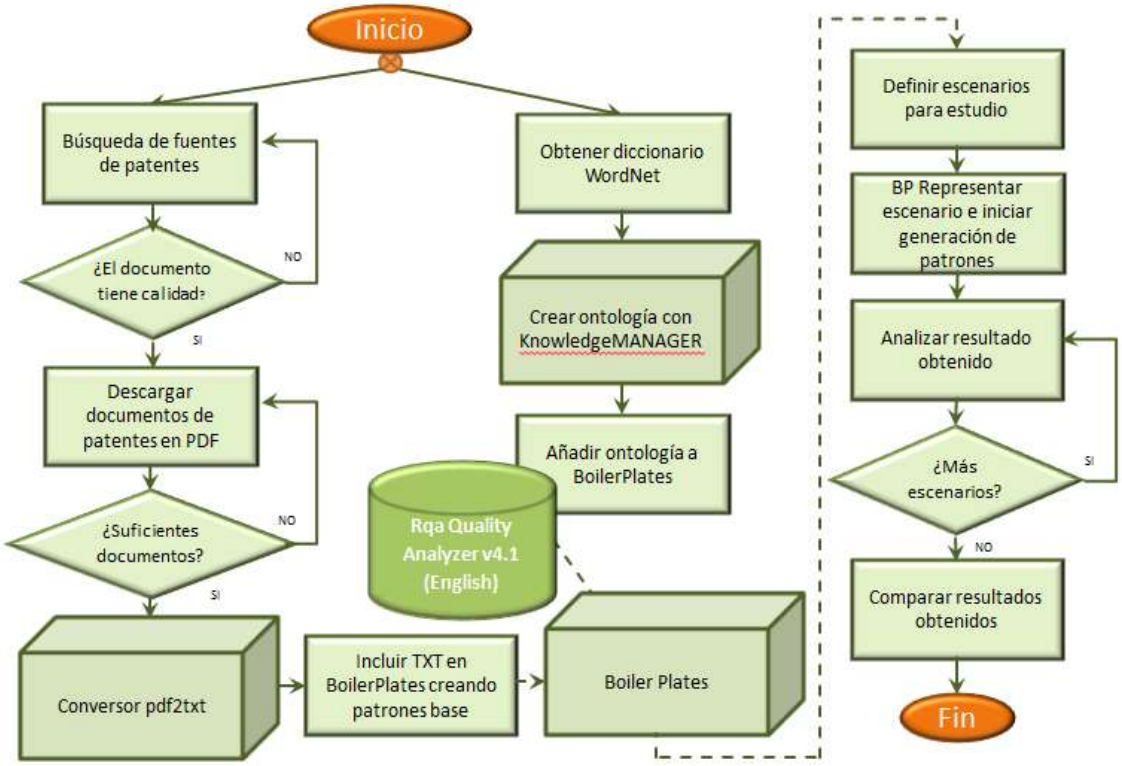
FASE 8: Los documentos TXT se incluirán uno a uno en la herramienta BoilerPlates, con este primer paso en la herramienta se estarán generando los patrones base.

FASE 9: Representar uno a uno escenario en la herramienta BoilerPlates e iniciar la generación de patrones.

FASE 10: Analizar los resultados obtenidos por escenario.

FASE 11: Analizar y comparar los resultados de todos los escenarios.

En el siguiente diagrama de flujo se representa la metodología a seguir en la investigación:



Gráfica 2. Diagrama flujo para la metodología a emplear

1.3 Requisitos de la investigación

En este proyecto se realiza un análisis sintáctico-semántico de una muestra de patentes registradas y hechas públicas, a través de una ontología basada en el lenguaje natural de palabras.

Para tener una mayor muestra de documentos de patentes a analizar, se ha decidido utilizar el inglés como lenguaje de análisis. Por ello todas las patentes que se utilicen en esta investigación estarán redactas en el idioma inglés.

Todas las patentes serán buscadas en Internet y el formato de los documentos tendrá que ser PDF.

No se fija ninguna temática en concreto ni ningún área de investigación en concreto, puesto que la investigación aquí desarrollada es válida para todas las temáticas.

Se van a diferenciar dos muestras de patentes, por una lado analizaremos documentos de la Oficina de Patentes y Marcas Registradas de Estados Unidos, de las que tenemos 359 documentos, y por otro lado analizaremos documentos de la Oficina Europea de Patentes, de las europeas disponemos de 379 documentos diferentes.

El estudio será realizado con más de 700 documentos de patentes, todos ellos serán analizados con la herramienta BoilerPlates.

La ontología que se incluye a la herramienta BoilerPlates, será gestionada con la herramienta KnowledgeMANAGER de REUSE Company. El vocabulario que formará la ontología es el que nos presta WordNet.

A priori, como no se centra la investigación en ninguna temática en concreto, la semántica de la que estará dotada la ontología es la proporciona la herramienta knowledgeMANAGER.

1.4 Motivación

La investigación aquí realizada es interesante porque podremos descubrir, a modo de patrones, cómo profesionales expertos redactan sus investigaciones, descubrimientos y estudios. Aquí se analiza el arte de documentar, tan importante es tener una idea como tan importante es saberla comunicar.

Los patrones que aquí se encuentren, podrán servir en un futuro para dirigir a los nuevos profesionales en el momento de la redacción.

Además, al contar con dos muestras diferentes, conoceremos las diferencias que puedan encontrarse en ambas.

1.5 Estructura del documento

Los pasos que se han seguido para la realización de este proyecto coinciden con el orden establecido en este documento, se resumen en los siguientes puntos:

- En el apartado 2 de este documento, se cuenta el estado del arte. Se explican los términos que van a ser necesarios conocer para que sea posible la realización de esta investigación.

- Tras la toma de requisitos y con el conocimiento obtenido en la investigación, se crea la planificación inicial del proyecto que puede verse en el apartado 3. Durante todo el periodo de desarrollo han surgido imprevistos que han hecho modificar la planificación, la planificación real, también la detallamos en el mismo apartado.

- Para la obtención de la muestra de patentes se analizan las fuentes de información que se detallan en el apartado 4 de este documento.

Son dos las fuentes de información seleccionadas y 738 documentos obtenidos. Es una cantidad muy buena para conocer la frecuencia de patrones utilizada en la publicación de patentes.

- Todos los documentos de patentes obtenidos están en formato PDF. La herramienta BoilerPlates sólo puede analizar documentos en formato TXT, con lo que tenemos la necesidad de convertir los documentos para que puedan ser analizados. Para ello, tras probar varios conversores online y software libre, se decide utilizar el creado por David Catalán, uno de los alumnos, mis agradecimientos a David.

  No se ha dedicado ningún apartado a esta parte del proyecto, por no contener información relevante. Pero si se ha incluido en el Anexo I la guía del conversor.

- En paralelo creamos una ontología con la base del diccionario de WordNet y con ayuda de la herramienta KnowledgeManager. El paso a paso realizado se puede consultar en el apartado 5 de este documento.

- Con todo lo anterior completado, llega el momento de usar la herramienta BoilerPlates, su funcionamiento se detalla en el apartado 6 de este documento. La ontología creada y mencionada en el anterior punto tres de esta sección, es adaptada dentro de esta herramienta, los detalles se pueden ver en el apartado "6.1.1. Rqa Quality Analyzer v4.1 (English)".

- En el apartado 7 se definen los requisitos del estudio.

- Los escenarios creados para cada grupo de patentes. Pueden verse en el apartado 8.

- Los documentos TXT se incluirán uno a uno en la herramienta BoilerPlates, generando patrones básicos por cada uno de ellos. Este proceso puede requerir de unas 12 horas para incluir todos los documentos obtenidos.

  Tras tener todos los patrones básicos creados, se procederá a crear los patrones con sus frecuencias. El tiempo para la creación de modelos de frecuencia dependerá de la cantidad mínima de frecuencia que se utiliza, las categorías gramaticales utilizadas, y si la diferenciación de patrones por su semántica está activado o no.

  Cuando se tengan todos los escenarios creados con BoilerPlaites, llega el momento de realizar el análisis de los patrones y las

frecuencias obtenidas, siendo éste el objetivo principal del proyecto. Los resultados se detallan en el apartado 9.

- Con todo el análisis realizado se llegan a las conclusiones comentadas en el apartado 10. Con la visión global de todos los resultados de los escenarios establecidos, se llegan a las conclusiones finales.

- En el apartado 11, para finalizar, se dan algunas recomendaciones e ideas para ampliar el estudio aquí realizado o para otros similares.

- Las referencias consultadas y utilizadas se listan en el apartado 12.

2. Estado del arte

En las siguientes líneas se explica el conocimiento base a tener en cuenta para la realización de este proyecto.

2.1 Procesamiento del Lenguaje Natural

El objetivo de este proyecto es el análisis de documentos de patentes, se van a analizar los contenidos de los documentos y para ello es necesario el procesamiento del lenguaje natural.

El procesamiento del lenguaje natural (PLN) es la ciencia estudiada por la lingüística computacional, con el objetivo de que una computadora pueda entender el lenguaje humano. Un contestador automático o un traductor de lenguajes, son un ejemplo claro de cómo una computadora procesa e interpreta el lenguaje natural.

Esta ciencia parece que comienza en el año 1950 con la publicación test de turing, como es conocida a día de hoy, y que es publicada por Alan Turing[1] bajo el título “Computing machinery and intelligence”. Con esta publicación en los años 50 ya se planteaban si las máquinas podrían pensar.

Inicialmente la poca memoria y la poca velocidad de los procesadores suponían un gran problema. Hoy en día ya no es un obstáculo. Sin embargo, existe otro obstáculo para que los programas puedan entender el lenguaje natural.

[1] Alan Mathison Turing; Londres, 1912-Wilmslow, Reino Unido, 1954. Matemático británico. http://www.biografiasyvidas.com/biografia/t/turing.htm

Nos encontramos con el problema es la ambigüedad, la misma expresión se puede interpretar de diferentes maneras. Dependiendo del contexto, una misma palabra puede formar parte de semánticas diferentes. Por ejemplo, la palabra "bota" se puede entender como un calzado o como un recipiente de cuero para guardar vino. Otra palabra con doble significado y que si pudiera estar dentro de los documentos de patentes, podría ser "corriente" que puede entenderse como luz eléctrica o como algo de mala calidad o de poco valor. Nos encontramos con infinidad de expresiones que tienen más de un significado.

No sólo nos encontramos palabras independientes ambiguas, también nos encontramos con frases difíciles de analizar semánticamente por encontrarlas más de un sentido; Por ejemplo en la frase "Diego come arroz con palillos", para un procesador no es claro si Diego come los palillos y come arroz o si los usa los palillo para comer el arroz, compárese con la frase "Diego come arroz con leche". No es difícil que un programa pueda reconocer todas las interpretaciones del texto, lo difícil es que sepa elegir la correcta.

Para el análisis que se realiza en este proyecto, veremos la diferencia que hay al analizar el texto como simples palabras, frente a realizar el análisis del texto diferenciando la semántica que contiene cada una de ellas.

En la mayoría de tareas de PLN es necesario obtener un corpus[2], a través de la utilización de recursos externos tales como: diccionarios, tesauros, ontologías, etc. Estos recursos proporcionan sus respectivas estructuras internas, interfaces, relaciones entre conceptos, etc.

En la metodología usada para la generación de patrones se distingue tres tipos de análisis:

a. **Análisis léxico**: El análisis léxico es el primer paso en la mayoría de aplicaciones para el procesamiento de texto, dónde el proceso consiste convertir un flujo de caracteres de entrada en un flujo de palabras o token[3]. Los token se identifican porque las palabras están segmentadas por espacios, comas, puntos, entre otros. Esos tokens son usados por el análisis sintáctico.

   Tokenización: El Tokenizado es una función aunque relativamente simple, muy importante. Según Peñas Padilla[4] "Un tokenizador sirve

[2] Corpus. Definición: Conjunto cerrado de textos o de datos destinado a la investigación científica.

[3] Token: aparición concreta de una palabra en un texto dentro de un contexto determinado.

[4] Anselmo Peñas Padilla. Profesor titular de Universidad. Experience from 1999 (2 quinquenios).

para separar oraciones, palabras y signos de puntuación de forma que sea posible su posterior tratamiento mediante herramientas como el analizador morfológico y el etiquetador de categorías gramaticales".

b. Análisis sintáctico: Se analiza la morfología del texto, permitiendo la búsqueda y recuento de palabras. El analizador sintáctico identifica la estructura gramatical de la oración. Realizando desde la descomposición de sus frases en su sintagma nominal y sintagma verbal hasta llegar a identificar la categoría gramatical adecuada para cada palabra.

Normalización: Para poder realizar el análisis sintáctico de cada token se aplica la normalización, y consiste en estandarizar todos los términos del texto.

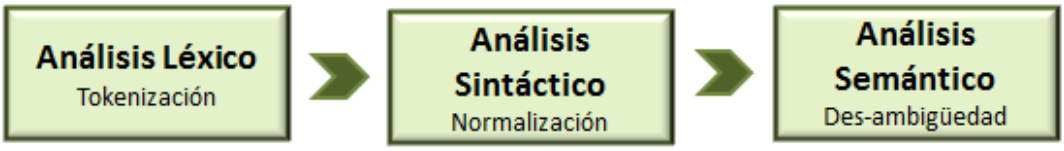
Consiste en homogeneizar todo el texto, por ejemplo el control de los términos en mayúscula o minúscula; el control de determinados parámetros como cantidades numéricas o fechas; el control de abreviaturas y acrónimos; cambiar el uso de los verbos en infinitivo, eliminar los plurales, entre otros.

En definitiva, determinar una forma única común a todas las posibles formas de una misma palabra.

c. Análisis semántico: Con él se busca el sentido a las palabras, se refiere a los aspectos del significado, sentido o interpretación del significado de un determinado elemento, símbolo, palabra, expresión o representación formal. El análisis semántico se realiza posteriormente al sintáctico, y es mucho más complicado de realizar por la interpretación ante ambigüedades.

Ambigüedad: Se da la ambigüedad en una palabra cuando una misma palabra admite dos o más significados distintos.

En el análisis semántico se empleará la des-ambigüedad para dar un significado semántico al token dependiendo del contexto en el que se encuentra.



Gráfica 3. Metodología en generación de patrones

2.2 Patrón básico

Los patrones básicos que obtendremos a través de la herramienta BoilerPlates serán los token que hemos definido anteriormente. Tendremos un token por cada una de las palabras o símbolos que contengan los documentos que se estén analizando.

Cada patrón básico está formado por un único token y vendrá identificado por su categoría gramatical y su semántica si la tuviera.

A un patrón básico se le ha realizado el análisis léxico, el análisis sintáctico y el análisis semántico. Más adelante se explicará con mayor detalle lo que nos da la herramienta BoilerPlates como patrón básico.

2.3 Patrón y Subpatrón

Definimos como patrón un grupo mínimo de dos palabras que son consecutivas dentro de un mismo texto. Se establecen cuatro tipos de patrones:

PATRON 1: palabra + palabra

PATRON 2: patrón + palabra

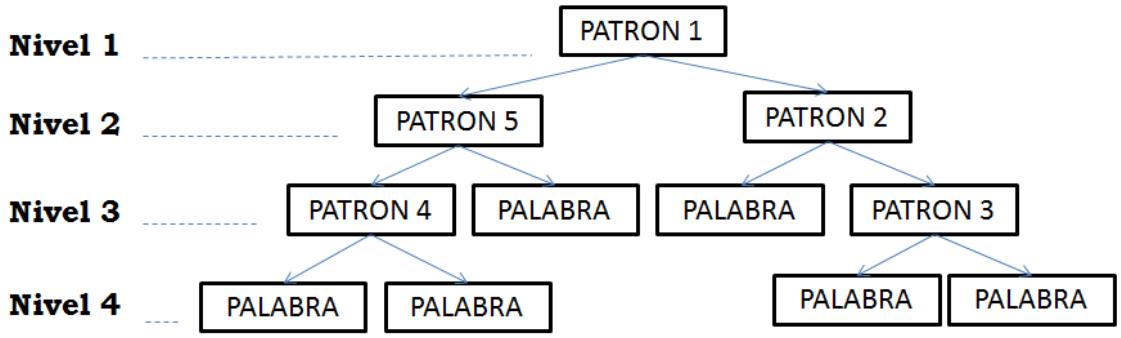
PATRON 3: palabra + patrón

PATRON 4: patrón + patrón

Véase que se está distinguiendo entre “patrón + palabra” y “palabra + patrón”, el orden dado es importante puesto que se está extrayendo el orden con el que está construida una frase.

Dentro de tres de los tipos de patrón se vuelve a incluir la expresión patrón, éstos serían los que se denominan subpatrón, será aquel que forme parte de otro patrón.

Como un patrón podría formar parte de otro patrón y éste último estar formado por otro patrón, sería lo que se denomina una jerarquía de patrones de varios niveles.



Gráfica 4. Jerarquía de patrones

En el ejemplo de la imagen vemos que el Patrón 1 tiene cuatro niveles de profundidad, la descomposición del patrón va mostrando subpatrones y palabras. La longitud final es de seis palabras, y son las que forman la frase. En definitiva tenemos que el patrón 1 es:

PATRON 1 = PALABRA + PALABRA + PALABRA + PALABRA + PALABRA + PALABRA.

Cada una de esas palabras pertenecerá a una categoría gramatical y pueden o no tener un significado semántico.

Más adelante se veremos como la descomposición de patrones nos da la longitud de un conjunto de categorías gramaticales como estructura de una oración.

2.4 Ontología

La ontología es un concepto filosófico de la rama metafísica que trata del ser en general y de sus propiedades transcendentales. “El estudio del ser” solo puede representar lo que existe.

La expresión ontología aparece por primera vez en 1606 por el filósofo alemán Jacob Lorhard en su obra *Ogdoas Scholastica.*

En la rama de la ciencia, las ontologías son clasificaciones. Es utilizada para la categorización y agrupación de la información en clases.

Las ontologías también son aplicadas en Inteligencia Artificial y en web semántica, representa la asimilación y la codificación del conocimiento, definiendo relaciones entre los conceptos.

Según Thomas Gruber[5] la ontología es definida como, "para apoyar el intercambio y la reutilización de conocimiento representado formalmente entre los sistemas de inteligencia artificial, es útil definir el vocabulario común en la que el conocimiento compartido está representado. Una especificación de un vocabulario de representación para un dominio compartido del discurso - definiciones de clases, relaciones, funciones y otros objetos - se llama una ontología." [3]

En definitiva, una ontología en su esencia es un esquema conceptual con conexiones, donde se moldea la realidad, la ontología define los términos que representan el conocimiento.

Algunos de los objetivos son, entre otros, permitir el intercambio de datos entre programas, simplificar las distintas representaciones incluso para diferentes idiomas.

Para este proyecto se va a crear una ontología basada en la recolección de vocabulario de WordNet.

2.4 WordNet

WordNet (WN) es una base de datos léxica del idioma Inglés. Agrupa palabras en inglés en conjuntos de sinónimos llamados *synsets*, proporcionando definiciones cortas y generales, y almacena las relaciones semánticas entre los conjuntos de sinónimos. Su propósito es doble: producir una combinación de diccionario y tesauro cuyo uso sea más intuitivo, y soportar análisis automático de texto y a aplicaciones de Inteligencia Artificial. La base de datos y las herramientas del software se han liberado bajo una licencia BSD y pueden ser descargadas[6] y usadas libremente. Además la base de datos puede consultarse en línea[7]. [5]

---

[5] Thomas Robert Gruber (nacido en 1959) es un estadounidense informático, es un innovador en tecnologías que aumentan la inteligencia humana, individual y colectivamente. La aplicación de las ideas de la Inteligencia Artificial, Ciencias Cognitivas y diseño, su trabajo ha explorado cómo conectar a la gente y las máquinas pueden fomentar la colaboración, el aprendizaje, el intercambio de conocimientos, y hacer las cosas.[4]

[6] Descarga de WN: http://wndomains.fbk.eu/download.html
http://hltdistributor.fbk.eu/index.php

[7] WN online: http://multiwordnet.fbk.eu/online/multiwordnet.php

Es uno de los recursos más utilizados en PLN, es utilizado en sus diferentes versiones e idiomas es WordNet (Fellbaum, 1998). Debido a su gran repercusión, otras herramientas tales como WordNet Domains (Magnini y Cavaglia, 2000), SUMO (Niles, 2001) o WordNet Affect (Esuliy Sebastiani, 2006) han sido desarrolladas basándose en las relaciones y estructuras internas de WordNet.

En ella se definen nombres, verbos, adjetivos y adverbios. La unidad básica de información en WN es el synset (synonym sets o conjuntos de sinónimos). Un synset representa un concepto de forma léxica (Ševčenko, 2003) y se codifica como un número único de ocho dígitos llamado offset. Dentro de la base de datos, cada synset representa un concepto distinto y entre cada synset existen conexiones que expresan relaciones semánticas, conceptuales o léxicas.

Actualmente, el desarrollo de tareas para la clasificación de documentos, discriminación de entidades o detección de autoría entre otros, ha hecho patente la necesidad de disponer de ciertos recursos semánticos que proporcionen información adicional a los contextos analizados: detección de subjetividad, dominio contextual, etc. El principal problema en el uso de estos recursos es su descentralización. A pesar de que la mayoría se basa en las relaciones internas de WN, no comparten una interfaz común que pueda proporcionar información de forma cohesionada.

Esta base de datos léxica se construye sobre la base de las categorías sintácticas de nombre, verbo, adjetivo y adverbio. Dichas categorías se organizan en distintas estructuras léxicas: los nombres en jerarquías léxicas sobre la base de relaciones de hiponimia y meronimia; los verbos en base a relaciones de implicación, y finalmente, los adjetivos y adverbios se organizan como hiperespacios N-dimensionales. Sin embargo, este tipo de organización produce una redundancia de información en los casos en que una unidad léxica pertenece a más de una categoría.

Para este trabajo, utilizamos WN como base para la ontología de recuperación de los datos, tendremos un lenguaje controlado general (no especializado por materias) y para la lengua inglesa.

La ontología WN será interpretada por la herramienta Knowledge Manager para este trabajo.

3. Planificación del proyecto

Para la realización de este proyecto es necesario tomar los roles de jefe de proyecto, analista, diseñador y programador.

En el diagrama gantt se representan las tareas a realizar y el rol que las realiza. Para cada una de ellas también se indica la fecha de inicio, la fecha de fin y la duración.

También queda representado el orden entre la tarea y las tareas precedentes si las tuviera.

3.1 Tiempos planificados

El proyecto se comienza el 26 de enero con la primera toma de contacto con el tutor. Se planifica finalizar con la defensa de este proyecto que será realizada antes de finalizar el mes de septiembre.

3.2 Costes

El coste de los trabajadores para el desarrollo de este proyecto se ha establecido en el siguiente listado de costes, para todos ellos se habla de su precio por hora:

Jefe de proyecto: 125 €/hora

Analista: 83 €/hora

Técnico: 35 €/hora

ID	PUNTOS DE FUNCION	DURACION (Horas)	RECURSOS		COSTE (€)
			EMPLEADO	HORAS	
3	Toma de requisitos y especificaciones del proyecto	68	Jefe de Proyecto	20,4	6.500,80
			Analista	47,6	
7	Búsqueda de fuentes de patentes	68	Jefe de Proyecto	6,8	5.929,60
			Analista	61,2	
9	Descarga documentación	112	Técnico	112	3.920,00
10, 11	Conversor de documentos	68	Analista	61,2	5.317,60
			Técnico	6,8	
12	Crear ontología	116	Analista	92,8	8.514,40
			Técnico	23,2	
17	Obtener frecuencia de patrones	44	Técnico	44	1.540,00
18, 19	Estudio de resultados	80	Técnico	80	2.800,00
20	Conclusiones	4	Analista	2	203,00
			Técnico	2	
21	Documentación	239	Jefe de Proyecto	11,95	9.595,00
			Analista	95,6	

ID	PUNTOS DE FUNCION	DURACION (Horas)	RECURSOS		COSTE (€)
			EMPLEADO	HORAS	
			Técnico	131,45	
TOTAL					44.320,40

Tabla 1. Costes del proyecto

3.3 Gantt inicial

La planificación inicial se muestra en el siguiente diagrama de Gantt.

ID	NOMBRE TAREA	DIAS	FECHA INICIO	FECHA FIN	PREVIAS	RECURSOS
1	**Proyecto Fin de Carrera**	**247**	**26/01/2015**	**30/09/2015**		
2	**Inicio del proyecto**	**0**	**26/01/2015**	**26/01/2015**		
3	**Toma de los requisitos**	**0**	**26/01/2015**	**26/01/2015**		Jefe de Proyecto (100%)
5	Especificaciones del proyecto	16	26/01/2015	11/02/2015		Jefe de Proyecto (30%) Analista (70%)
6	Reunión con los tutores	0	11/02/2015	11/02/2015		Jefe de Proyecto (30%) Analista (70%)
7	**Búsqueda de fuentes de patentes**	**17**	**11/02/2015**	**28/02/2015**	**3**	Analista (100%)
8	Elección de fuentes	0	28/02/2015	28/02/2015	7	Jefe de Proyecto (30%) Analista (70%)
9	**Descarga documentación**	**28**	**28/02/2015**	**28/03/2015**	**8**	Técnico (100%)
10	**Búsqueda de conversor de documentos**	**16**	**15/02/2015**	**03/03/2015**	**3**	Analista (100%)
11	**Convertir documentos a PDF**	**0**	**28/03/2015**	**28/03/2015**	**10**	Técnico (100%)
12	**Crear ontología**	**29**	**10/03/2015**	**30/04/2015**	**3**	
13	Selección fuente	21	10/03/2015	31/03/2015	3	Analista (100%)
14	Obtener vocabulario	2	01/04/2015	03/04/2015	13	Técnico (100%)
15	Obtener herramienta para gestión	5	03/04/2015	29/04/2015	3	Analista (100%)
16	Gestionar la ontología	1	29/04/2015	30/04/2015	15	Técnico (100%)
17	**Obtener frecuencia de patrones**	**11**	**01/09/2015**	**12/09/2015**	**10, 16**	Técnico (100%)
18	**Estudiar escenarios individualmente**	**9**	**04/09/2015**	**13/09/2015**	**17**	Técnico (100%)
19	**Analisis general de todos los escenarios**	**8**	**14/09/2015**	**22/09/2015**	**18**	Técnico (100%)
20	**Conclusiones**	**0**	**22/09/2015**	**22/09/2015**	**19**	Analista (50%) Técnico (50%)
21	**Memoria/Documentación**	**239**	**26/01/2015**	**22/09/2015**	**3**	Jefe de Proyecto (5%) Analista (40%) Técnico (55%)
22	**Fin de proyecto**	**0**	**22/09/2015**	**22/09/2015**		

Tabla 2. Gantt inicial

3.4 Gantt Final

Durante el desarrollo del proyecto surgen varios imprevistos que no permiten cumplir con la planificación inicial. Hubo que replanificar, la nueva y última planificación es la siguiente:

ID	NOMBRE TAREA	DIAS	FECHA INICIO	FECHA FIN	PREVIAS	RECURSOS
1	**Proyecto Fin de Carrera**	**266**	**26/01/2015**	**18/10/2015**		
2	**Inicio del proyecto**	**0**	**26/01/2015**	**26/01/2015**		
3	**Toma de los requisitos**	**0**	**26/01/2015**	**26/01/2015**		Jefe de Proyecto (100%)
5	Especificaciones del proyecto	16	26/01/2015	11/02/2015		Jefe de Proyecto (30%) Analista (70%)
6	Reunión con los tutores	0	11/02/2015	11/02/2015		Jefe de Proyecto (30%) Analista (70%)
7	**Búsqueda de fuentes de patentes**	**17**	**11/02/2015**	**28/02/2015**	**3**	Analista (100%)
8	Elección de fuentes	0	28/02/2015	28/02/2015	7	Jefe de Proyecto (30%) Analista (70%)
9	**Descarga documentación**	**28**	**28/02/2015**	**28/03/2015**	**8**	Técnico (100%)
10	**Búsqueda de conversor de documentos**	**16**	**15/02/2015**	**03/03/2015**	**3**	Analista (100%)
11	**Convertir documentos a PDF**	**0**	**28/03/2015**	**28/03/2015**	**10**	Técnico (100%)
12	**Crear ontología**	**29**	**10/03/2015**	**30/04/2015**	**3**	
13	Selección fuente	21	10/03/2015	31/03/2015	3	Analista (100%)
14	Obtener vocabulario	2	01/04/2015	03/04/2015	13	Técnico (100%)
15	Obtener herramienta para gestión	5	03/04/2015	29/04/2015	3	Analista (100%)
16	Gestionar la ontología	1	29/04/2015	30/04/2015	15	Técnico (100%)
17	**Obtener frecuencia de patrones**	**46**	**01/09/2015**	**17/10/2015**	**10, 16**	Técnico (100%)
18	**Estudiar escenarios individualmente**	**46**	**01/09/2015**	**17/10/2015**	**17**	Técnico (100%)
19	**Analisis general de todos los escenarios**	**1**	**17/10/2015**	**18/10/2015**	**18**	Técnico (100%)
20	**Conclusiones**	**1**	**18/10/2015**	**18/10/2015**	**19**	Analista (50%) Técnico (50%)
21	**Memoria/Documentación**	**266**	**26/01/2015**	**18/10/2015**	**3**	Jefe de Proyecto (5%) Analista (40%) Técnico (55%)
22	**Fin de proyecto**	**0**	**18/10/2015**	**18/10/2015**		

Tabla 3. Gantt final

4. Fuentes de la información

4.1 Fuentes de patentes.

Para la obtención de la muestra de patentes se analizan las siguientes fuentes de información:

A. Clasificación internacional de Patentes (OMPI)

Se analizan las patentes con el código "WO" que pertenecen a los registros internacionales de marcas en virtud del Arreglo de Madrid y el Protocolo de Madrid relativo al Registro Internacional de Marcas.

Este tipo de patentes también se pueden obtener en el buscador de google, aunque sin la posibilidad de realizar la descarga en formato pdf.

Las patentes aquí publicadas son patentes escaneadas guardadas como imagen, con lo que imposibilita su análisis.

El enlace a esta página puede encontrarse en la bibliografía [7]

B. Oficina de Patentes y Marcas Registradas de Estados Unidos (USPTO)

Todas estas patentes comienzan con el código "US" y las podemos obtener del buscador de google. Son documentos bastante completos y nos permite la extracción del texto para realizar el análisis.

El enlace a esta página puede encontrarse en la bibliografía [8]

C. Oficina Europea de Patentes (OEP)

Éstas patentes comienzan con el código "EP", obteniéndolas de la página oficial OEP y en el buscador de google. Los documentos también son completos y permite la extracción del texto.

El enlace a esta página puede encontrarse en la bibliografía [9]

D. Instituto Mexicano de la Propiedad Industrial (IMPI)

Los documentos de patentes que aquí se pueden obtener no están en el idioma elegido para el proyecto. Además los contenidos no son muy completos.

El enlace a esta página puede encontrarse en la bibliografía [10]

E. Oficina Española de Patentes y Marcas (OEPM).

A través de Latipat de búsqueda de patentes, encontramos patentes con contenidos pequeños, no suponen una muestra importante para analizar.

El enlace a esta página puede encontrarse en la bibliografía [11]

F. Otras fuentes revisadas

- Japan Patent Office (JPO)
  El enlace a esta página puede encontrarse en la bibliografía [12]
- Korean Intellectual Property Rights Information Service (KIPRIS)
  El enlace a esta página puede encontrarse en la bibliografía [13]
- State Intellectual Property Office of the P.R.C. (SIPO). El enlace a esta página puede encontrarse en la bibliografía [14]

Se decide utilizar como muestra las patentes de la Oficina de Patentes y Marcas Registradas de Estados Unidos (United States Patent and Trademark Office, USPTO) y de la Oficina Europea de Patentes (OEP). Son seleccionadas porque en ellas conseguimos documentos de patentes más completos y de mejor calidad.

Utilizando las patentes procedentes de USPTO y OEP, basamos el análisis en similitudes y diferencias de tipología y taxonomía que hay entre las patentes estadounidenses y europeas.

4.2 Fuentes de patentes seleccionadas

En esta sección escribimos un poco más de conocimiento sobre las fuentes elegidas para el proyecto.

4.2.1 Oficina Europea de Patentes (OEP).

La Patente Europea es un procedimiento unificado para la tramitación de Patentes entre los Estados contratantes del Convenio de Munich de 5 de Octubre de 1973 y que en España entró en vigor el 1 de Octubre de 1986.

El objeto del Convenio de la Patente Europea (CPE) es conseguir que la protección de las invenciones resulte más fácil, más barata, más fiable en los Estados Contratantes, mediante la creación de un procedimiento europeo único de concesión de patentes basado en una legislación uniforme de patentes.

La Oficina Europea de Patentes es uno de los dos órganos que forman parte de la Organización Europea de Patentes, y tiene su sede en Múnich, y delegaciones en La Haya, Berlín y Viena.

ESTADOS CONTRATANTES: Austria, Bélgica, Bulgaria, Chipre, República Checa, Suiza, Liechtenstein, Alemania, Dinamarca, Finlandia, España, Francia, Reino Unido, Grecia, Irlanda, Italia, Luxemburgo, Mónaco, Holanda, Portugal, Suecia, Malta, Polonia, Eslovaquia, Turquía, Islandia, Hungría, Estonia, Lituania, Letonia, Rumania y Eslovenia.

4.2.2 Oficina de Patentes y Marcas Registradas de Estados Unidos (USPTO)

La Oficina de Patentes y Marcas de Estados Unidos (conocida en inglés como la *United States Patent and Trademark Office*, con el acrónimo *PTO* o *USPTO*) es una agencia en el Departamento de Comercio de Estados Unidos que expide patentes a los inventores y las empresas para sus inventos, y registro de marcas para la identificación de productos y propiedades intelectuales.

La oficina está basada en Alexandria, Virginia, desde 2006.

La oficina coopera con la Oficina Europea de Patentes (OEP) y la Oficina Japonesa de Patentes (OPJ) como una de las "Oficinas Trilaterales de Patentes." La USPTO es también una oficina receptora, una administración encargada para búsqueda internacional, y una autoridad para examen preliminar internacional de solicitudes internacionales para patentes presentadas en virtud del tratado de cooperación en materia de patentes.

La misión de la Oficina de Patentes y Marcas es promover el progreso industrial y tecnológico en los Estados Unidos y fortalecer la economía nacional mediante:

- la administración de las leyes relativas a patentes y marcas;
- la provisión de asesoramiento al Secretario de Comercio, al Presidente, y a la administración sobre protección de patentes, marcas comerciales, y derechos de autor; y
- la provisión de asesoramiento sobre los aspectos de propiedad intelectual que se relacionan con el comercio.

Estos datos han sido inspirados en la referencia bibliográfica [6]

4.2.3 Buscador de patentes

Como fuente para obtener las patentes, se utiliza la página oficial de OEP y el buscador de Google.

En Google realizan constantemente recopilaciones importantes de información. Google patentes permite descubrir, buscar y leer online los millones de ideas que se han enviado a las oficinas de patentes europeas o estadounidenses.

Todos los documentos disponibles a través de Google patentes proceden de la Oficina de Patentes y Marcas Registradas de Estados Unidos

(United States Patent and Trademark Office, USPTO) y de la Oficina Europea de Patentes (OEP). Las solicitudes de patentes de Estados Unidos se remontan a 1790 y las de la OEP, a 1978.

5. Knowledge Manager

Knowledge Manager (KM) es una herramienta para la gestión del conocimiento que tiene como objetivo gestionar ontologías.

KM es una herramienta creada por el Knowledge Reuse Group que permite representaciones ontológicas y jerárquicas.

Para este trabajo utilizamos la herramienta para gestionar la ontología WN, ampliamos nuestro vocabulario base para realizar el análisis sintáctico-semántico a través de la herramienta Boilerplates.

5.1 Conexión a la base de datos

Para utilizar esta herramienta es necesario solicitar una clave a The REUSE Company. La clave podrá utilizarse durante 90 días.

Como se puede ver en la siguiente imagen, la conexión se realiza a la base de datos Access llamada CAKE V14.1 que ya viene incluida en la instalación.

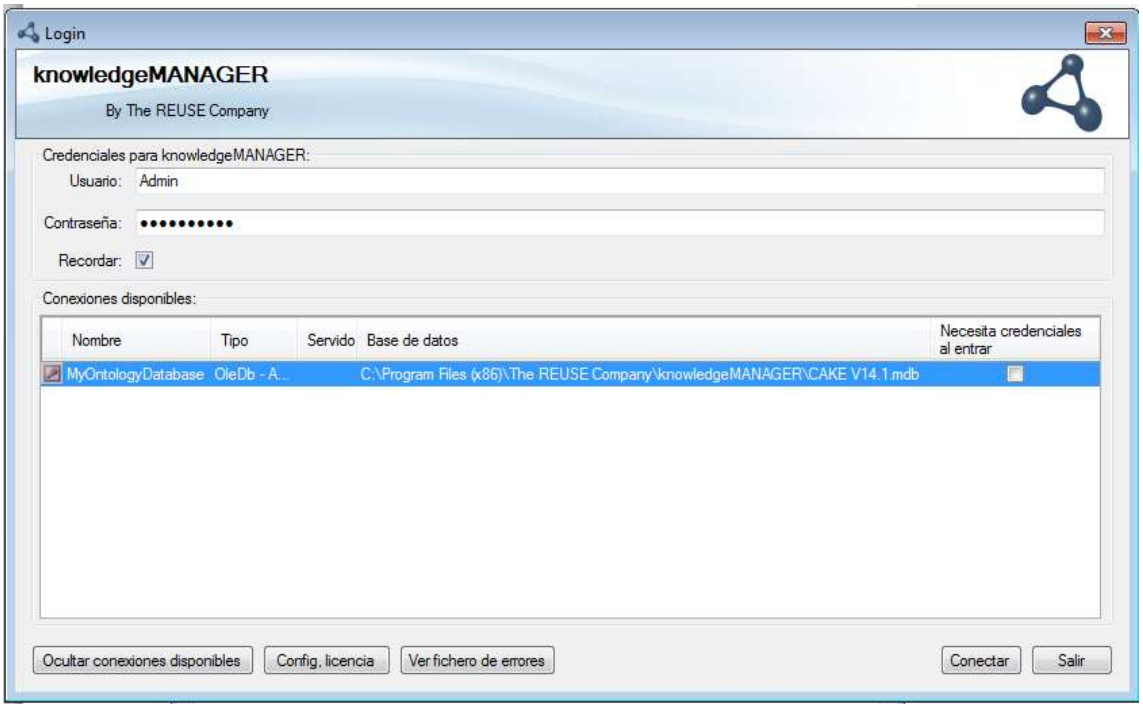


Imagen 1. KM. Conexión a Knowledge Manager

5.2 Nuevos términos

Utilizaremos el vocabulario obtenido de WN para ampliar los términos de esta base de datos. Para ello, partimos de ficheros de texto que serán importados en la herramienta KM.

Se incluye vocabulario de las categorías gramáticas nombres, verbos, adjetivos y adverbios.

En la imagen podemos ver el momento de importar el listado de los adverbios:

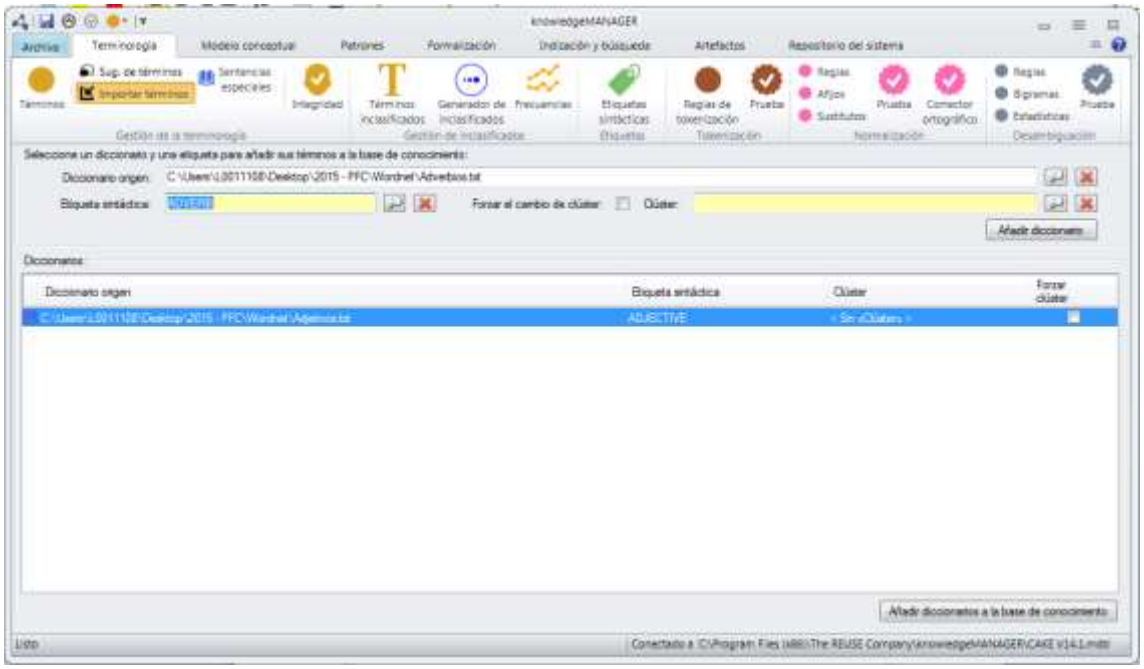


Imagen 2. KM. Incluyendo nuevos términos

La herramienta genera un informe tras añadir los nuevos términos, parte de los informes generados para este proyecto son los siguientes:

*INFORME DE IMPORTACIÓN DE DICCIONARIOS DEL FICHERO CON RUTA 'C:\Users\L0011108\Desktop\2015 – PFC\Wordnet**Nombres.txt**'*

117932 términos añadidos

1 términos no añadidos

*INFORME DE IMPORTACIÓN DE DICCIONARIOS DEL FICHERO CON RUTA 'C:\Users\L0011108\Desktop\2015 – PFC\Wordnet**Adverbios.txt***'*

762 términos añadidos

3590 términos no añadidos

*INFORME DE IMPORTACIÓN DE DICCIONARIOS DEL FICHERO CON RUTA 'C:\Users\L0011108\Desktop\2015 – PFC\Wordnet**Adjetivos.txt***'*

21205 términos añadidos

280 términos no añadidos

2 términos modificados

-'given': Cambiado la etiqueta sintáctica 'UNCLASSIFIED ADJECTIVE' por 'ADJECTIVE'

-'specified': Cambiado la etiqueta sintáctica 'UNCLASSIFIED ADJECTIVE' por 'ADJECTIVE'

*INFORME DE IMPORTACIÓN DE DICCIONARIOS DEL FICHERO CON RUTA 'C:\Users\L0011108\Desktop\2015 – PFC\Wordnet**Verbos.txt***'*

10841 términos añadidos

632 términos no añadidos

Los términos que no incluye la herramienta son debido a que ya existen en base de datos con la misma etiqueta gramatical.

Aquí ya se ve como está presente la ambigüedad de la que se ha hablado en el Estado del arte de este documento. Un mismo término está dentro de varias etiquetas, como por ejemplo el término *across* tiene la etiqueta de adverbio y de preposición, como se puede ver en la imagen:

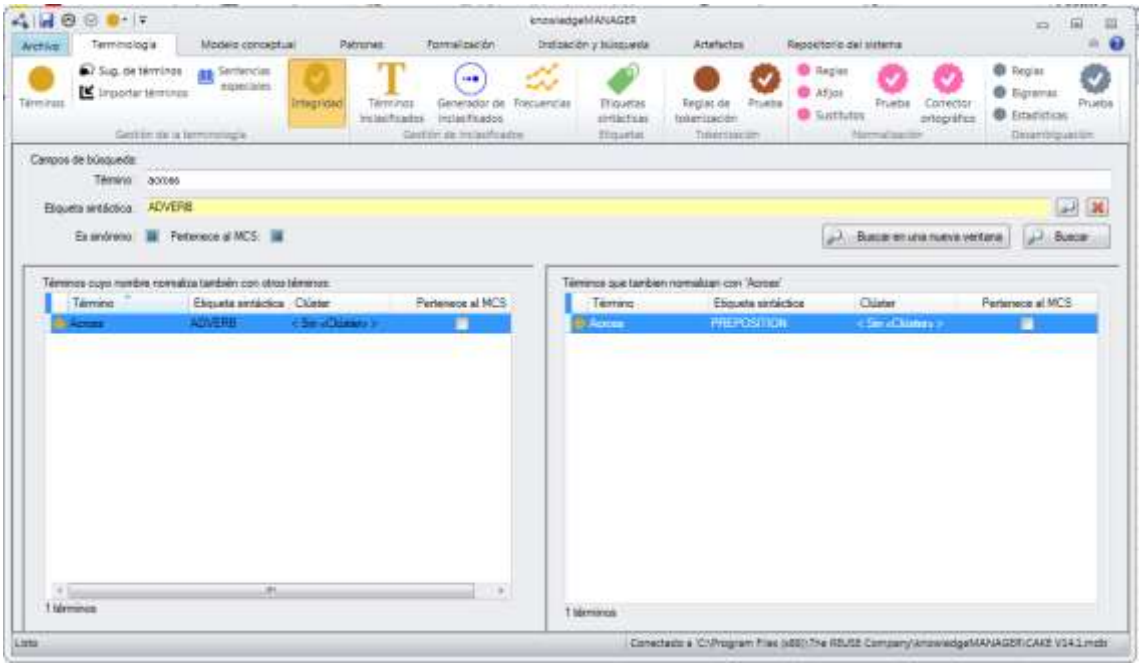


Imagen 3. KM. Ejemplo mismo término en varias etiquetas.

5.3 Reglas de tokenización

La propia herramienta tiene herramientas de tokenización para expresiones regulares como son el reemplazo por grupos, identificación de entidades, reemplazos simples, separación de símbolos, eliminación de múltiples espacios consecutivos…

La herramienta esta provista de 141 reglas de tokenización.

Con el siguiente ejemplo comprobamos como lleva a cabo la tokenización.

Ejemplo: Still the general belief among tumour immunologists is therefore that one of the best ways to eliminate tumours would be to induce a strong specific anti-tumour CTL response

En la siguiente tabla, podemos ver el resultado de KM de cómo realiza la tokenización:

Paso Nº	Id de regla	Regla	Resultado
1	-1	Initial Text	[still the general belief among tumour immunologists is therefore that 1 of the best ways to eliminate tumours would be to induce a strong specific anti-tumour ctl response]
2	131	(?<NUMBER>(^|\s)(-|\+|\+/-)?[0-9]*[\.|,]?[0-9]+([eE^][-+]?[0-9]+)?)	[still the general belief among tumour immunologists is therefore that] [1| (NUMBER)] [of the best ways to eliminate tumours would be to induce a strong specific anti-tumour ctl response]
3	132	\s(?<1>[/\\:.@(]+)(?<2>[a-za-z0-9+-]+)(?<3>[/\\:.@)]+)\s|\s(?<2>[a-za-z0-9+-]+)(?<3>[/\\:.@)]+)\s|\s(?<1>[/\\:.@(]+)(?<2>[a-za-z0-9+-]+)\s	[still the general belief among tumour immunologists is therefore that] [1| (NUMBER)] [of the best ways to eliminate tumours would be to induce a strong specific anti-tumour ctl response]
4	151	(?<1>[a-za-z]+)(?<2>[\-/])(?<3>[0-9a-za-z]+)	[still the general belief among tumour immunologists is therefore that] [1| (NUMBER)] [of the best ways to eliminate tumours would be to induce a strong specific anti - tumour ctl response]
5	165	(?<1>[,.'"“”+±\-/])	[still the general belief among tumour immunologists is therefore that] [1| (NUMBER)] [of the best ways to eliminate tumours would be to induce a strong specific anti - tumour ctl response]
6	167	[ ]+	[still the general belief among tumour immunologists is therefore that] [1| (NUMBER)] [of the best ways to eliminate tumours would be to induce a strong specific anti - tumour ctl response]

Tabla 4. KM. Aplicando reglas de tokenización

Las reglas utilizadas que aplican a este ejemplo son las mostradas en la siguiente tabla, los datos son obtenidos de la herrameinta KM:

Id de regla	Tipo de reemplazo
131	Expresiones regulares: Identificación de entidades
132	Expresiones regulares: Reemplazo por grupos
151	Expresiones regulares: Reemplazo por grupos
165	Expresiones regulares: Separación de símbolos
167	Expresiones regulares: Eliminación de múltiples espacios consecutivos

Tabla 5. KM. Ejemplo de reglas de tokenización

5.4 Reglas de normalización

Disponemos de las siguientes reglas para la normalización de los textos, los datos han sido extraídos de la herramienta KM:

Identificador	Descripción	Etiqueta sintáctica	Activada
1	INFINITIVE FROM GERUND TO BE FORM	VERB TO BE	True
2	to be verb	VERB TO BE	True
3	NORMALIZES AFIRMATIVE CONTRACTED FORMS	VERB	True
4	INFINITIVE FROM PRESENT INDICATIVE	VERB	True
5	INFINITIVE FROM PARTICIPLE/PRETERIT (ENGLISH)	VERB	True
6	INFINITIVE (OR ENG. NOUN) FROM A GERUND	VERB	True
7	NORMALIZES A COMPARATIVE ADJECTIVE	ADJECTIVE	True
8	INFINITIVE FROM PRESENT INDICATIVE (IRREGULAR)	VERB	True
9	SUBSTANTIVES (TO MASCULINE AND SINGULAR NOUNS)	NOUN	True
10	SUBSTANTIVES (TO MASCULINE AND PLURAL NOUNS)	NOUN	True
11	SUBSTANTIVES (TO FEMININE AND SINGULAR NOUNS)	NOUN	True
12	SUBSTANTIVES (TO FEMININE AND PLURAL NOUNS)	NOUN	True
13	SINGLE VERB REPLACEMENT	VERB	True
14	RETURNS ADJECTIVES FROM AN ADJECTIVE	ADJECTIVE	True
15	NORMALIZES NEGATIVE CONTRACTED FORMS	VERB	True
16	GENITIVE	NOUN	True

Tabla 6. KM. Reglas de normalización.

Utilizando el mismo ejemplo anterior, vemos el resultado tras haber aplicado la normalización:

Identificador	Frase	Texto normalizado	Etiqueta sintáctica	Semántica	Género	Número
145435	still	Still	NOUN	< Sin «Clúster» >	N/D	Invariante
2592	the	The	ARTICLE	< Sin «Clúster» >	N/D	Invariante
18405	general	General	ADJECTIVE	< Sin «Clúster» >	N/D	Invariante
53601	belief	Belief	NOUN	< Sin «Clúster» >	N/D	Invariante
1227	among	Among	PREPOSITION	< Sin «Clúster» >	N/D	Invariante
153613	tumour	Tumour	NOUN	< Sin «Clúster» >	N/D	Invariante
98352	immunologists	Immunologist	NOUN	< Sin «Clúster» >	N/D	Invariante
4991	is	Be	VERB TO BE	< Sin «Clúster» >	N/D	Invariante
518	therefore	Therefore	ADVERB	< Sin «Clúster» >	N/D	Invariante
2595	that	That	RELATIVE PRONOUN	< Sin «Clúster» >	N/D	Invariante
4949	1	1	NUMBER	< Sin «Clúster» >	N/D	Invariante
2878	of	Of	PREPOSITION OF	< Sin «Clúster» >	N/D	Invariante
2592	the	The	ARTICLE	< Sin «Clúster» >	N/D	Invariante
13059	best	Best	ADJECTIVE	< Sin «Clúster» >	N/D	Invariante
158103	ways	Ways	NOUN	< Sin «Clúster» >	N/D	Invariante
2567	to	To	PREPOSITION TO	< Sin «Clúster» >	N/D	Invariante
2205	eliminate	Eliminate	VERB	«Remove»	N/D	Invariante
153613	tumours	Tumour	NOUN	< Sin «Clúster» >	N/D	Invariante
2526	would	Would	MODAL VERB	«MODAL OPTIONAL»	N/D	Invariante
4991	be	Be	VERB TO BE	< Sin «Clúster» >	N/D	Invariante
2567	to	To	PREPOSITION TO	< Sin «Clúster» >	N/D	Invariante
10683	induce	Induce	VERB	< Sin «Clúster» >	N/D	Invariante
43646	a	A	NOUN	< Sin «Clúster» >	N/D	Invariante
27905	strong	Strong	ADJECTIVE	< Sin «Clúster» >	N/D	Invariante
143467	specific	Specific	NOUN	< Sin «Clúster» >	N/D	Invariante
48277	anti	Anti	NOUN	< Sin «Clúster» >	N/D	Invariante
10647	-	-	ARITHMETICOPERATOR	< Sin «Clúster» >	N/D	Invariante
153613	tumour	Tumour	NOUN	< Sin «Clúster» >	N/D	Invariante
N/D	ctl	< Sin Término >	< Sin ETIQUETA >	< Sin «Clúster» >	N/D	N/D
133110	response	Response	NOUN	< Sin «Clúster» >	N/D	Invariante

Tabla 7. KM. Aplicando reglas de normalización.

5.5 Patrones

Los patrones identificados para la frase utilizada como ejemplo en los puntos anteriores, se muestra en la siguiente imagen:

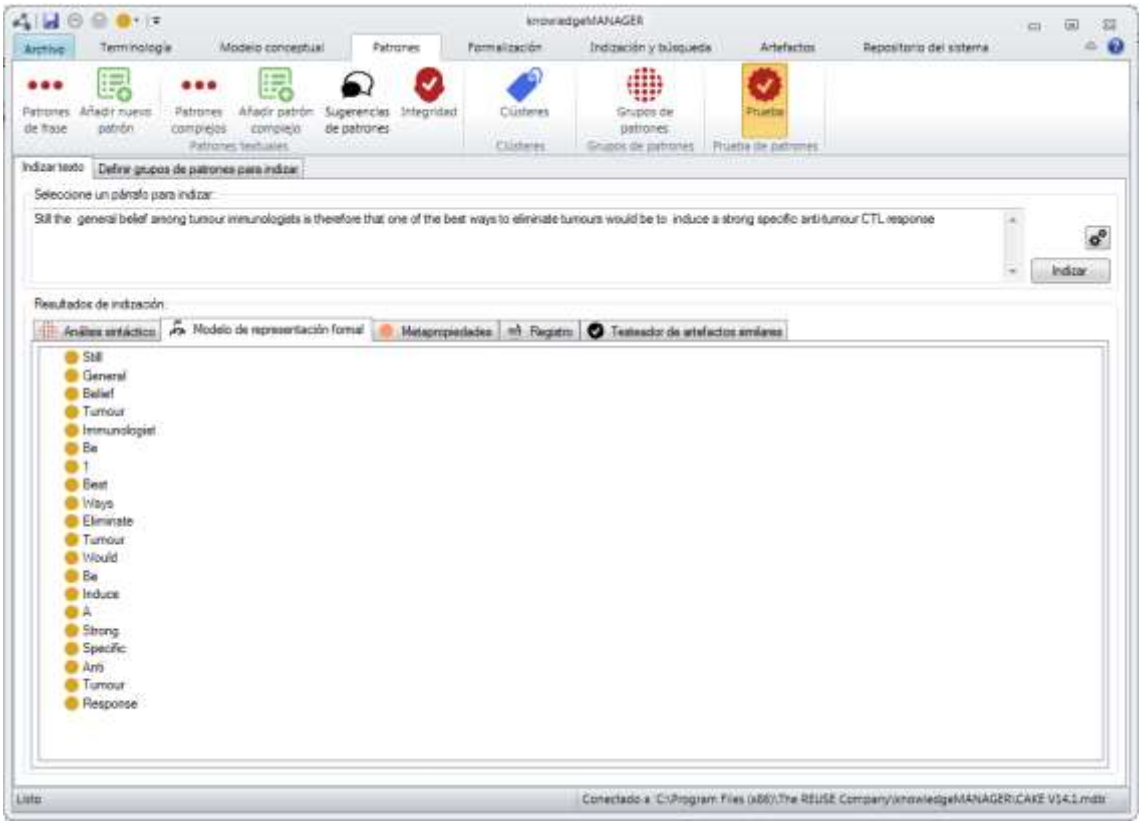


Imagen 4. KM. Ejemplo de patrones

La nueva ontología con el vocabulario de WN, con relaciones y reglas de tokenización y normalización, quedan dentro de la base de datos CAKE V14.1. Esta será la que utilicemos en BoilerPlates.

6. BoilerPlates

Es una herramienta creada por el equipo de investigación de la Universidad Carlos III de Madrid. Necesita dos bases de datos, que serán explicadas a continuación.

Con la herramienta BoilerPlates se analiza cada oración del contenido de los documentos. Internamente lo que se está realizando es un procesamiento del lenguaje natural, la herramienta es capaz de realizar un análisis léxico, análisis sintáctico y análisis semántico de cada

palabra. A cada palabra, definida como token en la herramienta, se le asigna su categoría gramatical y semántica. Además se incluye a un token dentro de un patrón si el conjunto mínimo de dos token consecutivos en el texto cumplen con la frecuencia indicada al inicio de análisis.

Como resultado tendremos palabras, patrones y jerarquía de patrones con su categoría gramatical y semántica si las tuviera. También tendremos la frecuencia de los patrones. También nos interesa conocer cuál es la longitud mayor obtenida con la jerarquía de patrones.

6.1 Bases de datos

6.1.1 Rqa Quality Analyzer v4.1 (English)

Esta base de datos Access consta de varias tablas, en la siguiente imagen se muestran algunas de las tablas principales y sus relaciones:

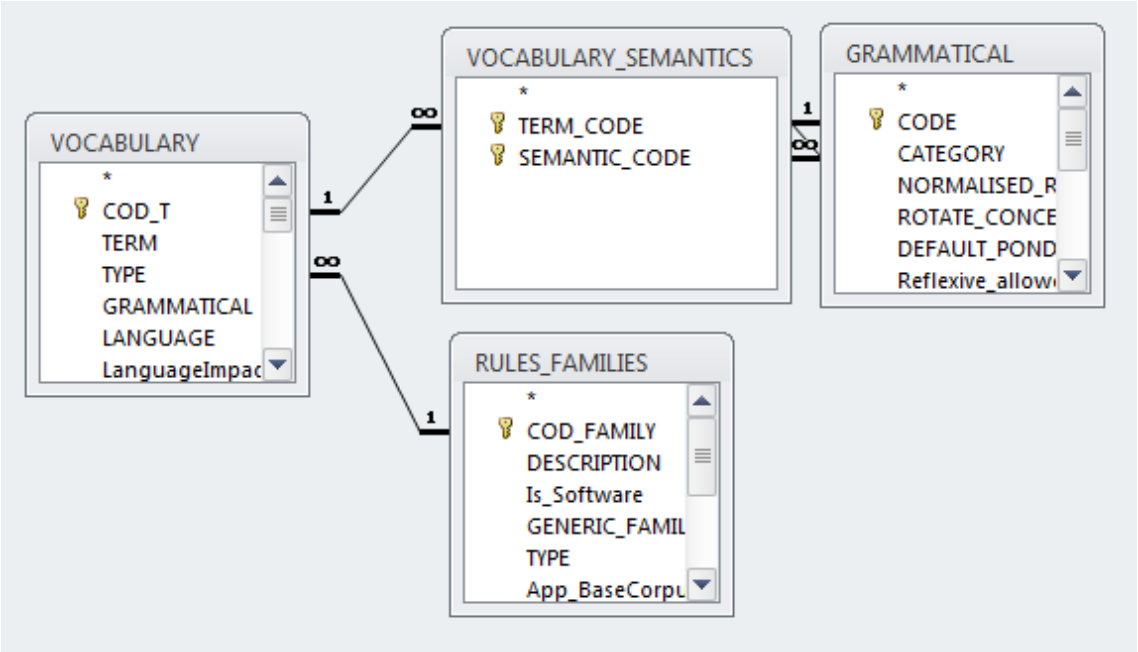


Imagen 5. BoilerPlates. RqaQualityAnalyzer v4.1 (English) – ER Simple

Esta base de datos se ha creado con la herramienta KM para este proyecto, se trata de la misma base de datos comentada anteriormente con el nombre CAKE V14.1. El procedimiento seguido para su creación se detalla en el apartado anterior.

La tabla más importante es *Vocabulary*, en la primera versión de esta base de datos, partimos con un vocabulario de 9.653 palabras con 49 familias. Con el diccionario de WordNet ampliamos el vocabulario a 155.988 palabras repartidas en 52 familias. La ontología tiene etiquetas creadas para 172 categorías gramaticales, de las que 52 de ellas tienen vocabulario asignado.

Una de las categorías que no tienen vocabulario asignado pero que tampoco tiene que tenerlo, es "UNCLASSIFIED NOUN". Todos los términos que no sean reconocidos por el vocabulario que forma la ontología se etiquetaran de esta manera.

El resto de las 73 categorías gramáticas creadas sin vocabulario asignado son las mostradas en la siguiente tabla:

ACRONYMS	INDEFINITE ARTICLE	PERIPHRASIS PARTICIPLE VERB 3
ADVERBIAL PHRASE	INITIALIZATION VERB	PHRASAL VERB
AGENTIVE VERB	INSTRUCTION VERB	PLACE
APOSTROPHE	INTERJECTION	PLUS
ASPECTUAL VERB	INTERNET URL	PREPOSED ADJECTIVE
ASTERISK	INTERROGATIVE ADVERB	PREPOSITION FOR
CARDINAL POINT	INTERROGATIVE PRONOUN	PROPER NOUN
CAUSATIVE VERB	LEFT SLASH	PUNCTUATION MARK
CLOSING ANGLE BRACKETS	LOCATION VERB	QUOTATION MARKS
CLOSING BRACE	MODAL PHRASE	RelationalOperator
CLOSING EXCLAMATION MARK	MONTH	RELATIVE ADVERB
CLOSING QUESTION MARK	MOVEMENT VERB	RELATIVE DETERMINER
CLOSING SQUARE BRACKETS	NATIONAL ADJECTIVE	SEMICOLON
COLON	NOT_Punctuation_Mark	SENTENCE BREAKER
CONECTOR PURPOSE/AIM	OBJECT PRONOUN	STOP
DASH	OBJECT-DATIVE PRONOUN	STOP WORD
DATE	OPEN ENUMERATION	UNCLASSIFIED ADVERB
DATIVE PRONOUN	OPENING ANGLE BRACKETS	UNCLASSIFIED PROPER NOUN
DAY	OPENING BRACE	UNCLASSIFIED VERB
DEFINITE ARTICLE	OPENING EXCLAMATION MARK	UNIT OF CURRENCY
DIVISION	OPENING QUESTION MARK	UNIT OF DISTANCE
ELLIPSIS	OPENING SQUARE BRACKETS	UNIT OF SURFACE
E-MAIL	PERCENTAGE	UNIT OF TEMPERATURE
ENDING VERB	PERIPHRASIS PARTICIPLE VERB 1	

EQUAL (Indexer)	PERIPHRASIS PARTICIPLE VERB 2	

Tabla 8. Categorías gramaticales vacías en ontología.

Las nuevas familias que se crean son:

PARTICLE. Se crea esta categoría grammatical para las partículas, tienen los siguientes términos: "meanwhile", "while", "when", "like", "just as", "so on", "so and so", "in case that", "whether or not", "whether", "in order that", "even though", "as", "only if", "of course", "not only", "as well as", "as long as", "as far as", "along with", "above all", "even so", "even if", "despite the fact that", "but rather", "but also" y "althoug".

PLACE ADVERBIAL PHRASE. Se tartan de las siguientes locuciones adverbiales: "up and down", "round and round", "straight down", "straight up", "straight ahead", "in and out", "how far", "hither and thither", "from side to side" y "back and forth".

POSSESIVE PRONOUN. Los pronombres posesivos que aquí se incluyen son: "hers", "his", "its", "mine", "ours", "theirs" y "yours".

REQUIREMENTS Domain. Los términos que aquí se incluyen son "level", "unit", "value", "period", "times", "rate" Y "interval".

ArithmeticOperator. Los operadores aritméticos son “+”, “-“ y “/”.

ARTICLE. Los artículos para esta categoría gramatical son: “the”, “and” y “a”.

PHRASAL VERB ABSOLUTE BASE. La base de verbos absolutos se define con “originate”, “write” y “belong”.

CAUSE. Los términos que aquí se incluyen son “that is why” y “because”.

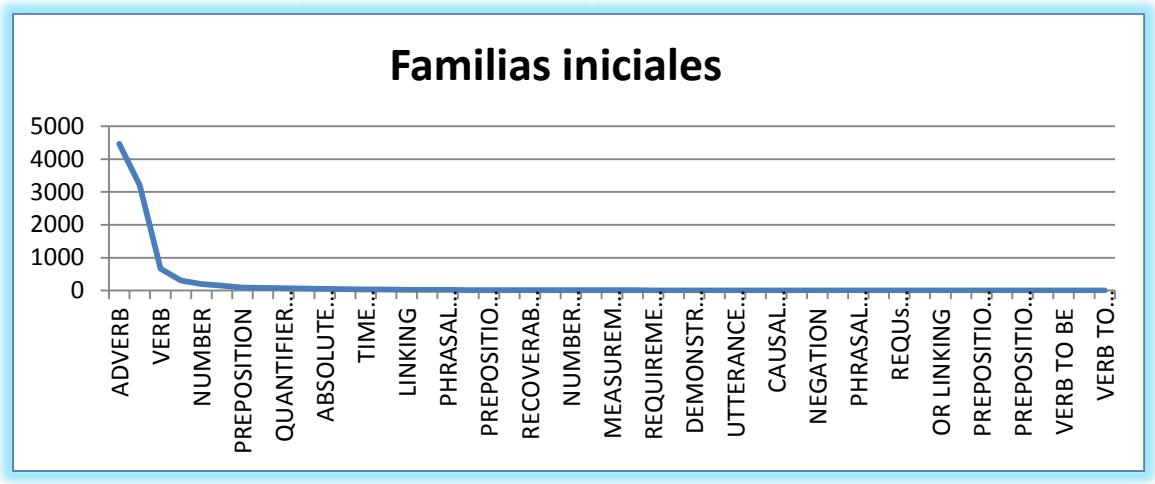
ABBREVIATION. Se crea esta familia para incluir la abreviatura de ejemplo: “e.g.”

OPENING ROUND BRACKETS. Creada para el paréntesis de apertura “(“.

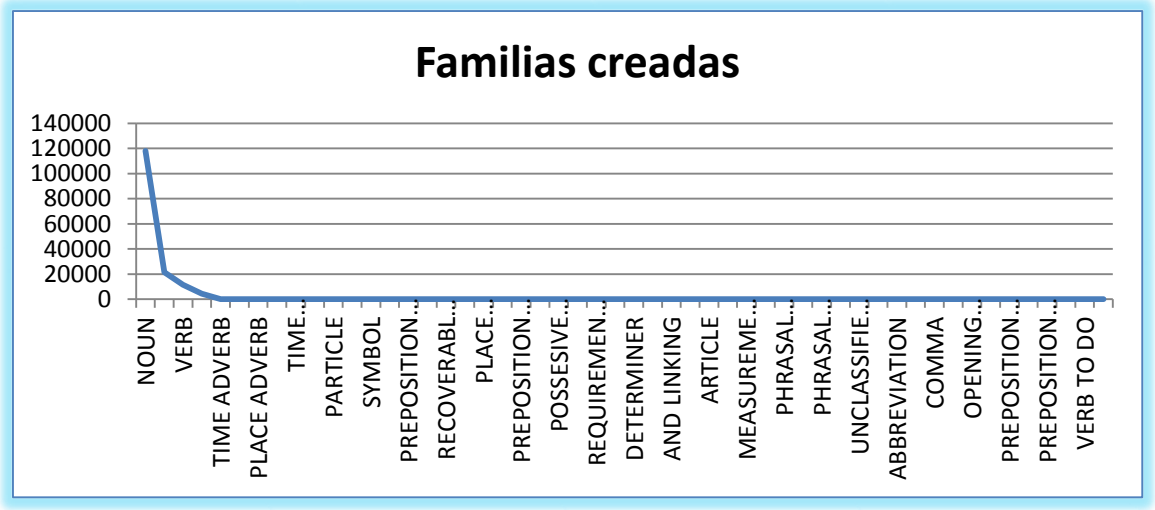
CLOSING ROUND BRACKETS. Creada para el paréntesis de cierre: “)”.

COMMA. Creada para la coma. “,”.

CONECTOR REQUIREMENT/CONDITION. Creado para el condicional “if”.



Gráfica 5. BoilerPlates. Familias iniciales



Gráfica 6. BoilerPlates. Familias creadas

En el Anexo III podemos ver el detalle de las familias que han sido creadas.

6.1.2 RequirementsClassification

También del sistema Access, esta es una segunda base de datos utilizada para conexión de la herramienta BoilerPlates, donde se generan los patrones base a partir de los documentos elegidos y se basa en las tablas de la base de datos anterior.

En la siguiente imagen se pueden ver las tablas que contiene y las relaciones entre ellas:

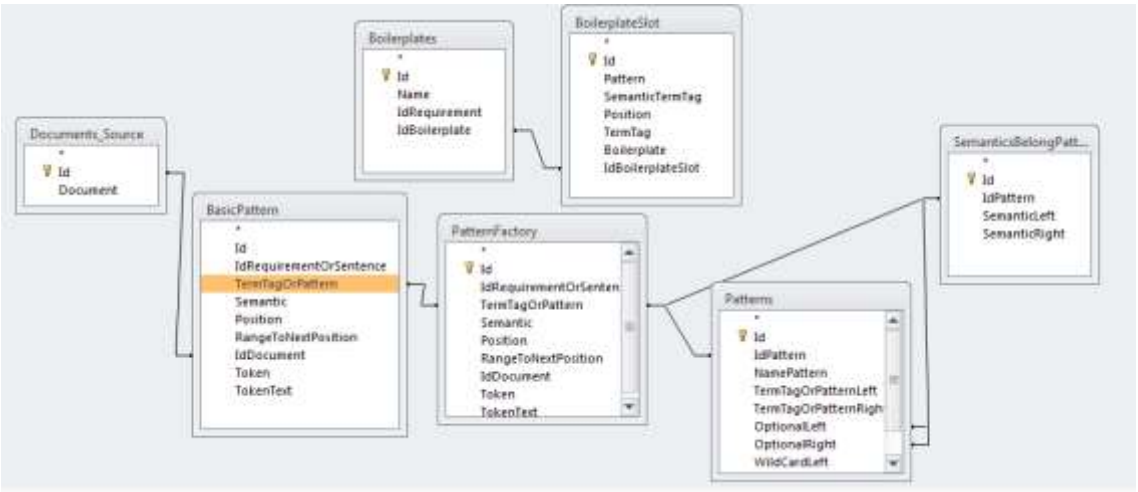


Imagen 6. BoilerPlates. RequirementsClassification – Entidad Relación

Antes de comenzar con el uso de la herramienta, todas las tablas de esta base de datos están vacías. Una vez se finalice con la generación de patrones, tendremos en ellas el resultado del análisis realizado.

Tras la generación de los patrones, se podrá analizar el contenido de las tablas dándolas forma para representar el resultado.

Su tabla principal es *BasicPattern*, donde quedarán creados todos los tokens con el vocabulario contenido en los documentos. Se puede saber de qué documento procede cada token, cuál es la frase a la que pertenece, a que gramática pertenece y que semántica tiene. La semántica la conocemos si se relaciona el campo Semantic con la tabla Grammatical de la base de datos RqaQualityAnalyzer v4.1 (English).

También se puede conocer la categoría gramatical si se cruza el campo TermTagOrPattern con la tabla Rules_Fmailies de RqaQualityAnalyzer v4.1 (English).

Todos los documentos que se analicen quedan registrados en la tabla Documents_Source.

La tabla Boilerplates contiene los boilerplates (patrones semánticos) que capturan cada frase de los documentos y los elementos que lo forman están dentro de la tabla BoilerplateSlot, con la que se puede saber la categoría gramatical y la semántica de cada uno. El análisis de esta parte queda fuera del alcance de esta investigación.

Todos los patrones generados por la herramienta se almacenan en la tabla Patterns. La relación de los patrones con los token está en la tabla PatternFactory. En la tabla PatternFactory también están los token que no forman parte de un patrón.

En Pattern se puede ver cómo está formado el patrón, si por jerarquía de patrones, por termtag o por mezcla de ambos.

En la tabla SemanticBelongPatterns están las semánticas asociadas a los elementos de los patrones.

6.2 Conexión a la base de datos

Primeramente hay que conectarse a la base de datos, como se puede ver en la imagen, la conexión se realiza a la base de datos Access explicada anteriormente, *RequirementsClassification*

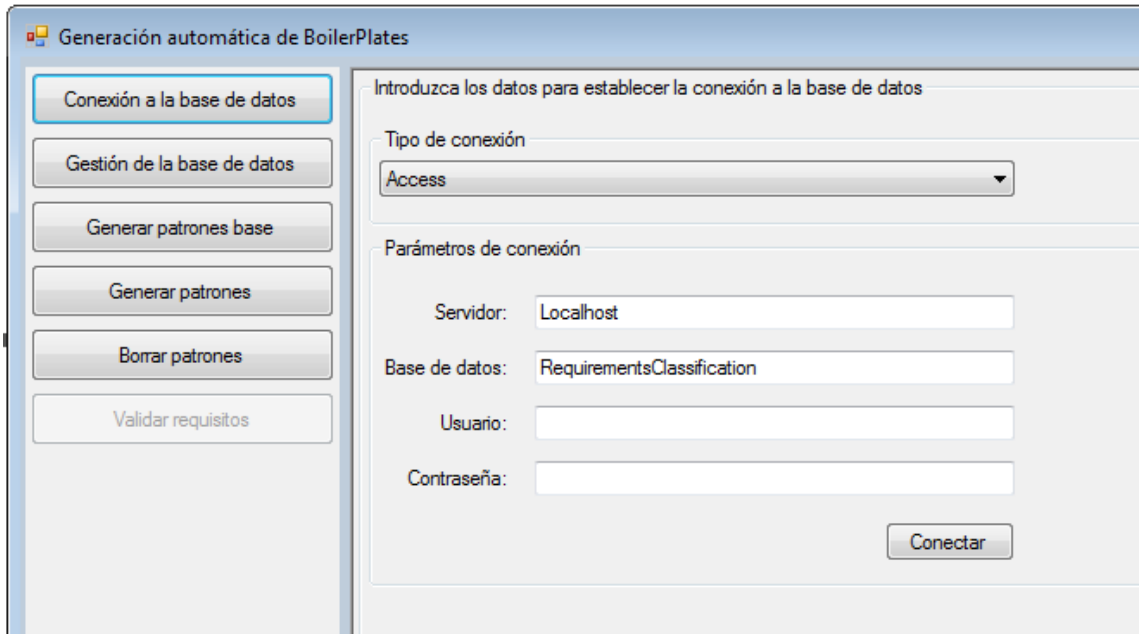


Imagen 7. BoilerPlates. Conexión a la base de datos.

Tras pulsar el botón de Conectar nos aparece el siguiente mensaje:

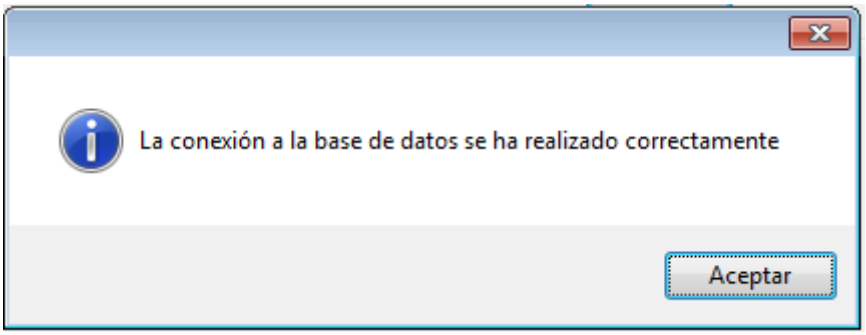


Imagen 8. BoilerPlates. Mensaje de conexión correcta.

6.3 Gestión de la base de datos

En esta sección podremos borrar todos los elementos que ya estuvieran creados anteriormente, de esta manera podremos empezar un nuevo análisis.

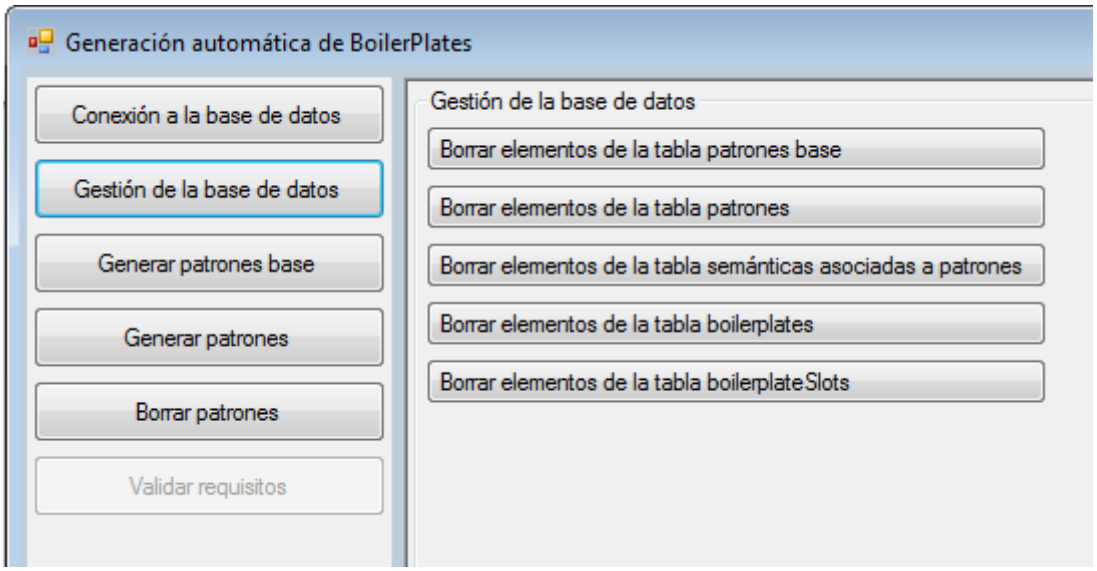


Imagen 9. BoilerPlates. Gestión de la base de datos.

6.4 Generar patrones base

En esta parte de la herramienta es donde incluiremos todos los documentos que se quieren analizar, seleccionando los documentos uno a uno y pulsando el botón *Iniciar proceso de generación de patrones* por cada uno de ellos.

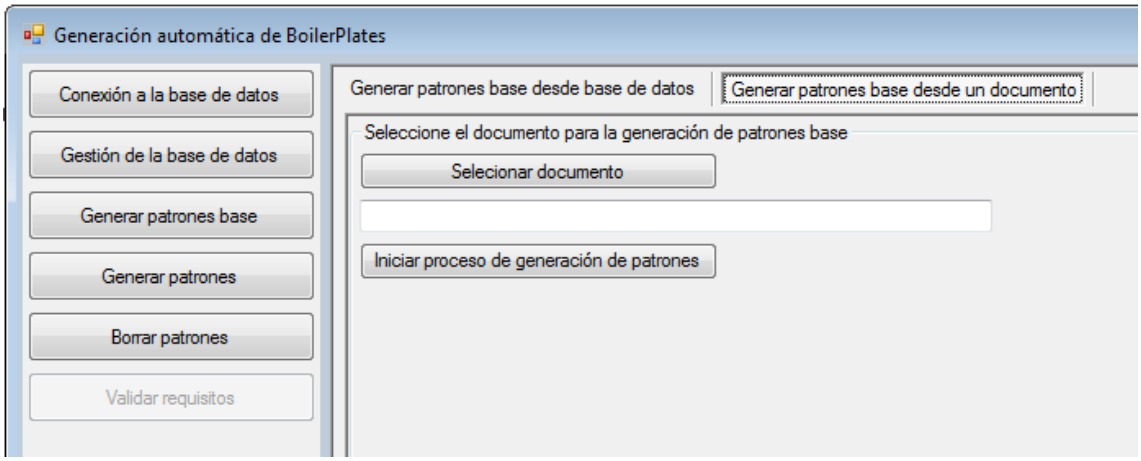


Imagen 10. BoilerPlates. Generar patrones base desde un documento

Al mismo tiempo que se incluye cada uno de los documentos se van creando sus patrones base, la herramienta extrae cada frase

reconociendo diferente frase cuando se encuentra con un "." Y extrae cada palabra que denominará token.

En la pestaña Generar patrones base desde base de datos tenemos todos los documentos que ya han sido incluidos en la herramienta.

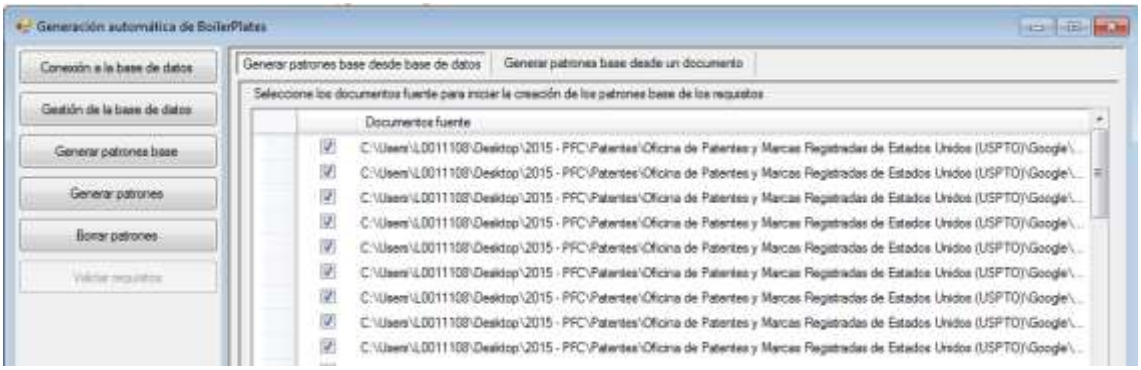


Imagen 11. BoilerPlates. Generar patrones base desde base de datos

6.5 Generar patrones

Una vez que los patrones básicos se han creado, se procede a crear los patrones. Esta parte de la herramienta nos permite seleccionar los patrones básicos que queremos analizar, las categorías gramaticales que se quieren identificar, la frecuencia de aparición de cada categoría gramatical que tendremos en cuenta y la posibilidad de diferenciar los patrones por su semántica.

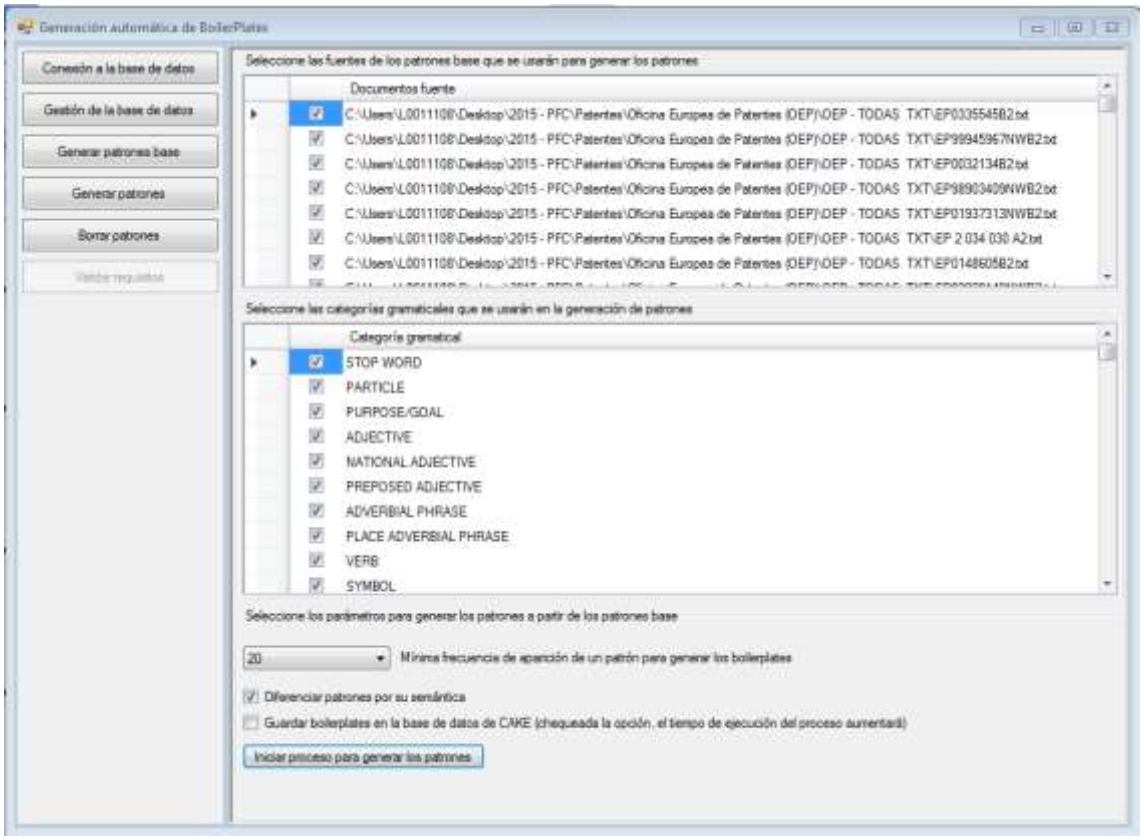


Imagen 12. BoilerPlates. Generar patrones.

Los escenarios que se establezcan en esta investigación serán representados en este momento.

La casilla "Guardar boilerplates en la base de datos de CAKE (chequeada la opción, el tiempo de ejecución del proceso aumentará)" no será chequeada en esta investigación porque estamos manejando un volumen muy algo de palabras.

6.6 Borrar patrones

Tras finalizar el proceso de generación de patrones, la herramienta nos permite administrar los mismos, sustituyendo los patrones por Slots opcionales o comodines.

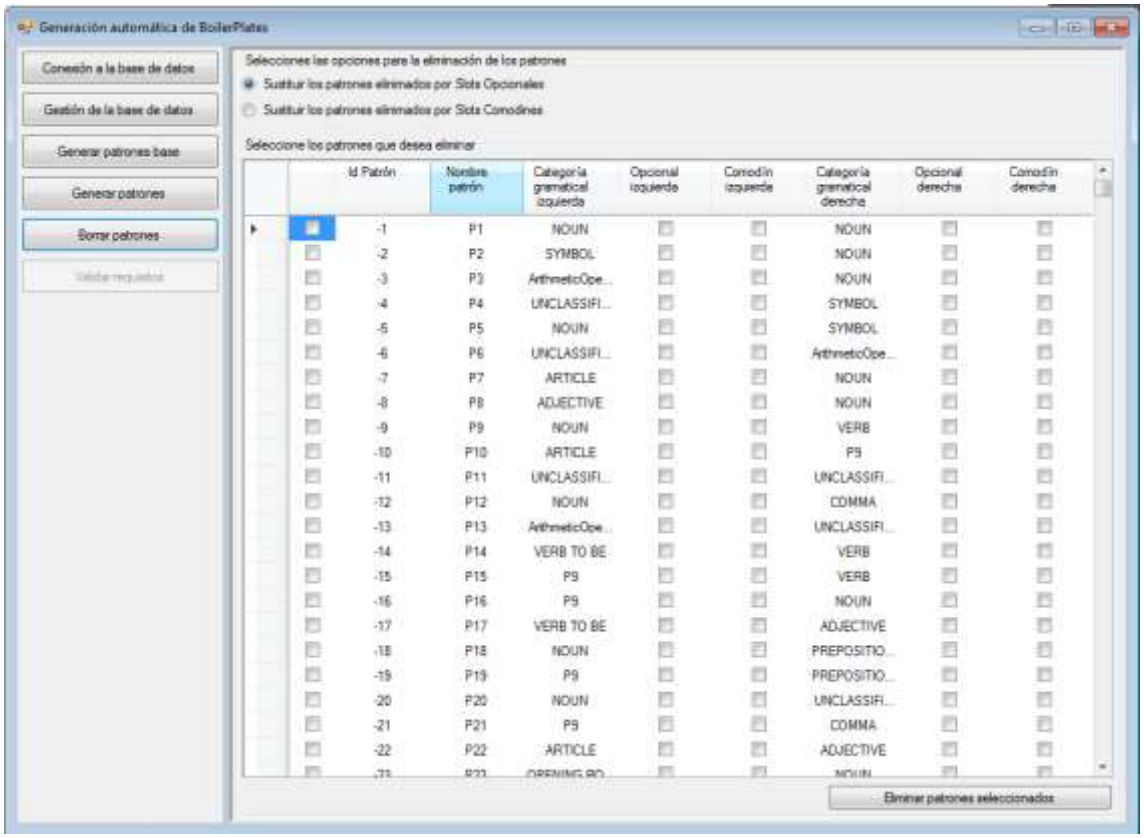


Imagen 13. BoilerPlates. Borrar patrones.

7. Requisitos del estudio

En este apartado se reflejan los requisitos establecidos para el estudio del proyecto, son los definidos en la siguiente tabla:

ID	NOMBRE	DESCRIPCION	OPCIONAL	PRIORIDAD	PRE CONDICIONES	POST CONDICIONES
#1	Idioma	Utilizar el idioma inglés como lenguaje de análisis.	NO	1	NO APLICA	NO APLICA
#2	Patentes	Los documentos de patentes tienen que ser de documentos de patentes registradas y públicas.	NO	2	NO APLICA	NO APLICA
#3	Documentos	Formato de los documentos en PDF.	NO	3	NO APLICA	NO APLICA
#4	Temática	No se especifica ninguna temática para los documentos.	SI	4	NO APLICA	NO APLICA

ID	NOMBRE	DESCRIPCION	OPCIONAL	PRIORIDAD	PRE CONDICIONES	POST CONDICIONES
#5	Muestras	Obtener al menos dos grupos de muestras para analizar.	SI	5	NO APLICA	NO APLICA
#6	Fuentes	La procedencia de los documentos no se especifica.	SI	6	NO APLICA	NO APLICA
#7	Análisis	Realizar análisis sintáctico-semántico con la herramienta BoilerPlates.	NO	7	NO APLICA	NO APLICA
#8	Ontología	Crear una nueva ontología utilizando el vocabulario de WordNet	NO	8	NO APLICA	NO APLICA
#9	Gestión ontología	Herramienta KnowledgeMANAGER, de REUSE Company, para la gestión de la ontología.	NO	9	NO APLICA	NO APLICA
#10	Escenario	Utilizar todas las categorías gramaticales disponibles en la ontología para el análisis.	NO	10	NO APLICA	NO APLICA
#11	Escenario	Utilizar la mínima frecuencia disponible de la herramienta BoilerPlates.	NO	11	NO APLICA	NO APLICA
#12	Escenario	Utilizar la máxima frecuencia disponible de la herramienta BoilerPlates.	NO	12	NO APLICA	NO APLICA
#13	Escenario	Utilizar otra frecuencia diferente a la máxima y la mínima disponible en BoilerPlates.	SI	13	NO APLICA	NO APLICA
#14	Escenario	No diferenciar patrones por su semántica en alguno de los escenarios.	SI	14	Deberá existir otro escenario con la misma frecuencia en el que si se diferencien los patrones por semántica	NO APLICA

Tabla 9. Requisitos del estudio.

Los requisitos obligatorios son: utilizar patentes que estén escritos en inglés (#1). Los documentos de patentes tienen que ser documentos de patentes registradas y públicas (#2). El formato de los documentos de patentes tiene que ser en PDF (#3).

La herramienta para realizar el análisis sintáctico-semántico debe ser BoilerPlates (#7).

Crear una nueva ontología del idioma inglés en la que se recoja el vocabulario que recoge WordNet (#8). La gestión de dicha ontología se realizará con la herramienta KnowledgeMANAGER, de REUSE Company (#9).

A la hora de comenzar el análisis en todos los escenarios que se establezcan, se deberán utilizar todas las categorías gramaticales disponibles (#10).

También tenemos obligatorio utilizar la frecuencia mínima (#11) y máxima (#12) disponible en la herramienta BoilerPlates, tendrá que formar parte en alguno de los escenarios.

Los requisitos opcionales serían: la elección de la fuente para obtener los documentos de patentes (#6), y la temática de los documentos es libre (#4).

Con la idea de disponer más patrones para comparar, se pueden obtener al menos dos grupos diferentes, ya puede ser por la temática, la procedencia u otras variables (#5).

Utilizar, además de la mínima y máxima disponible, otra frecuencia intermedia (#13).

No diferencias patrones por su semántica es opcional (#14), pero si se utiliza esta opción para alguno de los escenarios, deberá existir otro escenario con la misma frecuencia en el que si se diferencien los patrones por su semántica, con el fin de poder comparar los resultados obtenidos.

8. Escenarios

Como indica el requisito con identificador #7 se utiliza la herramienta BoilerPlates para realizar los diferentes análisis sintáctico-semánticos.

La ontología es creada con el vocabulario ofrecido por WordNet y su gestión se ha realizado con la herramienta knowledgeMANAGER como indican los requisitos #8 y #9. Los pasos realizados se han detallado en el apartado 5 de este documento.

Atendiendo al requisito opcional con identificador #5, se van a diferenciar dos muestras de patentes, por una lado analizaremos documentos de la Oficina de Patentes y Marcas Registradas de Estados Unidos (USPTO), de las que tenemos 359 documentos, y por otro lado

analizaremos documentos de la Oficina Europea de Patentes (OEP), de las europeas disponemos de 379 documentos diferentes.

Las muestras no estarán dirigidas a ninguna temática en concreto, como nos daba la libertad el requisito opcional con identificador #4.

Ambas muestras están formadas por documentos escritos en inglés de patentes registradas y públicas en formato PDF. Con éstos puntos cumplimos con los requisitos obligatorios #1, #2 y # 3.

Para todos los escenarios se tienen en cuenta todas las categorías gramáticas, así se obtendrá un resultado más completo y cubrimos el requisito con identificador #10.

Se han creado ocho escenarios diferentes, el primer escenario y los dos últimos son creados con el objetivo de conocer la diferencia, en los patrones finales, cuando se utiliza semántica. Con éstos tres escenarios estamos llevando a cabo el requisito opcional con identificador #14. Son los escenario 1, 7 y 8. Para cumplir con la precondición de este requisito, se han creado los escenarios 2, 5 y 6.

Las parejas 1-2, 5-7 y 6-8 son análogas en la elección de las categorías gramaticales, el mínimo de frecuencia y se diferencian en que uno de ellos se distingue por semántica mientras que en el otro no.

Los escenario 1 y 2 cumplen con el requisito obligatorio #11, al utilizar la frecuencia mínima disponible en la herramienta.

Utilizar la frecuencia máxima disponible en la herramienta, es otro de los requisitos obligatorios (identificador #12), y éste se cumple en los escenarios 5, 6, 7 y 8.

Los escenarios 3 y 4 utilizan otra frecuencia diferente a la máxima y la mínima, se trata de la frecuencia 20. Con estos dos escenarios se realiza el requisito opcional con identificador #13.

La pareja 3-4 son análogas en la elección de las categorías gramaticales, el mínimo de frecuencia y ambos diferencian patrones por su semántica.

En los escenarios 1, 2, 3, 5 y 7 se utiliza la muestra obtenida de patentes estadounidenses, mientras que para los patrones 4, 6 y 8 se generan para la muestra europea.

En modo resumen los escenarios que en este proyecto se establecen son:

Escenario 1:

- Muestra de patentes USPTO

- Se utilizan todas las categorías gramaticales disponibles.
- Utilizar un mínimo de frecuencia de 1 para crear patrones.
- Diferenciar patrones por su semántica está desactivado.

Escenario 2:

- Muestra de patentes USPTO
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 1 para crear patrones.
- Diferenciar patrones por su semántica está activado.

Escenario 3:

- Muestra de patentes USPTO
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 20 para crear patrones.
- Diferenciar patrones por su semántica está activado.

Escenario 4:

- Muestra de patentes OEP
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 20 para crear patrones.
- Diferenciar patrones por su semántica está activado.

Escenario 5:

- Muestra de patentes USPTO
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 100 para crear patrones.
- Diferenciar patrones por su semántica está activado.

Escenario 6:

- Muestra de patentes OEP
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 100 para crear patrones.
- Diferenciar patrones por su semántica está activado.

Escenario 7:

- Muestra de patentes USPTO
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 100 para crear patrones.
- Diferenciar patrones por su semántica está desactivado.

Escenario 8:

- Muestra de patentes OEP
- Utilizar todas las categorías gramaticales.
- Utilizar un mínimo de frecuencia de 100 para crear patrones.
- Diferenciar patrones por su semántica está desactivado.

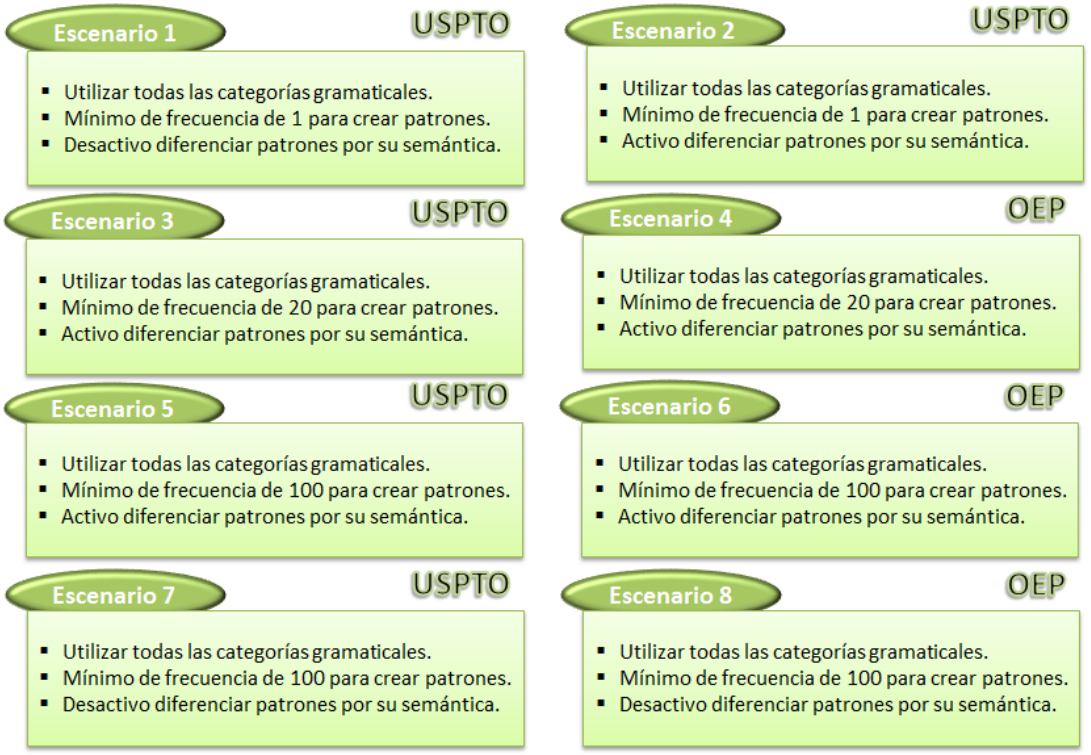


Tabla 10. Escenarios creados para el estudio.

Además de analizar cada uno de los escenarios por separado, se realizaran comparaciones entre los pares 1-2, 3-4, 5-6 y 7-8 para poder comparar las dos fuentes de información utilizadas.

También se realizará un análisis-comparativo de todos los escenarios en común para obtener conclusiones generales a todo lo analizado.

9. Detalle de los resultados obtenidos

En esta sección se va a explicar todos los resultados obtenidos de los escenarios creados. Con las diferentes opciones elegidas en BoilerPlaites descritas en el apartado anterior.

9.1 Patrones básicos

Debido a las limitaciones de la base de datos Access, nos hemos visto obligados a realizar la investigación con sólo cien de los documentos de la muestra OEP. A pesar de que para la muestra OEP tan sólo habíamos conseguido 20 patrones más que para la muestra de USPTO,

el volumen de palabras registradas en OEP casi dobla al volumen de palabras de USPTO. Si ha sido posible generar los patrones básicos para el total de las muestras conseguidas, pero no ha sido posible generar los patrones para la muestra completa de patentes europeas.

La solución tomada ha sido utilizar cien de los documentos europeos porque con ellos ya se igualaba el número de palabras con la muestra estadounidense y porque con ellos sí que ha sido posible generar los patrones que estábamos buscando.

Por lo tanto, hemos analizado los patrones básicos europeos en su totalidad y los patrones básicos europeos que se corresponden con los cien documentos seleccionados. Estos dos análisis los comparamos entre sí y para comparar con la muestra estadounidense utilizaremos los cien documentos europeos. Los resultados obtenidos los detallamos a continuación.

Los patrones básicos generados son los mismos para todos los escenarios creados de los documentos obtenidos de USPTO. Lo mismo ocurre con los patrones básicos obtenidos de los documentos procedentes de OEP, todos los escenarios dan el mismo resultado.

Esto es así porque los patrones básicos están formados por las palabras obtenidas directamente de los documentos.

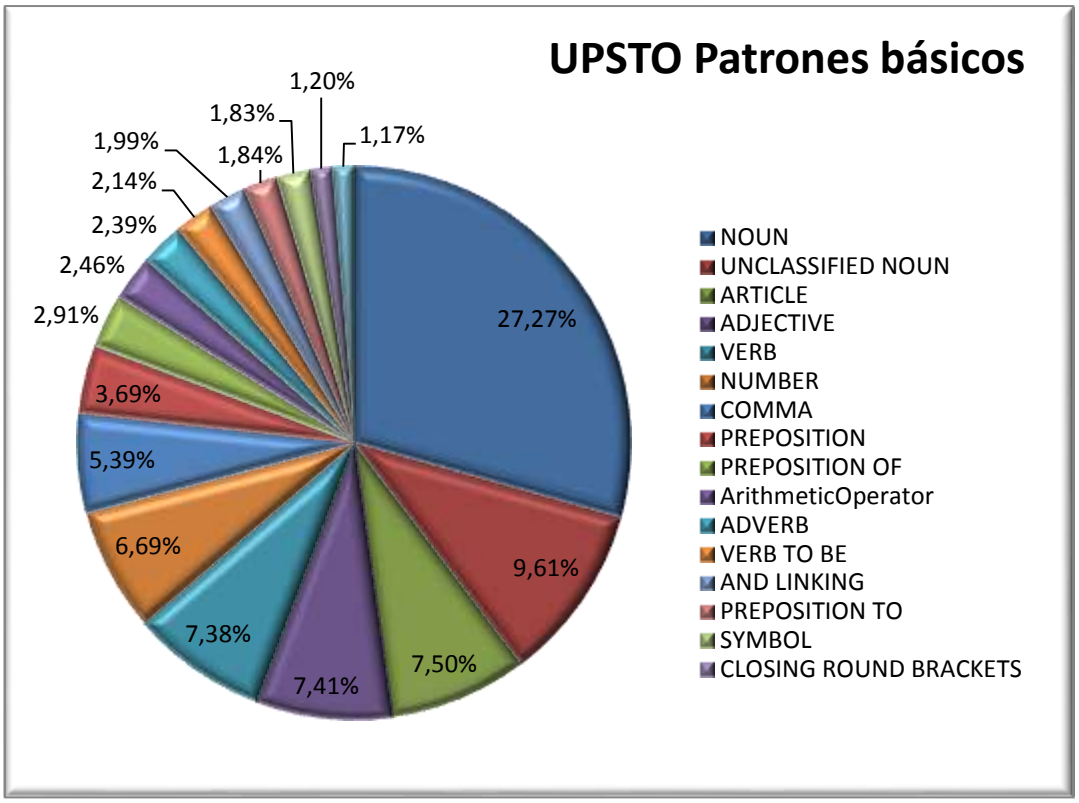
9.1.1 Patrones básicos USPTO

Para los 359 documentos de los que se dispone, se han extraído 2.089.157 palabras (o token para la herramienta) y hay un total de 130.382 frases.

9.1.1.1 Categorías gramaticales

Partimos con 54 categorías gramaticales en la ontología (denominadas TermTag en la herramienta BP) en las siguientes gráficas se muestran las categorías gramaticales que más aparen en las muestras.

Aparecen 53 categorías gramaticales. Considerando aquellas que superan el 1% de aparición, son 17 categorías gramaticales las más repetidas en la muestra de documentos de patentes de UPSTO y son las mostradas en la siguiente gráfica:



Gráfica 7. Patrones básicos. Categorías gramaticales UPSTO

El resto de categorías no están mostradas en la gráfica, porque su porcentaje de aparición en los documentos está por debajo del 1%, y porque el mostrarlas dejaría una gráfica con demasiados datos y muy poco visual.

En el apartado 8.1.3 puede verse en la tabla 12 todas las categorías gramaticales que aparecen en la muestra, con su número de frecuencia y el porcentaje que supone en toda ella.

9.1.1.2 Semántica

Tenemos un total 83.966 palabras con semántica reconocida por la ontología, es un 4% de todas las palabras que recogen las muestras USPTO. En la siguiente tabla se muestran los tipos de semántica que más aparecen. Son 19 los mostrados porque son los que superan el 1%.

Semántica	Frecuencia	%
RANGE <= (MAXIMUM)	16.846	20,06%

Semántica	Frecuencia	%
MODAL OPTIONAL	14.848	17,68%
RANGE >= MINIMUM	9.065	10,80%
Deny	3.609	4,30%
RANGE ALL	3.179	3,79%
Operation	3.073	3,66%
RANGE < MAXIMUM	2.355	2,80%
WHEN ACTIVATION	2.247	2,68%
RANGE ANY	2.187	2,60%
RANGE LITTLE-FEW-SOME	2.027	2,41%
RATE	1.927	2,29%
RANGE BETWEEN	1.849	2,20%
RANGE > MINIMUM	1.809	2,15%
MODAL FUTURE	1.794	2,14%
Provide	1.725	2,05%
CONSTRAINT	1.509	1,80%
IF ACTIVATION	1.365	1,63%
UNIT	1.234	1,47%
Verify	853	1,02%

Tabla 11. Patrones básicos. Semántica USPTO

Son tres tipos de semántica los que más destacan en toda la muestra, que para entenderlo mejor listamos las palabras que lo forman:

RANGE <= (MAXIMUM) son las palabras del tipo: in, into, within, before, maximum, last, inside, down, beneath y inferior.

MODAL OPTIONAL son las palabras del tipo: Tokentext, may, can, would y should

RANGE >= MINIMUM son las palabras del tipo: tokentext, from, over, beyond, minimum, besides, outside y atop

9.1.2 Patrones básicos OEP

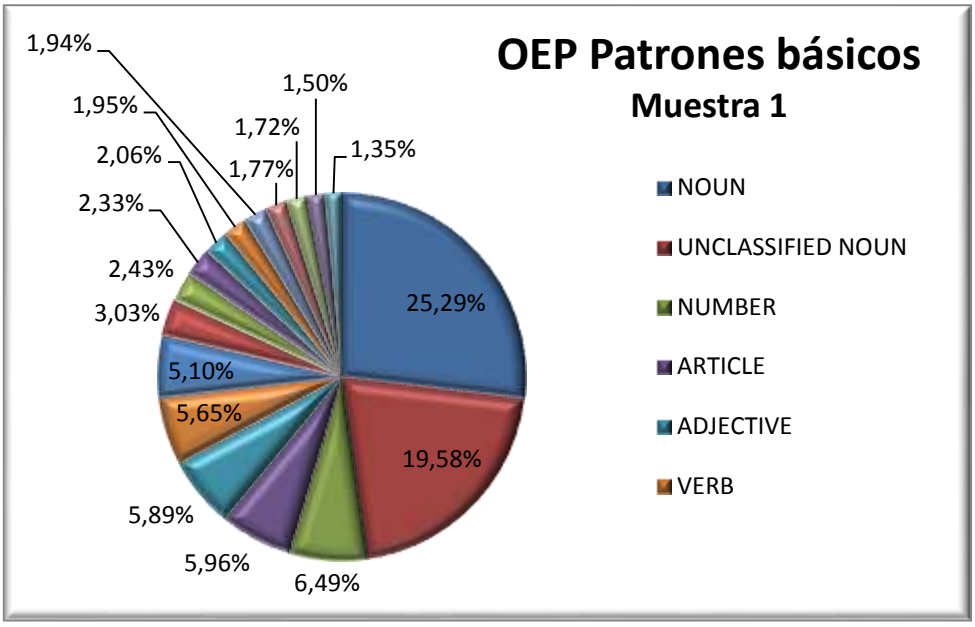
Para la primera muestra de OEP, de la que tenemos 379 documentos, la herramienta BP ha extraído 3.716.754 palabras de 186.711 frases.

En la segunda muestra de OEP, de la que tomamos 100 de los documentos, la herramienta BP ha extraído 2.050.851 palabras de 109.876 frases.

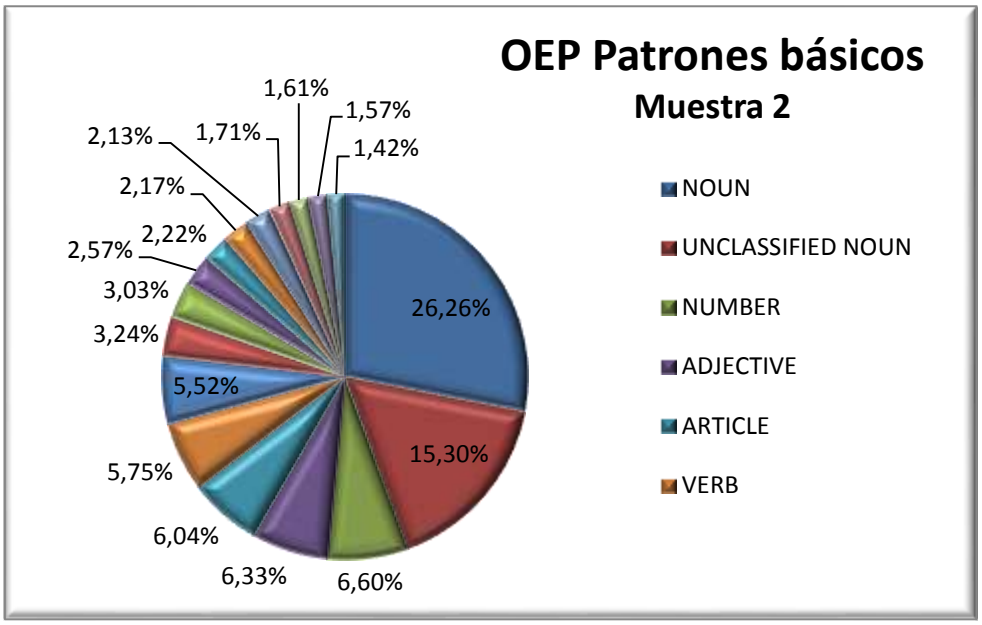
9.1.2.1 Categorías gramaticales

Partimos con 54 categorías gramaticales en la ontología (denominadas TermTag en la herramienta BP) en las siguientes gráficas se muestran las categorías gramaticales que más aparen en las muestras.

Para la muestra de documentos de patentes de OEP también son 53 categorías gramaticales las que están aquí presentes. Considerando las categorías gramaticales que superan el 1% de aparición, son 17 las que aquí más se repiten:

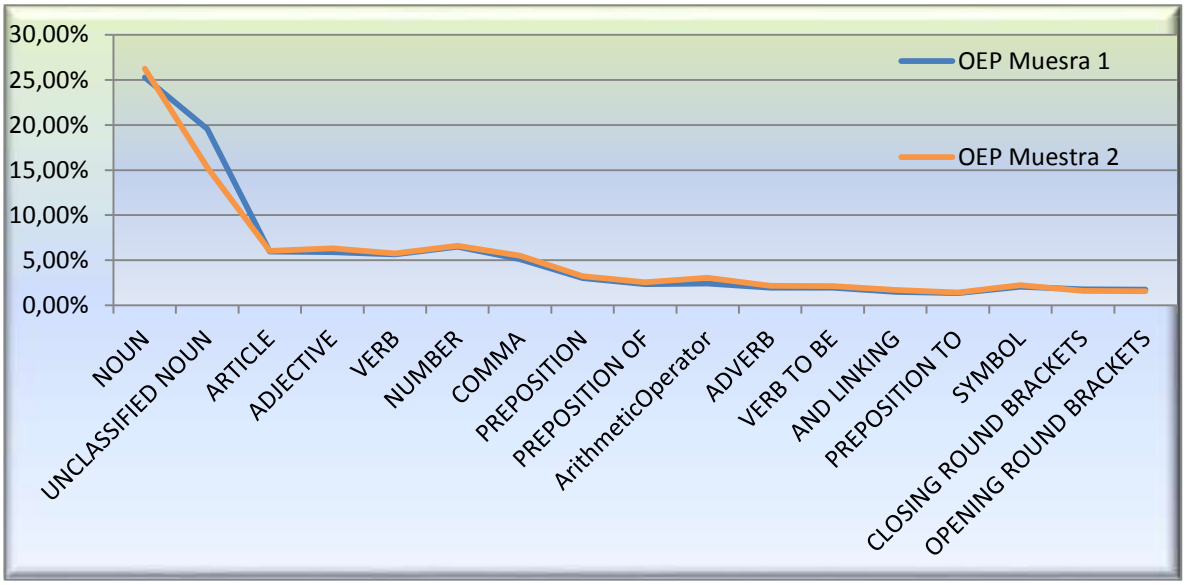


Gráfica 8. Patrones básicos. Categorías gramaticales OEP Muestra 1



Gráfica 9. Patrones básicos. Categorías gramaticales OEP Muestra 2

Se puede ver que para los dos casos hay muy poca diferencia entre las dos muestras, hay más número de palabras en la muestra 1 que en la muestra 2, pero no se aprecia en el porcentaje de aparición en cada una de las categorías gramaticales.



Gráfica 10. Patrones básicos. Categorías gramaticales OEP Muestra 1 vs. Muestra 2

En el apartado 8.1.3 puede verse en la tabla 12 todas las categorías gramaticales que aparecen en la muestra, con su número de frecuencia y el porcentaje que supone en toda ella.

8.1.2.2 Semántica

Para la muestra 1 tenemos un total 114.420 palabras con semántica reconocida por la ontología y para la muestra 2 tenemos 67.562 palabras. Para ambas muestras es un 3% del total de las palabras. En la siguiente tabla se muestran los tipos de semántica que más aparecen. Son 19 los mostrados porque son los que superan el 1%.

Casi la muestra 1 duplica el número de palabras a la muestra 2 pero en la siguiente tabla se puede ver que los porcentajes son muy similares para los tipos de semántica que aparece.

Semántica	Muestra 1	%	Muestra 2	%
RANGE <= (MAXIMUM)	29.982	26,20%	16.294	24,12%

Semántica	Muestra 1	%	Muestra 2	%
MODAL OPTIONAL	15.332	13,40%	9.461	14,00%
RANGE >= MINIMUM	12.774	11,16%	8.008	11,85%
Operation	4.735	4,14%	3.145	4,65%
RANGE ALL	5.053	4,42%	3.042	4,50%
Deny	4.939	4,32%	2.838	4,20%
RATE	3.870	3,38%	2.242	3,32%
WHEN ACTIVATION	3.430	3,00%	1.943	2,88%
MODAL FUTURE	2.893	2,53%	1.874	2,77%
RANGE > MINIMUM	2.652	2,32%	1.682	2,49%
RANGE ANY	2.390	2,09%	1.524	2,26%
RANGE LITTLE-FEW-SOME	2.279	1,99%	1.523	2,25%
IF ACTIVATION	1.973	1,72%	1.264	1,87%
RANGE BETWEEN	2.384	2,08%	1.157	1,71%
RANGE < MAXIMUM	1.998	1,75%	1.135	1,68%
UNIT	1.982	1,73%	966	1,43%
Provide	1.619	1,41%	919	1,36%
CONSTRAINT	1.701	1,49%	860	1,27%
RANGE MUCH-MANY	1.010	0,88%	679	1,01%

Tabla 12. Patrones básicos. Semántica OEP

Son tres tipos de semántica los que más destacan en las dos muestras en toda la muestra, que para entenderlo mejor listamos las palabras que lo forman:

RANGE <= (MAXIMUM) son las palabras del tipo: in, into, within, before, maximum, last, inside, down, beneath y inferior.

MODAL OPTIONAL son las palabras del tipo: Tokentext, may, can, would y should

RANGE >= MINIMUM son las palabras del tipo: tokentext, from, over, beyond, minimum, besides, outside y atop

9.1.3 USPTO vs. OEP

Los patrones básicos que aquí se comparan son los obtenidos de la muestra completa de USPTO y 100 documentos de la muestra OEP. De esta manera, al igualarse el número de tokens obtenidos, la comparación tiene mayor sentido al igualar las condiciones.

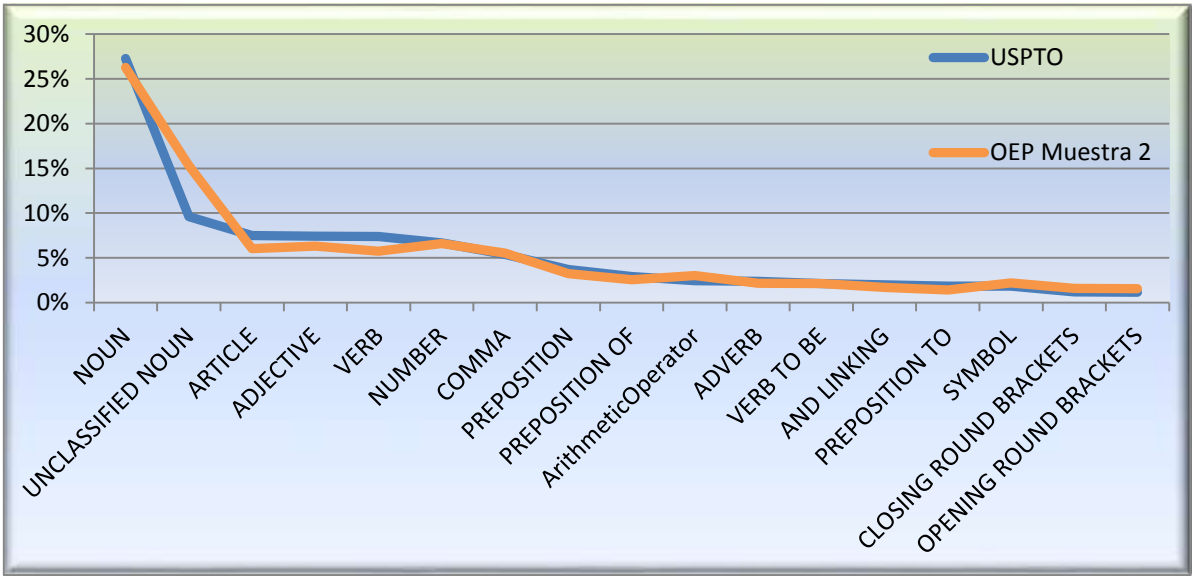
- ✓ UPTSO = 2.089.157 tokens

- ✓ OEP = 2.050.851 tokens

9.1.3.1 Categorías gramaticales

Comparando los dos resultados, vemos que el número de categorías gramaticales que superan el 1% es el mismo en ambas, pero con ligeras diferencias. En USPTO son los artículos la tercera categoría gramatical más repetida, mientras que en OEP son los números en esta posición. En ésta última, la repetición de patrones básicos sin clasificar es mucho mayor.

Vemos en la siguiente gráfica la representación comparativa de las 17 categorías gramaticales más repetidas:



Gráfica 11. Patrones básicos. Categorías gramaticales USPTO VS. OEP

Los resultados totales para todas las categorías gramaticales que hay dentro de los patrones básicos son los mostrados en la siguiente tabla.

CATEGORIA GRAMATICAL	USPTO		OEP	
NOUN	569.628	27,27%	939.989	26,26%
UNCLASSIFIED NOUN	200.745	9,61%	727.772	15,30%
ARTICLE	156.714	7,50%	221.464	6,04%
ADJECTIVE	154.839	7,41%	218.887	6,33%
VERB	154.225	7,38%	209.858	5,75%
NUMBER	139.800	6,69%	241.348	6,60%
COMMA	112.515	5,39%	189.688	5,52%
PREPOSITION	77.090	3,69%	112.606	3,24%
PREPOSITION OF	60.784	2,91%	86.491	2,57%

CATEGORIA GRAMATICAL	USPTO		OEP	
ArithmeticOperator	51.405	2,46%	90.168	3,03%
ADVERB	49.908	2,39%	72.351	2,17%
VERB TO BE	44.667	2,14%	71.975	2,13%
AND LINKING	41.644	1,99%	55.806	1,71%
PREPOSITION TO	38.449	1,84%	50.323	1,42%
SYMBOL	38.249	1,83%	76.430	2,22%
CLOSING ROUND BRACKETS	25.017	1,20%	65.666	1,61%
OPENING ROUND BRACKETS	24.537	1,17%	64.086	1,57%
MODAL VERB	17.536	0,84%	19.095	0,58%
RELATIVE PRONOUN	16.144	0,77%	23.958	0,66%
QUANTIFIER DETERMINER	15.094	0,72%	18.367	0,58%
ABSOLUTE VERB	13.843	0,66%	18.003	0,56%
PHRASAL VERB PARTICLE	13.655	0,65%	24.938	0,71%
OR LINKING	10.422	0,50%	14.169	0,48%
PREPOSITION BY	9.841	0,47%	15.326	0,46%
PARTICLE	7.611	0,36%	11.617	0,36%
VERB TO HAVE	6.514	0,31%	10.923	0,32%
DEMONSTRATIVE DETERMINER	6.076	0,29%	7.223	0,23%
REQUIREMENTS Domain	5.241	0,25%	9.762	0,28%
TIME ADVERB	4.280	0,20%	6.440	0,20%
PERSONAL PRONOUN	4.106	0,20%	8.068	0,25%
NEGATION	2.383	0,11%	2.998	0,08%
PLACE ADVERB	2.111	0,10%	1.672	0,05%
PHRASAL VERB BASE	2.004	0,10%	11.971	0,23%
RECOVERABLE PRONOUN	1.600	0,08%	3.029	0,08%
MEASUREMENT UNIT	1.391	0,07%	2.498	0,08%
CONECTOR REQUIREMENT/CONDITION	1.365	0,07%	1.973	0,06%
POSSESSIVE DETERMINER	1.111	0,05%	1.191	0,04%
PRONOUN	1.101	0,05%	1.536	0,04%
VERB TO DO	805	0,04%	1.242	0,03%
PREPOSITIONAL LOCATION	730	0,03%	675	0,01%
PREPOSITIONAL LINKING PHRASE	622	0,03%	824	0,02%
CAUSAL CONECTOR	588	0,03%	893	0,03%
PURPOSE/GOAL	533	0,03%	808	0,02%
PARTITIVE DETERMINER	427	0,02%	737	0,02%
UNCLASSIFIED ADJECTIVE	421	0,02%	356	0,01%
POSSESIVE PRONOUN	405	0,02%	564	0,02%
CAUSE	385	0,02%	438	0,01%
UTTERANCE DETERMINER	215	0,01%	81	0,00%
PHRASAL VERB ABSOLUTE BASE	180	0,01%	175	0,00%
TIME ADVERBIAL PHRASE	97	0,00%	184	0,00%

CATEGORIA GRAMATICAL	USPTO		OEP	
DETERMINER	41	0,00%	43	0,00%
PLACE ADVERBIAL PHRASE	41	0,00%	23	0,00%
Formula	22	0,00%	46	0,00%

Tabla 13. Categorías gramaticales. USPTO VS. OEP

Todas las categorías gramaticales que forman la ontología aparecen en ambas muestras, salvo una, la que ha sido etiquetada como "ABBREVIATION".

Ahora vamos a mostrar más detalle de algunas de las categorías gramaticales de ambas muestras.

Nombres

Los nombres son los TokenText que más se repiten en ambas muestras. Como se puede ver en gráficas anteriores, representan un 27% para las patentes estadounidenses y un 26% para las patentes europeas.

En ambas muestras el más repetido es "a", pero como carece de gran significado para el análisis, este nombre es obviado. Las que aparecen con mayor frecuencia en cada muestra son los siguientes, tienen nombres comunes en ambas muestras, se subrayan las palabras que son diferentes entre las más repetidas:

USPTO

Invention, example, acid, system, method, material, composition, figure, embodiment, agent, solution, temperature, water, protein, product, sequence, device, portion, weight, surface.

OEP

Invention, system, embodiment, device, method, example, data, material, user, application, claim, acid, member, portion, surface, information, module, component, power, led.

No clasificados

Los nombres no clasificados son los segundos termtag que más se repiten. Se tratan de símbolos, palabras incompletas, siglas…

En la muestra USPTO tenemos casi un 10% sin clasificar y en OEP tenemos un 15% sin clasificar. Para ambas muestras es un porcentaje muy elevado de aparición. A pesar de ello, no debe preocuparnos porque se tratan de palabras no recogidas en el diccionario de la lengua

inglesa y seguramente se deba a que los documentos de la muestra no son de alta calidad.

Adjetivos

Hay un 7% de aparición en los documentos USPTO y un 6% en los documentos de OEP, ambas muestras están bastante equilibradas.

Los adjetivos que están más presentes en los documentos son los siguientes: i, present, said, second, used, lower, such, described, based, preferred, solar, formed, about, first, suitable.

Verbos

El porcentaje de aparición en ambas muestras también es muy similar, 7% y 6%.

Los verbos que mas se repiten son: provide, comprise, show, use, claim, accord, form, group, illustrate, further, process, allow, substitute, pour, sequence, add, position, control, process, part y determine.

Verbos con categoría propia

- ✓ To be. Su presencia ronda el 2% en ambas muestras.
- ✓ Verbos modales (may, can, will, should, would, might, shall, must y ought)

Éstos son un 0,66% de la muestra estadounidense y un 0,48% de la muestra europea.

El verbo modal *ought* sólo aparece en las patentes europeas. El resto de estos verbos aparecen en ambos.

- ✓ Verbos absolutos. Un ejemplo de verbos modales que aparecen son: include, contain, receive, generate, characterize, produce, require, create, occur, correspond,a,hieve, obtain, utilize, communicate, appear, etc.

Éstos son un 0,10% de la muestra estadounidense y un 0,4% de la muestra europea.

- ✓ To have

El verbo tu have es representado con un 0,31% de la muestra estadounidense y un 0,29% de la muestra europea.

- ✓ To do

El verbo to do tan solo es un 0,04% en la muestra estadounidense y un 0,03% de la muestra europea.

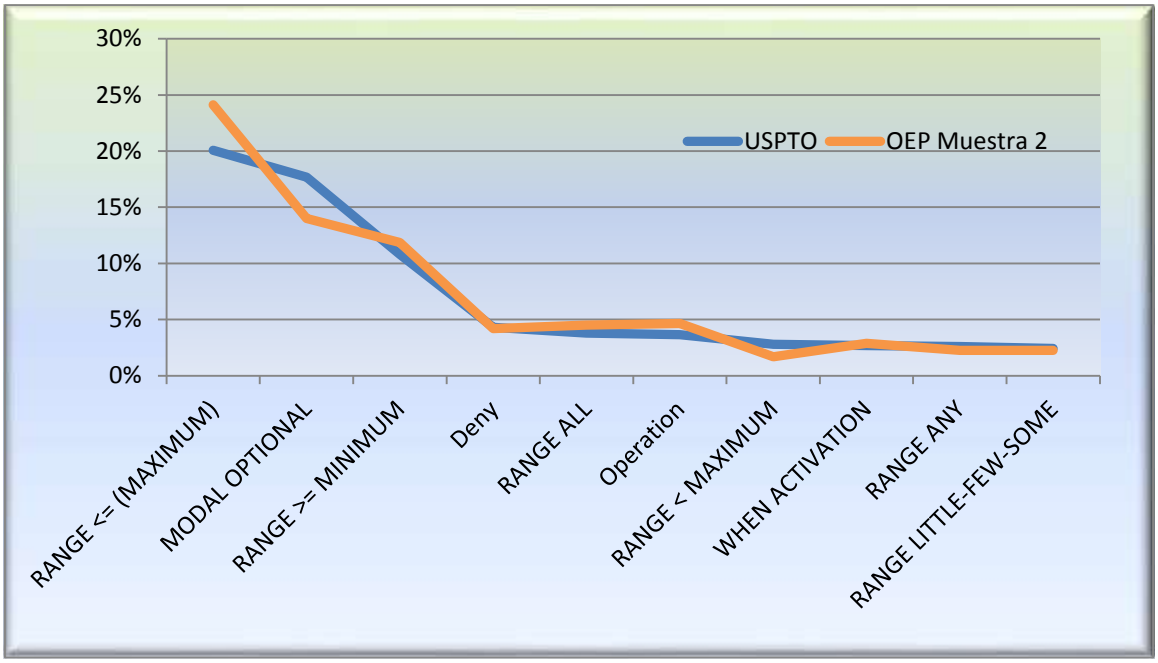
9.1.3.2 Semántica

La semántica está presente dentro de los patrones básicos pero de una manera muy reducida. Nos encontramos con un poco más de semántica dentro de las muestras estadounidenses.

	USPTO		OEP	
CON SEMÁNTICA	83.966	4%	67.562	3%
SIN SEMÁNTICA	2.005.191	96%	1.983.289	97%

Tabla 14. Patrones básicos. Semántica USPTO vs. OEP

En la siguiente gráfica se puede ver la semántica que más aparece en ambas muestras:



Gráfica 12. Patrones básicos. Semántica

Los 10 TokenText con gramática que más aparecen de la muestra estadounidense se encuentran los siguientes:

TokenText	Gramática	Semántica	USPTO
in	PREPOSITION	RANGE <= (MAXIMUM)	14%

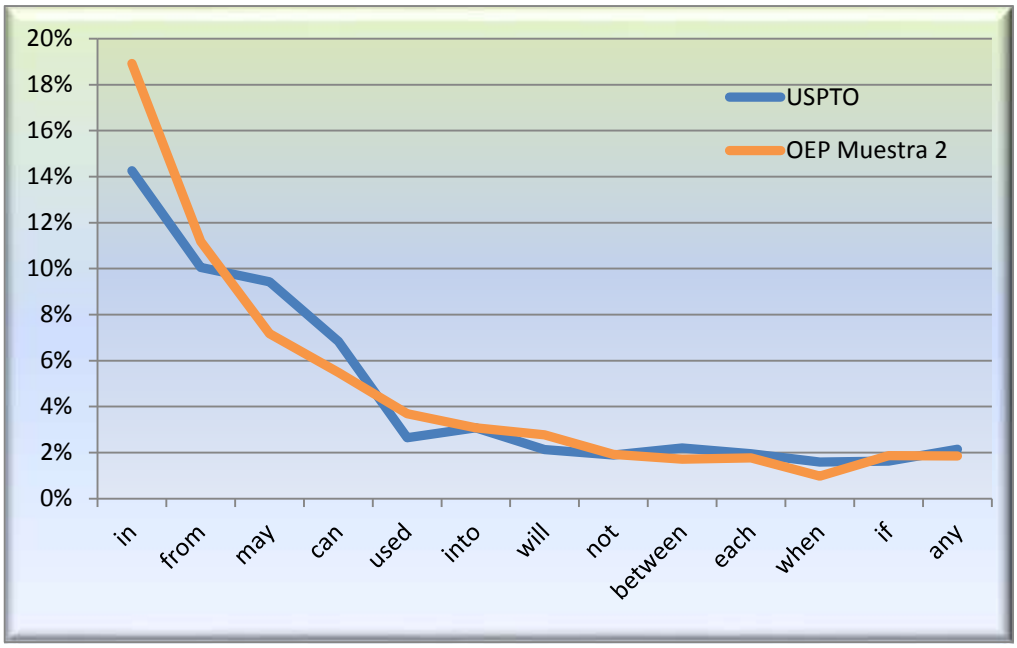
TokenText	Gramática	Semántica	USPTO
from	PREPOSITION	RANGE >= MINIMUM	10%
may	MODAL VERB	MODAL OPTIONAL	9%
can	MODAL VERB	MODAL OPTIONAL	7%
into	PREPOSITION	RANGE <= (MAXIMUM)	3%
used	ADJECTIVE	Operation	3%
between	PREPOSITION	RANGE BETWEEN	2%
any	QUANTIFIER DETERMINER	RANGE ANY	2%
will	MODAL VERB	MODAL FUTURE	2%
each	QUANTIFIER DETERMINER	RANGE ALL	2%

Tabla 15. Patrones básicos con semántica Top 10 USPTO

En la muestra europea, los TokenText *any, each y between* no están dentro de los 10 primeros y en su lugar tenemos la negación *not, when e if*.

TokenText	Gramática	Semántica	OEP
in	PREPOSITION	RANGE <= (MAXIMUM)	19%
from	PREPOSITION	RANGE >= MINIMUM	11%
may	MODAL VERB	MODAL OPTIONAL	7%
can	MODAL VERB	MODAL OPTIONAL	5%
used	ADJECTIVE	Operation	4%
into	PREPOSITION	RANGE <= (MAXIMUM)	3%
will	MODAL VERB	MODAL FUTURE	3%
not	NEGATION	Deny	2%
when	PARTICLE	WHEN ACTIVATION	2%
if	CONECTOR REQUIREMENT/CONDITION	IF ACTIVATION	2%

Tabla 16. Patrones básicos con semántica Top 10 OEP



Gráfica 13. Patrones básicos con semántica USPTO vs. OEP

9.2 Escenario 1

Características del escenario:

- Muestra de 359 patentes procedentes de USPTO.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 1 para crear patrones.
- No activamos el checkbox para no diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.089.157 (tabla BasicPattern)
- ✓ Patrones: 57.582 (tabla Pattern)
  - Formados por dos TermTag: 657
  - Formados por patrones: 45.740
  - Mixtos: 11.185
- ✓ Patrones con semántica: 4.632 (tabla SemanticsBelongPatterns)

Este es el primer escenario creado con las muestras de patentes estadounidenses, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 1 y no vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos serán iguales aunque pertenezcan a semánticas diferentes.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

Al ser una ejecución tan poco restrictiva, la ejecución de este escenario ha necesitado de 22,5 días para finalizar.

9.2.1 Patrones

Los patrones están compuestos por dos TermTag, por dos subpatrones o por un TermTag y un subpatrón.

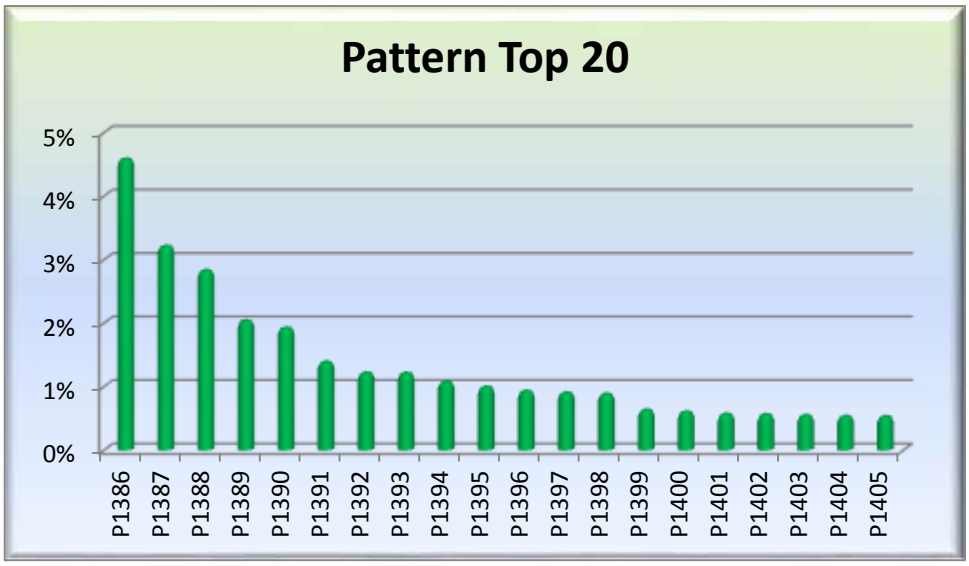
Para este escenario se han distinguido 57.582 patrones diferentes y un total de 1.007.385 repeticiones.

El patrón que más repeticiones tiene es P1386, su nivel de profundidad es infinito, esto es porque está compuesto por un artículo y por sí mismo.

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight	Repeticiones
P1386	ARTICLE	P1386	46.294
P1389	P1389	UNCLASSIFIED NOUN	20.629

Tabla 17. Escenario 1. Patrones de descomposición infinita

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 27% del total.



Gráfica 14. Escenario 1. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 657 que hay de este tipo. El total de este tipo representa el 6,77% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1401	NUMBER	NUMBER	5.829	0,58%
P1403	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	5.606	0,56%
P1416	NOUN	NUMBER	3.765	0,37%
P1419	OPENING ROUND BRACKETS	ADJECTIVE	3.005	0,30%
P1428	AND LINKING	VERB	2.734	0,27%
P1461	NOUN	ArithmeticOperator	1.588	0,16%
P1463	UNCLASSIFIED NOUN	NUMBER	1.504	0,15%
P1467	PREPOSITION	NUMBER	1.410	0,14%
P1474	ADVERB	PARTICLE	1.287	0,13%
P1487	PREPOSITION	ARTICLE	1.152	0,11%
P1498	AND LINKING	ADJECTIVE	1.058	0,11%
P1507	NUMBER	ARTICLE	907	0,09%

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1512	CLOSING ROUND BRACKETS	COMMA	878	0,09%
P1530	PREPOSITION OF	ARTICLE	786	0,08%
P1533	MODAL VERB	ABSOLUTE VERB	762	0,08%
P1538	AND LINKING	ADVERB	750	0,07%
P1549	ADVERB	QUANTIFIER DETERMINER	689	0,07%
P1554	OPENING ROUND BRACKETS	RECOVERABLE PRONOUN	670	0,07%
P1555	CLOSING ROUND BRACKETS	NOUN	664	0,07%
P1559	ARTICLE	SYMBOL	655	0,07%

Tabla 18. Escenario 1. Pattern - Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 45.740 combinaciones obtenidas, el total representa el 34,17% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P1392	P1389	P1389	12.377	1,23%
P1394	P1389	P1387	10.935	1,09%
P1396	P1387	P1386	9.446	0,94%
P1408	P1389	P1386	4.807	0,48%
P1409	P1392	P1386	4.479	0,44%
P1417	P1386	P1386	3.313	0,33%
P1427	P1394	P1394	2.819	0,28%
P1434	P1386	P1391	2.328	0,23%
P1447	P1392	P1392	1.816	0,18%
P1452	P1390	P1386	1.741	0,17%
P1458	P1396	P1386	1.629	0,16%
P1465	P1387	P1387	1.447	0,14%
P1469	P1392	P1387	1.358	0,13%
P1471	P1392	P1389	1.324	0,13%
P1477	P1388	P1386	1.255	0,12%
P1479	P1399	P1391	1.199	0,12%
P1480	P1392	P1398	1.199	0,12%
P1481	P1387	P1391	1.188	0,12%
P1482	P1392	P1396	1.184	0,12%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P1486	P1389	P1391	1.152	0,11%

Tabla 19. Escenario 1. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 5.411 patrones diferentes, suponen el 28,39% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1387	P1386	PREPOSITION OF	32.473	3,22%
P1388	P1386	VERB	28.601	2,84%
P1389	P1389	UNCLASSIFIED NOUN	20.629	2,05%
P1390	P1386	COMMA	19.502	1,94%
P1391	P1386	NUMBER	14.089	1,40%
P1395	P1387	NOUN	10.115	1,00%
P1399	P1391	PREPOSITION TO	6.462	0,64%
P1402	P1399	CLOSING ROUND BRACKETS	5.763	0,57%
P1404	P1389	NUMBER	5.483	0,54%
P1405	P1389	NOUN	5.451	0,54%
P1406	P1386	NOUN	5.313	0,53%
P1407	P1391	VERB	4.807	0,48%
P1411	P1386	ADJECTIVE	4.252	0,42%
P1414	P1389	ARTICLE	3.853	0,38%
P1426	P1387	UNCLASSIFIED NOUN	2.872	0,29%
P1430	P1419	CLOSING ROUND BRACKETS	2.580	0,26%
P1433	P1392	NOUN	2.332	0,23%
P1435	P1388	NOUN	2.321	0,23%
P1436	P1387	VERB	2.283	0,23%
P1437	P1398	NOUN	2.234	0,22%

Tabla 20. Escenario 1. Pattern – Top 20. Patrón + TermTag

Se encuentran 5.754 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un

30,67% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P1386	ARTICLE	P1386	46.294	4,60%
P1393	ADJECTIVE	P1386	12.319	1,22%
P1397	VERB	P1386	9.205	0,91%
P1398	PREPOSITION	P1386	8.991	0,89%
P1400	AND LINKING	P1386	6.186	0,61%
P1410	OPENING ROUND BRACKETS	P1403	4.330	0,43%
P1412	PREPOSITION OF	P1386	3.942	0,39%
P1413	NOUN	P1386	3.914	0,39%
P1415	PREPOSITION TO	P1386	3.776	0,37%
P1418	NOUN	P1391	3.283	0,33%
P1420	MODAL VERB	P1391	2.947	0,29%
P1421	SYMBOL	P1389	2.944	0,29%
P1422	ARTICLE	P1393	2.940	0,29%
P1423	ArithmeticOperator	P1386	2.928	0,29%
P1424	ADVERB	P1386	2.908	0,29%
P1425	PHRASAL VERB PARTICLE	P1386	2.874	0,29%
P1429	PREPOSITION TO	P1391	2.695	0,27%
P1431	ARTICLE	P1391	2.534	0,25%
P1432	UNCLASSIFIED NOUN	P1399	2.494	0,25%
P1439	UNCLASSIFIED NOUN	P1386	2.176	0,22%

Tabla 21. Escenario 1. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por otros dos patrones tienen mayor presencia en los documentos cuando la frecuencia de repetición es 1. La mayor frecuencia coincide con el mayor número de repeticiones. En la siguiente tabla mostramos el número de repeticiones que tiene cada tipo de patrón y el porcentaje que supone entre las repeticiones:

	Nº Patrones	Repeticiones
TermTag + TermTag	657	6,77%
Patrón + Patrón	45.740	34,17%
Patrón + TermTag	5.411	28,39%
TermTag + Patrón	5.774	30,67%
TOTAL	**57.582**	**100%**

Tabla 22. Escenario 1. Pattern. Repeticiones de los diferentes tipos

Para este escenario, tenemos que el patrón que menos repeticiones tiene es el simple, el formado por dos categorías gramaticales.

El número de patrones que está formado por un patrón a la izquierda y por una categoría gramatical a la derecha, es muy similar en número de patrones y en repeticiones al que está formado a la inversa, con gramática a la izquierda y patrón a la derecha.

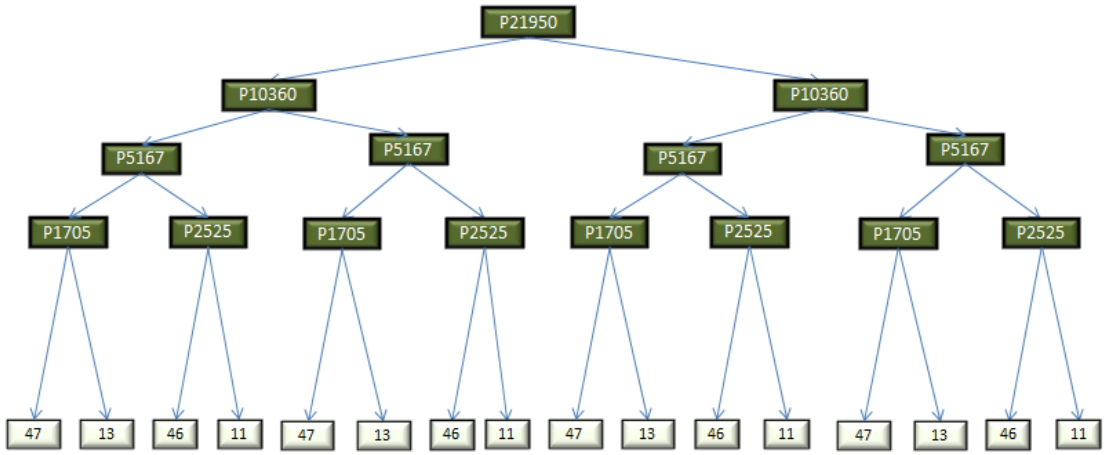
Longitud de patrones

Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

Los patrones más largos que se han generado con este escenario son P30776 y P36495, pero ambos están compuestos por gramática de nombres no clasificados y artículos de manera infinita. Esto se debe a que se han creado dos patrones que por su parte izquierda o por su parte derecha se sustituye por sí mismo de manera infinita, estos patrones son los mostrados en la primera tabla de este apartado.

Analizando los siguientes patrones con mayor longitud, descartando los que estén compuestos por los patrones de sustitución infinita, nos quedamos con 2.635 patrones, de los cuáles muchos son una sucesión de números, de símbolos, etc. Ninguno de ello parece mostrarnos un gran sentido gramatical.

En la siguiente gráfica mostramos un ejemplo con el patrón P21950, formado por 16 TermTag gramaticales:



Gráfica 15. Escenario 1. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P21950= OPENING ROUND BRACKETS + SYMBOL + CLOSING ROUND BRACKETS + VERB + OPENING ROUND BRACKETS + SYMBOL + CLOSING ROUND BRACKETS + VERB + OPENING ROUND BRACKETS + SYMBOL + CLOSING ROUND BRACKETS + VERB + OPENING ROUND BRACKETS + SYMBOL + CLOSING ROUND BRACKETS + VERB

Si simplificamos, lo que está representado es:

(SYMBOL) VERB (SYMBOL) VERB (SYMBOL) VERB (SYMBOL) VERB

Por regla general, las oraciones en inglés tienen que estar formadas al menos por un nombre y un verbo, en este escenario no encontramos un patrón finito que lo cumpla. En la frase anterior mostrada, tenemos un verbo pero no tenemos un nombre.

9.2.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 4.632 patrones los que tienen semántica, entre ellos hay muchos patrones repetidos pero con semántica diferente, nombres de patrones únicos hay 1.510, un ejemplo de patrón que puede tener semántica diferente, es el P1386 que indicábamos como más repetido en el apartado anterior. Y las diferentes semánticas que se encuentran entre los documentos para éste patrón son 10 tipos:

Pattern	SemanticLeft	SemanticRight
P1386	RANGE < MAXIMUM	NO SEMANTICA
P1386	AVAILABILITY	NO SEMANTICA
P1386	RANGE = EQUAL	NO SEMANTICA
P1386	Operation	NO SEMANTICA
P1386	RANGE > MINIMUM	NO SEMANTICA
P1386	Sustainability	NO SEMANTICA
P1386	RANGE <= (MAXIMUM)	NO SEMANTICA
P1386	SYSTEM FUNCTION	NO SEMANTICA
P1386	RANGE >= MINIMUM	NO SEMANTICA
P1386	DUTY ACTION	NO SEMANTICA

Tabla 23. Escenario 1. Semántica. Patrón más repetido con semántica

El artículo que forma el termtag de la izquierda puede adquirir la diferente semántica mostrada.

Como éste hay muchos patrones repetidos, al tener el mismo patrón tantas repeticiones no conocemos la frecuencia de aparición de cada uno de ellos entre los documentos.

El número de patrones con semántica a la izquierda y a la derecha es el siguiente:

SemanticLeft	SemanticRight	Nº Patrones
SI	SI	170
NO	SI	2.190
SI	NO	2.272
TOTAL		**4.632**

Tabla 24. Escenario 1. Totales patrones con semántica

Se puede ver que sólo semántica a la derecha o sólo semántica a la izquierda son los que más se dan. Con semántica a ambos lados del patrón son muchos menos.

Son casi 60mil patrones los que se han generado con este escenario y de ellos tan sólo 1500 llevan semántica. Hay muy poca presencia de patrones en la muestra completa.

9.3 Escenario 2

Características del escenario:

- Muestra de 359 patentes procedentes de USPTO.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 1 para crear patrones.
- Activamos el checkbox para diferenciar patrones por su semántica.

Este es el segundo escenario creado con las muestras de patentes estadounidenses, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 1 y vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos serán iguales aunque pertenezcan a semánticas diferentes. Esto es lo que le diferencia del escenario anterior.

Tampoco se elige la opción de guardar los boiler plates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

No ha sido posible contar con los resultados de este escenario porque el cálculo ha sido demasiado pesado para la herramienta. Lleva en ejecución más de 25 días. No es posible realizar un análisis de los patrones parciales porque la herramienta no guarda los cálculos realizados hasta que finaliza todo el proceso.

9.4 Escenario 3

Características del escenario:

- Muestra de 359 patentes procedentes de USPTO.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 20 para crear patrones.
- Activamos el checkbox para diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.089.157 (tabla BasicPattern)
- ✓ Patrones: 4.053 (tabla Pattern)
  - Formados por dos TermTag: 408
  - Formados por patrones: 1.753
  - Mixtos: 1.892
- ✓ Patrones con semántica: 678 (tabla SemanticsBelongPatterns)

Este es el tercer escenario creado con las muestras de patentes estadounidenses, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 20 y vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 20 veces entre los documentos. Además deberán tener la misma semántica para que puedan ser del mismo patrón.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

Es una ejecución más restrictiva que en los anteriores escenarios, la ejecución de este escenario ha necesitado 86,5 horas, casi 4 días, para finalizar.

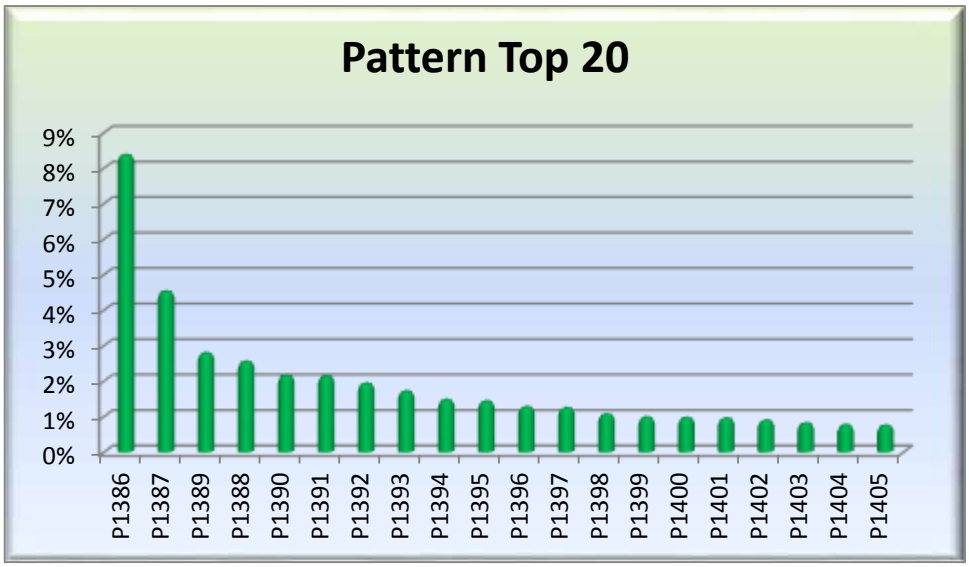
9.4.1 Patrones

Los patrones están compuestos por dos TermTag, por dos subpatrones o por un TermTag y un subpatrón.

Para este escenario se han distinguido 4.053 patrones diferentes y un total de 1.011.917 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, tiene el nombre P1386. Está formado por la categoría gramatical adjetivo por la izquierda y nombre por la derecha.

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 39% del total.



Gráfica 16. Escenario 3. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 408 que hay de este tipo. El total de este tipo representa el 30,60% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1386	ADJECTIVE	NOUN	84.797	8,38%
P1390	VERB	NOUN	21.869	2,16%
P1392	SYMBOL	UNCLASSIFIED NOUN	19.644	1,94%
P1394	VERB TO BE	ADJECTIVE	14.988	1,48%
P1395	NOUN	VERB	14.534	1,44%
P1396	VERB TO BE	VERB	12.937	1,28%
P1401	PREPOSITION	NOUN	9.700	0,96%
P1405	ADVERB	VERB	7.717	0,76%
P1406	PREPOSITION TO	VERB	7.500	0,74%
P1407	UNCLASSIFIED NOUN	VERB	7.471	0,74%
P1408	NOUN	NUMBER	6.615	0,65%
P1410	NUMBER	NUMBER	5.676	0,56%

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1411	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	5.584	0,55%
P1415	PREPOSITION	VERB	4.708	0,47%
P1421	VERB	VERB	4.160	0,41%
P1433	AND LINKING	VERB	3.207	0,32%
P1434	ADJECTIVE	PHRASAL VERB PARTICLE	2.993	0,30%
P1435	OPENING ROUND BRACKETS	ADJECTIVE	2.972	0,29%
P1436	COMMA	UNCLASSIFIED NOUN	2.948	0,29%
P1444	NOUN	ArithmeticOperator	2.208	0,22%

Tabla 25. Escenario 3. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 1.753 combinaciones obtenidas, el total representa el 17,73% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P1389	P1388	P1388	28.287	2,80%
P1409	P1389	P1389	5.979	0,59%
P1417	P1391	P1387	4.589	0,45%
P1429	P1388	P1387	3.272	0,32%
P1439	P1396	P1387	2.352	0,23%
P1440	P1388	P1386	2.265	0,22%
P1455	P1402	P1387	1.867	0,18%
P1463	P1409	P1409	1.692	0,17%
P1468	P1386	P1386	1.561	0,15%
P1482	P1391	P1386	1.277	0,13%
P1489	P1402	P1386	1.199	0,12%
P1494	P1389	P1388	1.137	0,11%
P1502	P1406	P1387	1.053	0,10%
P1507	P1396	P1386	988	0,10%
P1518	P1387	P1394	867	0,09%
P1519	P1387	P1386	853	0,08%
P1521	P1403	P1403	830	0,08%
P1528	P1404	P1387	790	0,08%
P1530	P1434	P1387	781	0,08%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P1532	P1426	P1388	769	0,08%

Tabla 26. Escenario 3. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 837 patrones diferentes, suponen el 26,62% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1388	P1388	UNCLASSIFIED NOUN	25.871	2,56%
P1391	P1387	PREPOSITION OF	21.678	2,14%
P1397	P1387	VERB	12.642	1,25%
P1398	P1388	NOUN	10.829	1,07%
P1399	P1386	COMMA	9.999	0,99%
P1400	P1386	VERB	9.902	0,98%
P1402	P1386	PREPOSITION OF	9.096	0,90%
P1403	P1387	NUMBER	8.329	0,82%
P1404	P1387	COMMA	7.812	0,77%
P1412	P1386	NOUN	5.266	0,52%
P1413	P1388	NUMBER	5.144	0,51%
P1416	P1405	CLOSING ROUND BRACKETS	4.645	0,46%
P1418	P1386	NUMBER	4.444	0,44%
P1422	P1388	COMMA	4.158	0,41%
P1424	P1388	ARTICLE	3.828	0,38%
P1426	P1391	NOUN	3.507	0,35%
P1430	P1388	VERB	3.266	0,32%
P1438	P1435	CLOSING ROUND BRACKETS	2.564	0,25%
P1442	P1396	PREPOSITION	2.230	0,22%
P1448	P1394	PREPOSITION TO	1.990	0,20%

Tabla 27. Escenario 3. Pattern – Top 20. Patrón + TermTag

Se encuentran 1.055 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un

25,06% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P1387	ARTICLE	P1386	45.896	4,54%
P1393	ADJECTIVE	P1386	17.394	1,72%
P1414	PREPOSITION	P1387	4.722	0,47%
P1419	ARTICLE	P1393	4.442	0,44%
P1420	OPENING ROUND BRACKETS	P1411	4.312	0,43%
P1423	PREPOSITION	P1386	4.118	0,41%
P1425	AND LINKING	P1386	3.780	0,37%
P1427	PREPOSITION TO	P1387	3.489	0,34%
P1428	ArithmeticOperator	P1386	3.275	0,32%
P1431	ARTICLE	P1392	3.265	0,32%
P1432	PREPOSITION OF	P1387	3.250	0,32%
P1437	MODAL VERB	P1394	2.794	0,28%
P1441	VERB	P1386	2.231	0,22%
P1443	AND LINKING	P1387	2.225	0,22%
P1445	UNCLASSIFIED NOUN	P1406	2.188	0,22%
P1446	UNCLASSIFIED NOUN	P1386	2.094	0,21%
P1449	ADVERB	P1387	1.990	0,20%
P1450	NUMBER	P1386	1.985	0,20%
P1451	VERB	P1387	1.979	0,20%
P1452	OPENING ROUND BRACKETS	P1426	1.935	0,19%

Tabla 28. Escenario 3. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es mucho menor al resto.

	Nº Patrones	Repeticiones
TermTag + TermTag	408	30,60%
Patrón + Patrón	1.753	17,73%
Patrón + TermTag	837	26,62%
TermTag + Patrón	1.055	25,06%
TOTAL	**4.053**	**100%**

Tabla 29. Escenario 3. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto por patrones a ambos lados, supone el mayor número de patrones diferentes, pero en cambio tiene el menor peso en cuestión de número de repeticiones totales.

Longitud de patrones

Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón con mayor longitud de descomposición dentro de este escenario es el patrón P2590. La equivalencia de los patrones sería la que se muestra en la siguiente tabla:

Nombre Patrón	TermTag Izquierda	TermTag Derecha
P2590	P1726	P1726
P1726	P1463	P1463
P1463	P1409	P1409
P1409	P1389	P1389
P1389	P1388	P1388
P1388	P1388	UNCLASSIFIED NOUN

Tabla 30. Escenario 3. Equivalencia patrón P2590

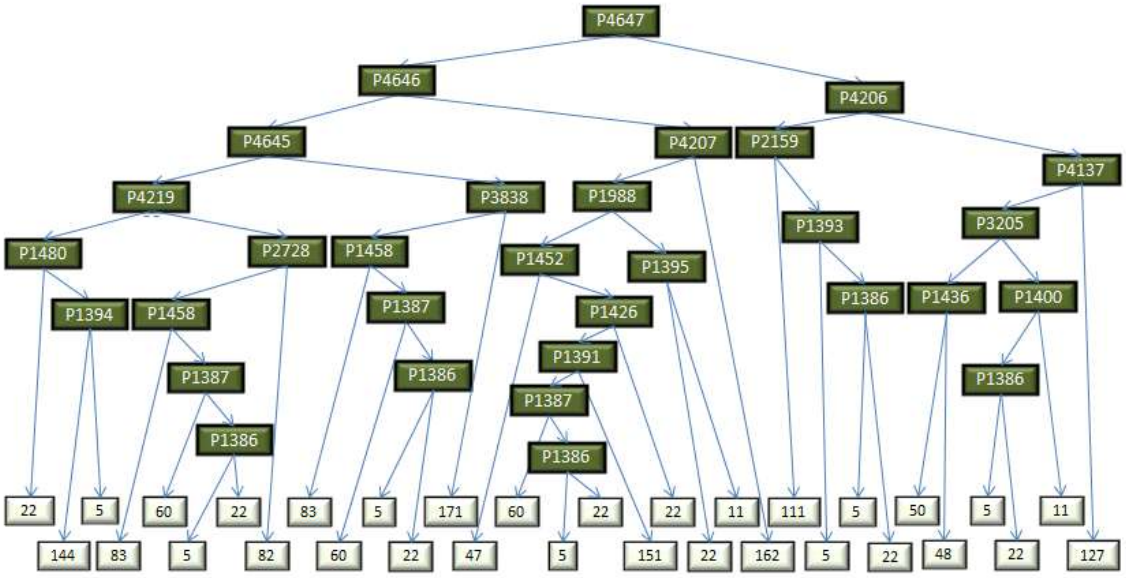
Descomponiendo todos los niveles llegamos a un punto, donde lo que tendría el patrón serían la concatenación de nombres no clasificados de manera infinita.

P2590=P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED

NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN + P1388 + UNCLASSIFIED NOUN

Este patrón es el más largo, pero no nos está aportando ningún valor, no nos da ninguna información útil que se pueda llevar a la práctica.

Un oración escrita en inglés, al menos debería tener un nombre y un verbo, con estas características, buscamos el siguiente patrón más largo que aporte información, nos encontramos con el patrón P4647 que forma una oración de 32 termtag gramaticales:



Gráfica 17. Escenario 3. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P4647= NOUN + VERB TO BE + ADJECTIVE + PHRASAL VERB PARTICLE + ARTICLE + ADJECTIVE + NOUN + ABSOLUTE VERB + PHRASAL VERB PARTICLE + ARTICLE + ADJECTIVE + NOUN + ArithmeticOperator + OPENING ROUND BRACKETS + ARTICLE + ADJECTIVE + NOUN + PREPOSITION OF + NOUN + NOUN + VERB + AND LINKING + RELATIVE PRONOUN + ADJECTIVE + ADJECTIVE + NOUN + COMMA + UNCLASSIFIED NOUN + ADJECTIVE + NOUN + VERB + PREPOSITION

9.4.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 677 patrones los que tienen semántica, suponen un 13,15% dentro de todos los patrones. Las siguientes tablas muestran diferentes combinaciones de los patrones con esta característica.

Los patrones que tienen mayor semántica se muestran en la siguiente tabla, donde se muestra por patrón el peso de las repeticiones dentro de los patrones con semántica (campo SEMANTICA) y el peso de las repeticiones teniendo en cuenta todos los patrones, con semántica y sin semántica (campo PATRONES):

Patrón	SemanticLeft	SemanticRight	FRECUENCIA	SEMANTICA	PATRONES
P1403	RANGE >= MINIMUM	NO SEMANTICA	8.329	6,26%	0,82%
P1405	RANGE > MINIMUM	NO SEMANTICA	7.717	5,80%	0,76%
P1409	RANGE <= (MAXIMUM)	NO SEMANTICA	5.979	4,49%	0,59%
P1425	NO SEMANTICA	RATE	3.780	2,84%	0,37%
P1426	MODAL OPTIONAL	NO SEMANTICA	3.507	2,64%	0,35%
P1433	RANGE LITTLE-FEW-SOME	NO SEMANTICA	3.207	2,41%	0,32%
P1435	NO SEMANTICA	Support	2.972	2,23%	0,29%
P1437	IF ACTIVATION	NO SEMANTICA	2.794	2,10%	0,28%
P1437	MODAL OPTIONAL	NO SEMANTICA	2.794	2,10%	0,28%
P1442	NO SEMANTICA	RANGE <= (MAXIMUM)	2.230	1,68%	0,22%
P1457	RANGE >= MINIMUM	NO SEMANTICA	1.812	1,36%	0,18%
P1460	Support	NO SEMANTICA	1.735	1,30%	0,17%
P1463	RANGE >= MINIMUM	NO SEMANTICA	1.692	1,27%	0,17%
P1465	RANGE <= (MAXIMUM)	NO SEMANTICA	1.675	1,26%	0,17%
P1465	MODAL OPTIONAL	NO SEMANTICA	1.675	1,26%	0,17%
P1466	MODAL OPTIONAL	NO SEMANTICA	1.670	1,25%	0,17%
P1474	NO SEMANTICA	RANGE >= MINIMUM	1.484	1,12%	0,15%
P1490	NO SEMANTICA	UNIT	1.183	0,89%	0,12%
P1491	RANGE <= (MAXIMUM)	NO SEMANTICA	1.169	0,88%	0,12%
P1498	RANGE <= (MAXIMUM)	NO SEMANTICA	1.097	0,82%	0,11%

Tabla 31. Escenario 3. TOP 20 - Patrones con semántica.

SemanticLeft	SemanticRight	Nº Patrones	PATRONES
SI	SI	17	0,23%
NO	SI	298	4,26%

SI	NO	363	8,89%
TOTAL		**678**	**13,15%**

Tabla 32. Escenario 3. Totales patrones con semántica

En la tabla anterior se ve la poca presencia de patrones con semántica que hay en la muestra completa. La semántica está más presente en el lado izquierdo, luego le sigue la semántica a la derecha y por último la semántica a ambos lados.

El patrón con semántica que más aparece es el P1403, sólo tiene semántica a la izquierda y se corresponde con "RANGE >= MINIMUM", dentro de esta categoría semántica están los siguientes token en la muestra: atop, beside, besides, beyond, from, mínimum, outside y over.

El patrón tiene tres niveles de profundidad y está compuesto por las siguientes categorías gramaticales:

ARTICLE + ADJETIVE + NOUN + NUMBER

La equivalencia del patrón y su gramática es la siguiente,

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1403	P1387	NUMBER

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1387	ARTICLE	P1386

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1386	ADJECTIVE	NOUN

Tabla 33. Escenario 3. Semántica. Patrón más repetido con semántica

9.5 Escenario 4

Características del escenario:

- Muestra de 379 patentes procedentes de OEP.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 20 para crear patrones.

- Activamos el checkbox para diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.050.851 (tabla BasicPattern)
- ✓ Patrones: 4.581 (tabla Pattern)
  - Formados por dos TermTag: 550
  - Formados por patrones: 2.062
  - Mixtos: 1.969
- ✓ Patrones con semántica: 368 (tabla SemanticsBelongPatterns)

Este es el primer escenario creado con las muestras de patentes europeas, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 20, el máximo disponible, y vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 20 veces entre los documentos. Además deberán tener la misma semántica para que puedan ser del mismo patrón.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

La ejecución es menos restrictiva a los anteriores escenarios, y su ejecución ha necesitado 86,5 horas para finalizar (más de 3 días y medio).

9.5.1 Patrones

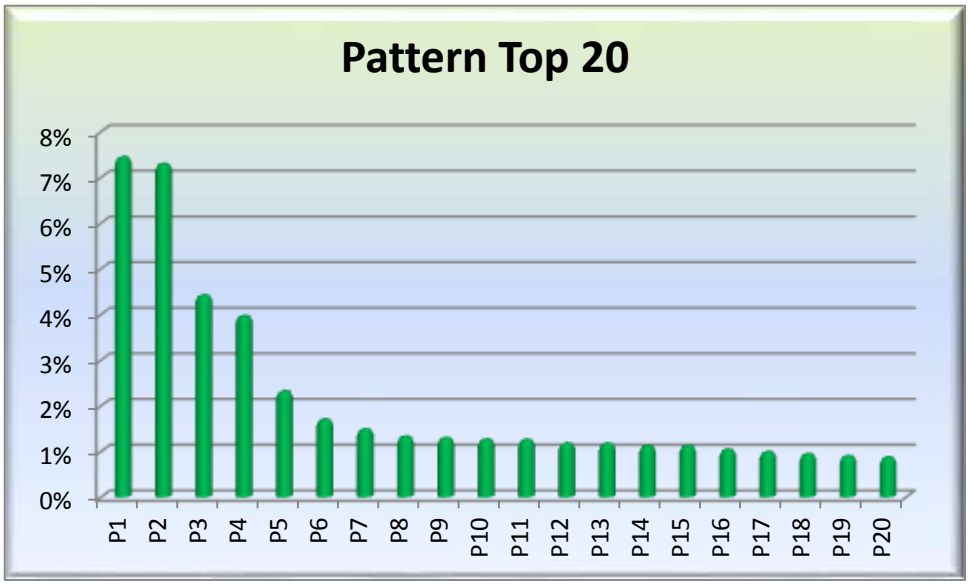
Los patrones están compuestos por dos TermTag, por dos subpatrones o por un TermTag y un subpatrón.

Para este escenario se han distinguido 4.581 patrones diferentes y un total de 1.518.865 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, es el patrón con nombre P1. Tanto por su lado izquierdo como por el derecho, se corresponde con un nombre no clasificado. El segundo patrón con más repeticiones y casi con el mismo número de repeticiones que el primero, es el patrón con nombre P2, éste también

tiene un único nivel de profundidad y tiene como TermTag a la derecha y a la izquierda un nombre; P2 = NOUN + NOUN

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 43% del total.



Gráfica 18. Escenario 4. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 550 que hay de este tipo. El total de este tipo representa el 59,68% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN	113.225	7,45%
P2	NOUN	NOUN	110.953	7,30%
P3	ADJECTIVE	NOUN	67.286	4,43%
P4	ARTICLE	NOUN	60.372	3,97%

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P5	UNCLASSIFIED NOUN	NOUN	35.352	2,33%
P6	NOUN	COMMA	25.939	1,71%
P7	UNCLASSIFIED NOUN	COMMA	22.628	1,49%
P10	UNCLASSIFIED NOUN	ArithmeticOperator	19.283	1,27%
P11	NUMBER	ArithmeticOperator	19.202	1,26%
P12	NUMBER	COMMA	17.992	1,18%
P13	NUMBER	NOUN	17.912	1,18%
P15	UNCLASSIFIED NOUN	SYMBOL	17.176	1,13%
P16	VERB TO BE	VERB	15.881	1,05%
P17	VERB	NOUN	15.164	1,00%
P18	NOUN	ArithmeticOperator	14.457	0,95%
P20	VERB TO BE	ADJECTIVE	13.374	0,88%
P23	NOUN	PREPOSITION OF	10.816	0,71%
P24	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	10.436	0,69%
P25	NOUN	VERB	9.650	0,64%
P26	NOUN	UNCLASSIFIED NOUN	9.528	0,63%

Tabla 34. Escenario 4. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 2.062 combinaciones obtenidas, el total representa el 12,58% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P9	P1	P1	19.764	1,30%
P55	P1	P5	3.717	0,24%
P77	P6	P6	2.445	0,16%
P93	P7	P7	2.044	0,13%
P95	P12	P12	2.005	0,13%
P102	P9	P9	1.874	0,12%
P104	P5	P1	1.809	0,12%
P111	P19	P4	1.622	0,11%
P123	P9	P1	1.440	0,09%
P131	P11	P10	1.356	0,09%
P138	P1	P26	1.276	0,08%
P139	P33	P24	1.258	0,08%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P140	P33	P27	1.244	0,08%
P148	P2	P2	1.177	0,08%
P149	P5	P5	1.171	0,08%
P152	P4	P3	1.148	0,08%
P167	P11	P7	1.043	0,07%
P175	P19	P2	979	0,06%
P183	P1	P2	939	0,06%
P184	P12	P11	934	0,06%

Tabla 35. Escenario 4. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 952 patrones diferentes, suponen el 12,66% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P19	P4	PREPOSITION OF	13.769	0,91%
P35	P4	VERB	6.273	0,41%
P36	P1	NOUN	5.933	0,39%
P43	P2	NOUN	4.879	0,32%
P45	P2	COMMA	4.454	0,29%
P48	P1	UNCLASSIFIED NOUN	4.217	0,28%
P57	P1	COMMA	3.581	0,24%
P58	P3	COMMA	3.509	0,23%
P67	P5	COMMA	2.909	0,19%
P68	P7	NOUN	2.875	0,19%
P69	P16	PREPOSITION	2.855	0,19%
P82	P3	VERB	2.309	0,15%
P85	P12	UNCLASSIFIED NOUN	2.289	0,15%
P86	P4	COMMA	2.269	0,15%
P87	P2	VERB	2.230	0,15%
P107	P6	NOUN	1.725	0,11%
P109	P3	PREPOSITION OF	1.691	0,11%
P112	P8	PREPOSITION OF	1.610	0,11%
P113	P8	VERB	1.591	0,10%
P114	P20	PREPOSITION TO	1.586	0,10%

Tabla 36. Escenario 4. Pattern – Top 20. Patrón + TermTag

Se encuentran 1.017 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un 15,08% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P8	ARTICLE	P3	20.258	1,33%
P14	ARTICLE	P2	17.186	1,13%
P21	SYMBOL	P15	12.724	0,84%
P22	ADJECTIVE	P2	12.347	0,81%
P29	UNCLASSIFIED NOUN	P2	7.662	0,50%
P31	NOUN	P3	7.210	0,47%
P47	OPENING ROUND BRACKETS	P27	4.335	0,29%
P56	PREPOSITION OF	P4	3.706	0,24%
P60	OPENING ROUND BRACKETS	P24	3.412	0,22%
P62	ARTICLE	P22	3.216	0,21%
P65	PREPOSITION	P2	2.926	0,19%
P74	PREPOSITION OF	P2	2.632	0,17%
P80	PREPOSITION OF	P8	2.372	0,16%
P83	ARTICLE	P5	2.296	0,15%
P97	MODAL VERB	P20	1.957	0,13%
P98	AND LINKING	P2	1.953	0,13%
P99	ArithmeticOperator	P2	1.953	0,13%
P101	NUMBER	P2	1.891	0,12%
P110	PREPOSITION	P4	1.652	0,11%
P118	NOUN	P1	1.523	0,10%

Tabla 37. Escenario 4. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es mucho menor al resto.

	Nº Patrones	Repeticiones
TermTag + TermTag	550	59,68%
Patrón + Patrón	2.062	12,58%

Patrón + TermTag	952	12,66%
TermTag + Patrón	1.017	15,08%
TOTAL	**4.581**	**100%**

Tabla 38. Escenario 4. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto por patrones a ambos lados, supone el mayor número de patrones diferentes, pero en cambio tiene el menor peso en cuestión de número de repeticiones totales.

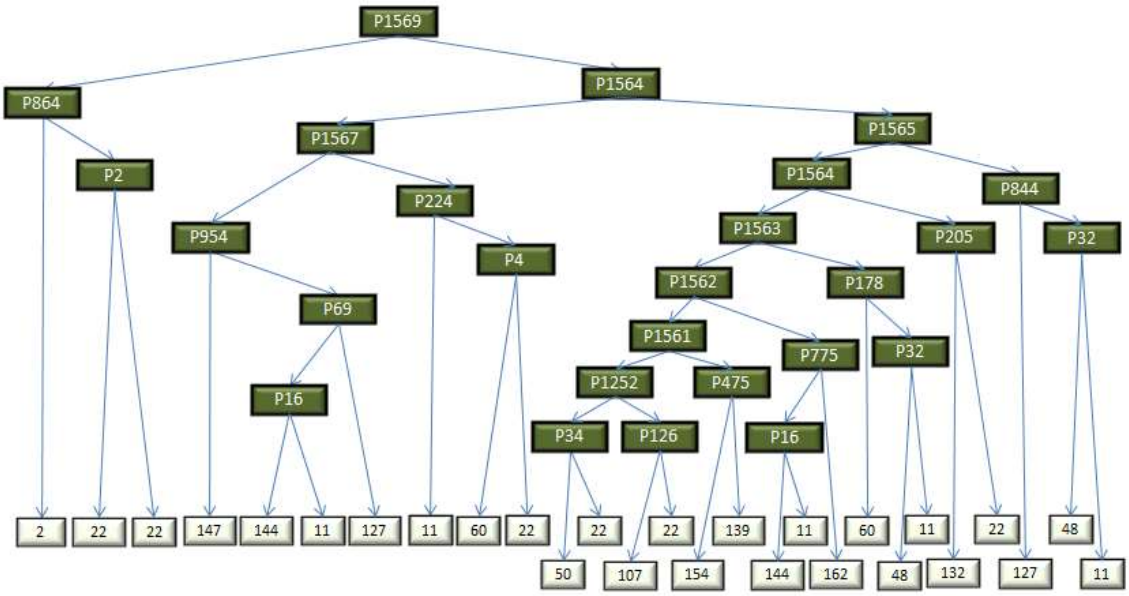
Longitud de patrones

Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón con mayor longitud de descomposición dentro de este escenario es el patrón P3539. Forma una oración de 56 termtag. Descomponiendo todos los niveles llegamos a una oración donde el termtag no clasificado aparece muchas veces. No nos está aportando ningún valor, no nos da ninguna información útil que se pueda llevar a la práctica.

P3539 = UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL

Una oración construida en inglés, al menos debería contener un nombre y un verbo para que dé sentido, con estas características buscamos el siguiente patrón más largo que aporte información, nos encontramos con el patrón P1569 que forma una frase de 27 termtag gramaticales. El padrón está formado por los siguientes subpatrones y TermTag:



Gráfica 19. Escenario 4. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P1569 = PARTICLE + NOUN + NOUN + VERB TO HAVE + VERB TO BE + VERB + PREPOSITION + VERB + ARTICLE + NOUN + COMMA + NOUN + OR LINKING + NOUN + MODAL VERB + NEGATION + VERB TO BE + VERB + AND LINKING + ARTICLE + UNCLASSIFIED NOUN + VERB + QUANTIFIER DETERMINER + NOUN + PREPOSITION + UNCLASSIFIED NOUN + VERB

9.5.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 368 patrones los que tienen semántica, suponen un 3,09% dentro de todos los patrones. Las siguientes tablas muestran diferentes combinaciones de los patrones con esta característica.

Los patrones que tienen mayor semántica se muestran en la siguiente tabla, donde se muestra por patrón el peso de las repeticiones dentro de los patrones con semántica (campo SEMANTICA) y el peso de las

repeticiones teniendo en cuenta todos los patrones, con semántica y sin semántica (campo PATRONES):

Patrón	SemanticLeft	SemanticRight	FRECUENCIA	SEMANTICA	PATRONES
P59	RANGE <= (MAXIMUM)	NO SEMANTICA	3491	7,44%	0,23%
P69	NO SEMANTICA	RANGE <= (MAXIMUM)	2855	6,08%	0,19%
P91	MODAL OPTIONAL	NO SEMANTICA	2072	4,41%	0,14%
P97	MODAL OPTIONAL	NO SEMANTICA	1957	4,17%	0,13%
P170	RANGE <= (MAXIMUM)	NO SEMANTICA	1018	2,17%	0,07%
P171	MODAL OPTIONAL	NO SEMANTICA	1009	2,15%	0,07%
P190	RANGE <= (MAXIMUM)	NO SEMANTICA	902	1,92%	0,06%
P198	NO SEMANTICA	RANGE >= MINIMUM	860	1,83%	0,06%
P201	NO SEMANTICA	Operation	847	1,80%	0,06%
P205	RANGE ALL	NO SEMANTICA	823	1,75%	0,05%
P214	MODAL OPTIONAL	NO SEMANTICA	788	1,68%	0,05%
P227	NO SEMANTICA	RANGE <= (MAXIMUM)	746	1,59%	0,05%
P260	NO SEMANTICA	RANGE >= MINIMUM	652	1,39%	0,04%
P279	RANGE <= (MAXIMUM)	NO SEMANTICA	582	1,24%	0,04%
P285	NO SEMANTICA	Deny	569	1,21%	0,04%
P299	NO SEMANTICA	Operation	534	1,14%	0,04%
P300	RATE	NO SEMANTICA	534	1,14%	0,04%
P303	NO SEMANTICA	RANGE >= MINIMUM	532	1,13%	0,04%
P316	NO SEMANTICA	RANGE >= MINIMUM	506	1,08%	0,03%
P327	Deny	NO SEMANTICA	478	1,02%	0,03%

Tabla 39. Escenario 4. TOP 20 - Patrones con semántica.

SemanticLeft	SemanticRight	Nº Patrones	PATRONES
SI	SI	8	0,04%
NO	SI	145	1,12%
SI	NO	215	1,93%
TOTAL		**368**	**3,09%**

Tabla 40. Escenario 4. Totales patrones con semántica

En la tabla anterior se ve la poca presencia de patrones con semántica que hay en la muestra completa. La semántica está más presente en el lado izquierdo, luego le sigue la semántica a la derecha y por último la semántica a ambos lados.

El patrón con semántica que más aparece es el P59, sólo tiene semántica a la izquierda y se corresponde con "RANGE <= (MAXIMUM)", dentro de esta categoría semántica están los siguientes token en la muestra: before, beneath, down, in, inferior, inside, into, last, máximum and within.

El patrón tiene un nivel de profundidad y está compuesto por las siguientes categorías gramaticales:

PREPOSITION + NOUN

La equivalencia del patrón y su gramática es la siguiente,

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha
P59	PREPOSITION	NOUN

Tabla 41. Escenario 4 – Semántica. Patrón más repetido con semántica

9.6 Escenario 5

Características del escenario:

- Muestra de 359 patentes procedentes de USPTO.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 100 para crear patrones.
- Activamos el checkbox para diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.089.157 (tabla BasicPattern)
- ✓ Patrones únicos: 1.027 (tabla Pattern)
  - Formados por dos TermTag: 161
  - Formados por subpatrones: 279
  - Mixtos: 587
- ✓ Patrones con semántica: 205 (tabla SemanticsBelongPatterns)

Este es el cuarto escenario creado con las muestras de patentes estadounidenses, al igual que en el resto de escenarios, se van a tener

en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 100, el máximo disponible, y vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 100 veces entre los documentos. Además deberán tener la misma semántica para que puedan ser del mismo patrón.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

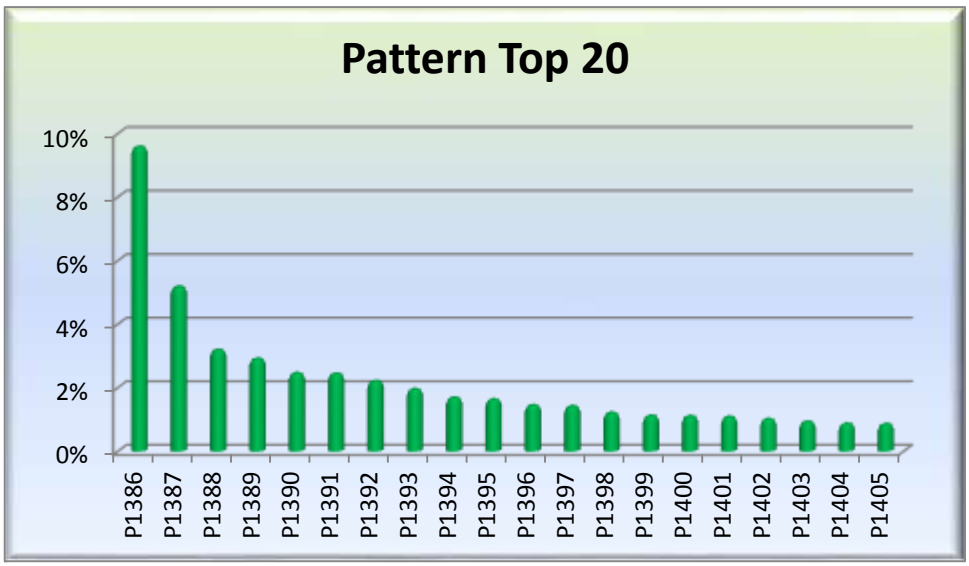
Es la ejecución más restrictiva que en los anteriores escenarios, la ejecución de este escenario ha necesitado 66 horas para finalizar.

9.6.1 Patrones

En la tabla Pattern de la base de datos RequirementsClassification, tenemos 1.027 patrones diferentes para este escenario y un total de 883.521 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, tiene el nombre P1386. Está formado por la categoría gramatical adjetivo por la izquierda y nombre por la derecha.

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 44,19% del total.



Gráfica 20. Escenario 5. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 161 que hay de este tipo. El total de este tipo representa el 33,81% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1386	ADJECTIVE	NOUN	84.797	9,60%
P1390	VERB	NOUN	21.869	2,48%
P1392	SYMBOL	UNCLASSIFIED NOUN	19.644	2,22%
P1394	VERB TO BE	ADJECTIVE	14.988	1,70%
P1395	NOUN	VERB	14.534	1,65%
P1396	VERB TO BE	VERB	12.937	1,46%
P1401	PREPOSITION	NOUN	9.700	1,10%
P1405	ADVERB	VERB	7.717	0,87%
P1406	PREPOSITION TO	VERB	7.500	0,85%
P1407	UNCLASSIFIED NOUN	VERB	7.471	0,85%
P1408	NOUN	NUMBER	6.615	0,75%
P1410	NUMBER	NUMBER	5.676	0,64%
P1411	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	5.584	0,63%
P1415	PREPOSITION	VERB	4.708	0,53%
P1421	VERB	VERB	4.160	0,47%
P1433	AND LINKING	VERB	3.207	0,36%
P1434	ADJECTIVE	PHRASAL VERB PARTICLE	2.993	0,34%
P1435	OPENING ROUND BRACKETS	ADJECTIVE	2.972	0,34%
P1436	COMMA	UNCLASSIFIED NOUN	2.948	0,33%
P1444	NOUN	ArithmeticOperator	2.208	0,25%

Tabla 42. Escenario 5. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 279 combinaciones obtenidas, el total representa el 13,51% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P1389	P1388	P1388	28.287	3,20%
P1409	P1389	P1389	5.979	0,68%
P1417	P1391	P1387	4.589	0,52%
P1429	P1388	P1387	3.272	0,37%
P1439	P1396	P1387	2.352	0,27%
P1440	P1388	P1386	2.265	0,26%
P1455	P1402	P1387	1.867	0,21%
P1463	P1409	P1409	1.692	0,19%
P1468	P1386	P1386	1.561	0,18%
P1482	P1391	P1386	1.277	0,14%
P1489	P1402	P1386	1.199	0,14%
P1494	P1389	P1388	1.137	0,13%
P1502	P1406	P1387	1.053	0,12%
P1507	P1396	P1386	988	0,11%
P1518	P1387	P1394	867	0,10%
P1519	P1387	P1386	853	0,10%
P1521	P1403	P1403	830	0,09%
P1528	P1404	P1387	790	0,09%
P1530	P1434	P1387	781	0,09%
P1532	P1426	P1388	769	0,09%

Tabla 43. Escenario 5. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 270 patrones diferentes, suponen el 27,66% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1388	P1388	UNCLASSIFIED NOUN	25.871	2,93%
P1391	P1387	PREPOSITION OF	21.678	2,45%
P1397	P1387	VERB	12.642	1,43%
P1398	P1388	NOUN	10.829	1,23%
P1399	P1386	COMMA	9.999	1,13%
P1400	P1386	VERB	9.902	1,12%

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1402	P1386	PREPOSITION OF	9.096	1,03%
P1403	P1387	NUMBER	8.329	0,94%
P1404	P1387	COMMA	7.812	0,88%
P1412	P1386	NOUN	5.266	0,60%
P1413	P1388	NUMBER	5.144	0,58%
P1416	P1405	CLOSING ROUND BRACKETS	4.645	0,53%
P1418	P1386	NUMBER	4.444	0,50%
P1422	P1388	COMMA	4.158	0,47%
P1424	P1388	ARTICLE	3.828	0,43%
P1426	P1391	NOUN	3.507	0,40%
P1430	P1388	VERB	3.266	0,37%
P1438	P1435	CLOSING ROUND BRACKETS	2.564	0,29%
P1442	P1396	PREPOSITION	2.230	0,25%
P1448	P1394	PREPOSITION TO	1.990	0,23%

Tabla 44. Escenario 5. Pattern – Top 20. Patrón + TermTag

Se encuentran 317 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un 25,01% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P1387	ARTICLE	P1386	45.896	5,19%
P1393	ADJECTIVE	P1386	17.394	1,97%
P1414	PREPOSITION	P1387	4.722	0,53%
P1419	ARTICLE	P1393	4.442	0,50%
P1420	OPENING ROUND BRACKETS	P1411	4.312	0,49%
P1423	PREPOSITION	P1386	4.118	0,47%
P1425	AND LINKING	P1386	3.780	0,43%
P1427	PREPOSITION TO	P1387	3.489	0,39%
P1428	ArithmeticOperator	P1386	3.275	0,37%
P1431	ARTICLE	P1392	3.265	0,37%
P1432	PREPOSITION OF	P1387	3.250	0,37%
P1437	MODAL VERB	P1394	2.794	0,32%
P1441	VERB	P1386	2.231	0,25%

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P1443	AND LINKING	P1387	2.225	0,25%
P1445	UNCLASSIFIED NOUN	P1406	2.188	0,25%
P1446	UNCLASSIFIED NOUN	P1386	2.094	0,24%
P1449	ADVERB	P1387	1.990	0,23%
P1450	NUMBER	P1386	1.985	0,22%
P1451	VERB	P1387	1.979	0,22%
P1452	OPENING ROUND BRACKETS	P1426	1.935	0,22%

Tabla 45. Escenario 5. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es mucho menor al resto.

	Nº Patrones	Repeticiones
TermTag + TermTag	161	33,81%
Patrón + Patrón	279	13,51%
Patrón + TermTag	270	27,66%
TermTag + Patrón	317	25,01%
TOTAL	**1.027**	**100%**

Tabla 46. Escenario 5. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto sólo por patrón a la derecha, supone el mayor número de patrones diferentes, y los compuestos por patrones a ambos lados tienen el menor peso en cuestión de número de repeticiones totales.

Longitud de patrones

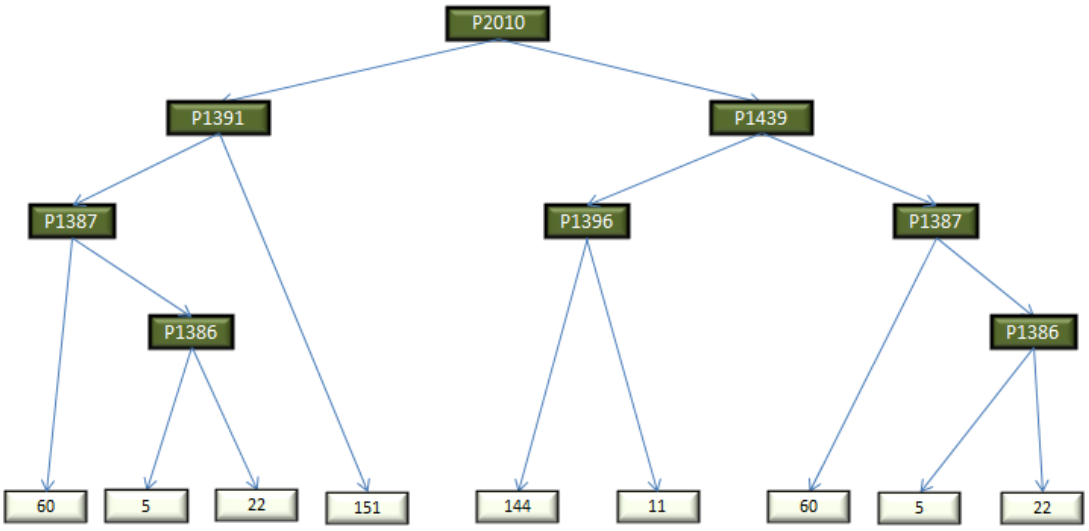
Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón con mayor longitud de descomposición dentro de este escenario es el patrón P1726. Forma una frase de infinitos termtag. Descomponiendo todos los niveles llegamos a una frase donde el

termtag no clasificado aparece infinitas veces. No nos está aportando ningún valor, no nos da ninguna información útil que se pueda llevar a la práctica.

P1726 = INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN + INFINITO + UNCLASSIFIED NOUN

Toda frase escrita en inglés, al menos tiene un nombre y un verbo. Buscamos el siguiente patrón más largo que aporte información, nos encontramos con el patrón P2010 que forma una frase de 9 termtag gramaticales y además tiene nombre y verbo:



Gráfica 21. Escenario 5. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P2010 = ARTICLE + ADJECTIVE + NOUN + PREPOSITION OF + VERB TO BE + VERB + ARTICLE + ADJECTIVE + NOUN

9.6.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 205 patrones los que tienen semántica, suponen un 12,49% dentro de todos los patrones. Las siguientes tablas muestran diferentes combinaciones de los patrones con esta característica.

Los patrones que tienen mayor semántica se muestran en la siguiente tabla, donde se muestra por patrón el peso de las repeticiones dentro de los patrones con semántica (campo SEMANTICA) y el peso de las repeticiones teniendo en cuenta todos los patrones, con semántica y sin semántica (campo PATRONES):

Pattern	SemanticLeft	SemanticRight	FRECUENCIA	SEMANTICA	PATRONES
P1403	RANGE >= MINIMUM	NO SEMANTICA	8.329	7,55%	0,94%
P1405	RANGE > MINIMUM	NO SEMANTICA	7.717	6,99%	0,87%
P1409	RANGE <= (MAXIMUM)	NO SEMANTICA	5.979	5,42%	0,68%
P1425	NO SEMANTICA	RATE	3.780	3,43%	0,43%
P1426	MODAL OPTIONAL	NO SEMANTICA	3.507	3,18%	0,40%
P1433	RANGE LITTLE-FEW-SOME	NO SEMANTICA	3.207	2,91%	0,36%
P1435	NO SEMANTICA	Support	2.972	2,69%	0,34%
P1437	IF ACTIVATION	NO SEMANTICA	2.794	2,53%	0,32%
P1437	MODAL OPTIONAL	NO SEMANTICA	2.794	2,53%	0,32%
P1442	NO SEMANTICA	RANGE <= (MAXIMUM)	2.230	2,02%	0,25%
P1457	RANGE >= MINIMUM	NO SEMANTICA	1.812	1,64%	0,21%
P1460	Support	NO SEMANTICA	1.735	1,57%	0,20%
P1463	RANGE >= MINIMUM	NO SEMANTICA	1.692	1,53%	0,19%
P1465	RANGE <= (MAXIMUM)	NO SEMANTICA	1.675	1,52%	0,19%
P1465	MODAL OPTIONAL	NO SEMANTICA	1.675	1,52%	0,19%
P1466	MODAL OPTIONAL	NO SEMANTICA	1.670	1,51%	0,19%
P1474	NO SEMANTICA	RANGE >= MINIMUM	1.484	1,34%	0,17%
P1490	NO SEMANTICA	UNIT	1.183	1,07%	0,13%
P1491	RANGE <= (MAXIMUM)	NO SEMANTICA	1.169	1,06%	0,13%
P1498	RANGE <= (MAXIMUM)	NO SEMANTICA	1.097	0,99%	0,12%

Tabla 47. Escenario 5. TOP 20 - Patrones con semántica.

SemanticLeft	SemanticRight	Nº Patrones	PATRONES
SI	SI	6	0,21%
NO	SI	77	3,65%
SI	NO	122	8,63%
TOTAL		**205**	**12,49%**

Tabla 48. Escenario 5. Totales patrones con semántica

En la tabla anterior se ve la poca presencia de patrones con semántica que hay en la muestra completa. La semántica está más presente en el lado izquierdo, luego le sigue la semántica a la derecha y por último la semántica a ambos lados.

El patrón con semántica que más aparece es el P1403, sólo tiene semántica a la izquierda y se corresponde con "RANGE >= MINIMUM", dentro de esta categoría semántica están los siguientes token en la muestra: atop, beside, besides, beyond, from, mínimum, outside y over.

El patrón tiene tres niveles de profundidad y está compuesto por las siguientes categorías gramaticales:

ARTICLE + ADJETIVE + NOUN + NUMBER

La equivalencia del patrón y su gramática es la siguiente,

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1403	P1387	NUMBER

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1387	ARTICLE	P1386

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P1386	ADJECTIVE	NOUN

Tabla 49. Escenario 5. Semántica. Patrón más repetido con semántica

9.7 Escenario 6

Características del escenario:

- Muestra de 379 patentes procedentes de OEP.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 100 para crear patrones.
- Activamos el checkbox para diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.050.851 (tabla BasicPattern)
- ✓ Patrones: 1.170 (tabla Pattern)
  - Formados por dos TermTag: 267
  - Formados por patrones: 338
  - Mixtos: 565
- ✓ Patrones con semántica: 86 (tabla SemanticsBelongPatterns)

Este es el segundo escenario creado con las muestras de patentes europeas, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 100, el máximo disponible, y vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 100 veces entre los documentos. Además deberán tener la misma semántica para que puedan ser del mismo patrón.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

Es la ejecución más restrictiva que en los anteriores escenarios, la ejecución de este escenario ha necesitado 27,5 horas para finalizar.

9.7.1 Patrones

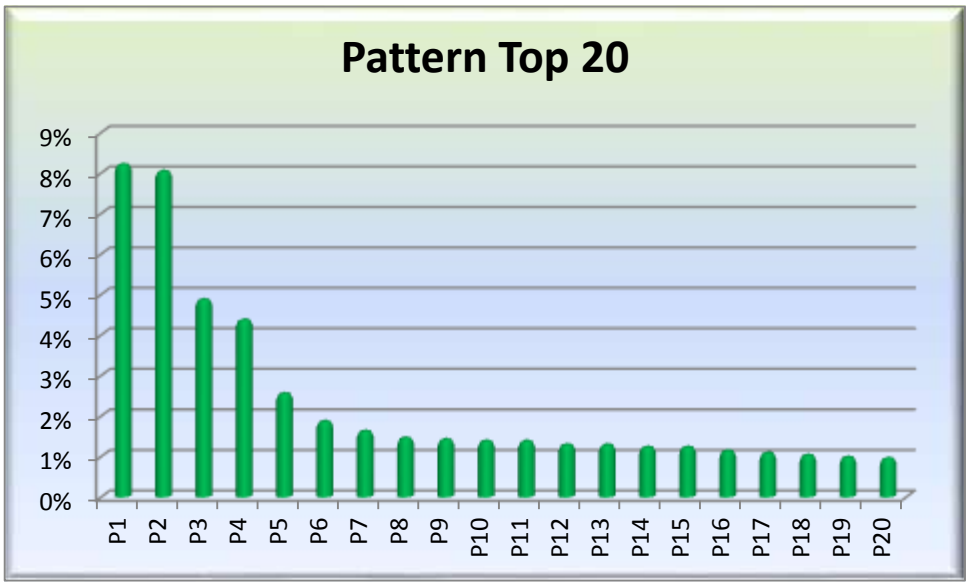
Los patrones están compuestos por dos TermTag, por dos subpatrones o por un TermTag y un subpatrón.

Para este escenario se han distinguido 1.170 patrones diferentes y un total de 1.375.981 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, es el patrón con nombre P1. Tanto por su lado izquierdo como por el derecho, se corresponde con un nombre no clasificado. El segundo patrón con más repeticiones y casi con el mismo número de

repeticiones que el primero, es el patrón con nombre P2, éste también tiene un único nivel de profundidad y tiene como TermTag a la derecha y a la izquierzda un nombre; P2 = NOUN + NOUN

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 48% del total.



Gráfica 22. Escenario 6. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 267 que hay de este tipo. El total de este tipo representa el 64,94% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN	113.225	8,23%
P2	NOUN	NOUN	110.953	8,06%
P3	ADJECTIVE	NOUN	67.286	4,89%
P4	ARTICLE	NOUN	60.372	4,39%

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P5	UNCLASSIFIED NOUN	NOUN	35.352	2,57%
P6	NOUN	COMMA	25.939	1,89%
P7	UNCLASSIFIED NOUN	COMMA	22.628	1,64%
P10	UNCLASSIFIED NOUN	ArithmeticOperator	19.283	1,40%
P11	NUMBER	ArithmeticOperator	19.202	1,40%
P12	NUMBER	COMMA	17.992	1,31%
P13	NUMBER	NOUN	17.912	1,30%
P15	UNCLASSIFIED NOUN	SYMBOL	17.176	1,25%
P16	VERB TO BE	VERB	15.881	1,15%
P17	VERB	NOUN	15.164	1,10%
P18	NOUN	ArithmeticOperator	14.457	1,05%
P20	VERB TO BE	ADJECTIVE	13.374	0,97%
P23	NOUN	PREPOSITION OF	10.816	0,79%
P24	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	10.436	0,76%
P25	NOUN	VERB	9.650	0,70%
P26	NOUN	UNCLASSIFIED NOUN	9.528	0,69%

Tabla 50. Escenario 6. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 338 combinaciones obtenidas, el total representa el 8,90% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P9	P1	P1	19.764	1,44%
P55	P1	P5	3.717	0,27%
P77	P6	P6	2.445	0,18%
P93	P7	P7	2.044	0,15%
P95	P12	P12	2.005	0,15%
P102	P9	P9	1.874	0,14%
P104	P5	P1	1.809	0,13%
P111	P19	P4	1.622	0,12%
P123	P9	P1	1.440	0,10%
P131	P11	P10	1.356	0,10%
P138	P1	P26	1.276	0,09%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P139	P33	P24	1.258	0,09%
P140	P33	P27	1.244	0,09%
P148	P2	P2	1.177	0,09%
P149	P5	P5	1.171	0,09%
P152	P4	P3	1.148	0,08%
P167	P11	P7	1.043	0,08%
P175	P19	P2	979	0,07%
P183	P1	P2	939	0,07%
P184	P12	P11	934	0,07%

Tabla 51. Escenario 6. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 278 patrones diferentes, suponen el 11,80% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P19	P4	PREPOSITION OF	13.769	1,00%
P35	P4	VERB	6.273	0,46%
P36	P1	NOUN	5.933	0,43%
P43	P2	NOUN	4.879	0,35%
P45	P2	COMMA	4.454	0,32%
P48	P1	UNCLASSIFIED NOUN	4.217	0,31%
P57	P1	COMMA	3.581	0,26%
P58	P3	COMMA	3.509	0,26%
P67	P5	COMMA	2.909	0,21%
P68	P7	NOUN	2.875	0,21%
P69	P16	PREPOSITION	2.855	0,21%
P82	P3	VERB	2.309	0,17%
P85	P12	UNCLASSIFIED NOUN	2.289	0,17%
P86	P4	COMMA	2.269	0,16%
P87	P2	VERB	2.230	0,16%
P107	P6	NOUN	1.725	0,13%
P109	P3	PREPOSITION OF	1.691	0,12%
P112	P8	PREPOSITION OF	1.610	0,12%
P113	P8	VERB	1.591	0,12%

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P114	P20	PREPOSITION TO	1.586	0,12%

Tabla 52. Escenario 6. Pattern – Top 20. Patrón + TermTag

Se encuentran 287 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un 14,36% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P8	ARTICLE	P3	20.258	1,47%
P14	ARTICLE	P2	17.186	1,25%
P21	SYMBOL	P15	12.724	0,92%
P22	ADJECTIVE	P2	12.347	0,90%
P29	UNCLASSIFIED NOUN	P2	7.662	0,56%
P31	NOUN	P3	7.210	0,52%
P47	OPENING ROUND BRACKETS	P27	4.335	0,32%
P56	PREPOSITION OF	P4	3.706	0,27%
P60	OPENING ROUND BRACKETS	P24	3.412	0,25%
P62	ARTICLE	P22	3.216	0,23%
P65	PREPOSITION	P2	2.926	0,21%
P74	PREPOSITION OF	P2	2.632	0,19%
P80	PREPOSITION OF	P8	2.372	0,17%
P83	ARTICLE	P5	2.296	0,17%
P97	MODAL VERB	P20	1.957	0,14%
P98	AND LINKING	P2	1.953	0,14%
P99	ArithmeticOperator	P2	1.953	0,14%
P101	NUMBER	P2	1.891	0,14%
P110	PREPOSITION	P4	1.652	0,12%
P118	NOUN	P1	1.523	0,11%

Tabla 53. Escenario 6. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es menor al resto.

	Nº Patrones	Repeticiones
TermTag + TermTag	267	64,94%
Patrón + Patrón	338	8,90%
Patrón + TermTag	278	11,80%
TermTag + Patrón	287	14,36%
TOTAL	**1.170**	**100%**

Tabla 54. Escenario 6. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto por patrones a ambos lados, supone el mayor número de patrones diferentes, pero en cambio tiene el menor peso en cuestión de número de repeticiones totales.

Longitud de patrones

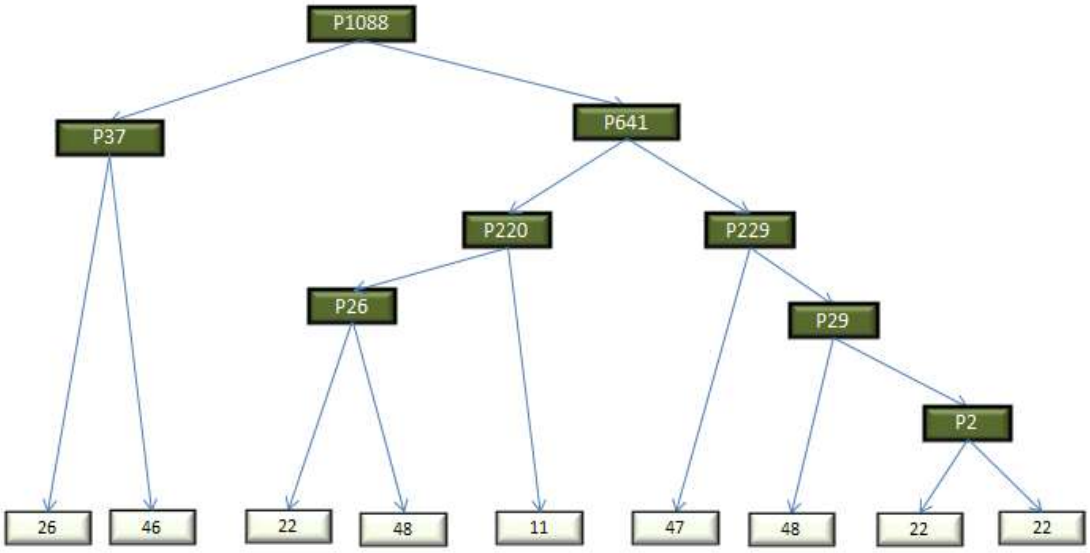
Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón más largo que se ha generado en este escenario es P954, pero está compuesto por 28 TermTag, pero muchos de ellos son no clasificados, símbolos y números. Aun siendo un patrón finito, no parece representar un oración con sentido.

P954 = CLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL

Las oraciones en inglés tienen que estar formadas, al menos, por un nombre y un verbo, teniendo en cuenta éstas características, encontramos el siguiente patrón más largo que lo cumple.

En la siguiente gráfica mostramos un ejemplo con el patrón P1088, formado por 9 TermTag gramaticales:



Gráfica 23. Escenario 6. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P1088 = NUMBER + CLOSING ROUND BRACKETS + NOUN + UNCLASSIFIED NOUN + VERB + OPENING ROUND BRACKETS + UNCLASSIFIED NOUN + NOUN + NOUN

No es un patrón de oración muy conciso al tener presente en tres ocasiones el nombre no clasificado.

9.7.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 86 patrones los que tienen semántica, suponen un 2,54% dentro de todos los patrones. Las siguientes tablas muestran diferentes combinaciones de los patrones con esta característica.

Los patrones que tienen mayor semántica se muestran en la siguiente tabla, donde se muestra por patrón el peso de las repeticiones dentro de los patrones con semántica (campo SEMANTICA) y el peso de las repeticiones teniendo en cuenta todos los patrones, con semántica y sin semántica (campo PATRONES):

Patrón	SemanticLeft	SemanticRight	FRECUENCIA	SEMANTICA	PATRONES
P59	RANGE <= (MAXIMUM)	NO SEMANTICA	3491	9,99%	0,25%

Patrón	SemanticLeft	SemanticRight	FRECUENCIA	SEMANTICA	PATRONES
P69	NO SEMANTICA	RANGE <= (MAXIMUM)	2855	8,17%	0,21%
P91	MODAL OPTIONAL	NO SEMANTICA	2072	5,93%	0,15%
P97	MODAL OPTIONAL	NO SEMANTICA	1957	5,60%	0,14%
P170	RANGE <= (MAXIMUM)	NO SEMANTICA	1018	2,91%	0,07%
P171	MODAL OPTIONAL	NO SEMANTICA	1009	2,89%	0,07%
P190	RANGE <= (MAXIMUM)	NO SEMANTICA	902	2,58%	0,07%
P198	NO SEMANTICA	RANGE >= MINIMUM	860	2,46%	0,06%
P201	NO SEMANTICA	Operation	847	2,42%	0,06%
P205	RANGE ALL	NO SEMANTICA	823	2,36%	0,06%
P214	MODAL OPTIONAL	NO SEMANTICA	788	2,26%	0,06%
P227	NO SEMANTICA	RANGE <= (MAXIMUM)	746	2,14%	0,05%
P260	NO SEMANTICA	RANGE >= MINIMUM	652	1,87%	0,05%
P279	RANGE <= (MAXIMUM)	NO SEMANTICA	582	1,67%	0,04%
P285	NO SEMANTICA	Deny	569	1,63%	0,04%
P299	NO SEMANTICA	Operation	534	1,53%	0,04%
P300	RATE	NO SEMANTICA	534	1,53%	0,04%
P303	NO SEMANTICA	RANGE >= MINIMUM	532	1,52%	0,04%
P316	NO SEMANTICA	RANGE >= MINIMUM	506	1,45%	0,04%
P327	Deny	NO SEMANTICA	478	1,37%	0,03%

Tabla 55. Escenario 6. TOP 20 - Patrones con semántica.

SemanticLeft	SemanticRight	Nº Patrones	PATRONES
SI	SI	2	0,03%
NO	SI	30	0,88%
SI	NO	54	1,63%
TOTAL		**86**	**2,54%**

Tabla 56. Escenario 6. Totales patrones con semántica

En la tabla anterior se ve la poca presencia de patrones con semántica que hay en la muestra completa. La semántica está más presente en el lado izquierdo, luego le sigue la semántica a la derecha y por último la semántica a ambos lados.

El patrón con semántica que más aparece es el P59, sólo tiene semántica a la izquierda y se corresponde con "RANGE <= (MAXIMUM)", dentro de esta categoría semántica están los siguientes token en la muestra: before, beneath, down, in, inferior, inside, into, last, máximum and within.

El patrón tiene un nivel de profundidad, la equivalencia del patrón y su gramática es la siguiente:

Nombre Patrón	TermTagOrPatternLeft	TermTagOrPatternRight
P59	PREPOSITION	NOUN

Tabla 57. Escenario 6. Semántica. Patrón más repetido con semántica

9.8 Escenario 7

Características del escenario:

- Muestra de 359 patentes procedentes de USPTO.
- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 100 para crear patrones.
- No activamos el checkbox para no diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.089.157 (tabla BasicPattern)
- ✓ Patrones únicos: 1.112 (tabla Pattern)
  - Formados por dos TermTag: 233
  - Formados por subpatrones: 305
  - Mixtos: 574
- ✓ Patrones con semántica: 1.752 (tabla SemanticsBelongPatterns)

Este es el quinto y último escenario creado con las muestras de patentes estadounidenses, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 100, el máximo disponible, y no vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 100 veces entre los documentos. No tendrán que cumplir la misma semántica para ser el mismo patrón. Esto es lo que le diferencia del escenario 5.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

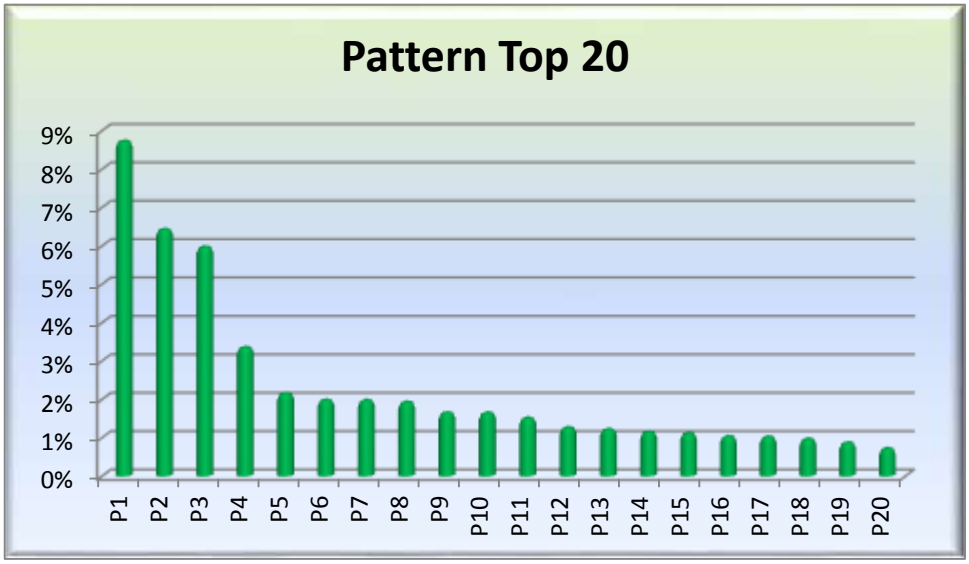
Es la ejecución más restrictiva que en los anteriores escenarios, la ejecución de este escenario ha necesitado 37 horas para finalizar.

9.8.1 Patrones

En la tabla Pattern de la base de datos RequirementsClassification, tenemos 1.112 patrones diferentes para este escenario y un total de 1.314.586 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, tiene el nombre P1. Está formado por la categoría gramatical nombre por la izquierda y nombre por la derecha.

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 47% del total.



Gráfica 24. Escenario 7. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 233 que hay de este tipo. El total de este tipo representa el 62,31% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1	NOUN	NOUN	114.978	8,75%
P2	ADJECTIVE	NOUN	84.797	6,45%
P3	ARTICLE	NOUN	78.710	5,99%
P4	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN	44.320	3,37%
P5	NOUN	COMMA	28.528	2,17%
P6	NUMBER	COMMA	26.265	2,00%
P8	NUMBER	ArithmeticOperator	25.585	1,95%
P9	UNCLASSIFIED NOUN	NOUN	21.928	1,67%
P10	VERB	NOUN	21.869	1,66%
P12	SYMBOL	UNCLASSIFIED NOUN	16.792	1,28%
P14	VERB TO BE	ADJECTIVE	15.303	1,16%
P16	NOUN	VERB	13.777	1,05%
P17	NOUN	PREPOSITION OF	13.623	1,04%
P18	VERB TO BE	VERB	12.937	0,98%
P19	UNCLASSIFIED NOUN	COMMA	11.636	0,89%
P20	PREPOSITION	NOUN	9.661	0,73%
P21	AND LINKING	NOUN	9.291	0,71%
P23	UNCLASSIFIED NOUN	VERB	8.252	0,63%
P24	NUMBER	NOUN	8.112	0,62%
P26	OPENING ROUND BRACKETS	NOUN	7.765	0,59%

Tabla 58. Escenario 7. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 205 combinaciones obtenidas, el total representa el 7,43% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P22	P8	P4	8.509	0,65%
P90	P5	P5	2.275	0,17%
P91	P22	P22	2.267	0,17%
P102	P13	P3	1.957	0,15%
P111	P6	P4	1.807	0,14%
P114	P20	P2	1.757	0,13%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P123	P19	P19	1.588	0,12%
P125	P6	P6	1.566	0,12%
P126	P8	P9	1.559	0,12%
P146	P63	P22	1.301	0,10%
P147	P3	P2	1.291	0,10%
P158	P17	P3	1.209	0,09%
P187	P4	P4	936	0,07%
P188	P32	P32	930	0,07%
P197	P29	P3	892	0,07%
P199	P2	P2	870	0,07%
P202	P17	P1	846	0,06%
P209	P17	P7	808	0,06%
P212	P10	P2	774	0,06%
P213	P111	P22	770	0,06%

Tabla 59. Escenario 7. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 247 patrones diferentes, suponen el 13,83% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P13	P3	PREPOSITION OF	16.203	1,23%
P25	P3	VERB	8.094	0,62%
P28	P6	UNCLASSIFIED NOUN	7.626	0,58%
P32	P8	UNCLASSIFIED NOUN	6.704	0,51%
P34	P2	VERB	6.125	0,47%
P39	P1	NOUN	5.516	0,42%
P43	P2	COMMA	4.718	0,36%
P46	P1	COMMA	4.653	0,35%
P48	P26	CLOSING ROUND BRACKETS	4.618	0,35%
P59	P4	NOUN	3.409	0,26%
P62	P13	NOUN	3.304	0,25%
P63	P28	ARTICLE	3.288	0,25%
P66	P2	PREPOSITION OF	3.214	0,24%
P69	P5	NOUN	3.171	0,24%
P72	P7	VERB	3.006	0,23%

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P74	P1	VERB	2.899	0,22%
P77	P6	NUMBER	2.743	0,21%
P81	P73	CLOSING ROUND BRACKETS	2.564	0,20%
P82	P3	NUMBER	2.556	0,19%
P83	P3	COMMA	2.483	0,19%

Tabla 60. Escenario 7. Pattern – Top 20. Patrón + TermTag

Se encuentran 327 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un 16,44% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P7	ARTICLE	P2	26.179	1,99%
P11	ARTICLE	P1	20.112	1,53%
P15	ADJECTIVE	P1	14.870	1,13%
P42	PREPOSITION OF	P3	4.737	0,36%
P49	OPENING ROUND BRACKETS	P37	4.312	0,33%
P50	PREPOSITION	P1	4.228	0,32%
P52	ARTICLE	P15	4.130	0,31%
P57	MODAL VERB	P14	3.518	0,27%
P60	PREPOSITION	P3	3.354	0,26%
P68	AND LINKING	P1	3.185	0,24%
P71	PREPOSITION TO	P10	3.086	0,23%
P75	ARTICLE	P12	2.781	0,21%
P84	SYMBOL	P9	2.411	0,18%
P94	MODAL VERB	P18	2.117	0,16%
P105	VERB TO BE	P27	1.870	0,14%
P107	UNCLASSIFIED NOUN	P30	1.858	0,14%
P109	PREPOSITION TO	P3	1.822	0,14%
P110	ArithmeticOperator	P1	1.810	0,14%
P112	OPENING ROUND BRACKETS	P51	1.779	0,14%
P113	ARTICLE	P9	1.764	0,13%

Tabla 61. Escenario 7. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mucha mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es menor al resto.

	Nº Patrones	**Repeticiones**
TermTag + TermTag	233	62,31%
Patrón + Patrón	305	7,43%
Patrón + TermTag	247	13,83%
TermTag + Patrón	327	16,44%
TOTAL	**1.112**	**100%**

Tabla 62. Escenario 7. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto sólo por patrón a la derecha, supone el mayor número de patrones diferentes, y los compuestos por patrones a ambos lados tienen el menor peso en cuestión de número de repeticiones totales.

Longitud de patrones

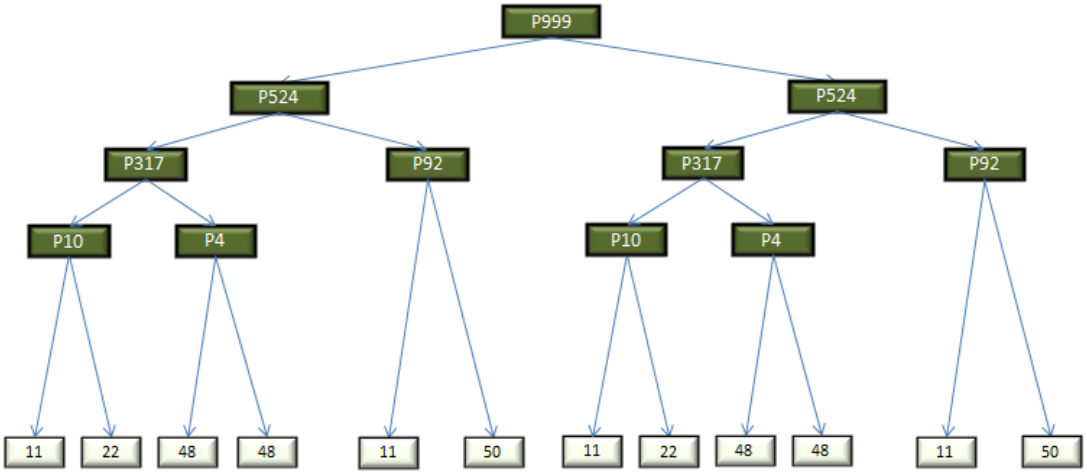
Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón más largo que se ha generado en este escenario es P902, pero está compuesto por 32 TermTag, pero muchos de ellos son no clasificados, operador aritmético y números. Aun siendo un patrón finito, no parece representar un oración con sentido, parece ser un patrón de funciones.

P902= NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NUMBER + ArithmeticOperator + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN

Las oraciones en inglés tienen que estar formadas, al menos, por un nombre y un verbo, teniendo en cuenta éstas características, encontramos el siguiente patrón más largo que lo cumple.

En la siguiente gráfica mostramos un ejemplo con el patrón P999, formado por 12 TermTag gramaticales:



Gráfica 25. Escenario 7. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P999 = VERB + NOUN + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + VERB + COMMA + VERB + NOUN + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + VERB + COMMA

No es un patrón de oración muy conciso al tener presente en tres ocasiones el nombre no clasificado.

9.8.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 1.752 patrones los que tienen semántica, entre ellos hay muchos patrones repetidos pero con semántica diferente, nombres de patrones únicos hay tan sólo 272, un ejemplo de patrón que puede tener semántica diferente, es el P2 y es el segundo más en toda la muestra. Está formado por un adjetivo a la izquierda y un nombre a la derecha.

Las diferentes semánticas que se encuentran entre los documentos para éste patrón son 10 tipos:

Pattern	SemanticLeft	SemanticRight
P2	RANGE < MAXIMUM	NO SEMANTICA

Pattern	SemanticLeft	SemanticRight
P2	AVAILABILITY	NO SEMANTICA
P2	RANGE = EQUAL	NO SEMANTICA
P2	Operation	NO SEMANTICA
P2	RANGE > MINIMUM	NO SEMANTICA
P2	Sustainability	NO SEMANTICA
P2	RANGE <= (MAXIMUM)	NO SEMANTICA
P2	SYSTEM FUNCTION	NO SEMANTICA
P2	RANGE >= MINIMUM	NO SEMANTICA
P2	DUTY ACTION	NO SEMANTICA

Tabla 63. Escenario 7. Semántica. Patrón más repetido con semántica

El adjetivo que forma el termtag de la izquierda puede adquirir la diferente semántica mostrada.

Como éste hay muchos patrones repetidos, al tener el mismo patrón tantas repeticiones no conocemos la frecuencia de aparición de cada uno de ellos entre los documentos.

El número de patrones (no únicos) con semántica a la izquierda y a la derecha es el siguiente:

SemanticLeft	SemanticRight	Nº Patrones
SI	SI	141
NO	SI	865
SI	NO	746
TOTAL		**1.752**

Tabla 64. Escenario 7. Totales patrones con semántica

Se puede ver que sólo semántica a la derecha o sólo semántica a la izquierda son los que más se dan. Con semántica a ambos lados del patrón son muchos menos.

Son 1.112 patrones los que se han generado con este escenario y de ellos tan sólo 272 llevan semántica. Cada uno de ellos adquiere diferente semántica.

9.9 Escenario 8

Características del escenario:

- Muestra de 379 patentes procedentes de OEP.

- Teniendo en cuenta todas las categorías gramaticales.
- Utilizando un mínimo de frecuencia de 100 para crear patrones.
- No activamos el checkbox para no diferenciar patrones por su semántica.

Resultados generales:

- ✓ Patrones básicos creados: 2.050.851 (tabla BasicPattern)
- ✓ Patrones: 1.137 (tabla Pattern)
  - Formados por dos TermTag: 237
  - Formados por patrones: 348
  - Mixtos: 552
- ✓ Patrones con semántica: 1.428 (tabla SemanticsBelongPatterns)

Este es el tercero y último escenario creado con las muestras de patentes europeas, al igual que en el resto de escenarios, se van a tener en cuenta todas las categorías gramaticales para que el estudio en conjunto de todos los escenarios pueda hacerse.

El mínimo de frecuencia elegido es 100, el máximo disponible, y no vamos a diferenciar patrones por su semántica, por lo que dos parejas de términos crearán un patrón cuándo aparezcan mínimo 100 veces entre los documentos. No tendrán que cumplir la misma semántica para ser el mismo patrón. Esto es lo que le diferencia del escenario 6.

Tampoco se elige la opción de guardar los boilerplates en la base de datos de CAKE, puesto que estamos manejando un gran número de documentos y esto retrasaría la ejecución del programa.

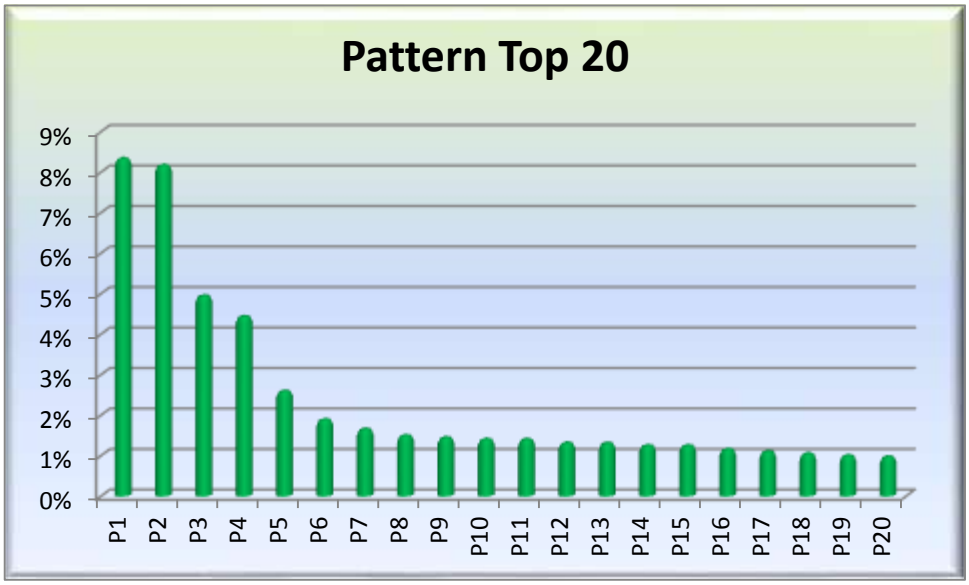
Es la ejecución más restrictiva que en los anteriores escenarios, la ejecución de este escenario ha necesitado 37 horas para finalizar.

9.9.1 Patrones

En la tabla Pattern de la base de datos RequirementsClassification, tenemos 1.137 patrones diferentes para este escenario y un total de 1.355.087 repeticiones.

El patrón que más repeticiones tiene, sólo tiene un nivel de profundidad, tiene el nombre P1. Está formado por un nombre no clasificado a ambos lados. El patrón con nombre P2, casi tiene el mismo número de repeticiones que el P1 y está formado por un nombre a ambos lados.

Los patrones con más repeticiones se muestran en la siguiente gráfica. Se incluyen los 20 primeros que representan el 49% del total.



Gráfica 26. Escenario 8. Pattern Top 20

Tipos de patrones

A continuación se muestran las diferentes combinaciones que hay dentro de los patrones generados por la herramienta.

Los TermTag son mostrados por su categoría gramatical correspondiente para que los resultados sean comprensibles por todos.

En la siguiente tabla tenemos los patrones que están compuestos por dos TermTag, se muestran los veinte con mayor frecuencia de los 237 que hay de este tipo. El total de este tipo representa el 64,45% de repeticiones.

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN	113.225	8,36%
P2	NOUN	NOUN	110.953	8,19%
P3	ADJECTIVE	NOUN	67.286	4,97%
P4	ARTICLE	NOUN	60.372	4,46%
P5	UNCLASSIFIED NOUN	NOUN	35.352	2,61%
P6	NOUN	COMMA	25.874	1,91%
P7	UNCLASSIFIED NOUN	COMMA	22.628	1,67%
P10	UNCLASSIFIED NOUN	ArithmeticOperator	19.283	1,42%
P11	NUMBER	ArithmeticOperator	19.202	1,42%

Nombre Patrón	Categoría gramatical Izquierda	Categoría gramatical Derecha	Repeticiones	%
P12	NUMBER	COMMA	17.992	1,33%
P13	NUMBER	NOUN	17.912	1,32%
P15	UNCLASSIFIED NOUN	SYMBOL	17.176	1,27%
P16	VERB TO BE	VERB	15.881	1,17%
P17	VERB	NOUN	15.164	1,12%
P18	NOUN	ArithmeticOperator	14.393	1,06%
P20	VERB TO BE	ADJECTIVE	13.374	0,99%
P23	NOUN	PREPOSITION OF	10.546	0,78%
P24	UNCLASSIFIED NOUN	CLOSING ROUND BRACKETS	10.436	0,77%
P25	NOUN	UNCLASSIFIED NOUN	9.441	0,70%
P26	NOUN	VERB	9.434	0,70%

Tabla 65. Escenario 8. Pattern – Top 20. TermTag + TermTag

Los patrones con subpatrones a ambos lados, en la siguiente tabla podemos ver los veinte que más se repiten del total 234 combinaciones obtenidas, el total representa el 9,35% de las repeticiones:

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P9	P1	P1	19.764	1,46%
P56	P1	P5	3.717	0,27%
P77	P6	P6	2.445	0,18%
P92	P7	P7	2.044	0,15%
P94	P12	P12	2.006	0,15%
P99	P9	P9	1.874	0,14%
P100	P5	P1	1.809	0,13%
P110	P19	P4	1.613	0,12%
P121	P9	P1	1.442	0,11%
P128	P11	P10	1.356	0,10%
P133	P1	P25	1.276	0,09%
P135	P33	P24	1.258	0,09%
P137	P33	P27	1.245	0,09%
P144	P4	P3	1.197	0,09%
P148	P5	P5	1.172	0,09%
P152	P2	P2	1.127	0,08%
P163	P11	P7	1.043	0,08%
P166	P19	P2	1.024	0,08%
P177	P1	P2	940	0,07%

Nombre Patrón	Patrón Izquierda	Patrón Derecha	Repeticiones	%
P179	P12	P11	934	0,07%

Tabla 66. Escenario 8. Pattern – Top 20. Patrón + Patrón

Patrones con subpatrón a la izquierda y TermTag a la derecha. Para éstos casos se han obtenido 273 patrones diferentes, suponen el 11,46% del total de repeticiones. En la siguiente tabla se muestran los veinte con mayor número de repeticiones:

Nombre Patrón	Patrón Izquierda	Categoría gramatical Derecha	Repeticiones	%
P19	P4	PREPOSITION OF	13.769	1,02%
P35	P4	VERB	6.273	0,46%
P36	P1	NOUN	5.925	0,44%
P43	P2	NOUN	4.842	0,36%
P45	P2	COMMA	4.448	0,33%
P48	P1	UNCLASSIFIED NOUN	4.199	0,31%
P57	P1	COMMA	3.581	0,26%
P58	P3	COMMA	3.555	0,26%
P64	P5	COMMA	2.909	0,21%
P66	P7	NOUN	2.874	0,21%
P74	P16	PREPOSITION	2.540	0,19%
P81	P12	UNCLASSIFIED NOUN	2.283	0,17%
P83	P4	COMMA	2.267	0,17%
P85	P3	VERB	2.237	0,17%
P103	P3	PREPOSITION OF	1.748	0,13%
P106	P6	NOUN	1.699	0,13%
P108	P8	PREPOSITION OF	1.634	0,12%
P109	P20	PREPOSITION	1.630	0,12%
P115	P2	VERB	1.543	0,11%
P117	P8	VERB	1.523	0,11%

Tabla 67. Escenario 8. Pattern – Top 20. Patrón + TermTag

Se encuentran 279 patrones diferentes que están formados por TermTag a la derecha y por subpatrón a la izquierda. Suponen un 14,74% del total de repeticiones. Mostramos los veinte que más se repiten:

Nombre Patrón	Categoría gramatical Izquierda	Patrón Derecha	Repeticiones	%
P8	ARTICLE	P3	20.508	1,51%
P14	ARTICLE	P2	17.186	1,27%
P21	SYMBOL	P15	12.724	0,94%
P22	ADJECTIVE	P2	12.347	0,91%
P29	UNCLASSIFIED NOUN	P2	7.668	0,57%
P31	NOUN	P3	7.125	0,53%
P46	OPENING ROUND BRACKETS	P27	4.334	0,32%
P55	PREPOSITION OF	P4	3.721	0,27%
P59	OPENING ROUND BRACKETS	P24	3.412	0,25%
P60	ARTICLE	P22	3.257	0,24%
P65	PREPOSITION	P2	2.903	0,21%
P70	PREPOSITION OF	P2	2.657	0,20%
P76	MODAL VERB	P20	2.454	0,18%
P78	PREPOSITION OF	P8	2.403	0,18%
P80	ARTICLE	P5	2.296	0,17%
P87	NUMBER	P2	2.154	0,16%
P90	AND LINKING	P2	2.082	0,15%
P91	ArithmeticOperator	P2	2.067	0,15%
P104	VERB	P2	1.743	0,13%
P107	PREPOSITION TO	P17	1.665	0,12%

Tabla 68. Escenario 8. Pattern – Top 20. TermTag + Patrón

Con esta información, observamos que los patrones formados por dos TermTag tienen mucha mayor presencia en los documentos, a pesar de que el número de patrones diferentes de este tipo es menor al resto.

	Nº Patrones	Repeticiones
TermTag + TermTag	237	64,45%
Patrón + Patrón	348	9,35%
Patrón + TermTag	273	11,46%
TermTag + Patrón	279	14,74%
TOTAL	**1.137**	**100%**

Tabla 69. Escenario 8. Pattern. Repeticiones de los diferentes tipos

El tipo de patrón que está compuesto sólo por patrón a la izquierda, supone el mayor número de patrones diferentes, y los compuestos por patrones a ambos lados tienen el menor peso en cuestión de número de repeticiones totales.

Longitud de patrones

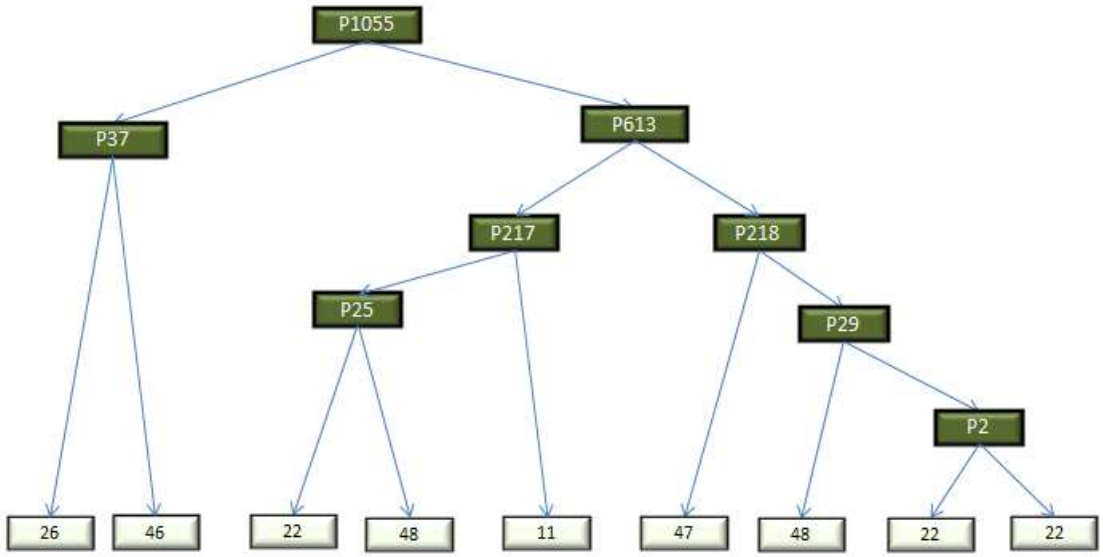
Para conocer el patrón más largo generado en la tabla Patterns de BoilerPlates, nos hemos creado un código en Shell Script para realizar las sustituciones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón. En el Anexo II se puede ver la guía del script.

El patrón más largo que se ha generado en este escenario es P955, está compuesto por 28 TermTag, pero muchos de ellos son nombres no clasificados, símbolos y números. Aun siendo un patrón finito, no parece representar un oración con sentido, parece ser un patrón de funciones.

P955= + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL + UNCLASSIFIED NOUN + RECOVERABLE PRONOUN + NUMBER + ARTICLE + SYMBOL + UNCLASSIFIED NOUN + SYMBOL

Como las oraciones en inglés tienen que estar formadas, al menos, por un nombre y un verbo, teniendo en cuenta éstas características, encontramos el siguiente patrón más largo que lo cumple.

En la siguiente gráfica mostramos un ejemplo con el patrón P999, formado por 12 TermTag gramaticales:



Gráfica 27. Escenario 8. Patrón más largo

Donde el resultado de toda la descomposición del patrón es:

P1055= NUMBER + CLOSING ROUND BRACKETS + NOUN + UNCLASSIFIED NOUN + VERB + OPENING ROUND BRACKETS + UNCLASSIFIED NOUN + NOUN + NOUN

9.9.2 Patrones con semántica

La semántica perteneciente a los patrones se puede ver en la tabla SemanticsBelongPatterns de BoilerPlates.

Son 1.428 patrones los que tienen semántica, entre ellos hay muchos patrones repetidos pero con semántica diferente, nombres de patrones únicos hay tan sólo 260. El patrón con semántica que tiene mayor presencia y con semántica diferente, es el P3, es el tercer patrón con mayor frecuencia de toda la muestra. Está formado por un adjetivo a la izquierda y un nombre a la derecha.

Las diferentes semánticas que se encuentran entre los documentos para éste patrón son 10 tipos:

Patrón	SemanticLeft	SemanticRight
P3	SYSTEM FUNCTION	NO SEMANTICA
P3	Sustainability	NO SEMANTICA
P3	Operation	NO SEMANTICA
P3	AVAILABILITY	NO SEMANTICA
P3	RANGE = EQUAL	NO SEMANTICA

Patrón	SemanticLeft	SemanticRight
P3	RANGE < MAXIMUM	NO SEMANTICA
P3	RANGE <= (MAXIMUM)	NO SEMANTICA
P3	RANGE >= MINIMUM	NO SEMANTICA
P3	RANGE > MINIMUM	NO SEMANTICA

Tabla 70. Escenario 8. Semántica. Patrón más repetido con semántica

El adjetivo que forma el termtag de la izquierda puede adquirir la diferente semántica mostrada.

Como éste hay muchos patrones repetidos, al tener el mismo patrón tantas repeticiones no conocemos la frecuencia de aparición de cada uno de ellos entre los documentos.

El número de patrones (no únicos) con semántica a la izquierda y a la derecha es el siguiente:

SemanticLeft	SemanticRight	Nº Patrones
SI	SI	109
NO	SI	655
SI	NO	664
TOTAL		**1428**

Tabla 71. Escenario 8. Totales patrones con semántica

Se puede ver que sólo semántica a la izquierda o sólo semántica a la derecha son los que más se dan. Con semántica a ambos lados del patrón son muchos menos.

Son 1.137 patrones los que se han generado con este escenario y de ellos tan sólo 260 llevan semántica. Cada uno de ellos adquiere diferente semántica.

10. Conclusiones

10.1 Tiempos de ejecución

Tras llevar a cabo todos los escenarios, podemos concluir que los escenarios cuanto mayor es el número de frecuencia elegido, menor es el tiempo de ejecución de la herramienta BP.

Nº de doc.	Tamaño doc.	Tokens	Tamaño BD	Frec.	Semántica	FINALIZA	Duración horas	Tamaño BD al terminar
359	10,3 MB	2.089.157	106 MB	1	NO	FINALIZA	540	243 MB
359	10,3 MB	2.089.157	106 MB	1	SI	EN EJECUCION	600	
359	10,3 MB	2.089.157	106 MB	20	SI	FINALIZA	73	241 MB
359	10,3 MB	2.089.157	106 MB	100	SI	FINALIZA	27,5	241 MB
359	10,3 MB	2.089.157	106 MB	100	NO	FINALIZA	37	241 MB
379	19,5 MB	3.716.754	189 MB	20	SI	SE CANCELA	8	423 MB
379	19,5 MB	3.716.754	189 MB	100	SI	SE CANCELA	6	423 MB
100	10,4 MB	2.050.851	107 MB	20	SI	FINALIZA	86,5	325 MB
100	10,4 MB	2.050.851	107 MB	100	SI	FINALIZA	66	325 MB
100	10,4 MB	2.050.851	107 MB	100	NO	FINALIZA	37	325 MB

Tabla 72. Tiempos empleados en la ejecución BP.

Como se muestra en la tabla anterior, tenemos dos ejecuciones fallidas para la muestra europea, el volumen de Tokens que contiene es tan elevado que los cálculos son demasiado pesados. Por este motivo eliminamos documentos de la muestra hasta quedarnos con un volumen de Tokens similar a la muestra estadounidense. De este modo, al equilibrar las condiciones, las comparaciones entre ambas muestras estarán más equilibradas.

La frecuencia 1 con diferenciación de semántica, lleva más de 25 días y sigue en ejecución. No podemos obtener resultados parciales puesto que la herramienta no guarda los patrones hasta que no finaliza todo el cálculo.

Vemos que la ejecución más corta es la realizada para la muestra estadounidense con frecuencia 1 y diferenciación de semántica.

10.2 Patrones básicos

En ambas muestras las categorías gramaticales obtenidas son muy similares. En las patentes europeas se han obtenido más termtag sin clasificar, que revisando los TokenText afectados se debe a palabras incompletas, mal escritas o símbolos.

10.3 Patrones y semántica

10.3.1 Escenario 1 vs escenario 2

Ambos escenarios son generados con las muestras de patentes estadounidenses. Para los dos se ha elegido la frecuencia mínima de 1. Lo que les distingue a ambos es la diferenciación de patrones por semántica, que sólo está activado para el escenario 2.

Para el escenario 1 obtenemos los resultados en 22 días y medio. Para el escenario 2 no hemos podido obtener resultados, tras más de 25 días la ejecución continua, al salirse de la planificación los resultados no pueden obtenerse. Por lo que no tenemos datos para conocer la diferencia que habría al utilizar o no la semántica.

Como la semántica no se ha tenido en cuenta la generación de patrones en el escenario 1, tenemos que un mismo patrón toma sentido en más de una semántica diferente.

10.3.2 Escenario 3 vs escenario 5

Ambos escenarios son generados con las muestras de patentes estadounidenses. La frecuencia mínima seleccionada es de 20 y 100 respectivamente.

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 3	408	1.753	837	1.055	4.053
Escenario 5	161	279	270	317	1.027
Diferencia	247	1.474	567	738	3.026
Pérdida	**60,54%**	**84,08%**	**67,74%**	**69,95%**	**74,66%**

Tabla 73. Escenario 3 vs escenario 5. Número de patrones.

Como era de esperar, se obtienen más patrones en el escenario 3 que en el escenario 5, esto es porque el escenario 3 es mucho más restrictivo en cuanto a la frecuencia mínima que tienen que cumplir los patrones.

Si en lugar de fijarnos en el número de repeticiones nos fijamos en la frecuencia que tenían los patrones, la pérdida es mucho menor.

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 3	30,60%	17,73%	26,62%	25,06%	1.011.917
Escenario 5	33,81%	13,51%	27,66%	25,01%	883.521
Diferencia	-3,21%	4,22%	-1,04%	0,05%	128.396
Pérdida					**12,69%**

Tabla 74. Escenario 3 vs escenario 5. Frecuencia patrones.

Todos los patrones que están en el escenario 5 están en el escenario 3. Los que están en el los dos escenarios tienen la misma frecuencia.

Los patrones que están en el escenario 3 y que no están en el escenario 5, son todos aquellos que su número de repeticiones está por debajo de 100. Éstos representan un 12,69% como muestra la tabla anterior.

Nos indica que los patrones en la muestra de documentos estadounidenses tienen muchas repeticiones. Al quedarnos con los que superan las 100 repeticiones, lo que más se pierde son los patrones que están formados por otros dos patrones y aumentan los formados por dos categorías gramaticales.

El patrón que más se repite es el mismo en los dos escenarios, es el formado por ADJETIVO + NOMBRE

En ambos escenarios tenemos que el patrón más largo es de sustitución infinita, por lo que no lo podemos considerar ninguno de los dos válidos. El segundo más largo lo encontramos en el escenario 3:

NOUN + VERB TO BE + ADJECTIVE + PHRASAL VERB PARTICLE + ARTICLE + ADJECTIVE + NOUN + ABSOLUTE VERB + PHRASAL VERB PARTICLE + ARTICLE + ADJECTIVE + NOUN + ArithmeticOperator + OPENING ROUND BRACKETS + ARTICLE + ADJECTIVE + NOUN + PREPOSITION OF + NOUN + NOUN + VERB + AND LINKING + RELATIVE PRONOUN + ADJECTIVE + ADJECTIVE + NOUN + COMMA + UNCLASSIFIED NOUN + ADJECTIVE + NOUN + VERB + PREPOSITION

Forma una oración de 32 termtag gramaticales.

El número de patrones con semántica es mucho menor en el escenario más restrictivo, supone un 30% menos de patrones con semántica. Pero que si miramos cuanto supone en el total de repeticiones de patrones en toda la muestra, tan sólo es un 0,66% lo que se obtiene de menos en el escenario 5.

Semantic Left	SemanticRight	Escenario 3		Escenario 5		Diferencia	
		Nº Patrones	% Repeticiones	Nº Patrones	% Repeticiones	Nº Patrones	% Repeticiones
SI	**SI**	17	0,23%	6	0,21%	11	0,02%
NO	**SI**	298	4,26%	77	3,65%	221	0,61%
SI	**NO**	363	8,89%	122	8,63%	241	0,26%
TOTAL		**678**	**13,15%**	**205**	**12,49%**	**473**	**0,66%**

Tabla 75. Escenario 3 vs escenario 5. Semántica

10.3.3 Escenario 4 vs escenario 6

Ambos escenarios son generados con las muestras de patentes europeas. La frecuencia mínima seleccionada es de 20 y 100 respectivamente.

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 4	550	2.062	952	1.017	4.581
Escenario 6	267	338	278	287	1.170
Diferencia	283	1.724	674	730	3.411
Pérdida	**51,45%**	**83,61%**	**70,80%**	**71,78%**	**74,46%**

Tabla 76. Escenario 4 vs escenario 6. Número de patrones.

Como era de esperar, se obtienen más patrones en el escenario 4 que en el escenario 6, esto es porque el escenario 6 es mucho más restrictivo en cuanto a la frecuencia mínima que tienen que cumplir los patrones.

Si en lugar de fijarnos en el número de repeticiones nos fijamos en la frecuencia que tenían los patrones, la pérdida es mucho menor.

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 4	59,68%	12,58%	12,66%	15,08%	1.518.865
Escenario 6	64,94%	8,90%	11,80%	14,36%	1.375.981
Diferencia	-5,26%	3,68%	0,86%	0,72%	142.884
Pérdida					**9,41%**

Tabla 77. Escenario 4 vs escenario 6. Frecuencia patrones.

Todos los patrones que están en el escenario 6 están en el escenario 4. Los que están en el los dos escenarios tienen la misma frecuencia.

Los patrones que están en el escenario 4 y que no están en el escenario 6, son todos aquellos que su número de repeticiones está por debajo de 100. Éstos representan un 9,41% como muestra la tabla anterior.

Nos indica que los patrones en la muestra de documentos europeas tienen muchas repeticiones. Al quedarnos con los que superan las 100 repeticiones, lo que más se pierde son los patrones que están formados por otros dos patrones y aumentan los formados por dos categorías gramaticales.

El patrón que más se repite es el mismo en los dos escenarios y está formado por dos nombres no clasificados, el siguiente con mayor frecuencia en ambos escenarios es el formado por dos nombres.

Encontramos el patrón más largo en el escenario 4 que forma una oración de 56 termtag gramaticales, pero no parece ser un patrón que pueda servir de guía, carece de verbo y nombre. Seleccionamos el siguiente que está formado por 27 termtag y también pertenece al mismo escenario.

PARTICLE + NOUN + NOUN + VERB TO HAVE + VERB TO BE + VERB + PREPOSITION + VERB + ARTICLE + NOUN + COMMA + NOUN + OR LINKING + NOUN + MODAL VERB + NEGATION + VERB TO BE + VERB + AND LINKING + ARTICLE + UNCLASSIFIED NOUN + VERB + QUANTIFIER DETERMINER + NOUN + PREPOSITION + UNCLASSIFIED NOUN + VERB

El número de patrones con semántica es mucho menor en el escenario más restrictivo, supone un 23% menos de patrones con semántica. Pero que si miramos cuanto supone en el total de repeticiones de patrones en toda la muestra, tan sólo es un 0,55% lo que se obtiene de menos en el escenario 6.

Semantic Left	Semantic Right	Escenario 4		Escenario 6		Diferencia	
		Nº Patrones	% Rep	Nº Patrones	% Rep	Nº Patrones	% Rep
SI	**SI**	8	0,04%	2	0,03%	6	0,01%
NO	**SI**	145	1,12%	30	0,88%	115	0,24%
SI	**NO**	215	1,93%	54	1,63%	161	0,30%
TOTAL		**368**	**3,09%**	**86**	**2,54%**	282	0,55%

Tabla 78. Escenario 4 vs escenario 6. Semántica

Podemos concluir que a mayor frecuencia obtenemos menos semántica para éstos escenarios.

10.3.4 Escenario 5 vs escenario 7

Ambos escenarios son generados con las muestras de patentes estadounidenses. La frecuencia mínima seleccionada es de 100. Lo que les diferencia es que para el escenario 5 se ha tenido en cuenta la semántica mientras que para el escenario 7 no se ha tenido en cuenta:

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 5	161	279	270	317	1.027

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 7	233	305	247	327	1.112
Diferencia	-72	-26	23	-10	-85
Pérdida	**-44,72%**	**-9,32%**	**8,52%**	**-3,15%**	**-8,28%**

Tabla 79. Escenario 5 vs escenario 7. Número de patrones.

Existe muy poca diferencia en cuanto al número de patrones que se obtienen en ambos escenarios. Se obtienen un poco más de patrones en el escenario 7 que en el escenario 5.

Si comparamos los escenarios por su número de repeticiones, vemos que el número de repeticiones obtenido en el escenario 7 es casi un 50% más que en el escenario 5.

Los patrones del escenario 7 tienen menos niveles de subpatrones, hay más presencia de patrones formados directamente por dos categorías gramaticales:

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	Repeticiones
Escenario 5	33,81%	13,51%	27,66%	25,01%	883.521
Escenario 7	62,31%	7,43%	13,83%	16,44%	1.314.586
Diferencia	-28,50%	6,08%	13,83%	8,57%	-431.065
Pérdida					**-48,79%**

Tabla 80. Escenario 5 vs escenario 7. Frecuencia patrones.

La fila de la tabla “Diferencia” muestra lo que tiene el escenario 5 menos el escenario 7.

Nos indica que los patrones generados sin diferenciación de semántica, pierden presencia de subpatrones y aumentan los que se forman por dos categorías gramaticales.

El patrón que más se repite es diferentes en los dos escenario, en ambos está formado por dos termtag gramaticales.

- ✓ Con gramática tenemos el patrón compuesto por ADJETIVO + NOMBRE
- ✓ Sin gramática tenemos el patrón compuesto por NOMBRE + NOMBRE

Teniendo en cuenta que las oraciones en inglés tienen que estar formadas, al menos, por un nombre y un verbo, encontramos en el escenario 5 el patrón más largo, con nueve TermTag gramaticales:

ARTICLE + ADJECTIVE + NOUN + PREPOSITION OF + VERB TO BE + VERB + ARTICLE + ADJECTIVE + NOUN

El número de patrones con semántica es mucho menor cuando se utiliza la diferenciación por semántica, pero revisando los patrones a los que se les ha asignado semántica al escenario 7, nos encontramos que tan sólo hay 128 patrones únicos con semántica. Esto se debe a que un mismo patrón puede contener semánticas diferentes.

Al hacer la diferenciación de semántica cuando se generan los patrones, obtenemos un patrón nuevo cuando la semántica cambia.

SemanticLeft	SemanticRight	Escenario 5	Escenario 7	Diferencia
		Nº Patrones	Nº Patrones	Nº Patrones
SI	**SI**	6	141	-135
NO	**SI**	77	865	-788
SI	**NO**	122	746	-624
TOTAL		**205**	**1.752**	**-1.547**

Tabla 81. Escenario 5 vs escenario 7. Semántica

10.3.5 Escenario 6 vs escenario 8

Ambos escenarios son generados con las muestras de patentes europeas. La frecuencia mínima seleccionada es de 100. Lo que les diferencia es que para el escenario 6 se ha tenido en cuenta la semántica y para el escenario 8 no se ha tenido en cuenta.

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	TOTAL
Escenario 6	267	338	278	287	1.170
Escenario 8	237	348	273	279	1.137
Diferencia	30	-10	5	8	33
Pérdida	**11,24%**	**-2,96%**	**1,80%**	**2,79%**	**2,82%**

Tabla 82. Escenario 6 vs escenario 8. Número de patrones.

En la tabla anterior se puede ver que existe muy poca diferencia en cuanto al número de patrones que se obtienen en ambos escenarios. Se obtienen un poco más de patrones en el escenario 6 que en el escenario 8.

Para la muestra europea, si se hace diferenciación de semántica se obtiene mayor número de patrones.

Si comparamos los escenarios por su número de repeticiones, vemos que el número de repeticiones obtenido es muy similar en ambos.

Los patrones del escenario 8 tienen menos niveles de subpatrones, hay más presencia de patrones formados directamente por dos categorías gramaticales y menos por los formados por dos subpatrones:

	TermTag + TermTag	Patrón + Patrón	Patrón + TermTag	TermTag + Patrón	Repeticiones
Escenario 6	64,94%	8,90%	11,80%	14,36%	1.375.981
Escenario 8	64,45%	9,35%	11,46%	14,74%	1.355.087
Diferencia	0,49%	-0,45%	0,34%	-0,38%	20.894
Pérdida					**1,52%**

Tabla 83. Escenario 6 vs escenario 8. Frecuencia patrones.

La fila de la tabla "Diferencia" muestra lo que tiene el escenario 6 menos lo que tiene el escenario 8.

Nos indica que los patrones generados sin diferenciación de semántica, pierden presencia de subpatrones y aumentan los que se forman por dos categorías gramaticales.

El patrón que más se repite es el mismo en los dos escenarios y está formado por dos nombres no clasificados, el siguiente con mayor frecuencia en ambos escenarios es el formado por dos nombres.

Teniendo en cuenta que las oraciones en inglés tienen que estar formadas, al menos, por un nombre y un verbo, encontramos el mismo patrón en ambos escenarios, los nombres son diferentes, pero los TermTag que los forman son iguales.

NUMBER + CLOSING ROUND BRACKETS + NOUN + UNCLASSIFIED
NOUN + VERB + OPENING ROUND BRACKETS + UNCLASSIFIED
NOUN + NOUN + NOUN

El número de patrones con semántica es mucho menor cuando se utiliza la diferenciación por semántica. En el escenario 8, para 127 patrones nos está dando 1.428 semánticas diferentes. Esto se debe a que un mismo patrón puede contener semánticas diferentes.

Al hacer la diferenciación de semántica cuando se generan los patrones, obtenemos un patrón nuevo cuando la semántica cambia.

SemanticLeft	SemanticRight	Escenario 6	Escenario 8	Diferencia
		Nº Patrones	Nº Patrones	Nº Patrones
SI	**SI**	2	109	-107
NO	**SI**	30	655	-625
SI	**NO**	54	664	-610
TOTAL		**86**	**1.428**	**-1.342**

Tabla 84. Escenario 6 vs escenario 8. Semántica

10.3.6 Conclusiones escenarios 1, 2, 3, 5 y 7

Los escenario 1, 2, 3, 5 y 7 son todos realizados con los documentos de patentes estadounidenses.

Para el escenario 2 no se han obtenido resultados por lo que explicábamos anteriormente.

Para el resto, podemos concluir para la muestra estadounidense se cumple lo siguiente:

- ✓ A mayor frecuencia mínima, menor número de patrones
- ✓ A mayor frecuencia mínima, menor es la semántica obtenida.
- ✓ Diferenciar los patrones por semántica es una mejor práctica, para conocer la semántica real que se está utilizando al escribir oraciones. De no ser así, para un mismo patrón, la semántica que puede adoptar podría ser cualquiera.
- ✓ No diferenciar por semántica da como resultado mayor número de patrones, pero con menor número de subpatrones que lo forman.

10.3.7 Conclusiones escenarios 4, 6 y 8

Los escenarios 4, 6 y 8 son realizados con documentos de la muestra de patentes europeas.

Podemos concluir, para la muestra europea se cumple lo siguiente:

- ✓ A mayor frecuencia mínima, menor número de patrones
- ✓ A mayor frecuencia mínima, menor es la semántica obtenida.

- ✓ Diferenciar los patrones por semántica es una mejor práctica, para conocer la semántica real que se está utilizando al escribir oraciones. De no ser así, para un mismo patrón, la semántica que puede adoptar podría ser cualquiera.
- ✓ No diferenciar por semántica no da como resultado mayor número de patrones, y el número de subpatrones es muy similar.

10.4 Conclusiones generales

Tras el análisis realizado de documentos de patentes estadounidense y de patentes europeas podemos concluir lo siguiente:

- ✓ Los patrones básicos que se obtienen son independientes a la frecuencia y a la selección de categorías gramaticales en la herramienta BoilerPlates. Todos los patrones básicos son comunes dentro de la misma muestra.

- ✓ Cuanto mayor es la frecuencia utilizada en la herramienta BoilerPlates, menor es el número de patrones que se obtienen y menor es el tiempo necesario para obtenerlos.

- ✓ Se ha hecho diferenciación de patrones por su semántica en las frecuencias mínimas de 1, 20 y 100 para las muestras estadounidenses, y de 20 y 100 para las muestras europeas. Para la frecuencia 1 no ha sido posible obtener resultados debido al gran volumen de información que hemos manejado. Tras más de 25 días en ejecución la herramienta, se ha tenido que desestimar la frecuencia 1 para el estudio. Sobre las otras dos frecuencias, podemos decir que a mayor frecuencia el número de patrones obtenidos es menor.

- ✓ Se calculan patrones sin hacer diferenciación de semántica para las frecuencias mínimas de 1 y 100 con la muestra estadounidense. También se calcula con la muestra europea para la frecuencia mínima de 100, sin hacer diferenciación de patrones por su semántica. Se puede concluir que se obtienen patrones iguales con diferentes semánticas.

- ✓ Al aumentar la frecuencia perdemos los patrones que tienen mayor profundidad de descomposición. Puesto que su número de repeticiones es menor. El patrón con mayor profundidad puede verse en la gráfica 17 del escenario 3.

- ✓ Tras utilizar diferentes frecuencias para generar patrones en BoilerPlates, podemos decir que la frecuencia intermedia es la que nos ha dado unos mejores resultados.

- ✓ En ambas muestras está muy presente los nombres sin clasificar.

- ✓ Los patrones obtenidos en todos los escenarios pueden resultar de ayuda a aquellas personas que necesiten redactar una patente.

11. Recomendaciones

Tras la investigación realizada, con el conocimiento ahora adquirido, podemos dar unas recomendaciones a quien se enfrente en un futuro a un estudio similar.

- ✓ La ontología puede ser mejorada, tiene 73 etiquetas para categorías gramaticales pendientes por definir su vocabulario. Para este proyecto no se han completado todas porque las palabras más importantes quedan cubiertas. Las gramáticas pendientes de definir son del tipo de signos de puntuación, fechas, email, símbolos aritméticos, acrónimos, etc. Estas categorías sin definir pueden verse en la tabla 8.

- ✓ Han sido muchos los token los que se han clasificado bajo la etiqueta “UNCLASSIFIED NOUN”. Para estos casos vemos tres planes de acción:

    - Se podrían analizar todas ellas y darlas una categoría gramatical si fuera posible, así la búsqueda de patrones sería más precisa.
    - Si no fuera posible asignarles una categoría más concreta, habría que mirar la posibilidad de eliminar todas las palabras y símbolos no clasificables.
    - A la hora de generar los patrones con la herramienta BoilerPlates, no considerar la etiqueta “UNCLASSIFIED NOUN” en el cálculo.
- ✓ Los documentos que se han utilizado en este análisis pueden ser mejorados en la conversión realizada de PDF a TXT, en este

proceso se ha perdido información. Los documentos con imágenes son los que más información han perdido.

- ✓ Es posible realizar el análisis con frecuencia mayor a 100, puesto que hemos obtenido patrones donde su frecuencia de repetición es mayor a 100. Pero antes de realizar estudios con una frecuencia mínima mayor, se recomienda no tener en cuenta las palabras que no correspondan a una gramática de la ontología.

12. Bibliografía

[1] Antecedentes teóricos de J. Technol. Manag. Innov. 2013, Volume 8, Special Issue ALTEC. http://www.scielo.cl/pdf/jotmi/v8s1/art65.pdf (último acceso 6/10/2015)

[2] Biografía Alan Mathison Turing; http://www.biografiasyvidas.com/biografia/t/turing.htm (último acceso 6/10/2015)

[3] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993. http://tomgruber.org/writing/ontolingua-kaj-1993.htm (último acceso 6/10/2015)

[4] Biografía Tomas Robert Gruber. http://tomgruber.org/bio/short-bio.htm (último acceso 6/10/2015)

[5] Definición de WordNet extraída de Wikipedia https://es.wikipedia.org/wiki/WordNet (último acceso 22/05/2015)

[6] Datos sobre USPTO. https://es.wikipedia.org/wiki/Oficina_de_Patentes_y_Marcas_de_Estados_Unidos (último acceso 6/10/2015)

[7] Clasificación internacional de Patentes (OMPI). http://patentscope.wipo.int (último acceso 28/02/2015)

[8] Oficina de Patentes y Marcas Registradas de Estados Unidos (USPTO). http://www.uspto.gov/trademarks/index.jsp (último acceso 10/09/2015)

[9] Oficina Europea de Patentes (OEP). http://www.epo.org/searching.html (último acceso 10/09/2015)

[10] Instituto Mexicano de la Propiedad Industrial (IMPI). http://siga.impi.gob.mx/ (último acceso 28/02/2015)

[11] Oficina Española de Patentes y Marcas (OEPM). http://www.oepm.es/es/invenciones (último acceso 28/02/2015)

[12] Japan Patent Office (JPO). http://www19.ipdl.inpit.go.jp/PA1/cgi-bin/PA1INIT?1337272515265 (último acceso 28/02/2015)

[13] Korean Intellectual Property Rights Information Service (KIPRIS). http://engpat.kipris.or.kr/engpat/searchLogina.do?next=MainSearch (último acceso 28/02/2015)

[14] State Intellectual Property Office of the P.R.C. (SIPO). http://59.151.93.237/sipo_EN/search/tabSearch.do?method=init (último acceso 28/02/2015)

[15] Buscador de patentes en google http://www.google.com/patents (último acceso 10/09/2015)

[16] A. Gelbukh, G. Sidorov. Procesamiento automático del español con enfoque en recursos léxicos grandes. Segunda edición, ampliada y revisada. IPN, 2010.

[17] S. N. Galicia Haro, A. Gelbukh. Investigaciones en análisis sintáctico para el español. IPN, 2007.

[18] Procesamiento de lenguajes naturales. https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales (último acceso 6/10/2015)

[19] Análisis léxico. Fundación del analizador léxico. http://www.galeon.com/shock/tareas.html (último acceso 6/10/2015)

[20] Análisis semántico http://es.slideshare.net/pepgonzalez/analisis-semantico (último acceso 6/10/2015)

[21] Análisis léxico y análisis sintáctico http://es.slideshare.net/angiepao1717/anlisis-lxico-y-anlisis-sintctico (último acceso 6/10/2015)

[22] Análisis léxico por Francisco José Moreno y Gonzalo Aranda de la universidad de Huelva en el temario de Procesadores de lenguajes. 2014/2015. http://www.uhu.es/francisco.moreno/gii_pl/doc s/Tema_2.pdf (último acceso 6/10/2015)

[23] A.V. Aho, R. Sethi, J.D. Ullman, "Compiladores: principios, técnicas y herramientas", Editorial Addison-Wesley, 1990.

[24] Zhifeng Yang (2002). "Applying Information Retrieval Technology to Incremental Knowledge Management". En: Engineering and Deployment of Cooperative Information Systems: First International Conference, EDCIS 2002, Beijing, China, September 17–20, 2002 : Proceedings. Yanbo Han (Red.) p.117-120

[25] Peñas Padilla A. Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe: Ambigüedad, variación terminológica y multilingüismo. Sociedad española para el Procesamiento del Lenguaje Natural 2003.

[26] A Gelbukh - 2010 - gelbukh.com. Procesamiento de lenguaje natural. http://www.gelbukh.com/CV/Publications/2010/Procesamiento%20de%20lenguaje%20natural%20-%20BORRADOR.pdf (último acceso 6/10/2015)

[27] Procesamiento del lenguaje natural por Juan Manuel Rodriguez. 2012 http://pdln.blogspot.com.es/2012/10/normalizacion-del-texto-tokenizacion.html (último acceso 6/10/2015)

[28] Oraciones y gramática en inglés http://www.scientificpsychic.com/grammar/gramatica-inglesa-1.html (último acceso 17/10/2015)

Anexo I. Conversor de PDF a TXT

NOMBRE DEL CONVERSOR: pdf2txt

AUTOR: David Catalán

CODIGO FUENTE

El código fuente se encuentra en la carpeta src. Consta de dos clases:

- PdfConverter: Con la clase propia. Tiene dos métodos de Parseado, el Estándar y el adaptado para el analizador semántico. En la versión incluida, se encuentra activa la versión adaptada y comentada la standard.
- PDFTextStripper_Own : Adaptación de la clase de pdfbox, para una mejor separación por párrafos

PAQUETES JAVA

- Librerías libres de APACHE descargadas
    - commons-logging-1.2.jar
    - fontbox-1.8.8.jar
    - pdfbox-1.8.8.jar
- pdf2txt.jar : Librería resultante de la compilación del código fuente arriba indicado

COMPILACION

Permite generar el pdf2txt.jar

- Eclipse: El paquete está preparado para ser importado en eclipse de forma automática. La compilación es en tiempo de ejecución
- Windows (Línea de comandos) : Script compile.cmd
- Unix: Script compile.sh

EJECUCION

Son necesarios 3 parámetros:

- Directorio de entrada : Que contendrá todos los ficheros pdf
- Directorio de salida : Donde se guardarán los txt
- Número de palabras. Es el número de palabras mínimo para considerar que un párrafo debe ser considerado válido. El resto, se descartan. Para la versión standard este parámetro no se usa

Scripts de ejecución

- Eclipse : Se incorpora el launcher “PdfConverter.launch”
- Windows (Línea de comandos) : Script pdf2txt.cmd
- Unix: Script pdf2txt.sh

Anexo II. Script patrones.sh

Se ha realizado un código en Shell Script para Linux, con el objetivo de realizar las sustituciones de patrones y subpatrones automáticamente. Con éste script se realizan las sustituciones de TermTagOrPatternLeft y TermTagOrPatternRight por su TermTag o por su Pattern, se realiza de manera recursiva hasta conseguir la jerarquía completa de cada patrón.

El código puede ejecutarse desde un terminal Linux. Para su ejecución necesita tres parámetros de entrada, los tres parámetros tienen que ser un fichero plano.

El script tiene un algoritmo de sustitución que descifra la jerarquía completa de cada uno de los patrones recibidos. Tiene un control de recursividad infinita porque puede haber patrones que son sustituidos por si mismos de manera infinita.

Parámetro 1: Fichero plano que contenga el identificador del patrón y sus valores en el campo TermTagOrPatternLeft y en el campo TermTagOrPatternRight todo ello en una línea y con el siguiente formato:

|-1386|=|5|22|=
|-1387|=|60|-1386|=

Parámetro 2: Fichero plano que contenga por cada línea el nombre del patrón y sus valores en el campo TermTagOrPatternLeft y en el campo TermTagOrPatternRight. El formato tiene que ser:

P1387=|60|-1386|

Parámetro 3: Fichero plano que por línea tiene que contener el identificador y el nombre de cada gramática. El formato del fichero tiene que ser:

|5|=|ADJECTIVE|=
|22|=|NOUN|=
|60|=|ARTICLE|=

Resultado. El resultado se devuelve en el fichero recibido en el segundo parámetro. Con los ejemplos mostrados, el resultado tras la ejecución de script obtendremos:

P1387=|ARTICLE|ADJECTIVE|NOUN|

Anexo III. Categorías gramaticales en la ontología.

Familias	Total
NOUN	117.933
ADJECTIVE	21.529
VERB	11.485
ADVERB	4.465
TIME ADVERB	77
PREPOSITION	74
PLACE ADVERB	61
ABSOLUTE VERB	53
TIME ADVERBIAL PHRASE	32
PRONOUN	31
PARTICLE	27
QUANTIFIER DETERMINER	26
SYMBOL	19
PARTITIVE DETERMINER	16
PREPOSITIONAL LOCATION	13
PHRASAL VERB BASE	12
RECOVERABLE PRONOUN	12
NUMBER	11
PLACE ADVERBIAL PHRASE	10
MODAL VERB	9
PREPOSITIONAL LINKING PHRASE	9
PERSONAL PRONOUN	8
POSSESIVE PRONOUN	7
POSSESSIVE DETERMINER	7
REQUIREMENTS Domain	7
DEMONSTRATIVE DETERMINER	4
DETERMINER	4
RELATIVE PRONOUN	4
AND LINKING	3
ArithmeticOperator	3
ARTICLE	3
CAUSAL CONECTOR	3
MEASUREMENT UNIT	3
NEGATION	3
PHRASAL VERB ABSOLUTE BASE	3
CAUSE	2
PHRASAL VERB PARTICLE	2
PURPOSE/GOAL	2
UNCLASSIFIED ADJECTIVE	2
UTTERANCE DETERMINER	2
ABBREVIATION	1
CLOSING ROUND BRACKETS	1
COMMA	1
CONECTOR REQUIREMENT/CONDITION	1
OPENING ROUND BRACKETS	1
OR LINKING	1
PREPOSITION BY	1
PREPOSITION OF	1
PREPOSITION TO	1
VERB TO BE	1
VERB TO DO	1
VERB TO HAVE	1

Tabla 85. Anexo III. Ontología - Categorías gramaticales

Anexo IV. Acrónimos

WN: WordNet

KM: knowledge MANAGER

USPTO: Oficina de Patentes y Marcas Registradas de Estados Unidos (United States Patent and Trademark Office)

OEP: Oficina Europea de Patentes

PLN: Procesamiento del Lenguaje Natural.