Universidad Carlos III de Madrid
www.uc3m.es

PROYECTO FINAL DE CARRERA

Ingeniería Técnica Informática de Gestión

EVALUACIÓN DE UN SISTEMA DE PROCESAMIENTO DEL LENGUAJE NATURAL DE LA BANCA

Autor: Nuria de la O Maestro

Tutores: Anabel Fraga y Valentín Moreno

Leganés, Octubre 2015

Índice

Índice de imágenes

Índice de tablas

1. Agradecimientos

Para poder realizar este proyecto fin de carrera he contado con el apoyo de mi familia, amigos, compañeros de trabajo que me han mostrado su apoyo y su ayuda.

Quisiera dar las gracias especialmente a mis padres, hermana y pareja por su apoyo, aguantando mis malos días, irritabilidades, mis nervios….

A mi tía que gracias a su ayuda económica durante los años de carrera, han permitido que continuara estudiando.

En último lugar pero no menos importante a Sergio, mi hijo, por esos ratos en los que no hemos podido compartir juegos, cuentos, porque mamá tenía que trabajar.

También quisiera agradecer el apoyo de los tutores, Anabel y Valentín que me ofrecieron su ayuda y apoyo para poder realizar este trabajo con las tutorías presenciales y esos correos a alta horas. Y como no podía ser de otra forma a Eugenio Parra, quien diseñó la herramienta que se utiliza en el desarrollo de este proyecto y que constantemente nos ha dado su ayuda para que el proyecto fuera más fácil de realizar, incluso modificando la herramienta durante la realización de este proyecto.

2. Introducción

Actualmente existe mucha información referente en este caso al dominio de la banca. Por ello es necesario analizar como presentan la información a sus clientes, es decir, saber cómo está estructurada para poder organizarla correctamente, permitiendo a los usuarios realizar búsquedas posteriores más rápidas y eficaces.

Debido al hecho de que esta información se crea con una alta frecuencia es necesario analizar cómo se presenta la información a los usuarios. A nivel de la estructura y con la ayuda de gráficos de algunos de los resultados que se pueden encontrar son: grupos de palabras que se utilizan (para determinar si proceden de un vocabulario específico), categorías gramaticales más comunes, palabras más repetidas en un dominio, los patrones encontrados, la frecuencia de los patrones encontrados. Este proyecto ha sido creado ayudar aquellas personas que se dedican a redactar la información que se muestra a los usuarios en las distintas páginas web de los bancos.

Para el estudio se han obtenido más de doscientos archivos procedentes de diferentes bancos en formato pdf.

Con la herramienta llamada Boilerplates; proporcionada para este proyecto creada por Eugenio Parra y reutilizada por el Grupo de Conocimiento de la Universidad en distintos proyectos. Se procesarán los documentos obtenidos permitiéndonos evaluar el sistema de procesamiento del lenguaje, estableciéndose patrones básicos y la frecuencia con las que aparecen las palabras basándonos en categorías gramaticales y la semántica financiera.

Una vez obtenidos los patrones, analizaremos los resultados y podremos conocer aquellos patrones más comunes en banca y podremos facilitar a los clientes de los bancos y usuarios en general las búsquedas que realicen al tener una información organizada de la mejor forma posible.

3. Estudio del arte

Con este proyecto se quiere conseguir principalmente dos aspectos:

- Ayudar a los profesionales de banca a mostrar la información a sus clientes.

- Facilitar al cliente una búsqueda más fácil y eficiente de la información.

Debido a la gran cantidad de documentación existente y el aumento de ésta, es necesario tener un buen sistema de almacenamiento y recuperación de la documentación.

3.1. Almacenamiento y recuperación de la información

El hombre siempre ha necesitado representar el mundo que le rodea y mostrar su evolución. La escritura ha sido el mecanismo utilizado.

La evolución ha facilitado diferentes representaciones de la escritura, hoy en día se representa digitalmente y es posible su almacenamiento y recuperación de forma simple y rápida a los usuarios.

El volumen de la información crece vertiginosamente y se representa de diversas formas como puede ser un periódico electrónico, una página web, o la información bancaria que se analiza en este proyecto.

Algunos investigadores llevan planteando desde hace tiempo un fenómeno, que lo denominan “sobre carga de la información” [1]. Esto se debe al gran volumen de información y la disponibilidad que hacen que los usuarios no cuentes con suficiente tiempo para utilizar todos los medios a su alcance.

Con la finalidad que todos los usuarios puedan utilizar la información disponible. Existe el área de Recuperación de Información (Information Retrieval RI), que estudia y propone soluciones al escenario presentado, planteando modelos, algoritmos y heurísticas.

La RI se viene desarrollando desde finales de los años 50, pero en la actualidad adquiere un papel mayor debido al valor que tiene la información. El éxito o fracaso de la una operación puede deberse a disponer o no de la información.

Algunos investigadores entienden por RI lo siguiente:

Para Ricardo Baeza-Yates y otros [2] “la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información”.

Unos años antes, Salton [3] propuso una definición amplia que plantea que el área de RI “es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información”.

Croft [4] define la recuperación de la información como el “conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc.”.

Korfhage [5] definió la RI como “la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta”

Los Sistemas de Recuperación de Información (SRI) son sistemas donde la información se encuentra estructurada y almacenada en base de datos. Éstas están compuestas por documentos que procesan las consultas de los usuarios facilitándoles el acceso a la información.

La recuperación de la información puede llevarse a cabo mediante distintas herramientas: bases de datos, Internet, tesauros, ontologías, mapas... Conocer y manejar estas herramientas contribuye a una recuperación de calidad.

En esta memoria vamos hacer hincapié en la recuperación de la información mediante el procesamiento del lenguaje natural utilizando tesauros, ontologías.

3.2. Procesamiento del Lenguaje Natural

El procesamiento del lenguaje natural (PLN) es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por ordenador de información expresada en lenguaje humano para determinadas tareas, como la traducción automática, los sistemas de diálogo interactivos, el análisis de opiniones, etc.

Aparece en la década de los años 60, como una subárea de la Inteligencia Artificial y la Lingüística. Tuvo gran aceptación y éxito en sus inicios pero al poner en práctica las aplicaciones en entornos no controlados y con un vocabulario genérico surgieron muchas dificultades, por ejemplo la polisemia y sinonimia.

Actualmente, las aportaciones que se han hecho en este dominio han mejorado sustancialmente, permitiendo el procesamiento de grandes cantidades de información. El procesamiento del lenguaje natural presenta múltiples aplicaciones como pueden ser las siguientes:

- Corrección de textos.
- Traducción automática.
- Recuperación de la información.
- Extracción de Información y Resúmenes.
- Búsqueda de documentos.
- Sistemas Inteligentes para la Educación y el Entrenamiento.

3.3. Tesauros y Ontologías

Un tesauro según la definición de la UNESCO es "*un instrumento de control terminológico utilizado para traducir a un lenguaje más estricto el idioma natural empleado en los documentos y por los indizadores*". [6]

Para AENOR (Asociación Española de Normalización y Certificación) un tesauro es "*Vocabulario de un lenguaje de indización controlado organizado formalmente con objeto* de hacer *explícitas las relaciones, a priori, entre conceptos*". [7]

Mientras que para la norma elaborada por la NISO (National Information Standars Organization), un tesauro es "*a controlled vocabulary of terms in natural language that are designed por postcoordination* " [8]

Si nos hacemos eco de los manuales de mayor difusión en este ámbito, hay que destacar a Aitchinson y Gilchrist que en 1987 [9] definen al tesauro como "*un vocabulario de un lenguaje controlado de indización, organizado formalmente de manera que las relaciones conceptuales son establecidas a priori, y que puede ser utilizado para la recuperación de la información.*"

Para Slype en 1991 [10], lo considera ***"****una lista estructurada de conceptos, destinados a representar de manera unívoca el contenido de los documentos y de las consultas dentro de un sistema documental determinado, y a ayudar al usuario en la indización de los documentos y de las consultas*"

Debido al auge de la tecnología se empiezan se crean nuevas posibilidades en cuanto a las pautas de diseño, elaboración, gestión y uso de las herramientas utilizadas para la representación y recuperación de información. Gracias a Lopez-Huertas (1997) [11], Shiri y Revie (2000) [12], Qin y Paling (2001) [13] han realizado aportaciones en el entorno digital de los tesauros:

- La estructura de los tesauros a partir de la hipertextualidad, traduciéndose en el establecimiento de hipervínculos, y también entre las distintas partes del tesauro.

- Reducción del coste de actualización y mantenimiento.

- Integrar al usuario en el proceso de creación, gestión y optimización de los tesauros, permitiendo crear herramientas específicas a los requerimientos de los usuarios.

- Se amplía posibilidad de reutilizar e interoperabilidad.

El concepto de ontología surge en filosofía para estudiar la esencia del ser y por otro lado las características de toda la realidad.

A partir de los años 90, las ontologías comienzan a usarse en la Inteligencia Artificia, en especial en la Ingeniería del conocimiento, el Procesamiento del Lenguaje Natural (PLN) y la Representación del conocimiento.

La ontología se define según Gruber en 1993 [14], como “una especificación explicita y formal sobre una conceptualización compartida”.

A lo que Studer et al, en 1998 [15] añade “donde la semántica de la información se hace explícita por medio de los objetos, sus relaciones y las propiedades que los caracterizan, en un lenguaje formal que sea entendible por los ordenadores”

La ontología se utiliza para definir el vocabulario de un dominio acotándolo mediante un conjunto de términos básicos y relaciones entre dichos términos. Es decir la ontología es una teoría que especifica un vocabulario relativo a un dominio.

A continuación mostramos las similitudes y diferencias entre tesauros y ontologías.

Las similitudes son:

- Proporcionar una representación del conocimiento de un dominio compartido con el fin de facilitar la comunicación eficiente.

- Sistemas basados en conceptos que representan conocimientos complejos.

- Son lenguajes controlados, ya que definen un vocabulario de un dominio específico mediante términos básicos y sus relaciones.

- Están relacionados con la terminología empleada para representar conceptos del dominio particular.

- Utilizan jerarquías para agrupar términos en categorías y subcategorías.

- Se utilizan para catalogar y organizar la información.

Las diferencias entre ellos son las siguientes:

- Los tesauros indexan documentos y recuperan información, mientras que las ontologías modelan el conocimiento y comparten el conocimiento consensuado.

- Los tesauros van dirigido a profesionales de la información y Documentación y buscadores de información. Las ontologías van dirigidas a los diseñadores y desarrolladores de herramientas para la representación del conocimiento.

Qin y Paling [16] entienden que las ontologías son superiores a los tesauros por los siguientes motivos:

- Presentan un nivel más alto de concepción y descripción del vocabulario. Además, las ontologías se caracterizan por una descripción del vocabulario.

- Se caracterizan por un desarrollo semántico más profundo para las relaciones del tipo clase/subclase y para las relaciones cruzadas, lo que supone la ampliación de éstas y un mayor cuidado en su descripción, por supuesto explícita.

- Destacan el uso de la lógica de descripción empleada en la descripción de situaciones.

- Son reusables.

- Posibilidad de trabajo en sistemas heterogéneos, al describir formalmente objetos en el mundo, sus propiedades, y las relaciones entre estos objetos.

Ding y Foo [17] creen que las diferencias fundamentales entre una ontología y un vocabulario de representación convencional se sitúan en el nivel de abstracción, en las relaciones entre conceptos, en la capacidad para que sea comprensible para las máquinas y, lo más importante, en la expresividad que pueden proporcionar. Así:

- Una ontología puede estar elaborada de acuerdo con diferentes requerimientos y, al mismo tiempo, puede funcionar como un esquema de base de datos, como una auténtica base de conocimiento, para definir varias tareas o aplicaciones.

- Una ontología potencia la comunicación entre humanos y ordenadores mientras que un vocabulario convencional, en el mundo de lo que ellos llaman Library Science, sólo permite la comunicación entre seres humanos.

- Una ontología promueve la normalización y reutilización de la representación de la información mediante la identificación del conocimiento común y compartido.

- Las ontologías añaden valor a los tesauros tradicionales a través de una semántica más profunda, así como desde un prisma conceptual, relacional e informático. De hecho una mayor profundidad semántica puede implicar niveles más profundos de jerarquía, unas enriquecidas relaciones entre clases y conceptos, así como la capacidad de formular reglas de inferencia, etc.

3.4. Patrones de diseño

Un patrón de diseño describe un problema que se repite y describe una solución genérica a ese problema, de forma que podemos utilizar la solución en todas las ocasiones en las que afrontemos el problema. (Definición de Alexander et al., 1977 [18])

Para Erich Gamma, et al. [19], definen los patrones de diseño como "*descriptores de la comunicación entre objetos y clases adaptadas a la resolución de problemas generales de diseño en un contexto particular*". En su libro "*Design Patterns: Elements of reusable Object Oriented Software*", establecen veintitrés patrones de diseño que describen soluciones simples y elegantes a problemas específicos en el diseño del software orientado a objetos.

Un patrón de diseño se debería componer de los siguientes nueve elementos esenciales [20]:

- Nombre del patrón: nombre descriptivo y único que identifica el problema y la solución.
- Problema: describe la intención del patrón, es decir que metas y objetivos se quieren alcanzar.
- Contexto: problema recurrente en el que es aplicable el patrón.
- Fuerzas: descripción de las fuerzas, los objetivos y restricciones relevantes del patrón y de cómo éstas interaccionan entre ellas o con las metas que se deseen alcanzar.
- Solución: describe los elementos necesarios para el diseño de la solución adaptada a un caso específico.
- Ejemplos: pueden ser visuales, que ayuden al lector a entender el uso y la aplicabilidad del patrón.

- Contexto resultante: indica el estado del sistema después de aplicar el patrón, incluye las consecuencias, tanto positivas como negativas de haber aplicado el patrón.

- Exposición razonada: definición de cómo funciona el patrón y por qué es útil

- Patrones relacionados: patrones que se pueden combinar con este, o es posible aplicar a partir del contexto resultante, o soluciones alternativas.

Para Chambers et al. [21], el uso de patrones permiten mejorar la productividad y la calidad de las soluciones software debido a:

- Reutilización del diseño.

- Formar un vocabulario común de diseño, los nombres de los patrones ayudan a los diseñadores a comunicarse más fácilmente.

- Mejorar la documentación, la utilización de un patrón permite conocer que solución se ha aplicado.

A medida que crece el número de patrones, es necesario organizarlos de forma que permita al diseñador encontrar fácilmente el patrón o patrones que mejor se adapten a su problema. Esta clasificación permite organizar los patrones en grupos que compartan las mismas propiedades y dependiendo de los criterios elegidos se pueden definir esquemas de clasificación consiguiendo reducir el tamaño de la búsqueda.

Erich Gamma, et al. [19], cataloga sus veintitrés patrones en dos criterios: propósito (de creación, estructurales y de comportamiento) y ámbito (clase y objeto). Para cada una de las intersecciones entre categorías de los dos criterios encontramos una serie de patrones que son aplicables.

A continuación mostramos los veintitrés patrones identificando su categoría:

Propósito de Creación:

Factory Method: Subclase de un objeto instanciado. (Ámbito de Clase)

Abstract Factory: Familias de objetos de productos. (Ámbito de Objeto)

Builder: Como se crea un objeto compuesto. (Ámbito de Objeto)

Prototype: Clase de un objeto que es instanciado. (Ámbito de Objeto)

Singleton: Instanciación de un objeto único. (Ámbito de Objeto)

Propósito Estructural:

Adapter: Interfaz de una clase. (Ámbito de Clase)

Adapter: Interfaz de un objeto. (Ámbito de Objeto)

Bridge: Implementación de un objeto. (Ámbito de Objeto)

Composite: Estructura y composición de un objeto. (Ámbito de Objeto)

Decorator: Responsabilidades de un objeto sin subclases. (Ámbito de Objeto)

Facade: Interfaz de un subsistema. (Ámbito de Objeto)

Flyweight: Coste de almacenamiento en objetos. (Ámbito de Objeto)

Proxy: Como un objeto es accedido. Su localización. (Ámbito de Objeto)

Propósito de Comportamiento:

Interpreter: Gramática e interpretación de un lenguaje. (Ámbito de Clase)

Template Method: Pasos para un algoritmo. (Ámbito de Clase)

Chain of Responsibility: Objeto rellenado en una petición. (Ámbito de Objeto)

Command: Cuando y como una petición es rellenada. (Ámbito de Objeto)

Iterator: Como un elemento agregado es accedido, trasversalmente. (Ámbito de Objeto)

Mediator: Cómo y qué objetos interactúan entre sí. (Ámbito de Objeto)

Memento: Qué información privada se almacena fuera del objeto y cuando. (Ámbito de Objeto)

Observer: Número de objetos que se relacionan, y como mantener actualizadas las relaciones. (Ámbito de Objeto)

State: Estado de un objeto. (Ámbito de Objeto)

Strategy: Un algoritmo. (Ámbito de Objeto)

Visitor: Operaciones que pueden ser aplicadas a un objeto sin cambiar su clase. (Ámbito de Objeto)

3.5. Abstract art studio

Two aspects are wanted to aim with this project:

- Help banking professionals to present information to their clients.

- Make easier and more efficient client's searchs.

Due to existing information in quantity and its increase, it is necessary to have a secure documents storage and recovery systems.

Some researches have been studied for a while a phenom denominated “information overloaded” [1]. This is due to huge information load and availability of that information. For this reason, users does not have enough time to use all the available resources.

The purpose is that all the users can use the available information. There is a recovery information area (Information Retrieval RI) which studies and proposes solutions to this scenario with models, algorithms and heuristics.

The Recovery Information Systems (RIS) are where the information is structured and stored in data bases (DBs). RIS are composed of documents and process the user queries making easy the information access.

The information recovery can be done with different tools: Dbs, Internet, thesaurus, ontologies, maps.... Knowing and managing this tools contribute to recovery of quality.

In order to make easier user's searches, we use the nature language processing (NLP). NLP with an own ontology over the domain in which the user is doing the search and over the patterns that are used to write domain documents.

NLP is the field that combine technologies of computer science (as artificial intelligence, automatic learning or statistic inference) with applied linguistics. The aim is to make possible the understanding and the assisted computer processing

of information, expressed in human language to determined tasks like automatic translation, interactive dialogue systems, opinion analysis,...etc.

The natural language processing has multiple applications:

- Correction texts
- Automatic translation
- Information recovery
- Information extraction and summaries
- Documents search
- Intelligent systems for education and training

The concept of ontology comes up phylosophy in order to study on one hand the essence being, and on the other hand, characteristics of all reality.

Since 90's, the ontologies begin to use at Artificial Intelligence, in special at Knowledge Engineering, Natural Language Processing (NPL) and Knowledge Representation.

The ontology is defined according to Gruber in 1993 [14], as “an explicit and formal especification about a shared conceptualization”.

In 1998[15], Studer et al added “where semantic information is being explicit by objects, its connexions and its properties, in a formal language understandable by computers”.

The ontology is used to define the domain vocabulary delimited with a set of basic terms and connexions among these terms. It means that ontology is a theory which specifies a relative vocabulary for a domain.

A design pattern describes a repeated problem and a general solution to this problem. On this way, we can use this solution in all occasions in which the problem appears. (Alexander et al.'s definition, 1997 [18])

4. Desarrollo del proyecto

Para el desarrollo del proyecto se han contado con la herramienta Boilerplates; proporcionada para este proyecto creada por Eugenio Parra y reutilizada por el Grupo de Conocimiento de la Universidad. Dicha herramienta genera los patrones sintácticos utilizados en los documentos que se analicen.

Esta herramienta ha sido utilizada en los siguientes proyectos anteriormente:

- Automatic Generation of Semantic Patterns using Techniques of Natural Language Processing. (Pablo Suarez en 2013 [22])

- Evaluation of a natural language processing system in public health. (Valeria Rodriguez Barberena en 2014 [23])

Los requisitos principales del proyecto son los siguientes:

1. Obtener información sobre el dominio del proyecto a tratar, en este caso el dominio bancario.

2. Procesar la información conseguida en el apartado anterior, consiguiendo la ontología del dominio bancario.

3. Incluir la ontología en la base de datos de la herramienta Boilerplates proporcionada para realizar este proyecto

4. Genera los patrones base de los documentos encontrados.

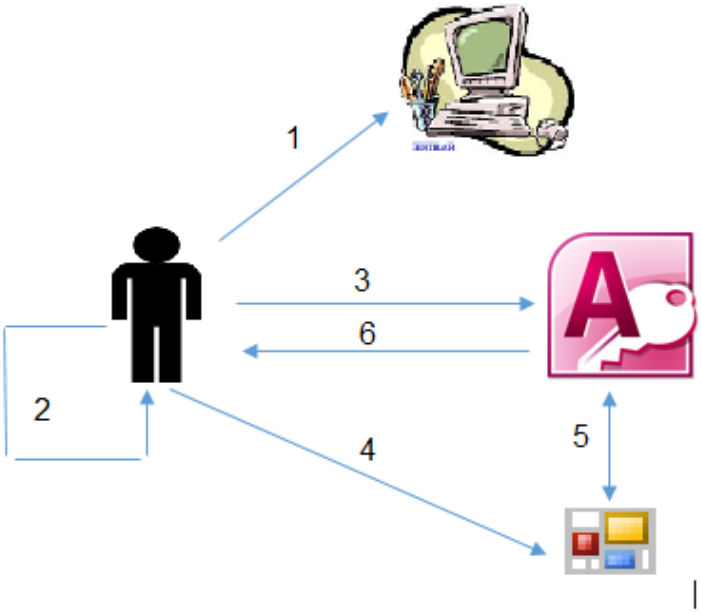
5. Generar los patrones finales de los dos escenarios propuestos:

    a. Todas las categorías sintácticas seleccionadas, con frecuencia mínima para crear patrón de uno y sin marcar el check de diferenciar por semántica.

b. Todas las categorías sintácticas seleccionadas, con frecuencia mínima para crear patrón de uno y marcando el check de diferenciar por semántica.

6. Analizar los resultados generados por la herramienta.

A continuación se mostrará un gráfico esquemático de la arquitectura del proyecto.



1. Búsqueda de la información.
2. Procesamiento de la información.
3. Insertar la ontología bancaria obtenida de los documentos en la base de datos.
4. Utilizar la herramienta Boilerplates.
5. La herramienta y la base de datos comparten información
6. Se obtiene los resultados de la base de datos para analizarlos.

Imagen 1 Diagrama de la arquitectura del proyecto

Seguidamente se explicarán los pasos que se han llevado a cabo para la realización del proyecto.

4.1. Búsqueda de documentación.

Se ha recopilado información bancaria de diversos Bancos españoles que mostraban en sus páginas web, sin necesidad de logarte en ellas, en formato pdf.

Los bancos son los siguientes:

- Banco de España
- Banco Pastor
- Bankia
- BBVA
- Inversis
- La Caixa
- Banco Sabadell

En total se disponen de doscientos cinco documentos.

4.2. Procesamiento de la información

Una vez obtenidos los documentos con formato pdf se ha procedido a convertirlos a formato txt, para poder ser procesados por la herramienta.

A continuación hemos creado la ontología, mediante las relaciones semánticas existentes en los documentos, según puede verse en el siguiente gráfico.

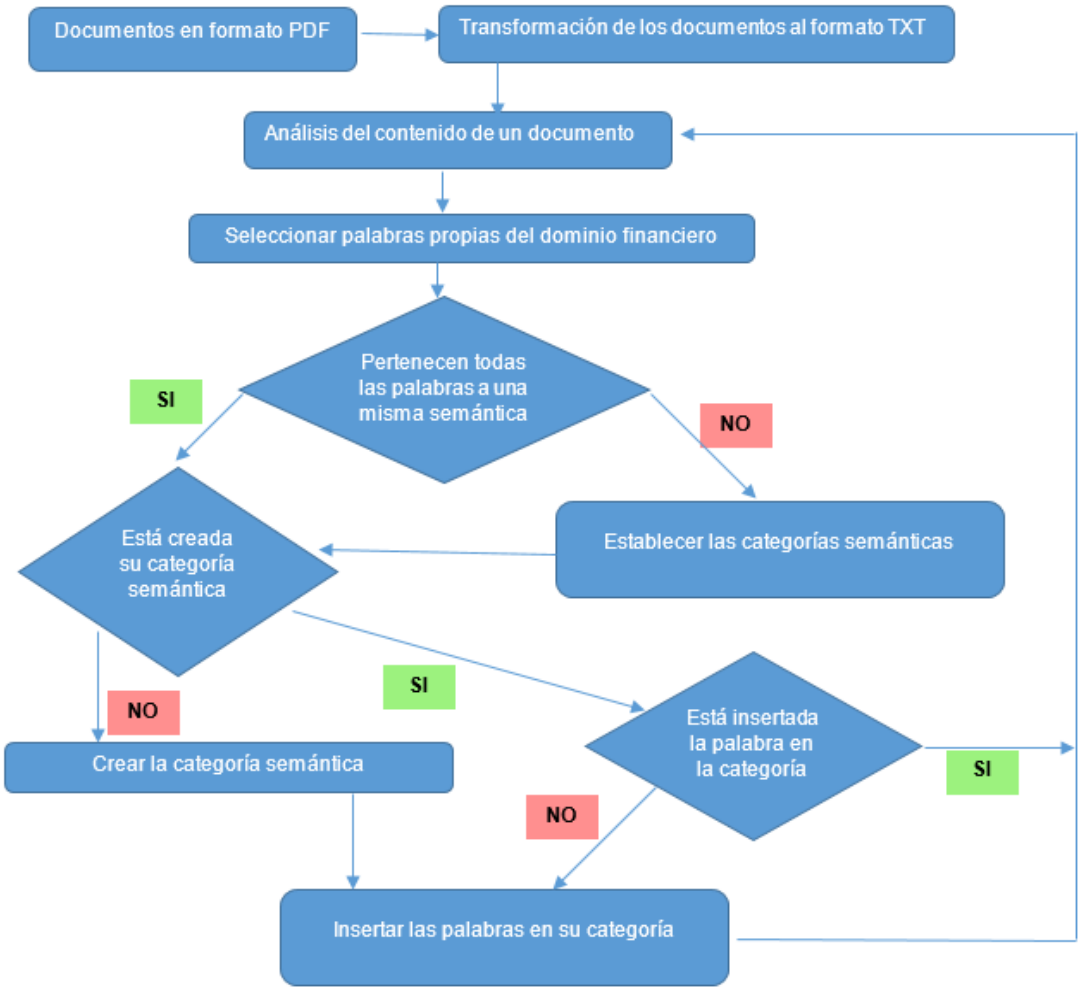


Imagen 2 Procesamiento de la información

Se han obtenido treinta y siete categorías semánticas, son las siguientes:

Nombres categorías	Descripción
Entidades	Organismos reguladores
Mercados Gestoras	Nombre de los mercados y gestoras
Banco	Distintas formas de referirse al banco
Composición Banco	Estructura bancaria
Persona	Datos necesarios de los clientes
Atención al cliente	Elementos propios de la atención al cliente
Normativas	Diferentes normativas aplicadas por el Banco
Conflictos de Interés	Política de conflictos de Interés del Banco
Delitos	Diferentes delitos que se pueden cometer en las finanzas
Países	Países
Moneda	Monedas y billetes
Comunicaciones	Forma de comunicación del banco a sus clientes
Formas de pago	Formas de pago al banco
Mercados	Tipo de mercado
Comisiones	Diferentes comisiones y gastos que el banco aplica a la utilización de sus servicios
Mifid elementos	Especificaciones normativa Mifid
Sepa elementos	Especificaciones SEPA
Mensajes swift	Distintos tipos de mensajes que se envían mediante Swift
Productos	Los diferentes productos que ofrece el banco
Tipo RF	Clasificación de renta fija
Tipo RV	Clasificación de renta variable
Tipo Cheque	Clasificación de cheques
Composición cuenta	Estructura de la cuenta
Tipo Cuenta	Clasificación de las cuentas
Tipo Remesa	Clasificación de las remesas
Tipo Tarjeta	Clasificación de las tarjetas
Gastos Transferencia	Los gastos de una transferencia SHA, OUR y BEN
Transferencia elementos	Composición de la transferencia
Tipo Avales	Clasificación de los avales
Tipo Fondo	Clasificación de los fondos
Fondos elementos	Composición de los fondos
Tipo Operaciones financieras	Clasificación de las operaciones financieras
Tipo Hipotecas	Clasificación de las hipotecas, préstamos, créditos

Nombres categorías	Descripción
Hipotecas elementos	Composición de las hipotecas, préstamos, créditos
Banco obtiene	Beneficios, pérdidas, incentivos…
Banco opera	Instrumentos financieros con los que opera el banco
Banco realiza	Acciones que puede realizar el banco

Tabla 1 Categorías semánticas del dominio de la banca

4.3. Utilización de la Herramienta Boilerplates

En este punto se va a explicar el uso de la herramienta Boilerplates.

4.3.1. Base de Datos

Esta herramienta se compone de dos bases de datos, implementadas en Access.

- RequirementsClassification.mdb, utilizada para generar los boilerplates.

- Rqa Quality Analyzer v4.1 (English).mdb, contiene la información de la ontología.

Se ha procedido a insertar en la base de datos Rqa Quality Analyzer v4.1 (English).mdb la ontología creada anteriormente, de la siguiente manera:

- Tabla Grammatical, en ella hemos introducido las treinta y siete categorías semánticas que hemos obtenido al procesar los documentos obtenidos.

- Tabla Vocabulary, en ella hemos introducido el vocabulario relacionándolo con la categoría semántica introducida en la tabla anterior y estableciendo una categoría semántica.

- Tabla Rules families, contiene las categorías sintácticas del vocabulario. En esta tabla no hemos insertado ningún registro ya que las categorías que estaban insertadas eran las necesarias.

A continuación mostraremos de manera gráfica las relaciones entre las tablas:

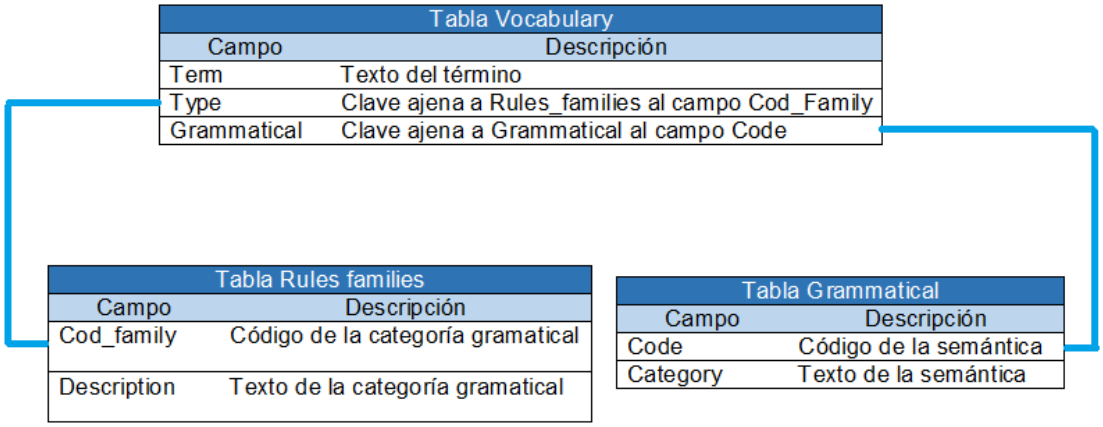


Imagen 3 Relación entre las tablas de la aplicación

Por ejemplo para la categoría semántica “Mensajes swift”:

- Se ha insertado un registro en la tabla Grammatical, para la categoría semántica “Mensajes swift”

Tabla Grammatical		
Campo	Descripción	Ejemplo
Code	Código de la semántica	1183
Category	Texto de la semántica	Mensajes swift

Tabla 2 Ejemplo de la tabla Grammatical

- Se ha insertado en un registro en la tabla Rules_ families para crear la categoría gramatical de “NOUN”

Tabla Rules families		
Campo	Descripción	Ejemplo
Cod_family	Código de la categoría gramatical	1119
Description	Texto de la categoría gramatical	NOUN

Tabla 3 Ejemplo de la tabla Rules_families

- Se ha insertado en la tabla Vocabulary un registro por cada tipo de mensaje Swift encontrado en los documentos analizados, en este caso los siguientes:

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt940
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 4 Ejemplo de la tabla Vocabulary 1

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt942
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 5 Ejemplo de la tabla Vocabulary 2

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt900
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 6 Ejemplo de la tabla Vocabulary 3

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt910
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 7 Ejemplo de la tabla Vocabulary 4

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt920
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 8 Ejemplo de la tabla Vocabulary 5

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	mt101
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1183

Tabla 9 Ejemplo de la tabla Vocabulary 6

4.3.2. Conexión a la base de datos

Una vez insertado la ontología, se procede a acceder a la herramienta Boilerplates, lo primero que tenemos que hacer es realizar la conexión con la base de datos RequirementsClassification.

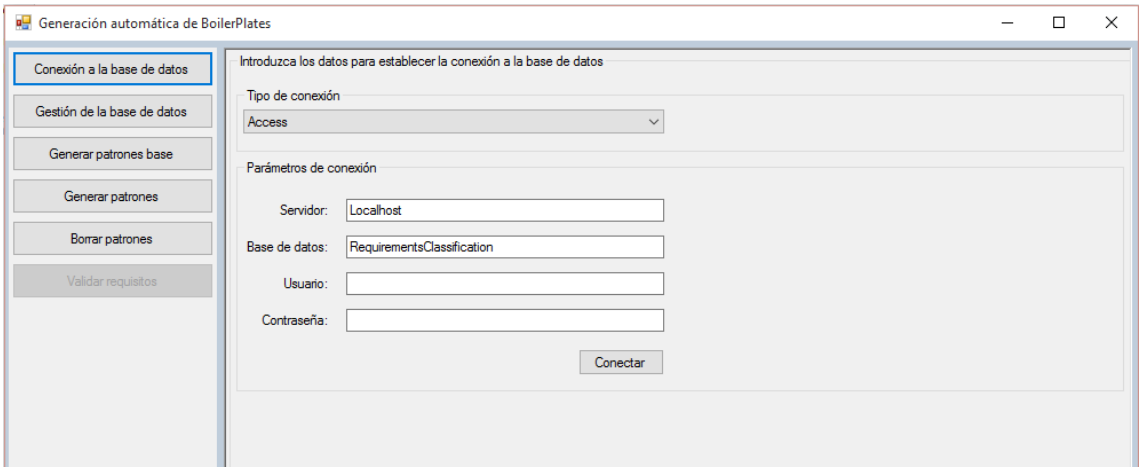


Imagen 4 Conexión a la base de datos

4.3.3. Crear patrones base

Una vez establecida la conexión a la base de datos, se procede a crear los patrones base, de los documentos obtenidos.

En la pestaña "*Generar patrones base desde un documentos*", se irán seleccionando uno a uno los documentos en formato txt que hemos seleccionado. E iniciaremos el proceso de generación de patrones.

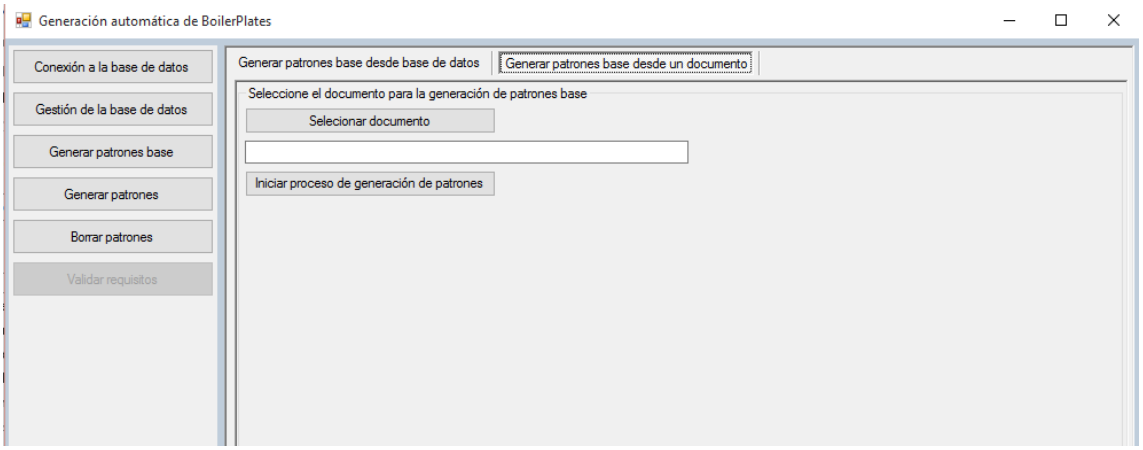


Imagen 5 Generar patrones base desde un documento

En la pestaña “*Generar patrones base desde base de datos*”, muestra la selección de documentos que se han utilizado para la generación de los patrones base.

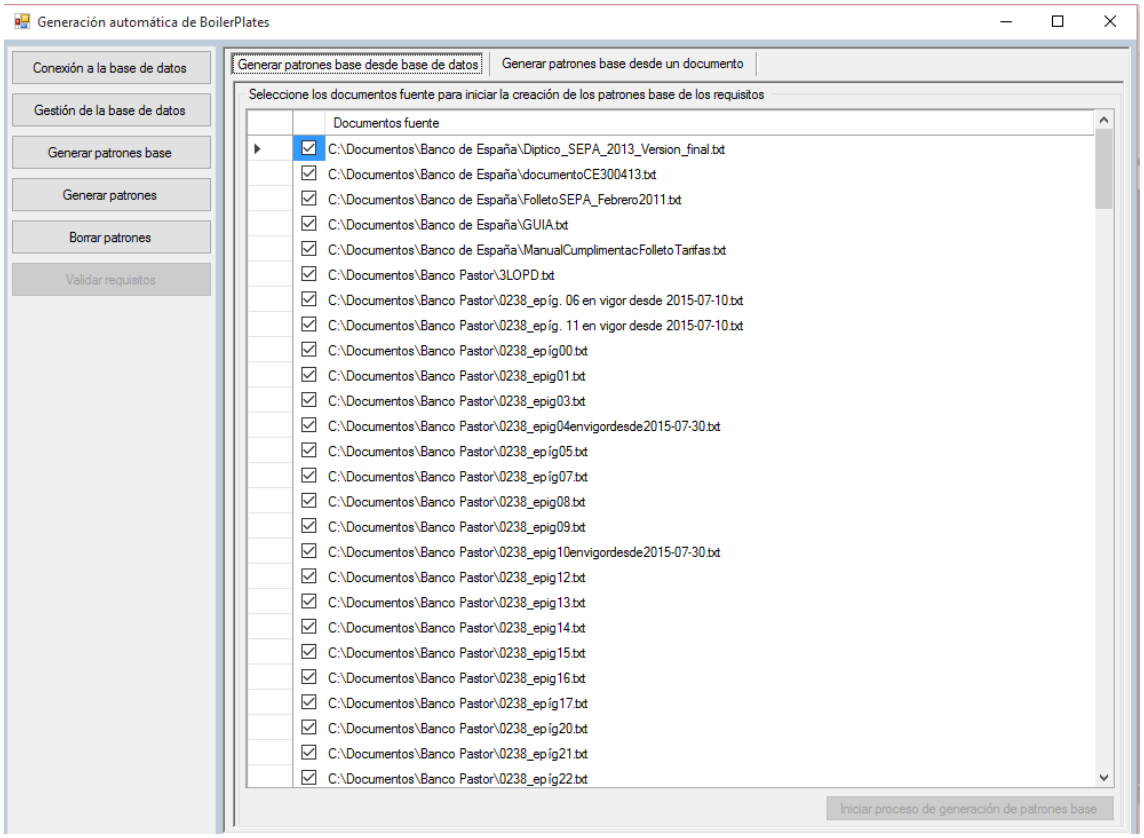


Imagen 6 Generar patrones base desde base de datos

4.3.4. Generación de patrones

Cuando se hayan generado los patrones base, el siguiente punto es generar los patrones específicos, para ello, seleccionamos “Generar patrones”

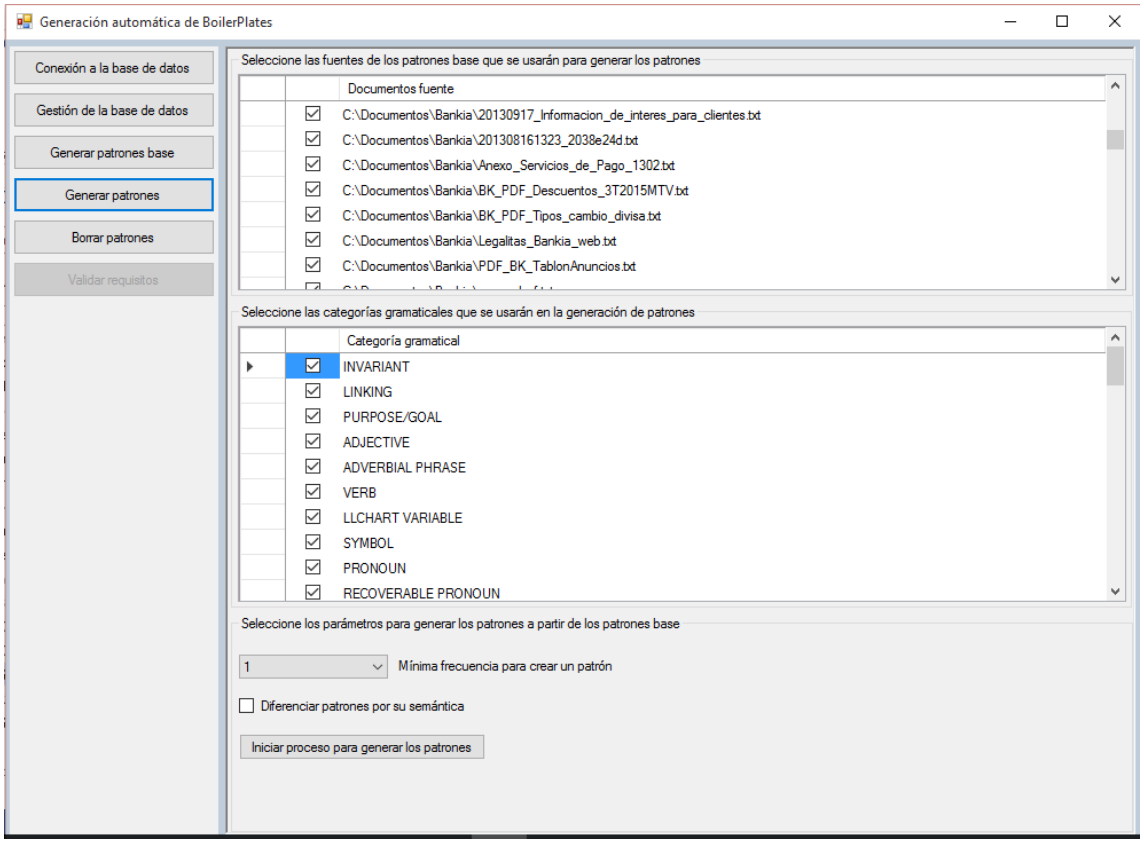


Imagen 7 Generar patrones

En la parte superior muestra los documentos con los que se va a trabajar. A continuación muestra las distintas categorías gramaticales que se van a usar en la generación de patrones. En la parte inferior, se puede especificar la frecuencia mínima para crear un patrón y si se quiere o no diferenciar por semántica.

4.3.5. Resto de opciones

El resto de opciones de la herramienta no se han utilizado para el desarrollo del proyecto, pero procedemos a explicar resumidamente cada opción.

4.3.5.1. Gestión de la base de datos

En esta opción se puede decidir eliminar patrones que ya se han creado, esto es útil cuando se quiere iniciar un nuevo análisis a partir de cero.

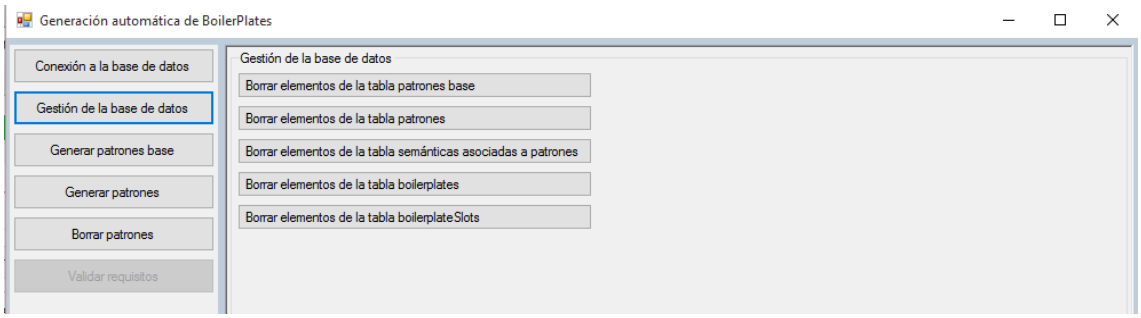


Imagen 8 Gestión de la base de datos

4.3.5.2. Borrado de patrones

En esta opción se pueden eliminar patrones y se sustituidos por un comodín.

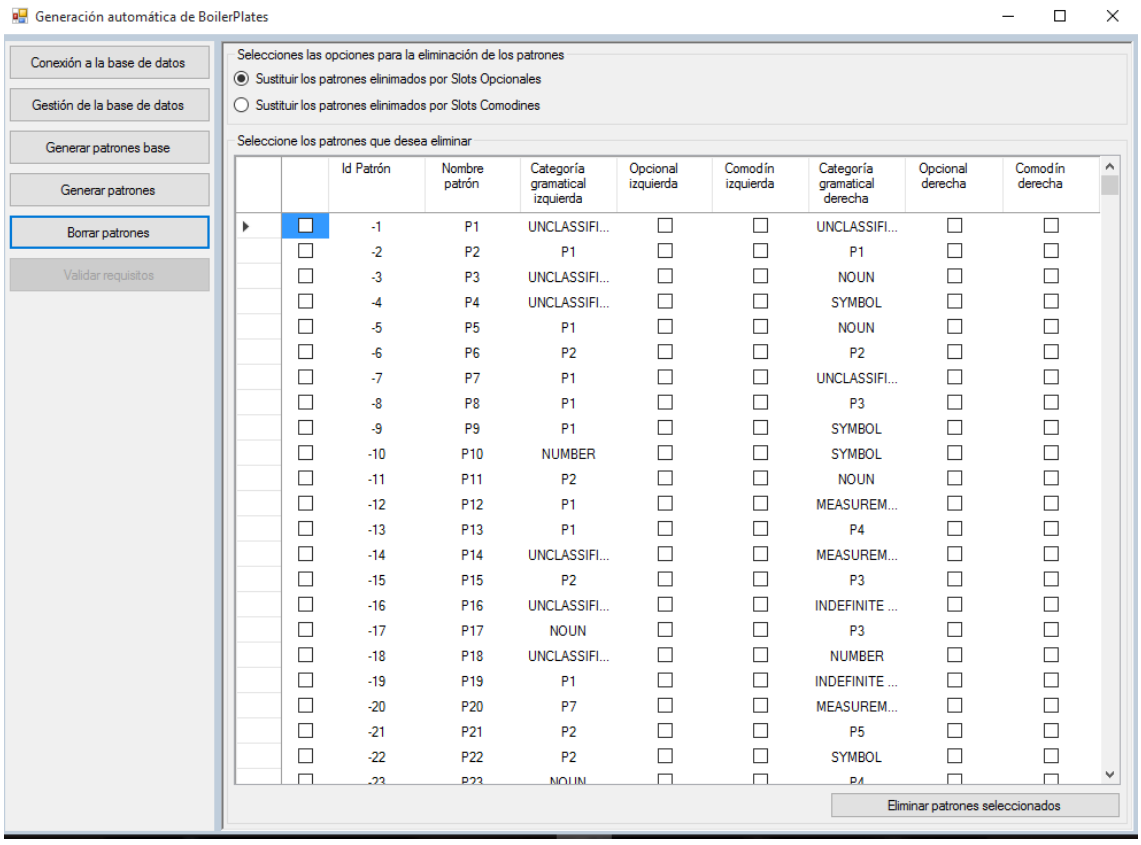


Imagen 9 Borrado de patrones

4.4. Funcionamiento de la herramienta

Boilerplates es la herramienta que analiza el dominio de documentos y creará patrones básicos y patrones de frecuencia. Analiza cada oración en cada uno de los documentos de texto.

La herramienta antes de analizar cada frase y crear los patrones tiene que realizar dos acciones previas:

- Tokenizar: El primer paso en la mayoría de las aplicaciones de procesamiento de texto es la segmentación del texto en palabras. Estas palabras están segmentados por espacios, comas, puntos, entre otros. En este caso, la separación de frases es a través del punto. Y cada palabra se reconoce con cada espacio en blanco.

- Normalizar: este paso estandariza todos los términos de todos los documentos. Esto significa, convierte palabras en plural al singular, cambiar los verbos a infinitivo, entre otros.

Una vez se han tokenizado y normalizado los documentos, se buscan los términos normalizados en la tabla Vocabulary para comenzar a crear patrones básicos.

Para la herramienta cada vez que encuentra un punto, lo establece como una nueva frase a segmentar.

4.4.1. Generación de patrones

El resultado de los patrones básicos se basa en las tablas Rules_Families, Vocabulary, Grammatical. Cuando la herramienta analiza cada una de las frases, cada una de las palabras normalizadas se busca en la tabla Vocabulary.

Si no se encuentra correspondencia, la herramienta le asigna la categoría gramatical Unclassified noun.

Si la encuentra, obtiene la categoría gramatical a la que pertenece mediante la relación con la tabla Rules_Families.

Un patrón se compone de categorías gramaticales. La composición de un modelo depende de las señales situadas en los documentos y las categorías gramaticales utilizados.

Ejemplos de patrones binarios serían:

Nombre del Patrón	Categoría gramatical 1	Categoría gramatical 2
P1	Noun	Verb
P2	Verb	Preposition

Tabla 10 Ejemplo de patrones binarios

Las categorías semánticas de las palabras (tokens) se obtienen gracias a la relación con la tabla Grammatical.

Si se quiere que la herramienta diferencie por semántica a la hora de generar los patrones no solamente se tiene en cuenta las categorías gramaticales de la tabla Rules_families, sino que también hay que tener en cuenta las categorías semánticas de la tabla Grammatical.

Por ejemplo si tenemos las categorías semánticas, CS1 y CS2, que pertenecen a la categoría gramatical NOUM, tendríamos los siguientes patrones:

Nombre del Patrón	Categoría gramatical 1	Categoría gramatical 2
P1.1	Noun CS1	Verb
P1.2	Noun CS2	Verb
P3	Verb	Preposition

Tabla 11 Ejemplo de patrones binarios con semántica

El patrón P1 formado por la categoría gramatical 1 Noun y la categoría gramatical 2 Verb al establecer la semántica se forman los patrones P1.1 y P1.2. en lugar de solamente el patrón P1 que se genera si no se diferencia semánticamente.

4.4.2. Patrones de frecuencia

Una vez creados los patrones básicos, se crean los patrones de frecuencia, patrones compuestos, en los que buscan el par más repetido en la tabla de BasicPattern y es sustituido por un patrón que se almacena en la base de datos. Estos patrones en base de datos pueden ser reconocidos porque tienen un número negativo como identificador y el nombre de la descripción incluye un prefijo "P".

El proceso de creación de patrones, utiliza la tabla PatternFactory que es un contenedor de patrones hasta generar los patrones finales que se almacenan en la tabla Patterns.

Para nuestro estudio solamente analizaremos los patrones finales que genera la herramienta.

Los patrones finales son creados por la búsqueda de la pareja más repetida de los patrones de la tabla BasicPattern. Los patrones de búsqueda pueden variar dependiendo de la frecuencia mínima establecida al generar los patrones la herramienta.

4.4.3. Ejemplos

La herramienta analiza frase por frase de los documentos, y los resultados tienen un termtag asociado dependiendo de las tablas Vocabulary y RulesFamilies. Los termtag son grabados en la tabla BasicPatterns.

A continuación mostraremos como se generan los patrones mediante unos ejemplos.

Si tenemos un documento que contiene las siguientes frases:

"El gato juega con la pelota. El perro come pienso".

Analicemos las frases:

Frase 1:

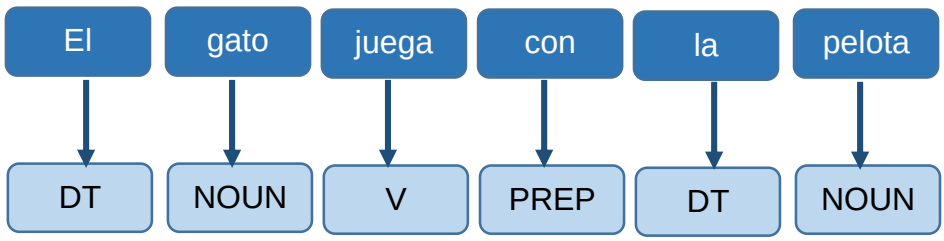


Imagen 10 Primera frase de ejemplo

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	DT	NOUN	V	PREP	DT	NOUM

Tabla 12 Primera frase de ejemplo

Frase 2:

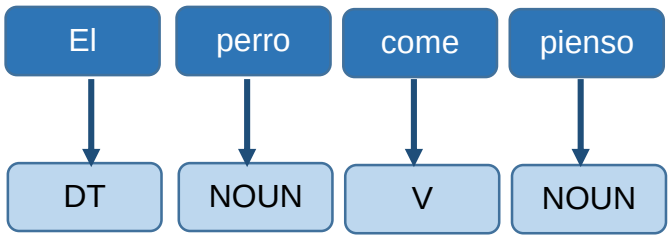


Imagen 11 Segunda frase de ejemplo

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase2	DT	NOUN	V	NOUN		

Tabla 13 Segunda frase de ejemplo

4.4.3.1. Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 1

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	DT	NOUN	V	PREP	DT	NOUN
Frase2	DT	NOUN	V	NOUN		

Tabla 14 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 1

Vemos que el patrón binario que más se repite es el formado por DT + NOUN, se crea el patrón P1 en la tabla Patterns y se sustituye el patrón binario DT + NOUN por el nuevo patrón P1.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P1	V	PREP	P1		
Frase2	P1	V	NOUN			

Tabla 15 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 1. Primeros patrones básicos

Vemos que el patrón binario que más se repite es el formado por P1 + V, se crea el patrón P2 en la tabla Patterns y se sustituye el patrón binario P1 + V por el nuevo patrón P2.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P2	PREP	P1			
Frase2	P2	NOUN				

Tabla 16 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 1. Patrones finales

4.4.3.2. Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1

Si hemos definido en la tabla Grammatical, las siguientes categorías gramaticales:

Tabla Grammatical		
Campo	Descripción	Ejemplo
Code	Código de la semántica	1
Category	Texto de la semántica	Animales

Tabla 17 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Categoría semántica animales

Tabla Grammatical		
Campo	Descripción	Ejemplo
Code	Código de la semántica	2
Category	Texto de la semántica	Juguetes

Tabla 18 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Categoría semántica juguetes

Tabla Grammatical		
Campo	Descripción	Ejemplo
Code	Código de la semántica	3
Category	Texto de la semántica	Comida

Tabla 19 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Categoría semántica comida

Y en el vocabulario tenemos definido:

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	gato
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1

Tabla 20 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Vocabulario gato

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	perro
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	1

Tabla 21 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Vocabulario perro

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	pelota
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	2

Tabla 22 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Vocabulario pelota

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	pienso
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	3

Tabla 23 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Vocabulario pienso

Las frases anteriores quedan definidas por:

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	DT	NOUN (animales)	V	PREP	DT	NOUN (juguetes)
Frase2	DT	NOUN (animales)	V	NOUN (comida)		

Tabla 24 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1.

Vemos que el patrón binario que más se repite es el formado por DT + NOUN, se crea el patrón P1 en la tabla Patterns y se sustituye el patrón binario DT + NOUN por el nuevo patrón P1.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P1	V	PREP	DT	NOUN (juguetes)	
Frase2	P1	V	NOUN (comida)			

Tabla 25 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Patrones básicos

Vemos que el patrón binario que más se repite es el formado por P1 + V, se crea el patrón P2 en la tabla Patterns y se sustituye el patrón binario P1 + V por el nuevo patrón P2.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P2	PREP	DT	NOUN (juguetes)		
Frase2	P2	NOUN (comida)				

Tabla 26 Ejemplo diferenciando por semántica y frecuencia mínima de patrón 1. Patrones finales

4.4.3.3. Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 2

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	DT	NOUN	V	PREP	DT	NOUN
Frase2	DT	NOUN	V	NOUN		

Tabla 27 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 2

Vemos que el patrón binario que más se repite es el formado por DT + NOUN, como se repite dos o más veces se crea el patrón P1 en la tabla Patterns y se sustituye el patrón binario DT + NOUN por el nuevo patrón P1.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P1	V	PREP	P1		
Frase2	P1	V	NOUN			

Tabla 28 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 2. Patrones básicos

Vemos que el patrón binario que más se repite es el formado por P1 + V, se repite dos veces por ello se crea el patrón P2 en la tabla Patterns y se sustituye el patrón binario P1 + V por el nuevo patrón P2.

Frase	Termtag 1	Termtag 2	Termtag 3	Termtag 4	Termtag 5	Termtag 6
Frase1	P2	PREP	P1			
Frase2	P2	NOUN				

Tabla 29 Ejemplo sin diferenciar por semántica y con frecuencia mínima de patrón 2. Patrones finales

4.4.3.4. Conclusiones

La creación de patrones depende del número de frecuencia mínima que elija en la herramienta y si se quiere o no diferenciar los patrones por la semántica. Esto significa que si elegimos una frecuencia mínima de 5, cuando la pareja más repetido de los patrones básicos se repite al menos cinco veces creará un patrón. En caso contrario, no va a crear una.

5. Análisis de los resultados

A continuación se explican los diferentes escenarios creados y cada uno de sus requisitos para este proyecto.

Se han creado dos escenarios diferentes:

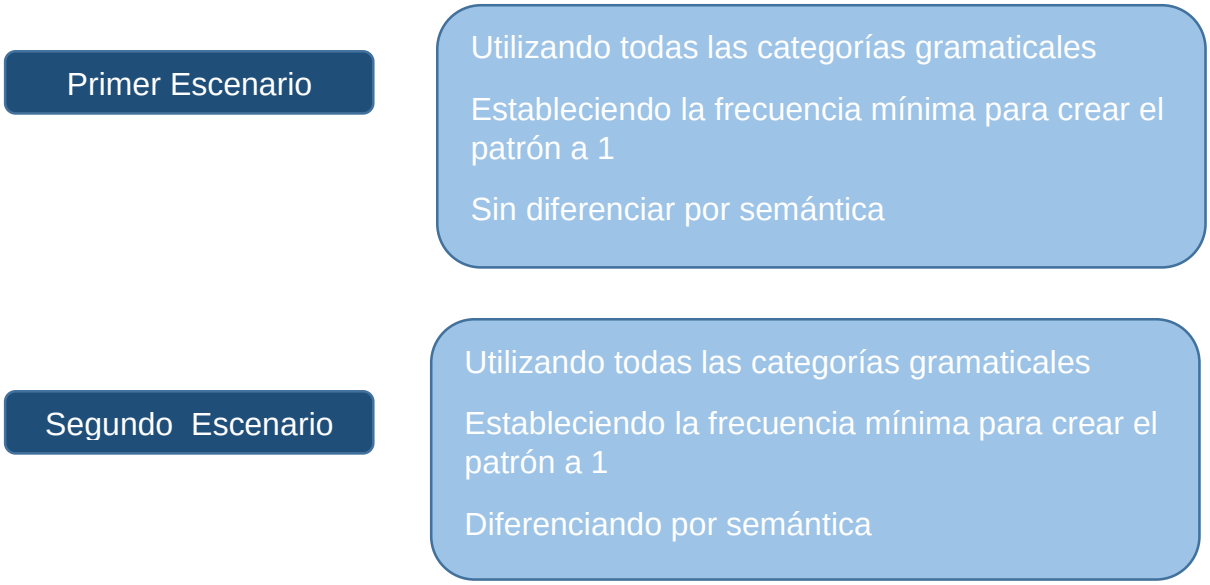


Imagen 12 Diferentes escenarios de prueba

Debido al elevado tiempo de procesamiento para generar los patrones, por lo que se ha reducido el número de documentos a procesar por la herramienta a ciento dieciséis, en lugar de los doscientos cinco.

Sin embargo para la generación de los patrones base se han utilizado los doscientos cinco documentos.

Los tiempos para los escenarios anteriores han sido:

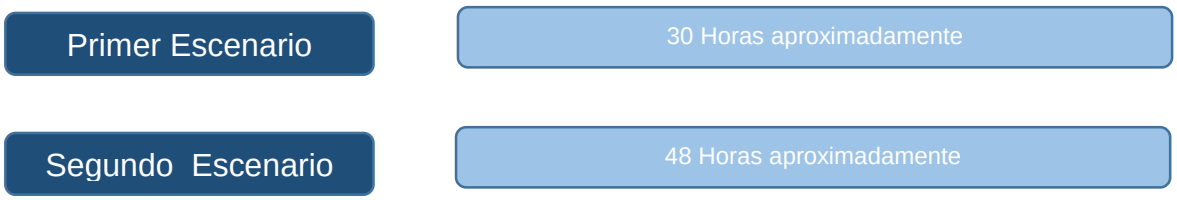


Imagen 13 Duración de los escenarios de prueba

5.1. Resultado primer escenario

Este escenario tiene las siguientes características:

- Generar todos los patrones básicos de los ciento dieciséis documentos de texto en la herramienta.

- Generar todos los patrones de estos documentos utilizando todas las categorías gramaticales ubicadas en la ficha Crear patrones en la herramienta.

- El check de diferenciar por semántica permanece desactivado.

- La frecuencia mínima para crear patrón es 1.

Analizando la base de datos, una vez procesados los ciento dieciséis documentos, podemos saber:

5.1.1. Patrones creados

Se han creado 10172 patrones binarios, a continuación mostraremos un gráfico con los 100 patrones más utilizados y la composición de estos patrones.

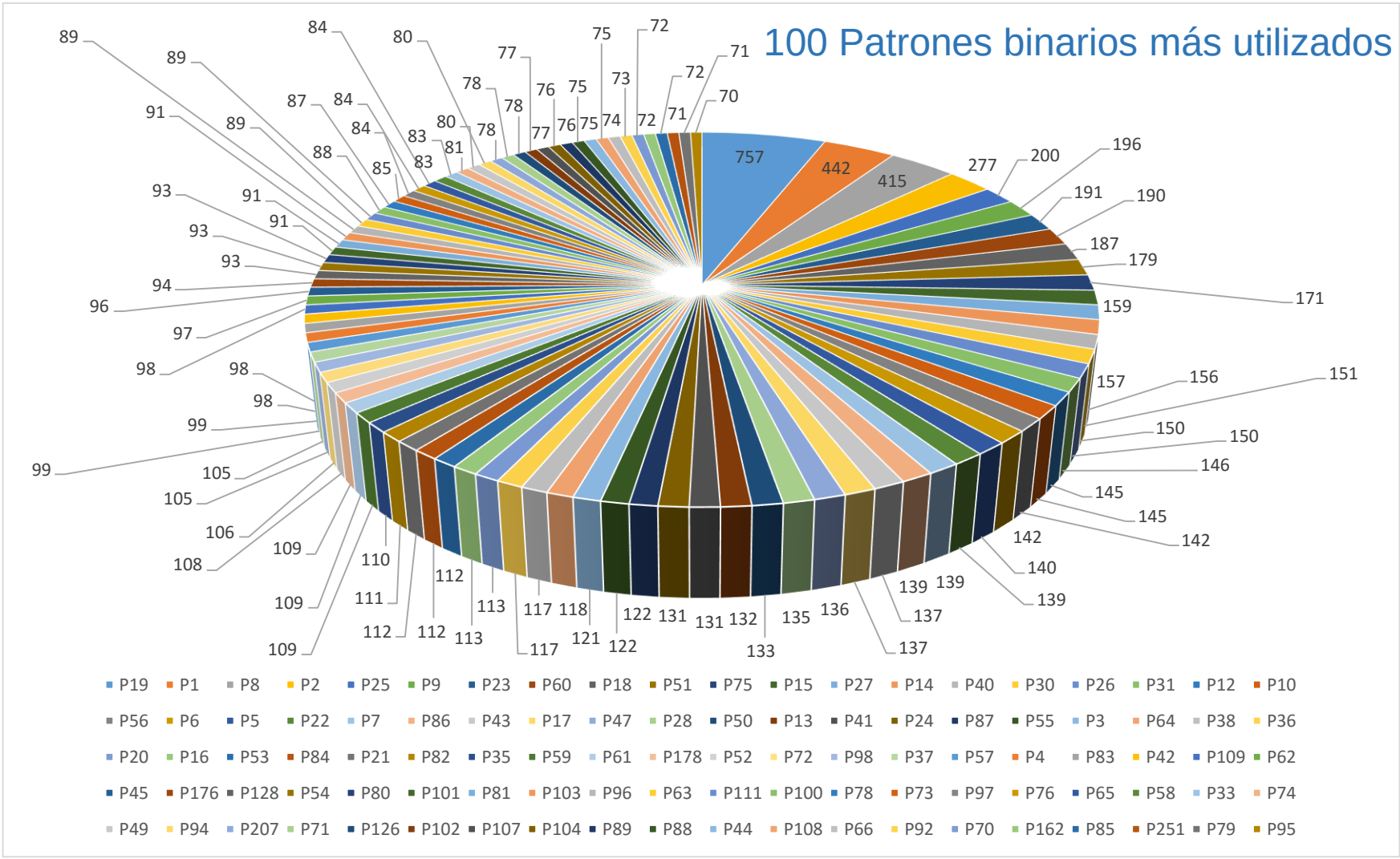


Imagen 14 Gráfico de los 100 patrones más usados del primer escenario

Patrón	Frecuencia	Termtag 1	Termtag 2
P19	757	UNCLASSIFIED NOUN	NUMBER
P1	442	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
P8	415	P1	UNCLASSIFIED NOUN
P2	277	P1	P1
P25	200	NOUN	UNCLASSIFIED NOUN
P9	196	P2	P2
P23	191	P4	P3
P60	190	P3	UNCLASSIFIED NOUN
P18	187	P2	P3
P51	179	P2	UNCLASSIFIED NOUN
P75	171	P3	P1
P15	159	P1	MEASUREMENT UNIT
P27	157	NOUN	P4
P14	156	NOUN	P2
P40	151	P2	P6
P30	150	P2	P4
P26	150	P3	P3
P31	146	P8	INDEFINITE ARTICLE
P12	145	NOUN	P3
P10	145	P1	SYMBOL
P56	142	NOUN	P1
P6	142	P1	P3
P5	140	UNCLASSIFIED NOUN	SYMBOL
P22	139	P1	INDEFINITE ARTICLE
P7	139	NOUN	NOUN
P86	139	P4	UNCLASSIFIED NOUN
P43	137	P2	SYMBOL
P17	137	P1	P5
P47	136	P2	P7
P28	135	P1	NUMBER
P50	133	P2	MEASUREMENT UNIT
P13	132	UNCLASSIFIED NOUN	MEASUREMENT UNIT
P41	131	P3	P7
P24	131	P8	MEASUREMENT UNIT

Patrón	Frecuencia	Termtag 1	Termtag 2
P87	122	P2	P1
P55	122	P2	INDEFINITE ARTICLE
P3	121	UNCLASSIFIED NOUN	NOUN
P64	118	P2	P5
P38	117	UNCLASSIFIED NOUN	ADVERB
P36	117	NOUN	P6
P20	113	NOUN	P5
P16	113	UNCLASSIFIED NOUN	INDEFINITE ARTICLE
P53	112	P3	P6
P84	112	MEASUREMENT UNIT	P10
P21	112	P2	NOUN
P82	111	P2	P13
P35	110	P3	P2
P59	109	NOUN	P10
P61	109	P3	P5
P178	109	P27	P90
P52	108	P8	NUMBER
P72	106	P2	P12
P98	105	P4	P1
P37	105	P4	P2
P57	99	P4	P7
P4	99	P1	NOUN
P83	98	P2	P16
P42	98	P1	ADVERB
P109	98	P2	P8
P62	97	P4	P6
P45	96	P3	NOUN
P176	94	NUMBER	NUMBER
P128	93	P6	UNCLASSIFIED NOUN
P54	93	P3	P4
P80	93	P3	P12
P101	91	P6	P6
P81	91	P2	P10
P103	91	P3	P8
P96	89	NOUN	P9
P63	89	P2	P14
P111	89	P4	P9
P100	88	P4	P10
P78	87	P6	P7
P73	85	P4	P4

Patrón	Frecuencia	Termtag 1	Termtag 2
P97	84	NOUN	P8
P76	84	P4	P5
P65	84	P3	SYMBOL
P58	83	P8	ADVERB
P33	83	NOUN	SYMBOL
P74	81	NOUN	P15
P49	80	P4	P12
P94	80	NOUN	P16
P207	78	P5	P19
P71	78	P3	P10
P126	78	P2	P22
P102	77	P3	P16
P107	77	P3	P9
P104	76	P2	P15
P89	76	P3	P15
P88	75	NOUN	P17
P44	75	P4	NOUN
P108	75	P2	NUMBER
P66	74	NOUN	P13
P92	73	P3	P13
P70	72	UNCLASSIFIED NOUN	1225
P162	72	P4	P8
P85	72	P4	P14
P251	71	P193	P241
P79	71	P1	1225
P95	70	P3	P14

Tabla 30 Los 100 patrones más usados del primer escenario

5.1.2. Patrones creados mismo termtags

De los 10172 patrones binarios, vemos que hay 3 patrones que están creados por el mismo termtags. A continuación mostramos los patrones creados.

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
P7	NOUN	NOUN
P176	NUMBER	NUMBER

Tabla 31 Patrones creados con el mismo termtags del primer escenario

5.1.3. Patrones creados con dos termtags

De los 10172 patrones binarios, hay 52 patrones que están compuestos por dos termtags distintos.

Los termtags más comunes que se encuentran a la izquierda (ordenados por frecuencia) son:

- Unclassified noun
- Noun
- Number
- Symbol
- Acronym
- Measurement unit
- Verb
- Adjective
- Indefinite article
- Adverb
- Verb to do
- Negation
- Absolute verb
- Not grouping noun

Los termtags más comunes que se encuentran a la derecha (ordenados por frecuencia) son:

- Unclassified noun
- Acronym
- Indefinite article
- Symbol
- Noun
- Negation
- Measurement unit
- Adverb
- Number
- Adjective
- Verb to do
- Personal pronoun
- Verb

A continuación mostramos una gráfica sobre el porcentaje de cada termtags, diferenciando si es el primer termtags o el segundo del termtags.

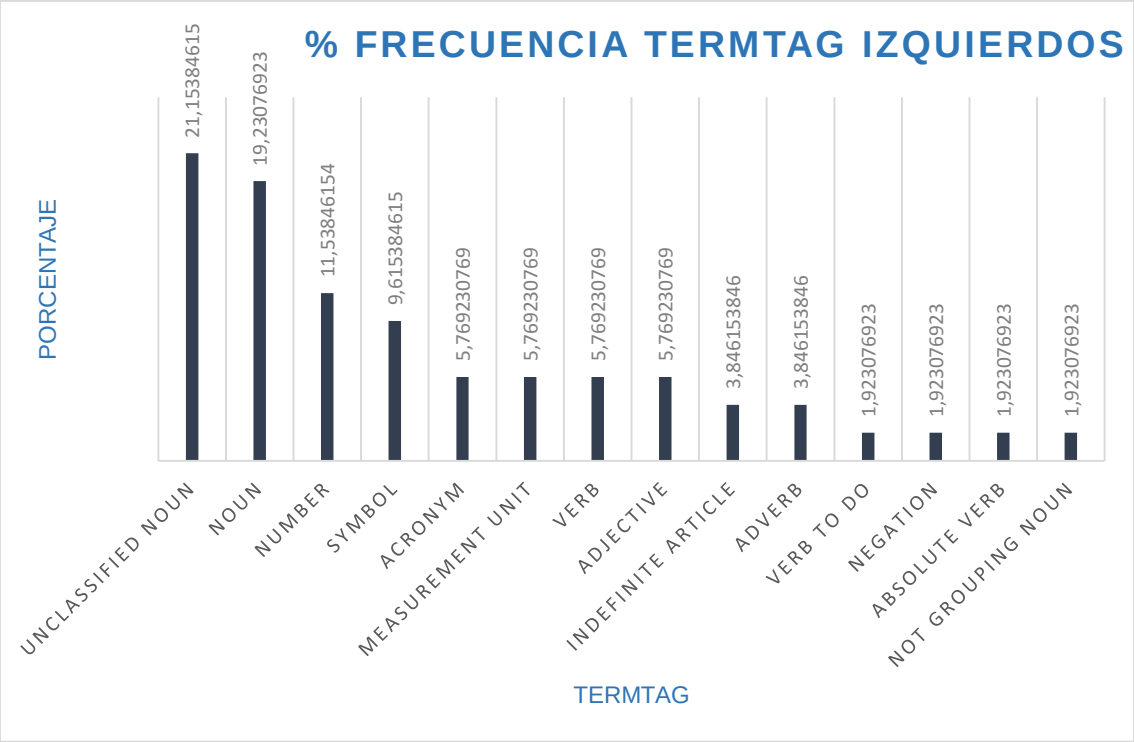


Imagen 15 Gráfico de frecuencia termtag izquierdo del primer escenario

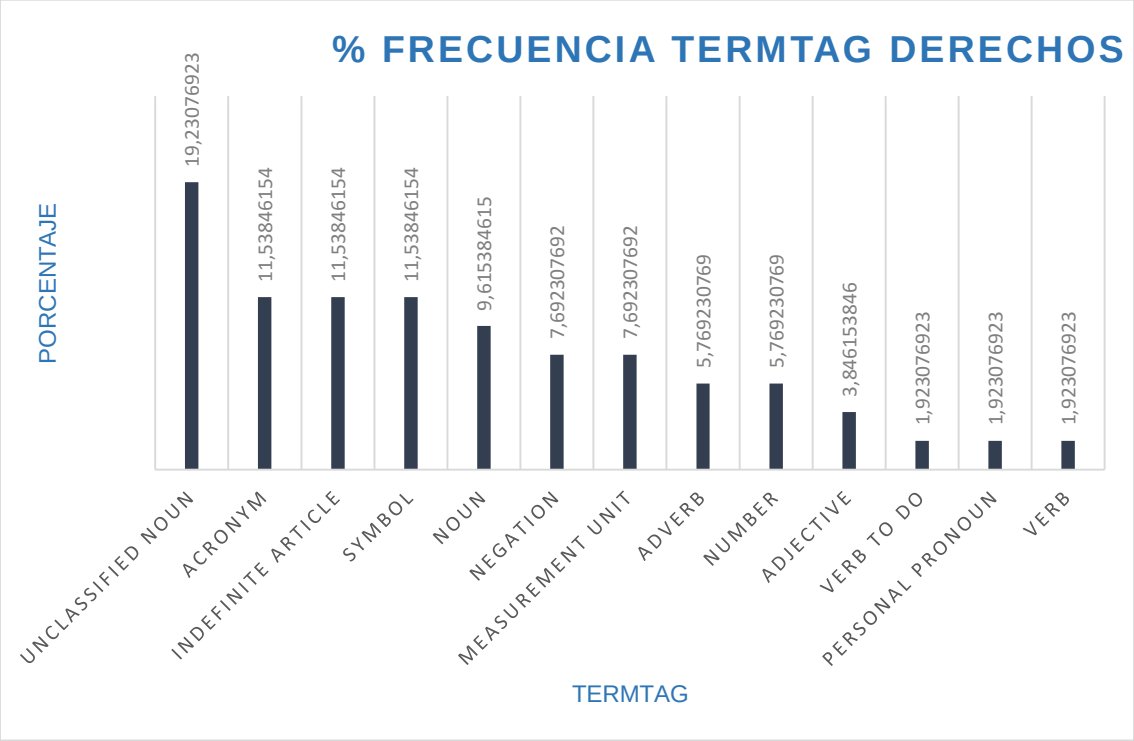


Imagen 16 Gráfico de frecuencia termtag derecho del primer escenario

5.1.4. Patrones creados con dos patrones o patrón y termtag

1. De los 10172 patrones binarios, hay 8766 patrones que están compuestos por dos patrones y 1351 patrones compuestos por un patrón y un termtags (independientemente del lugar donde aparezca el termtag a la derecha o a la izquierda del patrón).

Un ejemplo de patrón compuesto por un patrón y un termtag sería:

Nombre del Patrón	Termtag 1	Patrón
P20	NOUN	P5

Tabla 32 Ejemplo patrón 20 del primer escenario

Nombre del Patrón	Termtag 1	Termtag 2
P5	UNCLASSIFIED NOUN	SYMBOL

Tabla 33 Ejemplo patrón 5 del primer escenario

Por lo que el patrón 20 está formado por las categorías sintácticas:

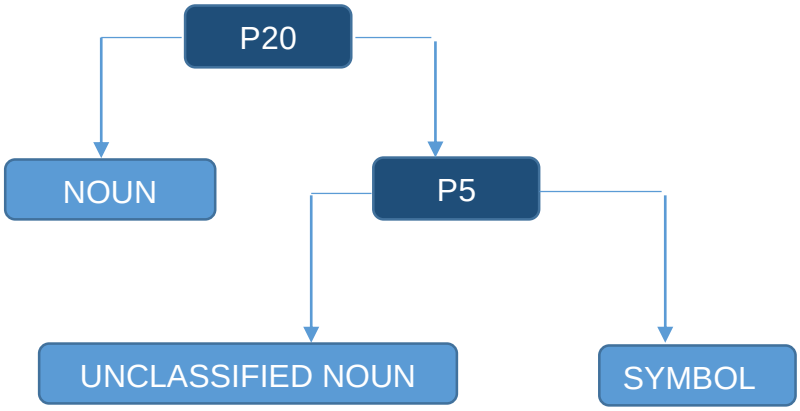


Imagen 17 Ejemplo patrón 20 del primer escenario

NOUN + UNCLASSIFIED NOUN + SYMBOL

Un ejemplo de patrón compuesto por dos patrones sería:

Nombre del Patrón	Patrón	Patrón
P17	P1	P5

Tabla 34 Ejemplo patrón 17 del segundo escenario

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN

Tabla 35 Ejemplo patrón 1 del primer escenario

Nombre del Patrón	Termtag 1	Termtag 2
P5	UNCLASSIFIED NOUN	SYMBOL

Tabla 36 Ejemplo patrón 5 del primer escenario

Por lo que el patrón 17 está formado por las categorías sintácticas:

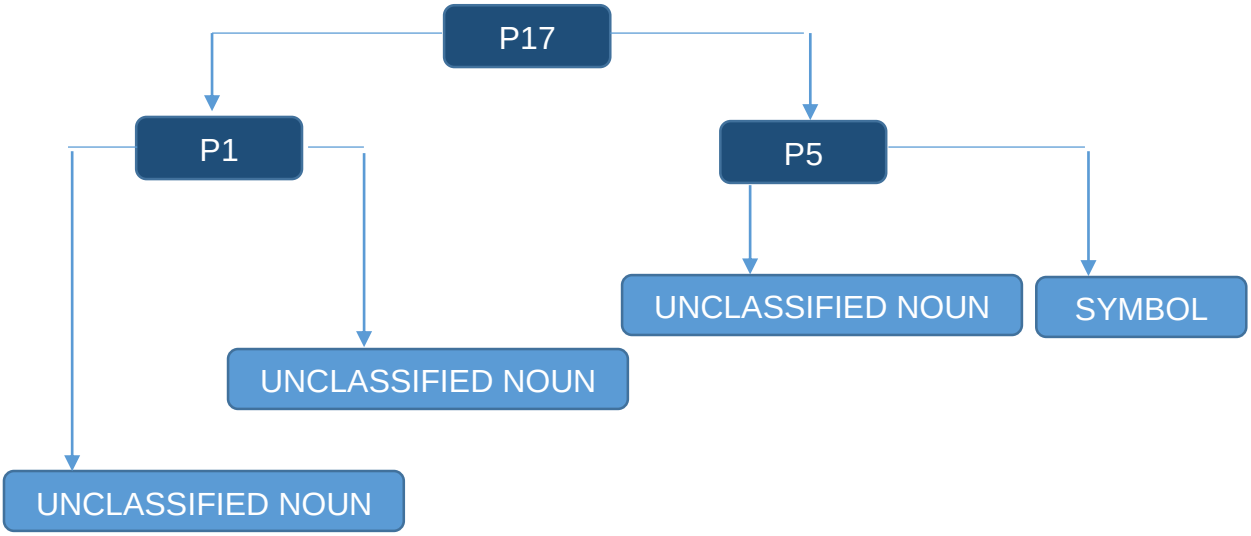


Imagen 18 Ejemplo del patrón 17 del primer escenario

UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + SYMBOL

Sabemos que un patrón se compone de dos termtags diferentes, uno a la izquierda y otro a la derecha. Hay casos en los que un patrón está compuesto por dos patrones que ya tienen termtags asignados.

En este tipo de patrones que podemos asumir que el modelo creado tiene 4 o más termtags relacionados. Desde un patrón de la izquierda tiene dos termtags asignados y un patrón a la derecha tiene dos termtags asignados. Puede haber un caso en el que también se crea uno de los patrones de la izquierda o la derecha desde otros patrones.

5.1.5. Patrones con semántica

De los 10172 patrones binarios, 176 patrones utilizan alguna de las categorías semánticas insertadas en la tabla Grammatical. Los 176 patrones están compuestos por dos categorías sintácticas o por un patrón y una categoría sintáctica (indistintamente de la posición que ocupen en el patrón).

Las categorías semánticas más utilizas en el lado izquierdo son:

- Deny (ok)
- Range <= (maximum) (ok)
- Range > minimum (ok)
- Modal optional (ok)
- Verify (ok)
- Access (ok)
- Document (ok)
- Specify (ok)
- Linguistic

Las categorías semánticas más utilizas en el lado derecho son:

- Deny (ok)
- Range <= (maximum) (ok)
- Range > minimum (ok)
- Specify (ok)
- Add (ok)
- Document (ok)
- Linguistic
- Access (ok)
- Operation (ok)

A continuación mostramos una gráfica sobre el porcentaje de cada categoría semántica, diferenciando si la categoría semántica está en el lado izquierdo o derecho del patrón.

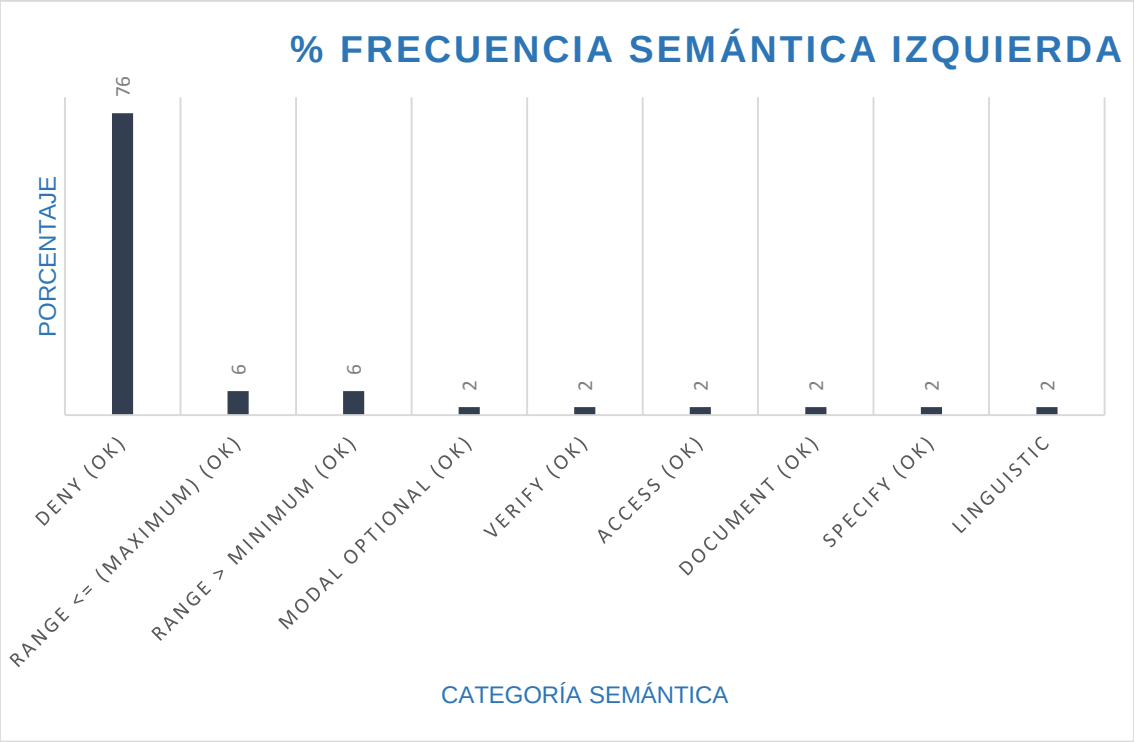


Imagen 19 Gráfico frecuencia semántica izquierdo del primer escenario

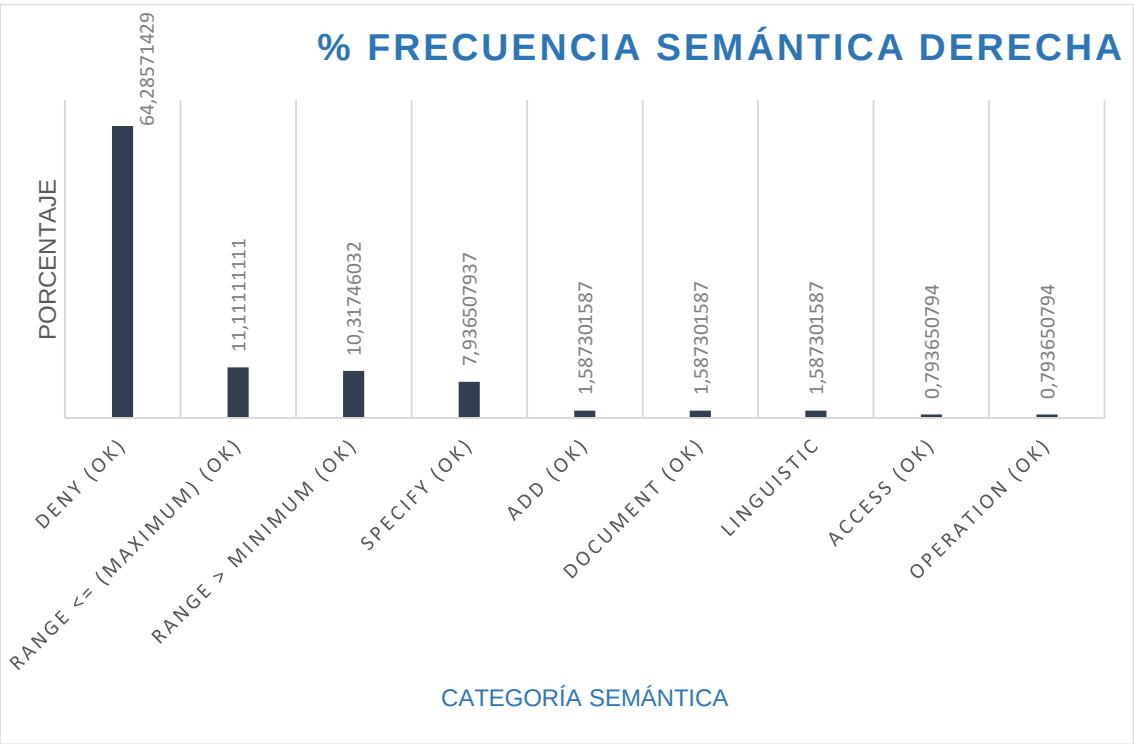


Imagen 20 Gráfico frecuencia semántica derecho del primer escenario

Los 176 patrones con categoría semántica son:

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P38	UNCLASSIFIED NOUN	ADVERB	Derecho	Deny (OK)
P42	P1	ADVERB	Derecho	Deny (OK)
P70	UNCLASSIFIED NOUN	NEGATION	Derecho	Deny (OK)
P79	P1	NEGATION	Derecho	Deny (OK)
P117	ADVERB	P10	Izquierdo	Deny (OK)
P123	P8	NEGATION	Derecho	Deny (OK)
P177	P2	ADVERB	Derecho	Deny (OK)
P216	NEGATION	P5	Izquierdo	Deny (OK)
P266	P2	NEGATION	Derecho	Deny (OK)
P272	NOUN	ADVERB	Derecho	Deny (OK)
P273	UNCLASSIFIED NOUN	VERB	Derecho	Add (OK)
P278	P5	ADVERB	Derecho	Deny (OK)
P284	P3	ADVERB	Derecho	Deny (OK)
P285	ADJECTIVE	INDEFINITE ARTICLE	Izquierdo	RANGE <= (MAXIMUM) (OK)
				RANGE > MINIMUM (OK)
P316	UNCLASSIFIED NOUN	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
				RANGE > MINIMUM (OK)
P332	SYMBOL	ADVERB	Derecho	Deny (OK)
P333	P51	NEGATION	Derecho	Deny (OK)
P338	P25	NEGATION	Derecho	Deny (OK)
P357	P1	VERB	Derecho	Specify (OK)
P364	P56	NEGATION	Derecho	Deny (OK)
P403	P1	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
				RANGE > MINIMUM (OK)
P404	ADVERB	P16	Izquierdo	Deny (OK)
P439	VERB	INDEFINITE ARTICLE	Izquierdo	Access (OK)
				Verify (OK)
				Specify (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P474	P17	ADVERB	Derecho	Deny (OK)
P477	P8	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
				RANGE <= (MAXIMUM) (OK)
P489	P4	ADVERB	Derecho	Deny (OK)
P491	P60	NEGATION	Derecho	Deny (OK)
P505	P51	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
				RANGE <= (MAXIMUM) (OK)
P515	VERB	MEASUREMENT UNIT	Izquierdo	Document (OK)
P538	ADJECTIVE	P166	Izquierdo	RANGE <= (MAXIMUM) (OK)
				RANGE > MINIMUM (OK)
P589	P10	ADVERB	Derecho	Deny (OK)
P614	ADVERB	UNCLASSIFIED NOUN	Izquierdo	Deny (OK)
P653	NEGATION	P6	Izquierdo	Deny (OK)
P686	P103	NEGATION	Derecho	Deny (OK)
P724	P75	NEGATION	Derecho	Deny (OK)
P781	NEGATION	P150	Izquierdo	Deny (OK)
P783	P133	ADVERB	Derecho	Deny (OK)
P786	P109	NEGATION	Derecho	Deny (OK)
P831	P97	NEGATION	Derecho	Deny (OK)
P872	ADJECTIVE	UNCLASSIFIED NOUN	Izquierdo	RANGE <= (MAXIMUM) (OK)
P889	P76	ADVERB	Derecho	Deny (OK)
P903	P45	ADVERB	Derecho	Deny (OK)
P935	VERB	P25	Izquierdo	Linguistic
P937	NOUN	NEGATION	Derecho	Deny (OK)
P984	P8	VERB	Derecho	Access (OK)
				Linguistic
				Add (OK)
				Specify (OK)
P1007	P64	ADVERB	Derecho	Deny (OK)
P1026	P20	ADVERB	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P1053	P3	NEGATION	Derecho	Deny (OK)
P1055	ADJECTIVE	SYMBOL	Izquierdo	RANGE > MINIMUM (OK)
P1087	P43	ADVERB	Derecho	Deny (OK)
P1251	ADVERB	P22	Izquierdo	Deny (OK)
P1257	NEGATION	UNCLASSIFIED NOUN	Izquierdo	Deny (OK)
P1263	P255	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P1374	NEGATION	P24	Izquierdo	Deny (OK)
P1442	P51	VERB	Derecho	Document (OK)
				Linguistic
				Specify (OK)
P1520	P2	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
				RANGE <= (MAXIMUM) (OK)
P1529	ADVERB	P23	Izquierdo	Deny (OK)
P1535	NEGATION	P1393	Izquierdo	Deny (OK)
P1536	NEGATION	P17	Izquierdo	Deny (OK)
P1555	P1089	NEGATION	Derecho	Deny (OK)
P1614	P87	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P1615	P86	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P1642	P59	ADVERB	Derecho	Deny (OK)
P1715	P12	ADVERB	Derecho	Deny (OK)
P1716	P12	NEGATION	Derecho	Deny (OK)
P1734	P6	ADVERB	Derecho	Deny (OK)
P1746	P2	VERB	Derecho	Document (OK)
P1765	NEGATION	P106	Izquierdo	Deny (OK)
P1766	NEGATION	P47	Izquierdo	Deny (OK)
P1767	NEGATION	P13	Izquierdo	Deny (OK)
P1768	NEGATION	P9	Izquierdo	Deny (OK)
P1843	P134	ADVERB	Derecho	Deny (OK)
P1877	P81	ADVERB	Derecho	Deny (OK)
P1901	P60	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P1933	P37	NEGATION	Derecho	Deny (OK)
P2032	P3	VERB	Derecho	Specify (OK)
P2049	NEGATION	P2	Izquierdo	Deny (OK)
P2084	P481	ADVERB	Derecho	Deny (OK)
P2153	P136	ADVERB	Derecho	Deny (OK)
P2300	P35	NEGATION	Derecho	Deny (OK)
P2442	ADVERB	P81	Izquierdo	Deny (OK)
P2444	ADVERB	P2	Izquierdo	Deny (OK)
P2461	NEGATION	P368	Izquierdo	Deny (OK)
P2463	ACRONYM	NEGATION	Derecho	Deny (OK)
P2576	P580	PREPOSITION	Derecho	RANGE <= (MAXIMUM) (OK)
P2582	P2581	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P2671	P237	ADVERB	Derecho	Deny (OK)
P2714	P165	ADVERB	Derecho	Deny (OK)
P2756	P128	VERB	Derecho	Specify (OK)
P2893	P63	NEGATION	Derecho	Deny (OK)
P2899	P61	ADVERB	Derecho	Deny (OK)
P2948	P44	ADVERB	Derecho	Deny (OK)
P3189	P4	NEGATION	Derecho	Deny (OK)
P3211	ADVERB	P135	Izquierdo	Deny (OK)
P3234	NEGATION	P421	Izquierdo	Deny (OK)
P3235	NEGATION	P18	Izquierdo	Deny (OK)
P3237	MODAL VERB	P707	Izquierdo	MODAL OPTIONAL (OK)
P3313	P2451	ADVERB	Derecho	Deny (OK)
P3722	P214	NEGATION	Derecho	Deny (OK)
P3732	P211	ADVERB	Derecho	Deny (OK)
P3764	P189	ADVERB	Derecho	Deny (OK)
P3774	P184	VERB	Derecho	Specify (OK)
P3821	P161	ADVERB	Derecho	Deny (OK)
P3844	P153	ADVERB	Derecho	Deny (OK)
P3864	P138	NEGATION	Derecho	Deny (OK)
P3947	P103	VERB	Derecho	Specify (OK)
P3979	P99	ADVERB	Derecho	Deny (OK)
P3990	P95	NEGATION	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P4065	P75	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P4329	P30	ADVERB	Derecho	Deny (OK)
P4384	P23	ADVERB	Derecho	Deny (OK)
P4651	NUMBER	NEGATION	Derecho	Deny (OK)
P4654	ADVERB	P700	Izquierdo	Deny (OK)
P4657	ADVERB	P30	Izquierdo	Deny (OK)
P4658	ADVERB	P17	Izquierdo	Deny (OK)
P4693	NEGATION	P238	Izquierdo	Deny (OK)
P4825	P4489	VERB	Derecho	Specify (OK)
P5082	P4043	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P5128	P5127	PHRASAL VERB BASE	Derecho	Operation (OK)
P5351	P5350	VERB	Derecho	Specify (OK)
P5481	P5480	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P5641	P3021	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P5746	P5745	VERB	Derecho	Specify (OK)
P5793	P2768	NEGATION	Derecho	Deny (OK)
P5896	P2580	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P6464	P6463	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P6664	P1265	ADVERB	Derecho	Deny (OK)
P6774	PUNCLASSIFIED NOUN	ADVERB	Derecho	Deny (OK)
P7047	P837	NEGATION	Derecho	Deny (OK)
P7360	P531	ADVERB	Derecho	Deny (OK)
P7761	P339	NEGATION	Derecho	Deny (OK)
P7919	P293	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P8127	P222	NEGATION	Derecho	Deny (OK)
P8128	P219	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P8152	P213	ADVERB	Derecho	Deny (OK)
P8224	P197	ADVERB	Derecho	Deny (OK)
P8481	P146	ADVERB	Derecho	Deny (OK)
P8647	P117	ADVERB	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P8705	P8704	PREPOSITION	Derecho	RANGE <= (MAXIMUM) (OK)
P8708	P110	ADVERB	Derecho	Deny (OK)
P8774	P101	ADVERB	Derecho	Deny (OK)
P9035	P73	ADVERB	Derecho	Deny (OK)
P9185	P57	NEGATION	Derecho	Deny (OK)
P9240	P52	NEGATION	Derecho	Deny (OK)
P9311	P45	NEGATION	Derecho	Deny (OK)
P9322	P44	NEGATION	Derecho	Deny (OK)
P9421	P37	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P9651	P21	ADVERB	Derecho	Deny (OK)
P9697	P18	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P9784	P14	NEGATION	Derecho	Deny (OK)
P9924	P7	NEGATION	Derecho	Deny (OK)
P10061	ADVERB	PPREPOSITION	Izquierdo	Deny (OK)
P10064	ADVERB	P156	Izquierdo	Deny (OK)
P10068	ADVERB	P51	Izquierdo	Deny (OK)
P10151	NEGATION	P9245	Izquierdo	Deny (OK)
P10152	NEGATION	P8768	Izquierdo	Deny (OK)
P10153	NEGATION	P8016	Izquierdo	Deny (OK)
P10154	NEGATION	P7832	Izquierdo	Deny (OK)
P10155	NEGATION	P7770	Izquierdo	Deny (OK)
P10156	NEGATION	P7030	Izquierdo	Deny (OK)
P10157	NEGATION	P5997	Izquierdo	Deny (OK)
P10158	NEGATION	P16	Izquierdo	Deny (OK)

Tabla 37 Patrones con semántica del primer escenario

5.1.6. Categorías sintácticas en los patrones

Observamos que la categoría sintácticas más utilizada es UNCLASSIFIED NOUN, esto es debido a que la herramienta establece esta categoría sintáctica si la palabra no está en la tabla Vocabulary y en este proyecto solamente se han incluido aquellas palabras propias del dominio de la banca, a continuación mostraremos los datos de frecuencia de cada categoría sintáctica.

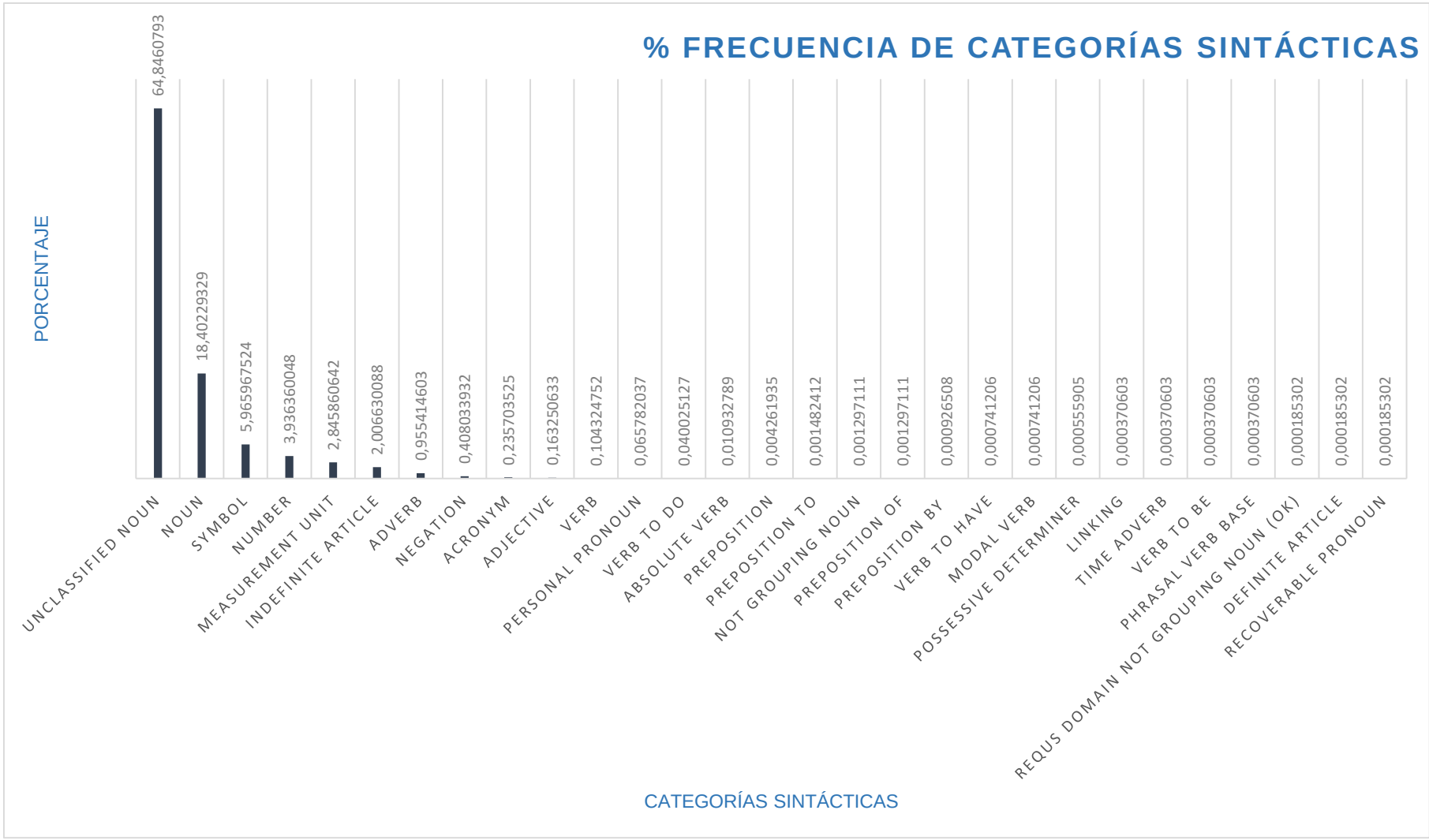


Imagen 21 Gráfico frecuencia categorías sintácticas del primer escenario

Categoría Sintáctica	Frecuencia	% Frecuencia
UNCLASSIFIED NOUN	349949	64,8460793
NOUN	99310	18,40229329
SYMBOL	32196	5,965967524
NUMBER	21243	3,936360048
MEASUREMENT UNIT	15358	2,845860642
INDEFINITE ARTICLE	10829	2,006630088
ADVERB	5156	0,955414603
NEGATION	2202	0,408033932
ACRONYM	1272	0,235703525
ADJECTIVE	881	0,163250633
VERB	563	0,104324752
PERSONAL PRONOUN	355	0,065782037
VERB TO DO	216	0,040025127
ABSOLUTE VERB	59	0,010932789
PREPOSITION	23	0,004261935
PREPOSITION TO	8	0,001482412
NOT GROUPING NOUN	7	0,001297111
PREPOSITION OF	7	0,001297111
PREPOSITION BY	5	0,000926508
VERB TO HAVE	4	0,000741206
MODAL VERB	4	0,000741206
POSSESSIVE DETERMINER	3	0,000555905
LINKING	2	0,000370603
TIME ADVERB	2	0,000370603
VERB TO BE	2	0,000370603
PHRASAL VERB BASE	2	0,000370603
REQUs Domain NOT GROUPING NOUN (OK)	1	0,000185302
DEFINITE ARTICLE	1	0,000185302
RECOVERABLE PRONOUN	1	0,000185302

Tabla 38 Categorías sintácticas del primer escenario

En la tabla Vocabulary se ha agregado el vocabulario de la ontología bancaria, la mayoría de los términos insertados son NOUN, el 18,40% de las palabras provienen del vocabulario propio del dominio de la banca, es decir si no hubiéramos añadido estos nombres a la tabla Vocabulary el porcentaje de la categoría sintáctica UNCLASSIFIED NOUN sería mayor, disminuyendo el número de patrones creados ya que los patrones creado hubieran sido solo con UNCLASSIFIED NOUN

Las siguientes categorías sintácticas más usadas son símbolos, números y unidades de medida, propias del dominio financiero.

5.2. Resultado segundo escenario

Este escenario tiene las siguientes características:

- Generar todos los patrones básicos de los ciento dieciséis documentos de texto en la herramienta.

- Generar todos los patrones de estos documentos utilizando todas las categorías gramaticales ubicadas en la ficha Crear patrones en la herramienta.

- El check de diferenciar por semántica permanece activado.

- La frecuencia mínima para crear patrón es 1.

La diferencia con el primer escenario es que se ha activado el check para que diferencie por semántica. Esto provoca que los resultados sean notablemente diferentes en cuanto a patrones y semántica.

Analizando la base de datos, una vez procesados los ciento dieciséis documentos, podemos saber:

5.2.1. Patrones creados

Se han creado 10210 patrones binarios, a continuación mostraremos un gráfico con los 100 patrones más utilizados y la composición de estos patrones.

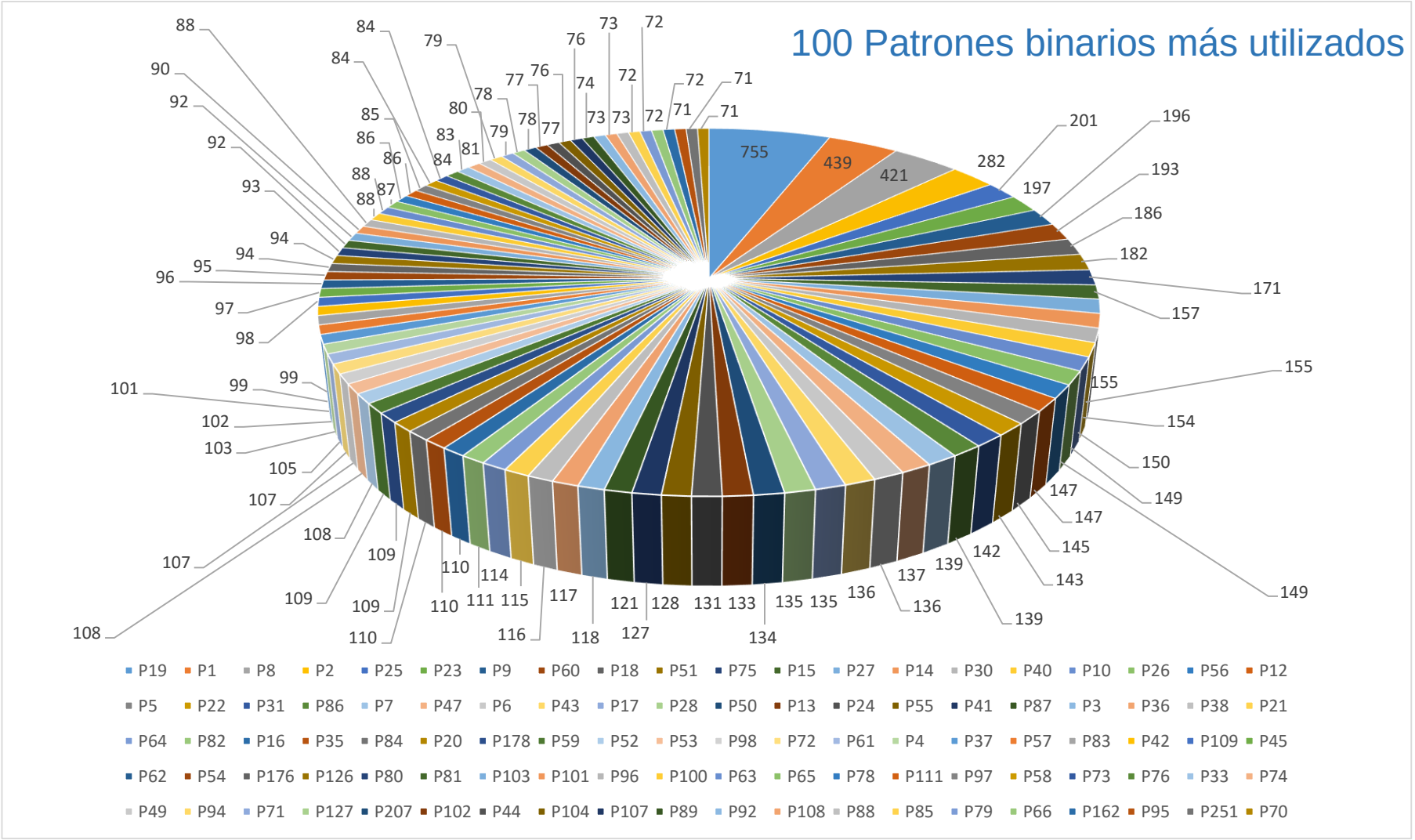


Imagen 22 Gráfico de los 100 patrones más usados del segundo escenario

Patrón	Frecuencia	Termtag 1	Termtag 2
P19	755	UNCLASSIFIED NOUN	NUMBER
P1	439	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
P8	421	P1	UNCLASSIFIED NOUN
P2	282	P1	P1
P25	201	NOUN	UNCLASSIFIED NOUN
P23	197	P4	P3
P9	196	P2	P2
P60	193	P3	UNCLASSIFIED NOUN
P18	186	P2	P3
P51	182	P2	UNCLASSIFIED NOUN
P75	171	P3	P1
P15	157	P1	MEASUREMENT UNIT
P27	155	NOUN	P4
P14	155	NOUN	P2
P30	154	P2	P4
P40	150	P2	P6
P10	149	P1	SYMBOL
P26	149	P3	P3
P56	147	NOUN	P1
P12	147	NOUN	P3
P5	145	UNCLASSIFIED NOUN	SYMBOL
P22	143	P1	INDEFINITE ARTICLE
P31	142	P8	INDEFINITE ARTICLE
P86	139	P4	UNCLASSIFIED NOUN
P7	139	NOUN	NOUN
P47	137	P2	P7
P6	136	P1	P3
P43	136	P2	SYMBOL
P17	135	P1	P5
P28	135	P1	NUMBER
P50	134	P2	MEASUREMENT UNIT
P13	133	UNCLASSIFIED NOUN	MEASUREMENT UNIT
P24	131	P8	MEASUREMENT UNIT

Patrón	Frecuencia	Termtag 1	Termtag 2
P55	128	P2	INDEFINITE ARTICLE
P41	127	P3	P7
P87	121	P2	P1
P3	118	UNCLASSIFIED NOUN	NOUN
P36	117	NOUN	P6
P38	116	UNCLASSIFIED NOUN	ADVERB
P21	115	P2	NOUN
P64	114	P2	P5
P82	111	P2	P13
P16	110	UNCLASSIFIED NOUN	INDEFINITE ARTICLE
P35	110	P3	P2
P84	110	MEASUREMENT UNIT	P10
P20	109	NOUN	P5
P178	109	P27	P90
P59	109	NOUN	P10
P52	108	P8	NUMBER
P53	108	P3	P6
P98	107	P4	P1
P72	107	P2	P12
P61	105	P3	P5
P4	103	P1	NOUN
P37	102	P4	P2
P57	101	P4	P7
P83	99	P2	P16
P42	99	P1	ADVERB
P109	98	P2	P8
P45	97	P3	NOUN
P62	96	P4	P6
P54	95	P3	P4
P176	94	NUMBER	NUMBER
P126	94	P6	UNCLASSIFIED NOUN
P80	93	P3	P12
P81	92	P2	P10
P103	92	P3	P8
P101	90	P6	P6
P96	88	NOUN	P9
P100	88	P4	P10
P63	88	P2	P14
P65	87	P3	SYMBOL

Patrón	Frecuencia	Termtag 1	Termtag 2
P78	86	P6	P7
P111	86	P4	P9
P97	85	NOUN	P8
P58	84	P8	ADVERB
P73	84	P4	P4
P76	84	P4	P5
P33	83	NOUN	SYMBOL
P74	81	NOUN	P15
P49	80	P4	P12
P94	79	NOUN	P16
P71	79	P3	P10
P127	78	P2	P22
P207	78	P5	P19
P102	77	P3	P16
P44	77	P4	NOUN
P104	76	P2	P15
P107	76	P3	P9
P89	74	P3	P15
P92	73	P3	P13
P108	73	P2	NUMBER
P88	73	NOUN	P17
P85	72	P4	P14
P79	72	P1	NEGATION
P66	72	NOUN	P13
P162	72	P4	P8
P95	71	P3	P14
P251	71	P193	P241
P70	71	UNCLASSIFIED NOUN	NEGATION

Tabla 39 Los 100 patrones más usados del segundo escenario

5.2.2. Patrones creados mismo termtags

De los 10210 patrones binarios, vemos que hay 4 patrones que están creados por el mismo termtags. A continuación mostramos los patrones creados.

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
P7	NOUN	NOUN
P176	NUMBER	NUMBER
P10015	VERB	VERB

Tabla 40 Patrones creados con el mismo termtags del segundo escenario

5.2.3. Patrones creados con dos termtags

De los 10210 patrones binarios, hay 59 patrones que están compuestos por dos termtags distintos.

Los termtags más comunes que se encuentran a la izquierda (ordenados por frecuencia) son:

- Unclassified noun
- Noun
- Number
- Symbol
- Verb
- Adjective
- Acronym
- Measurement unit
- Adverb
- Indefinite article
- Verb to do
- Negation
- Absolute verb
- Not grouping noun

Los termtags más comunes que se encuentran a la derecha (ordenados por frecuencia) son:

- Unclassified noun
- Indefinite article
- Acronym
- Symbol
- Adverb
- Noun
- Negation
- Measurement unit
- Number
- Adjective
- Verb
- Verb to do
- Personal pronoun

A continuación mostramos una gráfica sobre el porcentaje de cada termtags, diferenciando si es el primer termtags o el segundo del termtags.

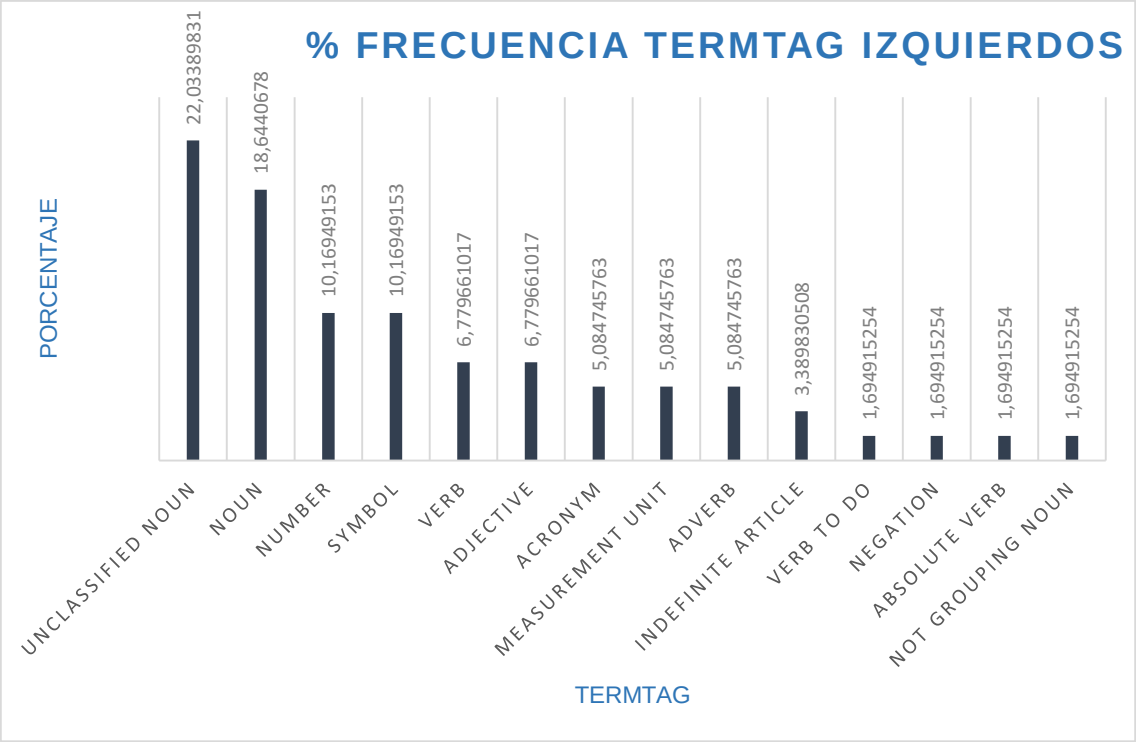


Imagen 23 Gráfico de frecuencia termtag izquierdo del segundo escenario

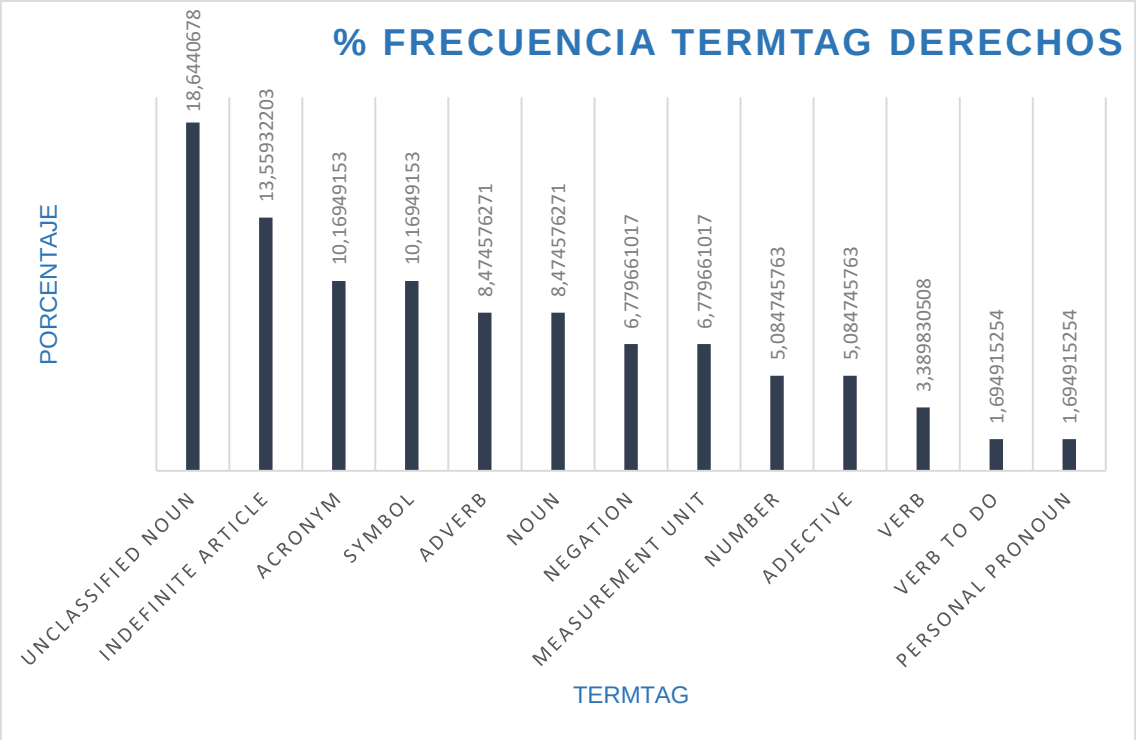


Imagen 24 Gráfico de frecuencia termtag derecho del segundo escenario

5.2.4. Patrones creados con dos patrones o patrón y termtag

De los 10210 patrones binarios, hay 8742 patrones que están compuestos por dos patrones y 1405 patrones compuestos por un patrón y un termtags (independientemente del lugar donde aparezca el termtag a la derecha o a la izquierda del patrón).

Un ejemplo de patrón compuesto por un patrón y un termtag sería:

Nombre del Patrón	Patrón	Termtag 2
P9645	P23	NOUN

Tabla 41 Ejemplo patrón 9645 del segundo escenario

Nombre del Patrón	Patrón	Patrón
P23	P4	P3

Tabla 42 Ejemplo patrón 23 del segundo escenario

Nombre del Patrón	Patrón	Termtag 2
P4	P1	NOUN

Tabla 43 Ejemplo patrón 4 del segundo escenario

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN

Tabla 44 Ejemplo patrón 1 del segundo escenario

Nombre del Patrón	Termtag 1	Termtag 2
P3	UNCLASSIFIED NOUN	NOUN

Tabla 45 Ejemplo patrón 3 del segundo escenario

Por lo que el patrón P9645 está formado por las categorías sintácticas:

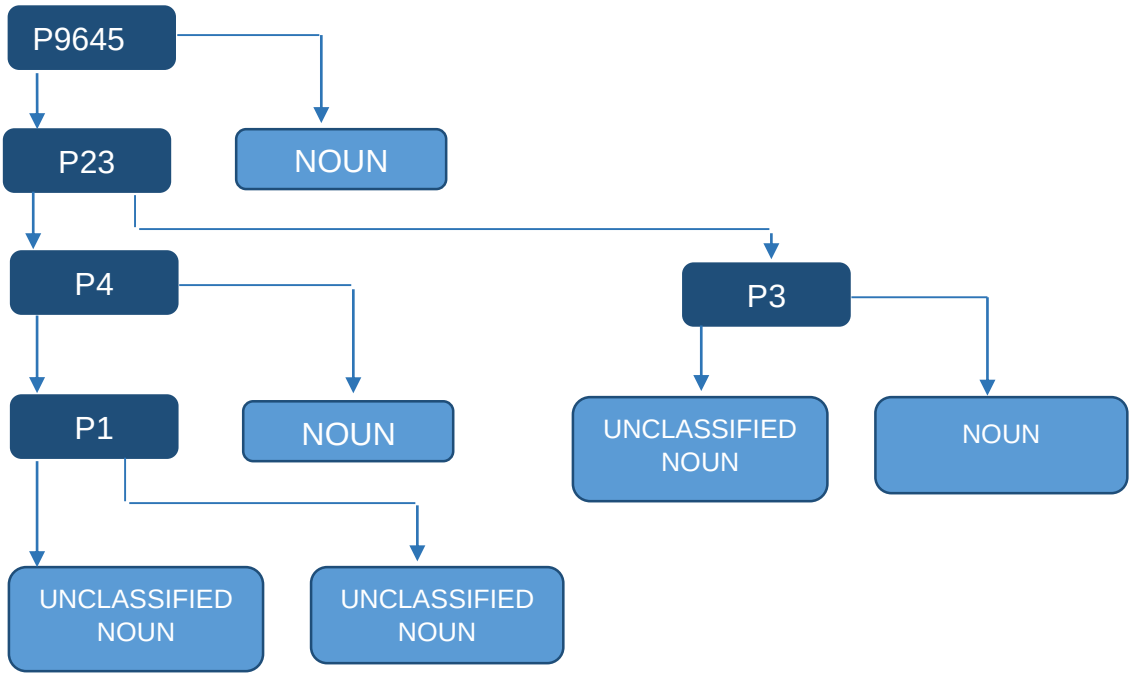


Imagen 25 Ejemplo patrón 9645 del segundo escenario

UNCLASSIFIED NOUN + UNCLASSIFIED NOUN + NOUN + UNCLASSIFIED NOUN + NOUN + NOUN

Un ejemplo de patrón compuesto por dos patrones sería:

Nombre del Patrón	Patrón	Patrón
P71	P3	P10

Tabla 46 Ejemplo patrón 71 del segundo escenario

Nombre del Patrón	Termtag 1	Termtag 2
P3	UNCLASSIFIED NOUN	NOUN

Tabla 47 Ejemplo patrón 3 del segundo escenario

Nombre del Patrón	Patrón	Termtag 2
P10	P1	SYMBOL

Tabla 48 Ejemplo patrón 10 del segundo escenario

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN

Tabla 49 Ejemplo patrón 1 del segundo escenario. Ejemplo 2

Por lo que el patrón 71 está formado por las categorías sintácticas:

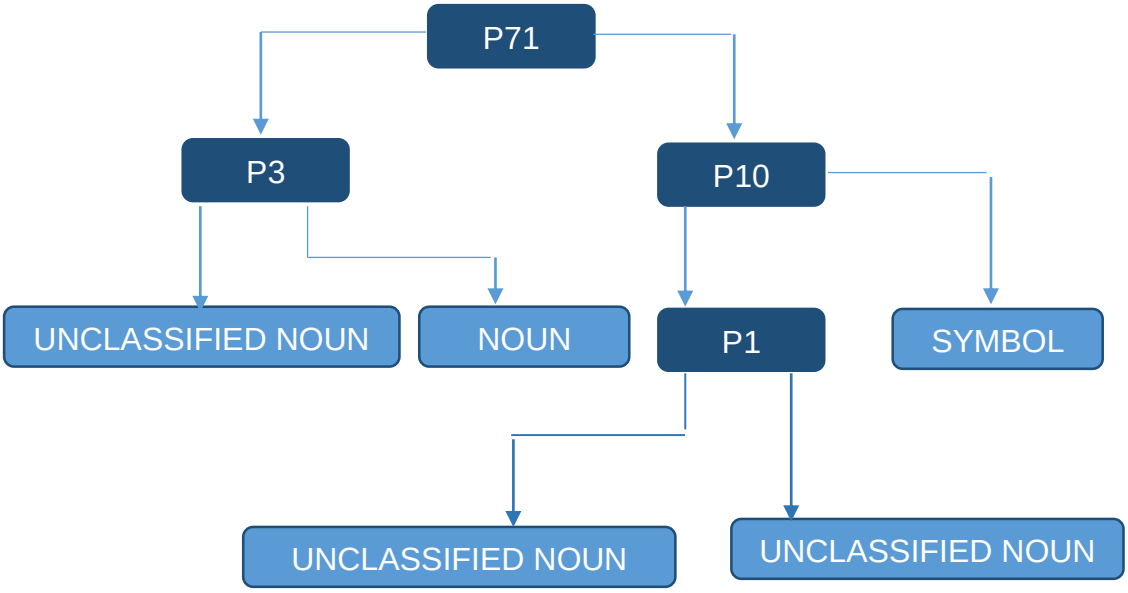


Imagen 26 Ejemplo patrón 71 del segundo escenario

UNCLASSIFIED NOUN + NOUN + UNCLASSIFIED NOUN+ UNCLASSIFIED NOUN + SYMBOL

Sabemos que un patrón se compone de dos termtags diferentes, uno a la izquierda y otro a la derecha. Hay casos en los que un patrón está compuesto por dos patrones que ya tienen termtags asignados.

En este tipo de patrones que podemos asumir que el modelo creado tiene 4 o más termtags relacionados. Desde un patrón de la izquierda tiene dos termtags asignados y un patrón a la derecha tiene dos termtags asignados. Puede haber un caso en el que también se crea uno de los patrones de la izquierda o la derecha desde otros patrones.

5.2.5. Patrones con semántica

De los 10210 patrones binarios, 160 patrones utilizan alguna de las categorías semánticas insertadas en la tabla Grammatical. Los 160 patrones están compuestos por dos categorías sintácticas o por un patrón y una categoría sintáctica (indistintamente de la posición que ocupen en el patrón).

Las categorías semánticas más utilizas en el lado izquierdo son:

- Deny (ok)
- Range <= (maximum) (ok)
- Range > minimum (ok)
- Specify (ok)
- Modal optional (ok)
- Access (ok)

Las categorías semánticas más utilizas en el lado derecho son:

- Deny (ok)
- Range <= (maximum) (ok)
- Range > minimum (ok)
- Specify (ok)
- Add (ok)
- Operation (ok)
- Document (ok)

A continuación mostramos una gráfica sobre el porcentaje de cada categoría semántica, diferenciando si la categoría semántica está en el lado izquierdo o derecho del patrón.

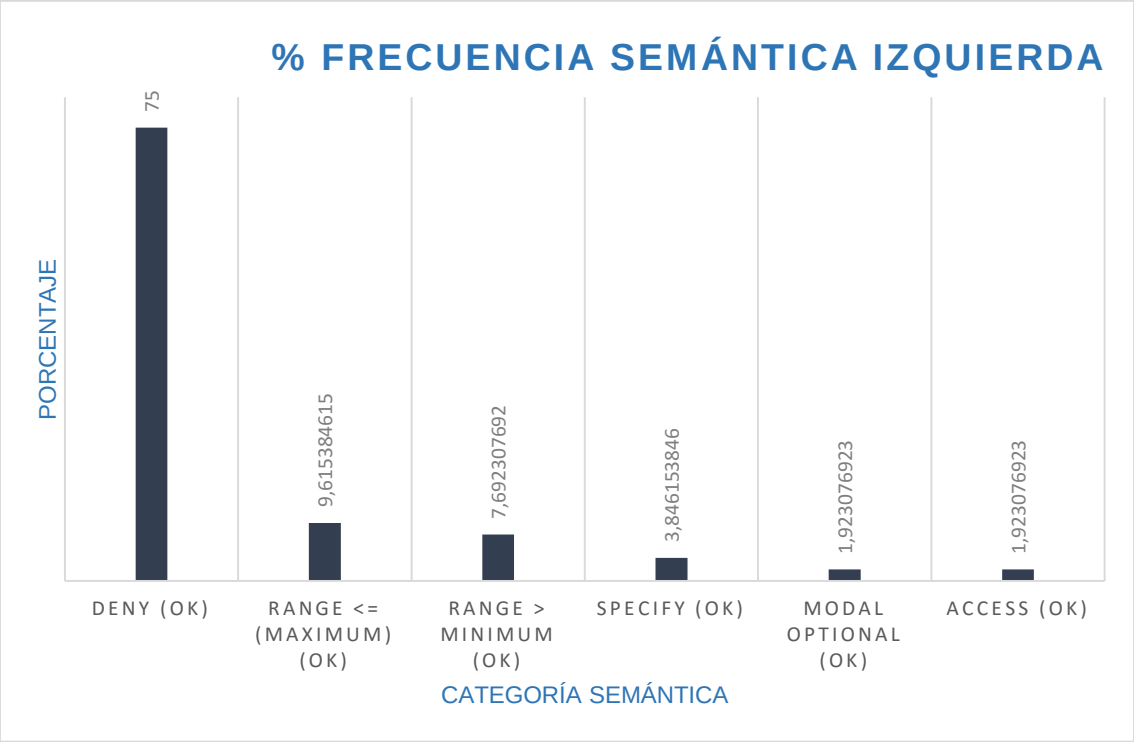


Imagen 27 Gráfico frecuencia semántica izquierdo del segundo escenario

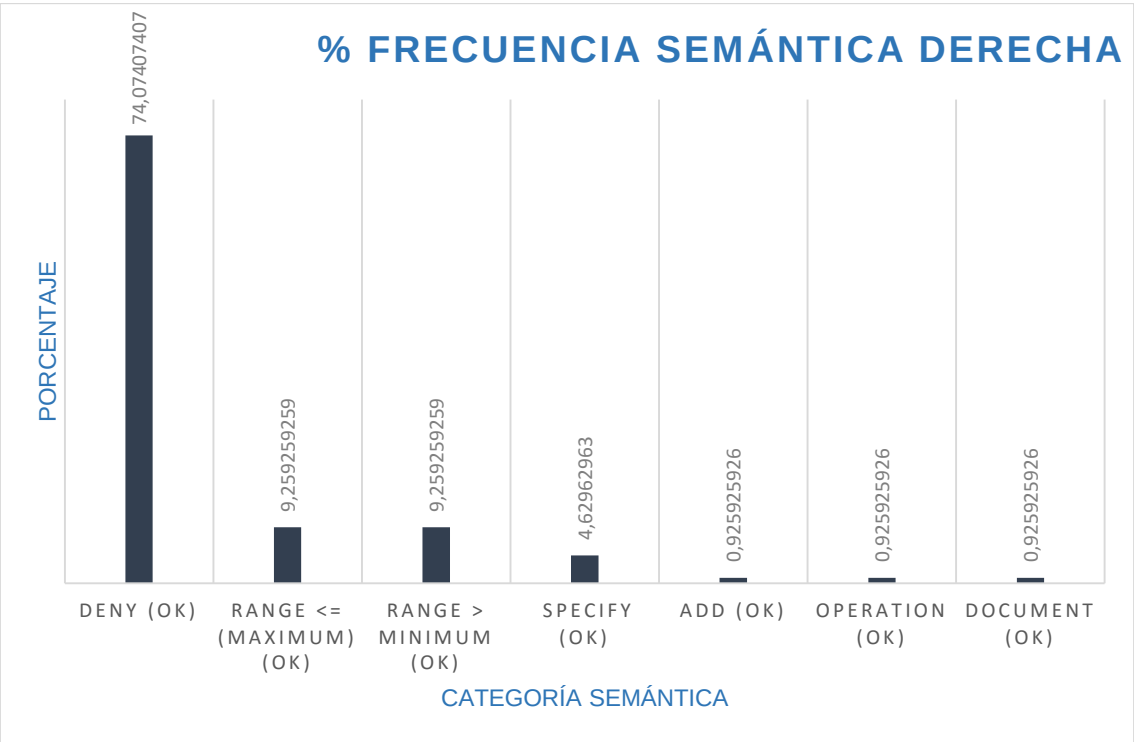


Imagen 28 Gráfico frecuencia semántica derecho del segundo escenario

Los 160 patrones con categoría semántica son:

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P70	UNCLASSIFIED NOUN	NEGATION	Derecho	Deny (OK)
P79	P1	NEGATION	Derecho	Deny (OK)
P123	P8	NEGATION	Derecho	Deny (OK)
P216	NEGATION	P5	Izquierdo	Deny (OK)
P266	P2	NEGATION	Derecho	Deny (OK)
P267	NOUN	ADVERB	Derecho	Deny (OK)
P284	P3	ADVERB	Derecho	Deny (OK)
P285	ADJECTIVE	INDEFINITE ARTICLE	Izquierdo	RANGE > MINIMUM (OK)
P333	P51	NEGATION	Derecho	Deny (OK)
P339	P25	NEGATION	Derecho	Deny (OK)
P365	P56	NEGATION	Derecho	Deny (OK)
P425	ADJECTIVE	P166	Izquierdo	RANGE > MINIMUM (OK)
P496	P60	NEGATION	Derecho	Deny (OK)
P578	ADJECTIVE	INDEFINITE ARTICLE	Izquierdo	RANGE <= (MAXIMUM) (OK)
P597	ADVERB	UNCLASSIFIED NOUN	Izquierdo	Deny (OK)
P612	P17	ADVERB	Derecho	Deny (OK)
P657	NEGATION	P6	Izquierdo	Deny (OK)
P692	P103	NEGATION	Derecho	Deny (OK)
P729	P75	NEGATION	Derecho	Deny (OK)
P787	NEGATION	P150	Izquierdo	Deny (OK)
P789	P133	ADVERB	Derecho	Deny (OK)
P792	P109	NEGATION	Derecho	Deny (OK)
P838	P97	NEGATION	Derecho	Deny (OK)
P941	P5	ADVERB	Derecho	Deny (OK)
P942	P4	ADVERB	Derecho	Deny (OK)
P947	NOUN	NEGATION	Derecho	Deny (OK)
P1065	P3	NEGATION	Derecho	Deny (OK)
P1068	1110	ADVERB	Derecho	Deny (OK)
P1152	ADVERB	P613	Izquierdo	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P1159	NEGATION	UNCLASSIFIED NOUN	Izquierdo	Deny (OK)
P1251	P6	ADVERB	Derecho	Deny (OK)
P1259	ADJECTIVE	P166	Izquierdo	RANGE <= (MAXIMUM) (OK)
P1387	NEGATION	P24	Izquierdo	Deny (OK)
P1517	P8	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P1545	NEGATION	P1406	Izquierdo	Deny (OK)
P1546	NEGATION	P17	Izquierdo	Deny (OK)
P1565	P1101	NEGATION	Derecho	Deny (OK)
P1610	P110	ADVERB	Derecho	Deny (OK)
P1727	P12	NEGATION	Derecho	Deny (OK)
P1773	NEGATION	P106	Izquierdo	Deny (OK)
P1774	NEGATION	P47	Izquierdo	Deny (OK)
P1775	NEGATION	P13	Izquierdo	Deny (OK)
P1776	NEGATION	P9	Izquierdo	Deny (OK)
P1807	P693	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P1936	P37	NEGATION	Derecho	Deny (OK)
P2037	ADJECTIVE	P19	Izquierdo	RANGE > MINIMUM (OK)
P2055	NEGATION	P2	Izquierdo	Deny (OK)
P2089	P483	ADVERB	Derecho	Deny (OK)
P2160	P136	ADVERB	Derecho	Deny (OK)
P2311	P35	NEGATION	Derecho	Deny (OK)
P2428	P1	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P2429	P1	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P2457	ADVERB	P16	Izquierdo	Deny (OK)
P2476	NEGATION	P371	Izquierdo	Deny (OK)
P2478	1257	NEGATION	Derecho	Deny (OK)
P2591	P582	PREPOSITION	Derecho	RANGE <= (MAXIMUM) (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P2597	P2596	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P2622	P420	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P2637	P359	ADVERB	Derecho	Deny (OK)
P2682	P237	ADVERB	Derecho	Deny (OK)
P2726	P165	ADVERB	Derecho	Deny (OK)
P2730	P161	ADVERB	Derecho	Deny (OK)
P2909	P63	NEGATION	Derecho	Deny (OK)
P2967	P45	ADVERB	Derecho	Deny (OK)
P2974	P43	ADVERB	Derecho	Deny (OK)
P3083	P20	ADVERB	Derecho	Deny (OK)
P3158	P10	ADVERB	Derecho	Deny (OK)
P3209	P4	NEGATION	Derecho	Deny (OK)
P3212	P2	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P3214	ADJECTIVE	P211	Izquierdo	RANGE > MINIMUM (OK)
P3236	UNCLASSIFIED NOUN	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P3237	ADVERB	P135	Izquierdo	Deny (OK)
P3259	NEGATION	P424	Izquierdo	Deny (OK)
P3260	NEGATION	P18	Izquierdo	Deny (OK)
P3262	1240	P747	Izquierdo	MODAL OPTIONAL (OK)
P3324	P2558	ADVERB	Derecho	Deny (OK)
P3441	P1194	ADVERB	Derecho	Deny (OK)
P3465	P963	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P3670	P302	ADVERB	Derecho	Deny (OK)
P3746	P214	NEGATION	Derecho	Deny (OK)
P3751	P213	ADVERB	Derecho	Deny (OK)
P3755	P211	ADVERB	Derecho	Deny (OK)
P3787	P189	ADVERB	Derecho	Deny (OK)
P3866	P153	ADVERB	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P3886	P138	NEGATION	Derecho	Deny (OK)
P3990	P101	ADVERB	Derecho	Deny (OK)
P4001	P99	ADVERB	Derecho	Deny (OK)
P4011	P95	NEGATION	Derecho	Deny (OK)
P4402	P23	ADVERB	Derecho	Deny (OK)
P4524	P12	ADVERB	Derecho	Deny (OK)
P4623	P2	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P4626	ADJECTIVE	P31	Izquierdo	RANGE <= (MAXIMUM) (OK)
P4631	VERB	INDEFINITE ARTICLE	Izquierdo	Specify (OK)
P4673	1123	NEGATION	Derecho	Deny (OK)
P4676	ADVERB	P706	Izquierdo	Deny (OK)
P4679	ADVERB	P30	Izquierdo	Deny (OK)
P4680	ADVERB	P23	Izquierdo	Deny (OK)
P4715	NEGATION	P238	Izquierdo	Deny (OK)
P4778	P4777	ADVERB	Derecho	Deny (OK)
P4843	P4842	ADVERB	Derecho	Deny (OK)
P4845	P4509	VERB	Derecho	Specify (OK)
P4869	P4868	ADVERB	Derecho	Deny (OK)
P4878	P4877	ADVERB	Derecho	Deny (OK)
P5049	P4165	VERB	Derecho	Specify (OK)
P5104	P4064	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P5157	P5156	PHRASAL VERB BASE	Derecho	Operation (OK)
P5377	P5376	VERB	Derecho	Specify (OK)
P5664	P3041	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P5767	P5766	VERB	Derecho	Specify (OK)
P5819	P2781	NEGATION	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P5855	P5854	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P5922	P2595	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P6168	P2129	VERB	Derecho	Document (OK)
P6422	P6421	ADVERB	Derecho	Deny (OK)
P6791	P1161	ADVERB	Derecho	Deny (OK)
P6794	P1158	ADVERB	Derecho	Deny (OK)
P7029	P884	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P7074	P844	NEGATION	Derecho	Deny (OK)
P7111	P798	ADVERB	Derecho	Deny (OK)
P7288	P7287	ADJECTIVE	Derecho	RANGE <= (MAXIMUM) (OK)
P7391	P533	ADVERB	Derecho	Deny (OK)
P7455	P488	VERB	Derecho	Specify (OK)
P7785	P340	NEGATION	Derecho	Deny (OK)
P8146	P222	NEGATION	Derecho	Deny (OK)
P8242	P197	ADVERB	Derecho	Deny (OK)
P8562	P134	ADVERB	Derecho	Deny (OK)
P8728	P8727	PREPOSITION	Derecho	RANGE <= (MAXIMUM) (OK)
P8895	P86	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P9044	P76	ADVERB	Derecho	Deny (OK)
P9045	P75	ADJECTIVE	Derecho	RANGE > MINIMUM (OK)
P9063	P73	ADVERB	Derecho	Deny (OK)
P9176	P61	ADVERB	Derecho	Deny (OK)
P9213	P57	NEGATION	Derecho	Deny (OK)
P9267	P52	NEGATION	Derecho	Deny (OK)
P9337	P45	NEGATION	Derecho	Deny (OK)

Patrón	TermTag1	TermTag2	Lado del Patrón	Semántica
P9348	P44	NEGATION	Derecho	Deny (OK)
P9682	P21	ADVERB	Derecho	Deny (OK)
P9813	P14	NEGATION	Derecho	Deny (OK)
P9952	P7	NEGATION	Derecho	Deny (OK)
P10008	ADJECTIVE	P1343	Izquierdo	RANGE <= (MAXIMUM) (OK)
P10009	ADJECTIVE	P19	Izquierdo	RANGE <= (MAXIMUM) (OK)
P10010	VERB	P885	Izquierdo	Access (OK)
P10015	VERB	VERB	Izquierdo	Specify (OK)
P10091	UNCLASSIFIED NOUN	VERB	Derecho	Add (OK)
P10096	ADVERB	P1231	Izquierdo	Deny (OK)
P10097	ADVERB	P1132	Izquierdo	Deny (OK)
P10103	ADVERB	P82	Izquierdo	Deny (OK)
P10104	ADVERB	P64	Izquierdo	Deny (OK)
P10106	ADVERB	P51	Izquierdo	Deny (OK)
P10107	ADVERB	P22	Izquierdo	Deny (OK)
P10108	ADVERB	P17	Izquierdo	Deny (OK)
P10109	ADVERB	P15	Izquierdo	Deny (OK)
P10189	NEGATION	P9271	Izquierdo	Deny (OK)
P10190	NEGATION	P8789	Izquierdo	Deny (OK)
P10191	NEGATION	P8035	Izquierdo	Deny (OK)
P10192	NEGATION	P7855	Izquierdo	Deny (OK)
P10193	NEGATION	P7797	Izquierdo	Deny (OK)
P10194	NEGATION	P7058	Izquierdo	Deny (OK)
P10195	NEGATION	P6024	Izquierdo	Deny (OK)
P10196	NEGATION	P16	Izquierdo	Deny (OK)

Tabla 50 Patrones con semántica del segundo escenario

5.2.6. Categorías sintácticas en los patrones

Observamos que la categoría sintácticas más utilizada es UNCLASSIFIED NOUN, esto es debido a que la herramienta establece esta categoría sintáctica si la palabra no está en la tabla Vocabulary y en este proyecto solamente se han incluido aquellas palabras propias del dominio de la banca, a continuación mostraremos los datos de frecuencia de cada categoría sintáctica.

Categoría Sintáctica	Frecuencia	% Frecuencia
UNCLASSIFIED NOUN	349949	64,846079
NOUN	99310	18,402293
SYMBOL	32196	5,9659675
NUMBER	21243	3,93636
MEASUREMENT UNIT	15358	2,8458606
INDEFINITE ARTICLE	10829	2,0066301
ADVERB	5156	0,9554146
NEGATION	2202	0,4080339
ACRONYM	1272	0,2357035
ADJECTIVE	881	0,1632506
VERB	563	0,1043248
PERSONAL PRONOUN	355	0,065782
VERB TO DO	216	0,0400251
ABSOLUTE VERB	59	0,0109328
PREPOSITION	23	0,0042619
PREPOSITION TO	8	0,0014824
NOT GROUPING NOUN	7	0,0012971
PREPOSITION OF	7	0,0012971
PREPOSITION BY	5	0,0009265
VERB TO HAVE	4	0,0007412
MODAL VERB	4	0,0007412
POSSESSIVE DETERMINER	3	0,0005559
LINKING	2	0,0003706
TIME ADVERB	2	0,0003706
VERB TO BE	2	0,0003706
PHRASAL VERB BASE	2	0,0003706
REQUs Domain NOT GROUPING NOUN (OK)	1	0,0001853
DEFINITE ARTICLE	1	0,0001853
RECOVERABLE PRONOUN	1	0,0001853

Tabla 51 Categorías sintácticas del segundo escenario

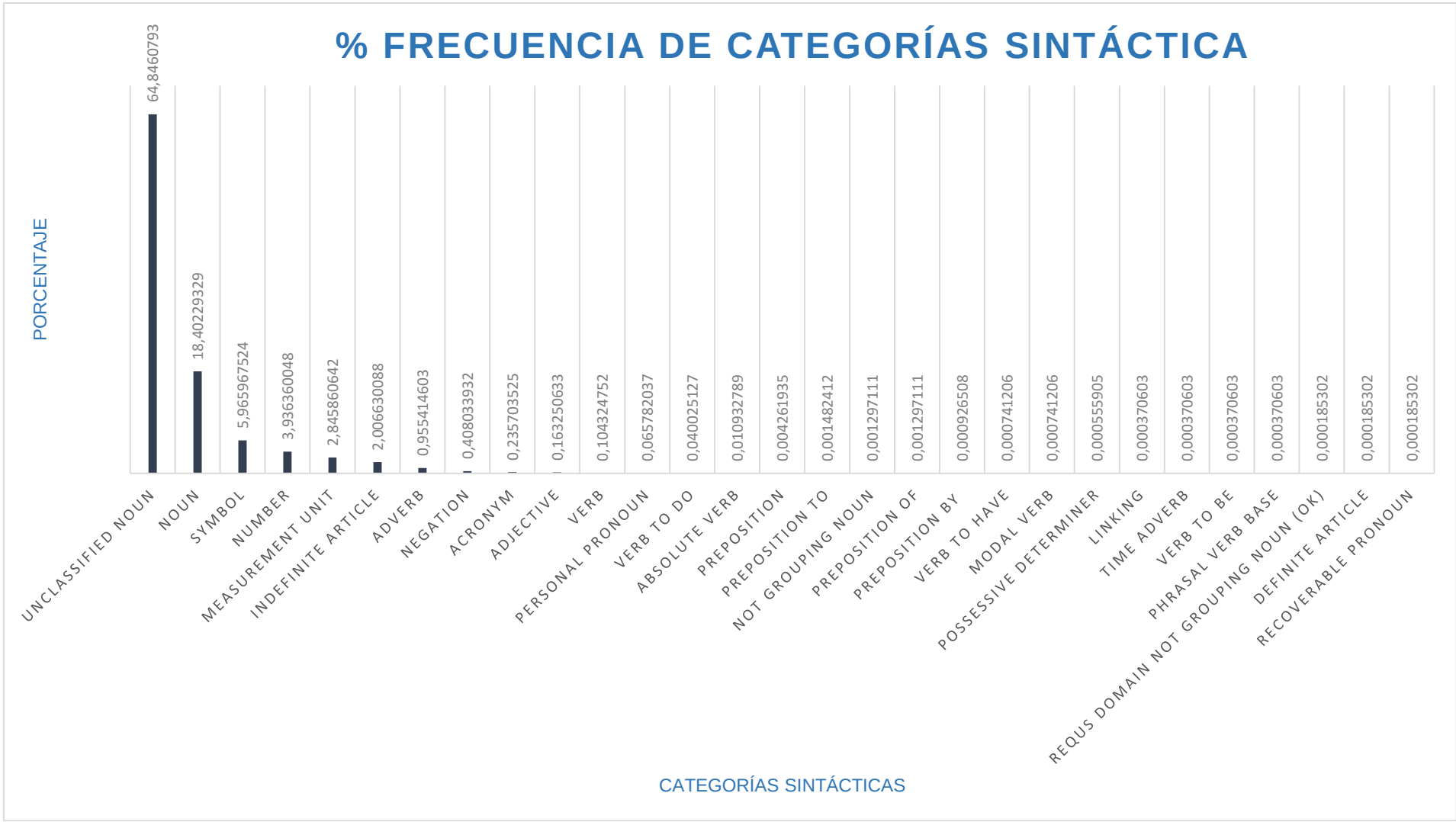


Imagen 29 Gráfico frecuencia categorías sintácticas del segundo escenario

En la tabla Vocabulary se ha agregado el vocabulario de la ontología bancaria, la mayoría de los términos insertados son NOUN, el 18,40% de las palabras provienen del vocabulario propio del dominio de la banca, es decir si no hubiéramos añadido estos nombres a la tabla Vocabulary el porcentaje de la categoría sintáctica UNCLASSIFIED NOUN sería mayor, disminuyendo el número de patrones creados ya que los patrones creado hubieran sido solo con UNCLASSIFIED NOUN

Las siguientes categorías sintácticas más usadas son símbolos, números y unidades de medida, propias del dominio financiero.

5.3. Conclusiones

Para los dos escenarios se han utilizado los mismos patrones básicos ya que en ambos escenarios se han procesado los mismos documentos.

Una vez creados los patrones básicos, se analizan todas las frases de los cientos dieciséis documentos y para cada una de las palabras (en base de datos se registran como token) se le asigna un termtag, en la base de datos RequirementsClassification, con la ayuda de las tablas Rules Families y Vocabulary de la base de datos Rqa Quality Analyzer v4.1 (English).

5.3.1. Patrones creados

Los patrones creados en ambos escenarios son diferentes en función de los requisitos con los que se ejecute la herramienta. En esta memoria hay dos escenarios, su diferencia está en diferenciar los patrones por semántica o no. Sin embargo la frecuencia mínima para crear el patrón en ambos casos es uno.

Los patrones generados en el escenario dos son mayores que los generados en el patrón uno, ya que tienen que diferenciar por semántica.

Patrones creados	
Escenario 1	Escenario 2
10172	10210

Tabla 52 Relación patrones creados ambos escenarios

Si nos fijamos en los resultados, de los patrones creados, que nos facilita la herramienta:

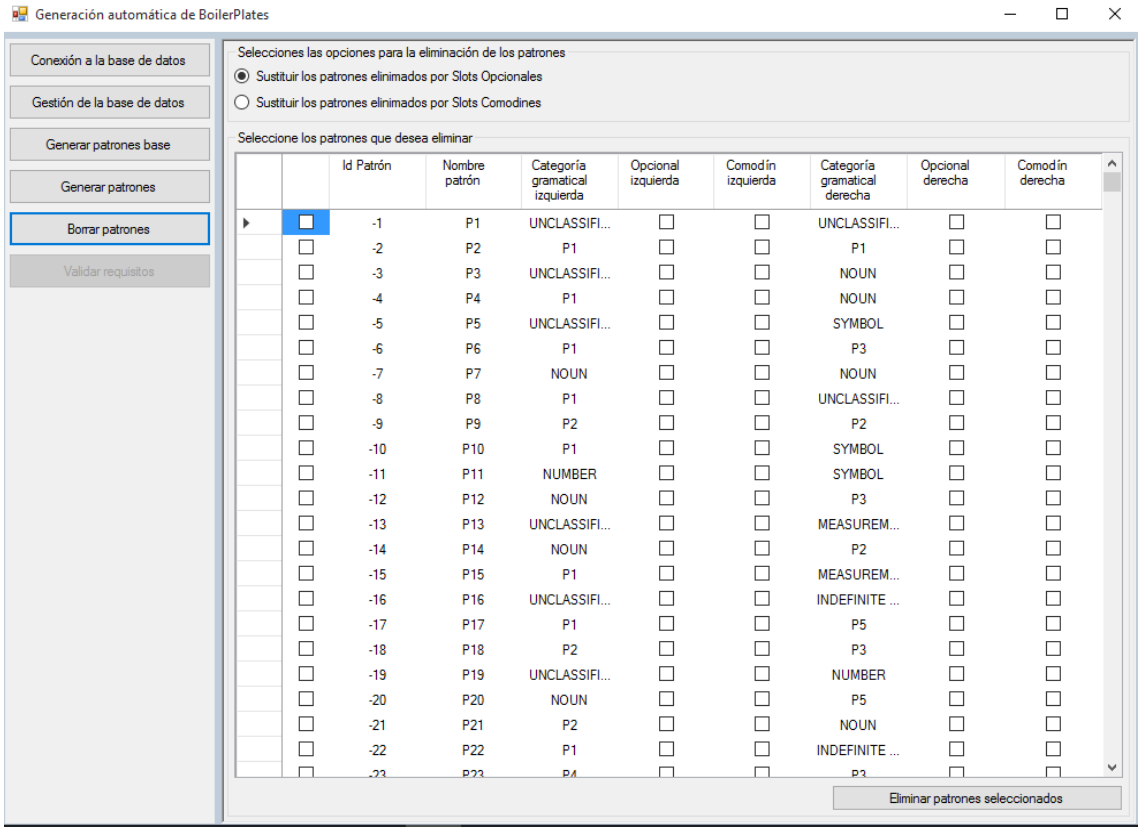


Imagen 30 Patrones finales

Los ciento veinticinco primeros patrones creados son idénticos en los escenarios analizados, a partir de aquí son totalmente distintos.

5.3.2. Patrones creados mismo termtags

En el primer escenario se han creado tres patrones con el mismo termtags en ambos lados. En el segundo escenario además de los tres del primer escenario se ha creado otro patrón más.

Los patrones comunes a ambos escenarios son:

Nombre del Patrón	Termtag 1	Termtag 2
P1	UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
P7	NOUN	NOUN
P176	NUMBER	NUMBER

Tabla 53 Patrones creados mismo termtags en ambos escenarios

El patrón que se ha creado en el segundo escenario es el siguiente:

Escenario 2		
Nombre del Patrón	Termtag 1	Termtag 2
P10015	VERB	VERB

Tabla 54 Patrón creado mismo termtags en escenario segundo

El patrón P1, Unclassified Noun - Unclassified Noun, se repite considerablemente en los dos escenarios, ya que está categoría se asigna cuando la palabra no se encuentra en la tabla Vocabulary. Algunas palabras clasificadas como Unclassified Noun en realidad pueden ser abreviaturas, símbolos…. Que no son propias del dominio bancario y no se ha añadido en la tabla Vocabulary.

5.3.3. Patrones creados con dos termtags

Los patrones más repetidos en ambos escenarios que están compuestos por dos termtags distintos. A continuación mostramos los termtags más comunes de este tipo de patrón.

Termtags más utilizados	
Termtags izquierdos	Termtag derechos
UNCLASSIFIED NOUN	UNCLASSIFIED NOUN
NOUN	INDEFINITE ARTICLE
NUMBER	ACRONYM
SYMBOL	SYMBOL
ACRONYM	ADVERB
MEASUREMENT UNIT	NOUN
VERB	NEGATION
ADJECTIVE	MEASUREMENT UNIT
INDEFINITE ARTICLE	NUMBER
ADVERB	ADJECTIVE
VERB TO DO	VERB
NEGATION	VERB TO DO

Tabla 55 Patrones creados con dos termtags más comunes en ambos escenarios

En ambos escenarios, la categoría sintáctica más utiliza es Unclassified Noun, esto es debido a que la herramienta establece esta categoría sintáctica a cada palabra que no esté incluida en la tabla Vocabulary, indistintamente de la categoría sintáctica propia de la palabra. Para este proyecto únicamente se han añadido a la tabla Vocabulary, aquellas palabras de los documentos que son propias del domino de la banca.

5.3.4. Patrones creados con dos patrones o patrón y termtag

Se han creado en ambos escenarios patrones formados por dos patrones o por un patrón y un termtag (indistintamente del lugar en el que se encuentren en el patrón).

Patrones creados con dos patrones	
Escenario 1	Escenario 2
8766	8742

Tabla 56 Relación patrones creados a partir de dos patrones en ambos escenarios

Patrones creados con patrón y termtag	
Escenario 1	Escenario 2
1351	1405

Tabla 57 Relación patrones creados a partir de un patrón y un termtag en ambos escenarios

5.3.5. Patrones con semántica

Algunos de los patrones creados en ambos escenarios tienen semántica asignada en el lado izquierdo o derecho del patrón y otros patrones que no tienen.

En el escenario uno, hay patrones con diferentes semántica pero tienen el mismo identificador y nombre, esto es porque la check para diferenciar por semántica no está activado. En el escenario dos al estar activado el check para diferenciar por semántica no se repiten patrones y su identificador es único.

5.3.6. Categorías sintácticas en los patrones

Las categorías sintácticas más utilizadas en ambos escenarios son:

- Nombres
- Símbolos
- Números
- Unidades de medida

Estas categorías sintácticas son muy utilizadas dentro del dominio de la banca.

5.4. Methodology and development

Thanks to Boilerplates, a tool provided by Eugenio Parra and used by “Grupo de Conocimiento” of University, it has been possible to develop this project.

The steps of the project will be explained below:

1. At first, bank information in pdf format, from several spanish bank webs without having to log in, has been collected.

2. Next, these documents are converted to txt format in order to be able processed by the tool

3. To continue, the ontology with semantic connections among the documents has created.

4. The previous antology has inserted in DB Rqa Quality Analyzer v4.1(English).mdb

5. Thanks to the provided tool, the basic patterns from the 205 documents are created. These patterns are created only once.

6. With basic patterns generated, we generate final patterns establishing the minimum frecuency to generate pattern of one without semantic differences.

7. We regenerated final patterns establishing the minimum frecuency to generate pattern of one with semantic differences.

The patterns creation depends on the number of minimum frecuency chosen by the tool and if they are differenced by semantic or not.

Later, with final patterns generated, the results has been analyzed taking into account the following points:

- The most repeated hundred patterns with their content and a graph showing the number of times that these patterns appear.

- Patterns created by the same *termtag* in both pattern sides.

- Patterns are created by two different *termtags.* The *termtags* are different in the side of the pattern in which appear. Also, two graphs has been created, one for each side of the pattern, which show the *termtags* frecuency.

- Patterns are created from two patterns or one pattern and one termtag (without order in the pattern). There is an example for each type of pattern.

- Patterns with its own semantic. It shows the final patterns with the semantic and a graph for each side of the pattern with the semantic category and their percent.

- Syntactic categories of patterns and a graph which shows the syntactic categories and their frecuencies.

6. Conclusiones finales y nuevas líneas de trabajo

El uso de un dominio de documentos relacionados con la banca y haciendo diferentes escenarios para crear patrones es posible concluir lo siguiente:

1. Los patrones básicos se crean al analizar todos los documentos de texto. Después de tener los resultados, es evidente que hay termtags comunes entre todos los documentos que son Unclassified Noun, Noun, Number…

2. A pesar de que los documentos son creados por diferentes autores y pertenecen a distintos bancos, los patrones creados son comunes entre todos los escenarios. Los patrones pueden tener un nombre diferente, pero la composición de ellos es consistente en las diferentes pruebas realizadas en este proyecto.

3. Se ha utilizado una frecuencia mínima para crear patrones de uno, así conseguimos que mayor número de patrones ya que basta con que aparezca una única vez para que se cree el patrón.

4. Si la frecuencia mínima hubiera sido mayor, el número de patrones hubiera descendido, ya que para crearse patrón debe repetirse al menos el número de veces que se indique en la frecuencia mínima al ejecutarse la herramienta.

5. Con la ayuda de la herramienta boilerplates la creación de patrones ha sido un éxito. El procesamiento de doscientos cinco documentos a la vez para patrones básicos no fue problema para la herramienta. Posteriormente, los patrones de frecuencia se crearon dos veces diferentes debido a los diferentes escenarios y procesando ciento dieciséis documentos para que no hubiera ningún error de procesamiento.

6. Después de terminar todos los escenarios y analizar los resultados, se puede concluir que los autores escriben artículos sobre un tema (en este caso, bancario) tienen similitud en la forma en que escriben. Más de un 18% de las palabras de los ciento dieciséis documentos son Noun procedentes del dominio bancario y están incluidos en la tabla Vocabulary que la herramienta utiliza para crear patrones. Esto mejoró la creación de patrones de frecuencia para observar las composiciones de los Noun del dominio bancario y cualquier otra categoría gramatical.

7. Los autores de los escritos bancarios utilizar un vocabulario similar y términos apropiados que hace la lectura más fácil porque podemos tener un glosario propio de entender el contenido del documento.

8. El estudio de los patrones facilitará la búsqueda de documentos en los motores de búsqueda o bases de datos.

Las nuevas líneas de trabajo pueden ser las siguientes:

1. Para conocer el número máximo de patrones creados, y evitando los Unclassified Noun, se recomienda ampliar el vocabulario existente en la base de datos. Más ontologías se pueden incluir, símbolos, lenguajes argot, palabras diferentes idiomas, entre otros. Ampliando la tabla Vocabulary en la base de datos dependerá del tema y la cantidad de los documentos que desea analizar.

2. Al crear patrones, es preferible diferenciar los patrones por su semántica, ya que esto evitará tener patrones con el mismo ID y semántica diferente. Es mejor tener nombres distintos para los patrones y esto maximizará la creación y diferentes composiciones de ellos.

3. Se podrían realizar más pruebas utilizando diferentes escenarios como los que se mencionan en este proyecto. Utilizando diferentes frecuencias mínimas en el momento de la creación de patrones ayudará a comparar y analizar resultados. Para proyectos futuros, escenarios de la utilización de una frecuencia mínima de 5, 10, 50, 100, 150, 200 se pueden aplicar para buscar aquella frecuencia mínima que haga no se creen patrones.

6.1. Final conclusions and new lines of work

The use of a domain formed by documents related to banking and different scenarios to create patterns, are the base to conclude the following points:

1. The basic patterns are created while all text documents are analyzed. After the results appear, the evidence is that there are common termtags among the documents as Unclassified Noun, Noun, Number, ...

2. Although the documents are created by diffetent authors and they belong to distinct banks, the created patterns are common in all scenarios. The patterns can be named in a different way, but the composition is consistent in different tests.

3. It has used a minimum frecuency in order to create patterns of one. On this way, we create more number of patterns considering that it is enough with it appears once in order to the pattern is created.

4. If the minimum frecuency had been greater, the number of patterns would have descend. This occurs because, in order to create a pattern, it should be repeated at least the number of times determined by the minimum frecuency.

5. Boilerplates help us to create patterns successfully. The process of two hundred and five documents at the same time is possible thanks to the chosen tool. Next step is create frecuency patterns twice due to different scenarios. For these patterns, it has been procesed a hundred and sixteen documents in order to not had been any errors during processing.

6. After finishing all the scenarios and analyzing the results, it can conclude that articles about a theme (in this case banking) are similar formally. More than a 18% of the words of a hundred and sixteen documents are “Noun” from banking domain. They are included in “Vocabulary” table which is used by the tool in order to create patterns. This fact improves the frecuency pattern creation to observe the compositions of the “Noun” (banking domain) and others grammatical categories.

7. Due to the authors of the bancking documents use similar vocabulary and appropiate terms, the reading is easier because we can have a own glosary to understand the document contents.

8. The study of the patterns will make the documents search easier in search engines or data bases.

The new lines of work could be the followings:

- In order to know the maximum number of created patterns avoiding the termtags “Unclassified Noun”, it is recommended to increase the existing vocabulary in the database. It could be included, for instance, more ontologies, symbols, argot and words in different languages. Expanding Vocabulary table in database depends on theme and quantity of documents to analyze.

- It is preferable in order to create patterns, distinguish the patterns by semantic to avoid patterns with the same ID and diffetent semantic. It is better to have distinct names for the patterns. This maximise the number of patterns and the distinct compositions of them.

- More tests using diffent scenarios as the examples mentioned in this project could be made. Using different minimum frecuencies at the pattern creation, helps to compare and analyze results. In future projects, scenarios with minimum frecuency like 5, 10, 50, 100, 150, 200 could be used. This is good idea to find the minimum frecuency with the pattern creation is impossible.

7. Planificación y coste del proyecto

7.1. Planificación

A continuación incluimos la planificación del proyecto. Con la ayuda de Microsoft Project, obtenemos datos como la duración de cada una de las tareas, la duración del proyecto final, así como el diagrama de Gantt.

La lista de tareas es la siguiente:

Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Definición del proyecto	1 día	mié 24/06/15	mié 24/06/15	
Obtener la documentación	3 días	lun 20/07/15	mié 22/07/15	1
Transformar a txt los 205 documentos	17 días	jue 23/07/15	vie 14/08/15	2
Obtener la ontología de los documentos	6 días	lun 17/08/15	lun 24/08/15	3
Insertar la ontología en la base de datos	1 día	mar 25/08/15	mar 25/08/15	4
Crear los patrones base	1 día	mié 26/08/15	mié 26/08/15	5
Generar los patrones finales del escenario 1	3 días	mié 02/09/15	vie 04/09/15	6
Analizar los resultados del escenario 1	3 días	lun 07/09/15	mié 09/09/15	7
Generar los patrones finales del escenario 2	6 días	jue 10/09/15	jue 17/09/15	8
Analizar los resultados del escenario 2	3 días	vie 18/09/15	mar 22/09/15	9
Realización de la memoria	17 días	mié 23/09/15	jue 15/10/15	10
Realización de la presentación	2 días	vie 16/10/15	lun 19/10/15	11
Estudio de la herramienta	38 días	mié 25/08/15	jue 15/10/15	4

Tabla 58 Planificación de proyecto

El diagrama de Gantt es el siguiente:

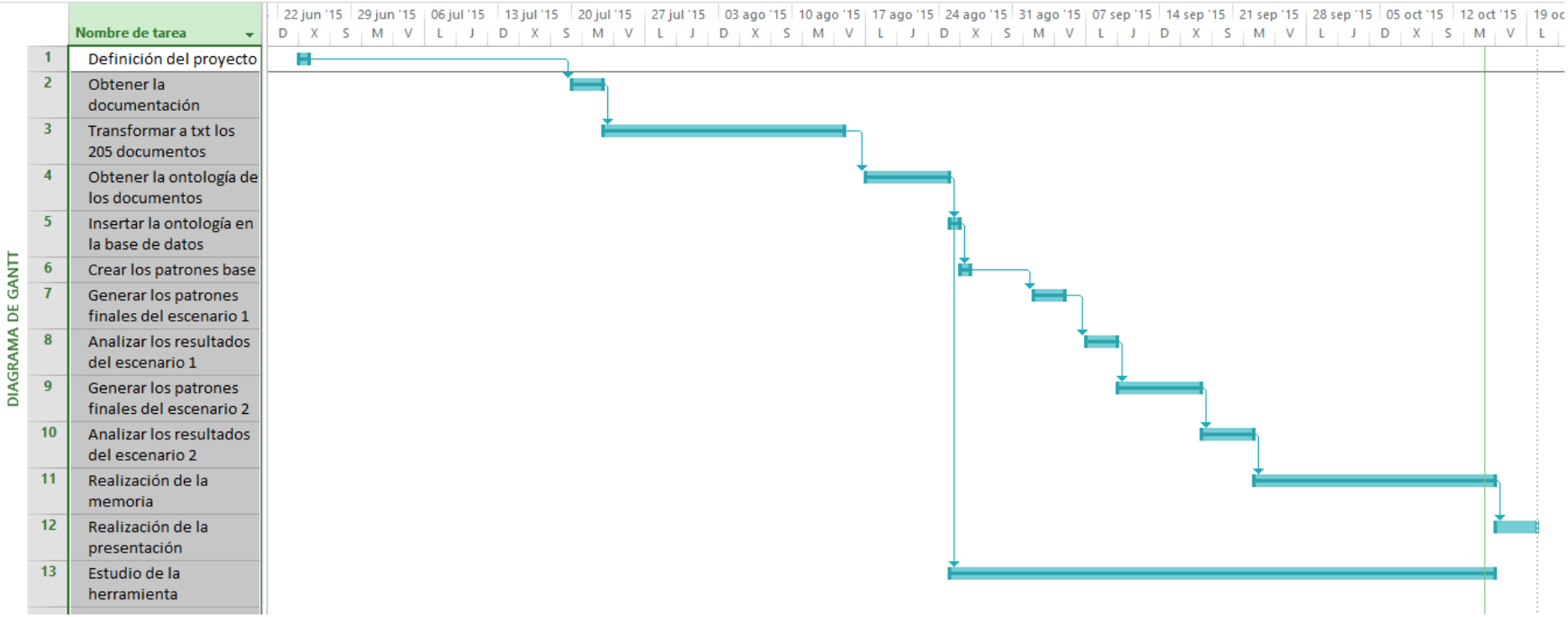


Imagen 31 Diagrama de Gantt

7.2. Coste

El coste total del proyecto se ha estimado en base a los gastos básicos, salario para los trabajadores, y otros gastos.

Los gastos se describen a continuación:

1. Recursos humanos. Los datos de ingeniero en la realización del proyecto han sido los siguientes:

   Duración del proyecto 63 días laborables.
   Jornada de trabajo de 4 horas.
   Semana de trabajo de 5 días.

Gastos del personal			
Concepto	Horas	Precio	Total
Ingeniero	252	20 €/hora	5.040 €
Tutores	10	50 €/hora	500 €
Total			5.540 €

Tabla 59 Gastos del personal

2. Gastos Hardware. Para el desarrollo del proyecto ha sido necesario adquirir un ordenador portátil. El ordenador tiene las siguientes características: Ordenador Lenovo con procesador Intel® Core™ i-5500U CPU @ 2,40GHz, con una memoria RAM de 8,00GB

Gastos Hardware			
Concepto	Unidades	Precio	Total
Ordenador portátil	1	650 €	650 €
Total			650 €

Tabla 60 Gastos Hardware

3. Gastos Software. Para el desarrollo del proyecto hemos necesitado el paquete de Microsoft Office 2013 Proffesional y Microsoft Project 2013 Proffesional.

Gastos Software			
Concepto	Unidades	Precio	Total
Microsoft Office 2013	1	611 €	611 €
Microsoft Project 2013	1	590 €	590 €
Total			1201 €

Tabla 61 Gastos Software

4. Gastos adicionales. Para elaborar el presupuesto final, también hay que tener en cuenta los gastos asociados con el desarrollo del proyecto, tales como los consumibles y gastos de comunicaciones.

Gastos Adicionales		
Concepto	Precio	Total
Folios y encuadernación presentación	100 €	100 €
Cd's memoria	20 €	20 €
Internet	70 €/mes * 4 meses	280 €
Total		400 €

Tabla 62 Gastos Adicionales

A continuación mostramos una tabla con los gastos resumidos:

Gastos	
Concepto	Total
Gastos del personal	5.540 €
Gastos Hardware	650 €
Gastos Software	1.201 €
Gastos Adicionales	400 €
Total	7.791 €

Tabla 63 Gastos resumen

8. Bibliografía

[1] Maes, P. “Agents that Reduce Work and Information Overload”. Communications of the ACM, Vol. 37, Nro. 7, págs. 30-40. 1994.

[2] Baeza-Yates, R. y Ribeiro-Neto, B. “Modern Information Retrieval”. ACM Press. Addison Wesley. 1999.

[3] Salton, G. Y Mc Gill, M.J. “Introducttion to Modern Information Retrieval”. New York. Mc Graw-Hill Computer Series. 1983.

[4] Croft, W.B. “Approaches to intelligent information retrieval.” Information Proccesing & Management, 23, 4, pp. 249-254. 1987.

[5] Korfhage, R. R. “Information Storage and Retrieval”. New York. Wiley Computer Publishing. 1997.

[6] UNESCO, http://databases.unesco.org/thessp/ [Última visita: 15-10-2015]

[7] UNE 50-106-90, 1990: 5. Las normas UNE, editadas en España por la AENOR, son la traducción al castellano de las normas publicadas por la ISO (International Organization for Standarization).

[8] ANSI/NISO Z39.19-2003, 2003: 1. La NISO es una institución que desarrolla, mantiene y publica normas técnicas en el ámbito de la gestión de información a nivel de los Estados Unidos de América.

[9] Aitchison, J and Gilchrist, A. "Thesaurus construction". 2 ed. London: Aslib, 1987.

[10] Slype, G. Van (1991). Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. Madrid, Salamanca: Fundación Germán Sánchez Ruipérez.

[11] López-Huertas, M. J. (1997). "Thesaurus structure design: a conceptual approach for improved interaction". EN: Journal of Documentation 53(2). 139-177

[12] Shiri, A. A.; Revie, C. (2000). "Thesauri on the Web: current developments and trends". EN: Online Information Review 24(4). 273-279. http://arizona.openrepository.com/arizona/bitstream/10150/105440/1/thesauri.pdf [Última visita: 15-10-2015]

[13] Qin, J.; Paling, S. (2001). "Converting a controlled vocabulary into an ontology: the case of GEM". EN: Information Research 6(2). http://informationr.net/ir/6-2/paper94.html [Última visita: 15-10-2015]

[14] Gruber T (1993) A translation approach to portable ontologies. Knowledge Acquisition

[15] R. Studer, R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 25(1–2):161–198, 1998

[16] Qin, J.; Paling, S. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 2000-01, vol. 6, nº 2. Disponible en: http://informationr.net/ir/6-2/paper94.html [Última visita: 15-10-2015]

[17] Qin, J.; Paling, S. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 2000-01, vol. 6, nº 2. Disponible en: http://informationr.net/ir/6-2/paper94.html [Última visita: 15-10-2015]

[18] Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I. y Angel, S. (1977). A Pattern Language: Towns, Buildings, Construction. New York: Oxford University Press

[19] Erich Gamma, Richard Helm, Ralph Johnson y John Vlissides. “Design Patterns: Elements of Reusable Object-Oriented Software” 1994

[20] Appleton, B. (2000). Patterns and Software: Essential Concepts and Terminology.

[21] C. Chambers, B. Harrison, y J. Vlissides. A debate on language and tool support for design patterns. In Proceedings of the 27th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, pages 277–289. ACM Press, 2000.

[22] Pablo Suarez (2013). Sky 2013 “Automatic Generation of Semantic Patterns using Techniques of Natural Language Processing.”

[23] Valeria Rodriguez Barberena (Octubre 2014). Proyecto final de master “Evaluation of a natural language processing system in public health”. Tutores Anabel Fraga y Valentín Moreno.