UNIVERSIDAD CARLOS III DE MADRID
CAMPUS DE COLMENAREJO

EXTRACCIÓN DE PATRONES SEMÁNTICOS EN DOMINIOS ESPECÍFICOS

Trabajo de Fin de Grado
Grado en Ingeniería Informática

Autor:

Javier García Lozano

Tutores:

Valentín Moreno Pelayo

Anabel Fraga Vázquez

Agradecimientos

Llegados ya al punto final de este proyecto, me gustaría echar la vista atrás y agradecer de todo corazón a todos los que me apoyaron, directa e indirectamente, a lo largo de todo el trayecto durante estos tiempos de cambio e incertidumbre para mí.

Quiero dar las gracias a mi familia, por estar conmigo hasta el final, dándome el apoyo y la fuerza necesaria cuando todo parecía venirse abajo y la voluntad para seguir adelante con la misma ilusión y determinación.

Asimismo, me gustaría agradecer a mis compañeros de la Universidad las experiencias, victorias y frustraciones que hemos vivido juntos. Cada uno de vosotros ha refinado mi forma de ser.

Por último, quiero expresar mi agradecimiento a mis tutores, Valentín Moreno Pelayo y Anabel Fraga Vázquez, por su constante apoyo y su aparente infinita paciencia, así como la puntual ayuda de Eugenio Parra Corredor cuando la tecnología a emplear parecía de otro mundo.

Contenido

Índice de figuras

1. Introducción

1.1. Objetivo

En este proyecto se plantea el siguiente problema: ¿es posible que un ordenador sea capaz de interpretar patrones lingüísticos en un documento? Por ejemplo, si se lee:

"*The patient had a temperature of 102F at Saturday*"

"Determinante + Nombre + Verbo + Preposición + Nombre + Preposición + Nombre + Preposición + Nombre"

¿Sería posible reconocer la secuencia sintáctica y semántica de la frase? Y, una vez adquirido este conocimiento, ¿es posible generar patrones que permitan reconocer secuencias similares a esta a lo largo del texto a analizar?

A partir de este problema, se ha realizado el presente Trabajo de Fin de Grado en la Universidad Carlos III de Madrid. Existen dos objetivos fundamentales a perseguir:

- Adquisición de la terminología de un dominio específico.
- Realización del proceso de aprendizaje de patrones lingüísticos más frecuentes a partir de una serie de documentos del dominio de forma automática.

El dominio en el que se enfocará este proyecto será el de la **sordera genética**, y el idioma empleado será el **inglés**.

1.2. Motivación

Este proyecto tiene dos motivaciones fundamentales:

- **Extracción eficaz de información a partir de un documento de texto**. Cuando un humano lee una frase, extrae una parte de información que considera útil. Véase el ejemplo anterior: *"The patient had a temperature of 102F at Saturday"*. Al leer esta frase, el lector retiene el conocimiento de "*102 grados Fahrenheit*" y "*estaba así el sábado*".

  Para un ordenador, la extracción de información relevante es más difícil. No existe un algoritmo trivial que, de la frase que acaba de leer, adquiera el conocimiento de la temperatura del paciente y del día.

  El aprendizaje previo de patrones lingüísticos podría ayudar a resolver este problema. Si el ordenador fuese capaz de reconocer la frase de ejemplo como un patrón que conoce, entonces podría reconocer que la séptima palabra determina la temperatura ("*102F*") y que la última determina el día ("*Saturday*").

  Por tanto, el análisis de patrones ayudaría al sistema a adquirir la información necesaria a partir de un documento de longitud variable. En el dominio de estudio, esto podría contribuir a la generación de diagnósticos de forma rápida y eficiente.

- **Generación de manuales de estilo**. La medicina es un campo en donde la cantidad de información disponible está en constante crecimiento. Además de conocer la información que se presentará al lector de un documento, también es necesario conocer su estructura, con el fin de facilitar la búsqueda y adquisición de datos específicos.

  Por este motivo, es necesario plantearse cómo se escribe un documento asociado al dominio de estudio. ¿Qué familias de palabras son más frecuentes? ¿Qué secuencias lingüísticas se emplean en otros documentos? Estas preguntas y otras más deben ser contestadas para un redactor que vaya a escribir acerca de la sordera genética.

  Por tanto, el estudio de los patrones lingüísticos en este proyecto podría contribuir al aprendizaje del estilo esperado en los documentos del dominio, permitiendo que la información quede lo más organizada y precisa posible.

1.3. Metodología de trabajo

La metodología se basa en el uso de una serie de conceptos específicos al dominio de la sordera genética para reconocer los patrones más frecuentes en una serie de documentos relacionados con el dominio, enfrentándose a los principales problemas que se aprecian en el Procesamiento del Lenguaje Natural (NLP).

En este proyecto se hará uso de la herramienta **BoilerPlates**, desarrollada por el equipo de investigación de la Universidad Carlos III de Madrid. Esta herramienta permitirá el estudio de los patrones lingüísticos dado una serie de conceptos y categorías gramaticales aprendidas previamente.

Por tanto, la metodología se divide en los siguientes pasos:

1. **Adquisición y procesamiento de los documentos del dominio**. Para realizar un estudio acerca de los patrones lingüísticos del dominio, es necesario adquirir documentos que tengan alguna relación con la sordera genética. A continuación, se procesarán de forma que su contenido sea reconocible por la herramienta BoilerPlates.

2. **Extracción de la terminología relacionada con el dominio**. Se realizará consultas en tesauros oficiales relacionadas con las ciencias de la vida para obtener los conceptos asociados a los conceptos “Sordera” y “Genética”. Una vez hecho esto, se procederá a la extracción de la terminología correspondiente.

3. **Adaptación de la terminología a BoilerPlates**. Una vez adquirida la terminología del dominio, se procederá a agrupar los conceptos en semánticas. Después, se preparará toda la información para su inserción en las bases de datos que emplea la herramienta para funcionar.

4. **Uso de la herramienta BoilerPlates**. Una vez insertada la ontología del dominio en BoilerPlates y procesados los documentos sobre la sordera genética, se procederá a realizar el proceso de extracción de patrones. Para ello se identificarán ocho escenarios, cada uno con una configuración distinta de la herramienta. Por cada escenario se obtendrá un conjunto de resultados individual.

5. **Extracción de los resultados**. Una vez terminado el uso de la herramienta, se procederá a adaptar la información contenida para su posterior análisis.

6. **Exposición y análisis de los resultados**.

1.4. Requisitos del trabajo

Para la correcta extracción de los patrones en un dominio específico, es imprescindible el uso de documentación o textos que estén relacionados con el propio dominio.

Ya que se está trabajando sobre la sordera genética, los textos a analizar deberán provenir de fuentes oficiales en el campo de la medicina, como la documentación de hospitales o centros de investigación médica. Asimismo, la terminología que el ordenador empleará para la extracción de patrones debe encontrarse en bibliotecas o tesauros certificados.

Durante el proceso de análisis, se considerará la influencia que podría tener la semántica de las categorías gramaticales que componen un patrón lingüístico. Para ello, se establecerán diferentes escenarios de estudio, en los cuales se realizará la extracción de patrones con la herramienta BoilerPlates utilizando una configuración distinta en cada uno. Se probará a modificar los siguientes parámetros:

- Considerar uso o no uso de semántica.
- Variar la frecuencia mínima de repeticiones para que se considere una secuencia lingüística como un patrón.

Por último, a la hora de extraer los resultados, se obtendrá el número total de patrones generados, así como listas de los 100 patrones más frecuentes que empleen terminología del dominio o no. También se estudiará los conceptos del dominio que más se repiten en los patrones.

La especificación formal de los requisitos del proyecto está adjuntada en el anexo C.

1.5. Estructura del documento

El presente documento está estructurado en apartados, que serán definidos a continuación:

- **Estado del arte**. En esta sección se define la situación actual de la investigación y tecnología correspondiente al contexto de este proyecto.
- **Herramientas empleadas para el proyecto**. En esta sección se describen las herramientas que han sido utilizadas para la elaboración

de este proyecto junto a las condiciones legales a las que están sometidas.

- **Entorno socio-económico y presupuesto.** En esta sección se justifica la viabilidad del proyecto y se calcula el presupuesto asociado.

- **Planificación del proyecto**. En esta sección se define la planificación de las tareas que engloban el proyecto, así como el diagrama de Gantt correspondiente.

- **Arquitectura del proyecto**. En esta sección se plantea el diseño del proyecto.

- **Desarrollo del proyecto**. En esta sección se explica detalladamente el proceso de desarrollo completo de todo el proyecto, así como el uso de las herramientas empleadas para dicho fin.

- **Experimentación**. En esta sección se exponen los resultados obtenidos en los diferentes escenarios de estudio.

- **Conclusiones**. En esta sección se resumen y justifican los resultados de la experimentación.

- **Conclusiones finales**. En esta sección se definen las conclusiones adquiridas al finalizar el proyecto, así como futuras líneas de trabajo.

2. Estado del arte

2.1. Recuperación de la información

A partir de los comienzos de la era digital, la cantidad de información que está a disposición del hombre ha sufrido un crecimiento exponencial. Y, a día de hoy, dicha cantidad sigue aumentando vertiginosamente. Joe Tucci, ex-CEO de EMC, declaró en la EMC World 2014 que los gigabytes quedarán obsoletos como medida ya que el mundo manejará en 2020 un total de 44 Zettabytes[1].

Toda esta información se presenta al usuario de diferentes formas, tales como periódicos digitales, páginas web y libros e informes electrónicos, entre otros. Sin embargo, esto nos plantea un problema: con esta enorme cantidad de datos a nuestra disposición, ¿cómo se puede realizar una extracción eficaz de la información que necesitamos?

Esto nos lleva al problema planteado como "sobrecarga de información" ("*information overload*"). La enorme cantidad de datos y la disponibilidad de las herramientas disponibles para su extracción hacen muy difícil obtener la información que necesitamos obtener. [2,3]

Sin embargo, es un proceso fundamental en nuestra sociedad, tal y como Lewis (1996) afirma: *"Professional and personal survival in modern society clearly depends on our ability to take on board vast amounts of new information. Yet that information is growing at an exponential rate."* [4]

A raíz de este problema surge el concepto de la Recuperación de la Información (*Information Retrieval*, IR), que engloba las técnicas que permiten realizar una extracción eficiente de la información necesaria situadas en bases de datos o cualquier fuente digital.

La IR tiene sus orígenes a finales de los años 40, con los primeros ordenadores capaces de realizar búsquedas de información. En los años 60 se encontrarían estos sistemas en programas de índole comercial o de paso de inteligencia. A día de hoy ha cobrado mayor relevancia gracias al crecimiento de Internet. [5]

Manning, Raghavan y Schütze (2008) definen la IR como "*finding material of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*" [6]

Greengrass (2000) definiría de una forma similar a la IR: "*The discipline that deals with retrieval of unstructured data, especially textual documents, in response to a query or topic statement, which may itself be unstructured.*" [7]

Para Korfhage (1997), la IR es "*la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta.*" [8]

Salton (1983), propondría una definición más abierta. La IR es "*un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información.*" [9]

Por último, cabe mencionar los Sistemas de Recuperación de Información (*Information Retrieval Systems,* IRS). Las IRS son sistemas que almacenan la información de una forma estructurada por medio de documento. A partir de las consultas realizadas por el usuario, es posible extraer dicha información. Algunos ejemplos típicos de IRS serían bases de datos u ontologías.

2.2. Procesamiento de lenguajes naturales

El Procesamiento de Lenguajes Naturales (*Natural Language Processing*, NLP) es una disciplina de la Inteligencia Artificial que engloba la investigación y desarrollo de mecanismos computacionales que permitan establecer una vía de comunicación eficaz entre los ordenadores y el lenguaje humano [10]. Algunos ejemplos del uso del NLP serían los traductores automáticos de internet, reconocimiento de voz o resúmenes automáticos de textos, entre otros.

Los primeros pasos de esta ciencia surgieron en 1950, con la publicación del conocido "Test de Turing", publicado por Alan Turing con el nombre "*Computing machinery and intelligence*", en donde se planteaba la posibilidad de los ordenadores de "tener inteligencia" [11]. En 1956, se presentó el "experimento Georgetown", una colaboración entre IBM y la Universidad de Georgetown que consiguió realizar una traducción muy precisa de 60 oraciones rusas al inglés [12].

Los sorprendentes resultados hicieron despertar un gran interés por parte de los gobiernos a invertir en esta ciencia. Sin embargo, conforme la complejidad de los problemas que planteaba la traducción automática crecía, añadiendo problemas como la polisemia, sinonimia o ambigüedad, así como las reducidas capacidades de los ordenadores de la época, el proceso de investigación se ralentizó de tal manera que los fondos destinados al desarrollo de esta ciencia decrecieron considerablemente [13].

En los últimos años, las aportaciones realizadas han mejorado las expectativas para el NLP, permitiendo procesar grandes cantidades de texto con un grado de eficacia considerable. Además del conocido desarrollo de "Google Translate", existen varios proyectos actuales para mejorar los sistemas de traducción automática, como el proyecto "Molto" de la Unión Europea. [14]

Algunos de los principales desafíos a los que se enfrenta fundamentalmente el NLP son los siguientes:

- **La ambigüedad** [15]. En el lenguaje humano, es frecuente ver una palabra con varios significados. Por ejemplo, “banco” puede entenderse tanto como una empresa comercial como un asiento para personas. Asimismo, las oraciones también pueden presentar problemas de ambigüedad, como sería por ejemplo la frase “Javier bebe un refresco con una pajilla”. El ordenador puede interpretar frases de muchas maneras distintas (en este ejemplo, Javier puede estar “bebiendo” la pajilla y el refresco, o bien Javier podría estar usando la pajilla para beber su refresco), pero es difícil que la interpretación seleccionada sea la correcta.

- **La detección de separación de las palabras** [16]. Esto es un problema que se puede reproducir tanto en el lenguaje hablado como en el escrito. En el lenguaje hablado, las separaciones entre las palabras pueden definir cuál es el sentido de la frase. Adicionalmente, en algunos idiomas escritos, como el chino mandarín, no hay separaciones entre palabras.

2.3. Tesauro

Tesauro es un término derivado del latín que significa “tesoro”. De acuerdo a la definición proporcionada por la UNESCO, un tesauro es el “*Lenguaje documental controlado y dinámico que contiene términos relacionados semántica y genéricamente que abarcan de manera exhaustiva una esfera concreta del conocimiento*” [17].

Para la norma elaborada por la NISO (National Information Standards Organization) un tesauro es “*a controlled vocabulary arranged in a known order and structured so that the various relationships among terms and displayed clearly and identified by standardized relationship indicators.*” [18]

De acuerdo al autor Georges van Slype (1991), un tesauro es “*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*” [19].

2.4. Ontología

De acuerdo a la definición ofrecida por la RAE, una ontología es la “*parte de la metafísica que trata del ser en general y de sus propiedades trascendentales*” [20].

En el ámbito informático, las ontologías son especificaciones de conceptos y sus relaciones entre ellos siguiendo una jerarquía. Las ontologías empezaron a ser aplicadas en campos como la Ingeniería del Conocimiento y el NLP.

Gruber (1993) definiría una ontología como "*una especificación explicita y formal sobre una conceptualización compartida*" [21].

Studer et al (1998) añadirían a la definición de Gruber: "*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*" [22].

El uso de las ontologías en la Ingeniería del Software fue estudiado por varios autores. En concreto, Uschold y Gruninger (1996) determinaron las siguientes utilidades [23]:

- **En la comunicación**: las ontologías permiten mitigar el problema de la ambigüedad, permitiendo una comunicación más eficaz.

- **En interoperabilidad**: permite un traspaso de información más eficiente entre diferentes usuarios.

- **En la ingeniería de sistemas**: el uso de las ontologías puede ayudar al desarrollo de sistemas software, especialmente en la comunicación entre diferentes componentes del sistema.

De acuerdo a Grunninger y Lee (2002), las ontologías pueden ser de utilidad en: [24]

- **En la comunicación**: las ontologías facilitan la intercomunicación entre sistemas computaciones, entre humanos y entre un humano y un sistema computacional.

- **En la inferencia computacional**: las ontologías pueden ser útiles en la representación de planes y el análisis de algoritmos.

- **En la reutilización y organización del conocimiento**: el uso de la ontología permite estructurar la información de la planificación.

Según Guarino (1998), al analizar el uso de una ontología en un sistema se deben considerar tanto el *tiempo en que son utilizadas* y su *aspecto estructural* [25]:

- **Respecto al momento en que son utilizadas**: las ontologías pueden ser creadas durante el desarrollo del sistema ("desarrollo dirigido por ontología) o bien en tiempo de ejecución ("con conciencia de ontología").

- **Según su aspecto estructural**: además del uso frecuente de las ontologías en bases de datos, también pueden aplicarse con éxito en interfaces de usuario y programas de aplicación.

2.5. Biblioteca Nacional de Medicina de Estados Unidos

Con orígenes que datan de 1836, la Biblioteca Nacional de Medicina de Estados Unidos (National Library of Medicine, NLM) es la biblioteca de terminología médica más grande del mundo.

La NLM tiene su sede en Bethesda, Maryland desde 1962, en donde se puede encontrar aproximadamente 19 millones de libros, revistas y manuscritos entre otras fuentes de información sobre la medicina, incluyendo las obras más antiguas de la historia [26]. Su acceso es libre tanto para científicos y médicos como al público general.

Con el avance de la tecnología, la NLM se ha adaptado para facilitar el acceso a sus fuentes de información por medio de tesauros y buscadores en línea. Uno de sus productos más destacados es la base de datos MEDLINE.

MEDLINE es una adaptación digital de más de 26 millones de registros de artículos de revistas de medicina desde 1946 hasta hoy. Esta base de datos se puede consultar y extraer información a partir del motor de búsqueda PubMed, que fue presentado al público por primera vez en enero de 1996.

2.5.1 MeSH

"MeSH" (**Me**dical **S**ubject **H**eadings) es un tesauro de vocabulario terminológico empleado para la indización de artículos relacionados con las ciencias de la vida. Fue introducido en 1960 por la NLM y fue actualizándose cada año en papel hasta 2007 [27].

A día de hoy, MeSH se puede consultar gratuitamente a partir de la página web PubMed y es posible descargarlo en diferentes formatos sin necesidad de registrarse en la página web. Si bien MeSH ha sido traducido en varios idiomas, para este proyecto se empleará la versión en inglés.

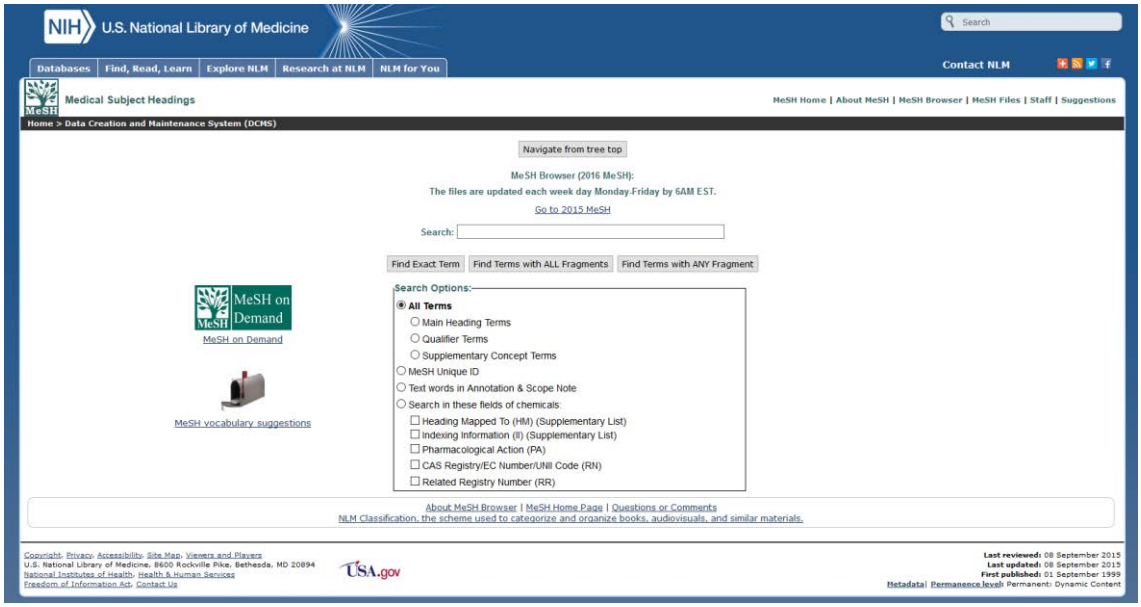

Figura 1: Página de entrada a MeSH 2016 [28]

En la versión de 2016, la empleada para este proyecto, existen un total de 27.883 términos, conocidos como *descriptores*, así como más de 87.000 *términos* asociados a los descriptores. Los descriptores se ordenan mediante el uso de árboles jerárquicos. En la raíz del árbol se encuentran los descriptores más genéricos y, a medida que se avanza por el árbol, se encuentran descriptores más específicos. En total existen 16 árboles de terminología, designados con las letras de la 'A' a la 'N', incluyendo V y Z.

MeSH Tree Structures - 2016

Return to Entry Page

1. + **Anatomy [A]**
2. + **Organisms [B]**
3. + **Diseases [C]**
4. + **Chemicals and Drugs [D]**
5. + **Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]**
6. + **Psychiatry and Psychology [F]**
7. + **Phenomena and Processes [G]**
8. + **Disciplines and Occupations [H]**
9. + **Anthropology, Education, Sociology and Social Phenomena [I]**
10. + **Technology, Industry, Agriculture [J]**
11. + **Humanities [K]**
12. + **Information Science [L]**
13. + **Named Groups [M]**
14. + **Health Care [N]**
15. + **Publication Characteristics [V]**
16. + **Geographicals [Z]**

Return to Entry Page

Figura 2: Árboles jerárquicos en la página web de MeSH [29]

Los descriptores presentan un identificador que permiten establecer su ubicación en el árbol. Los identificadores siguen la siguiente estructura:

<a>XXX.YYY.ZZZ...

Donde:

- <a> define el árbol donde se sitúa el descriptor.
- XXX, YYY y ZZZ definen la sucesión de nodos por la que se ha avanzado en el árbol hasta obtener el descriptor. En este caso, <a>XXX.YYY.ZZZ es nodo hijo de <a>.XXX.YYY y nodo nieto de <a>XXX.

La siguiente figura muestra un ejemplo del árbol jerárquico empezando por el término A03 (Digestive System):

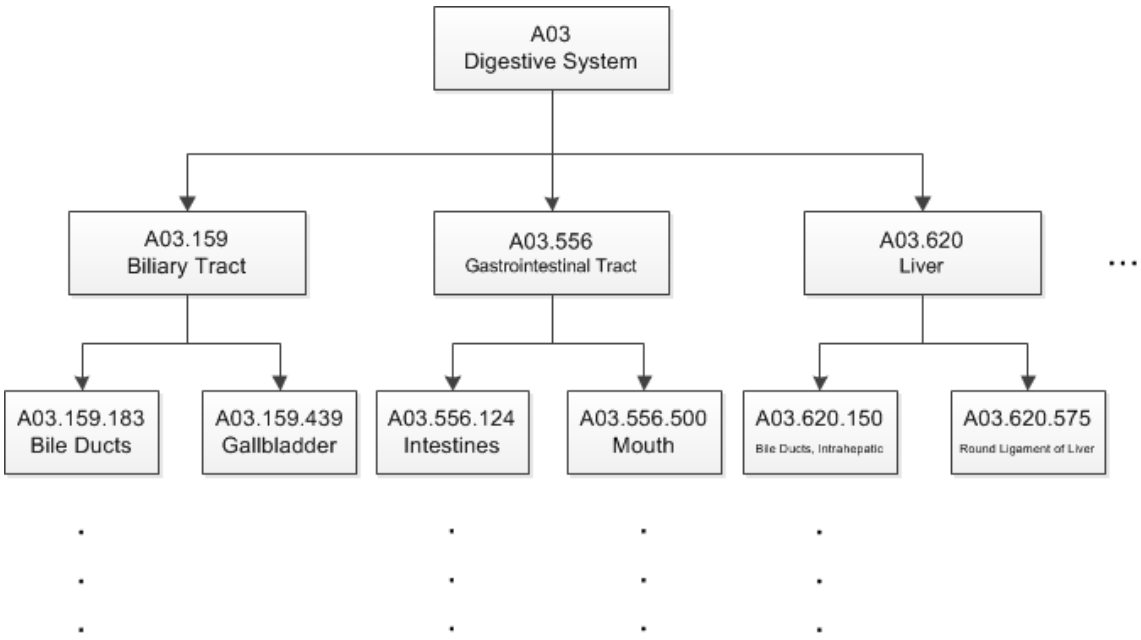

Figura 3. Árbol de conceptos con raíz en A03 (Digestive System)

Cada descriptor contiene la siguiente información, si está disponible:

- **Descriptor** (*MeSH heading*): nombre del descriptor.

- **Identificador** (*Tree Number*): identificador que define la localización del descriptor. Es posible que un descriptor tenga varios identificadores; esto es debido a que aparece en diferentes árboles jerárquicos.

- **Anotaciones** (*Annotation*): muestra las observaciones realizadas en el descriptor.

- **Nota** (*Scope Note*): muestra las observaciones realizadas sobre el propio concepto que define el descriptor.

- **Términos** (Entry Term): Define los términos disponibles para el descriptor.

- **Calificadores** (*Allowable Qualifiers*): MeSH emplea una serie de calificadores que se emplean junto con la jerarquía de árbol para la indización de artículos.

- **Indización anterior** (*Previous Indexing*): define descriptores que anteriormente eran nodos padres del descriptor actual.

- **Véase** (*See also*): presenta otros descriptores que guardan relación con éste, aunque no estén en la misma jerarquía.

- Adicionalmente, se muestran otros campos empleados en la gestión interna del MeSH y, por tanto, no son relevantes para el proyecto.

National Library of Medicine - Medical Subject Headings

2016 MeSH

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Hearing Loss, Sudden
Tree Number	C09.218.458.341.900
Tree Number	C10.597.751.418.341.900
Tree Number	C23.888.592.763.393.341.900
Scope Note	Sensorineural hearing loss which develops suddenly over a period of hours or a few days. It varies in severity from mild to total deafness. Sudden deafness can be due to head trauma, vascular diseases, infections, or can appear without obvious cause or warning.
Entry Term	Deafness, Sudden
Entry Term	Sudden Deafness
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Previous Indexing	Deafness (1966-1978)
Previous Indexing	Deafness, Sudden (1979-2002)
History Note	2003 (1979)
Date of Entry	19780522
Unique ID	D003639

Figura 4: Entrada para el término "Hearing Loss, Sudden"[30]

3. Herramientas empleadas para el proyecto

En esta sección se describen las herramientas utilizadas para el desarrollo de este proyecto junto las posibles condiciones legales a las que estén sometidas.

3.1. Entorno de programación: herramienta Eclipse

Eclipse es un Entorno de Desarrollo Integrado (IDE) de código abierto desarrollado por IBN. Si bien es más conocido por el soporte del lenguaje de programación Java, también ha sido empleado en otros lenguajes como C, C++ o PHP. La versión empleada para este proyecto es Eclipse LUNA.

Eclipse se comenzó a desarrollar en noviembre de 1998. IBN desarrolló una IDE para Java a partir de sus recursos de Object Technology International (OTI). Al principio no fue fácil recibir financiación de sus partners, al no tener certeza de que el producto final fuese eficaz. Por tanto, IBM decidió en noviembre de 2001 adoptar la licencia de código abierto para ayudar a publicitar la herramienta. En 2003, las primeras versiones de Eclipse fueron un éxito y un número considerable de desarrolladores ya lo estaban utilizando para sus programas. En 2004 en adelante, se fundó Eclipse Foundation, una fundación independiente de IBM destinada a continuar el desarrollo y mantenimiento de Eclipse. [31]

La base de Eclipse está basada en la Plataforma de Cliente Enriquecido (Rich Client Platform, RCP). Eclipse destaca por el empleo de un sistema de módulos (*plug-ins*), que permite al usuario utilizar las funcionalidades que desee en lugar de todas las posibles como sucedería en un sistema monolítico. [32]

La siguiente ilustración muestra un esquema simplificado de los subsistemas de la plataforma Eclipse. Cada plataforma está implementada con el uso de al menos un módulo:

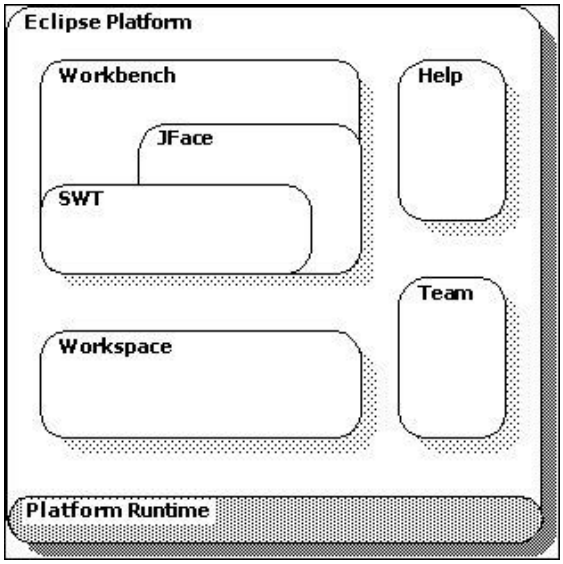

Figura 5. Subsistemas de la plataforma Eclipse

3.1.1 Condiciones de uso

Si bien Eclipse es considerado software libre, no emplea la misma licencia que la Licencia Pública General de GNU (GNU GPL). En su lugar, adopta la Eclipse Public License (EPL) [33].

La EPL y la GNU GPL son similares entre sí. La mayor diferencia se puede apreciar en el tratamiento de juicios por infracción de patentes [34]. En el segundo párrafo de la séptima sección del GNU GPL se puede leer lo siguiente:

"If Recipient institutes patent litigation against a Contributor with respect to a patent applicable to software (including a cross-claim or counterclaim in a lawsuit), then any patent licenses granted by that Contributor to such Recipient under this Agreement shall terminate as of the date such litigation is filed."

Este párrafo fue omitido en EPL por considerarse perjudicial para el crecimiento de Eclipse [35].

3.1.2 Lenguaje de programación: Java

Todas las implementaciones realizadas para este proyecto se han realizado mediante el uso del lenguaje de programación Java.

Como referencia, se ha empleado la guía de estilo propuesta por Oracle Technology Network: *Java Code Conventions* (1997) [36]. El uso de un estándar permite realizar una programación más robusta y facilitar el entendimiento del código para terceros. Si bien esta guía de estilo está planteada para la programación en Java, también es aplicable a otros lenguajes.

A continuación, se exponen algunas recomendaciones de estilo:

- Evitar las líneas de código de tamaño superior a 80 caracteres.

  Si una expresión no cabe en una línea, se deben separar después de una coma o un operador. Las nuevas líneas se comienzan a escribir al mismo nivel que el inicio de la expresión que se está insertando.

- Los métodos deben ir acompañados de comentarios para facilitar su comprensión. Además, es posible introducir breves comentarios en su código correspondiente para mejorar más el entendimiento.

  Sin embargo, también es importante evitar el uso excesivo o la redundancia de comentarios, ya que es frecuente que los comentarios incensarios se queden desactualizados.

- Es recomendable realizar una declaración de variable por línea. Está desaconsejado utilizar una sola línea para declarar múltiples variables.

  No se debe declarar una variable y un método en la misma línea bajo ningún concepto, ni tampoco declarar variables de distinto tipo.

  Es altamente recomendable inicializar variables en el momento de su inicialización.

- Los ficheros Java contendrán una sola clase o interfaz.

3.2. Suite ofimática: Microsoft Office

Microsoft Office es una suite de aplicaciones pensada para su uso en el ámbito de oficina o negocios. Cada aplicación tiene una funcionalidad específica, y tienen la capacidad de comunicarse con el resto de programas del conjunto.

Los comienzos de Office se remontan a noviembre de 1990, cuando MS Office 1.0 salió al mercado para ser empleado en Windows 2.0. En esta época Office sólo disponía de las versiones más básicas de Word, Excel y PowerPoint. Desde entonces, Office siguió creciendo, incorporando nuevas aplicaciones y mejoras a las antiguas. A día de hoy, esta suite de aplicaciones es el software más empleado en Occidente. [37]

En este proyecto se ha empleado la versión 2013 de MS Office. De los programas ofrecidos por el paquete, se han empleado los siguientes:

- **Microsoft Word**. Permite la creación, visualización y edición de documentos escritos. Este programa fue empleado para la elaboración de informes de control y la presente memoria.

- **Microsoft Excel**. Permite la creación, visualización y edición de hojas de cálculo, con las cuales se realizó el estudio estadístico de los resultados obtenidos.

- **Microsoft Access**. Este programa se emplea para la elaboración y uso de una base de datos. Access fue utilizado para realizar la inserción de la terminología en BoilerPlates, así como la extracción de los resultados al completar el análisis.

- **Microsoft Visio**. Este programa permite la creación de diagramas y planos de varias índoles. Visio se utilizó para varios esquemas, visibles en esta memoria.

- **Microsoft Project**. Este programa se emplea para la administración de tiempo y recursos de un proyecto. Project fue utilizado para la planificación y seguimiento de este proyecto.

3.2.1 Condiciones de uso de MS Office

Los programas empleados para este proyecto están sujetos a las condiciones de uso propuestas por el Contrato de Licencia del Software de Microsoft. [38]

Los términos de licencia comercial distinguen dos posibles escenarios dependiendo de cómo se adquirió Office 2013:

- **Términos de Licencia OEM**: estos términos son aplicables si el ordenador en el cual se emplea el software tenía Office preinstalado.

- **Términos de Licencia Comercial**: estos términos son aplicables si el ordenador en el cual se emplea el software no tenía Office preinstalado, por lo que fue adquirido en formato físico o descargado.

El ordenador en el cual se realizó este proyecto no disponía de MS Office preinstalado y fue adquirido en una descarga gratuita gracias a un acuerdo entre la Universidad Carlos III de Madrid y Microsoft que permite adquirir 15 licencias de Office Pro para los alumnos matriculados. Por tanto, este proyecto está sometido a los Términos de Licencia Comercial de MS Office 2013, es decir, el segundo escenario.

3.3. Herramientas internas de la Universidad

3.3.1 BoilerPlates

La herramienta BoilerPlates es una implementación de la patente [39] desarrollada en el lenguaje vb.Net. En la primera concepción de la implementación se tenía como objetivo la generación automática de patrones sobre especificaciones de requisitos de ingeniería. El proceso se desarrollaba teniendo en cuenta todas las categorías sintácticas y semánticas contenidas en la ontología a la que se conectaba.

Con el fin de mejorar el conjunto de patrones resultado tras realizar el análisis, se amplió la funcionalidad de la herramienta permitiendo seleccionar las categorías sintácticas que formaban parte en el proceso. Además, se parametrizó la condición de parada y se permitió la opción de diferenciar de patrones por su semántica. Todas estas medidas permitieron experimentar las mejores condiciones para conseguir conjuntos se patrones más reducidos pero que conseguían representar de igual forma la información de los requisitos.

Para permitir el uso de la herramienta en dominios externos al ámbito de requisitos de ingeniería, se modificó la herramienta para obtener como fuente de entrada documentos en formato de texto plano.

En la versión más reciente de la herramienta se añadió la sustitución de patrones borrados por elementos opcionales y comodines, y una interfaz de la presentación de los resultados.

Esta herramienta ha permitido realizar parte de la experimentación de una tesis doctoral [40], así como de proyectos fin de carrera y trabajos fin de máster [41, 42, 43].

La descripción de la funcionalidad y uso de la herramienta BoilerPlates en este proyecto se describe con más profundidad en las secciones 7.4, 7.5, 7.6 y 7.7.

3.3.2 pdf2txt

Esta herramienta es un programa escrito en Java que permite la conversión de un fichero en formato pdf a un fichero de texto plano (txt).

A partir de uno o varios ficheros en formato pdf, el programa reconoce los textos contenidos en ellos y los inserta en ficheros finales de texto plano, ignorando imágenes o formatos de texto. Se genera un fichero txt por cada fichero pdf.

Puede obtener más detalles sobre su uso en el Anexo D.

3.3.3 Condiciones de uso

Las herramientas BoilerPlates y pdf2txt son herramientas internas desarrolladas por el grupo de investigación de la Universidad Carlos III de Madrid. Su utilización está autorizada para trabajos dirigidos por miembros del mismo, con previa autorización.

4. Entorno socio-económico y presupuesto

4.1. Entorno socio-económico

Una dificultad frecuente en la programación y mantenimiento de sistemas software es la adquisición de la información inicial por medio de ficheros de entrada.

La solución frecuente a este problema es el establecimiento de formatos y plantillas de ficheros, de forma que el ordenador puede adquirir los datos de entrada sin necesidad de “entender” lo que está recibiendo. El principal problema de esta metodología es que es necesario procesar la información inicial de forma que se adapte a las especificaciones del software, lo cual consume más tiempo y esfuerzo.

Además, ya que el proceso de adaptación de la información debe ser realizada por un humano, siempre existe la posibilidad cometer errores durante la tarea de extracción de datos. Por ejemplo, si el documento inicial tiene una extensión de 200.000 palabras, es de esperar que se pierda información relevante durante el proceso, especialmente si el tiempo destinado a dicha tarea es limitado.

Por estos motivos, proporcionar al software la capacidad de “leer” automáticamente los ficheros de entrada y adquirir la información sin necesidad de una extracción previa por parte del usuario puede ser de gran utilidad en diversos campos. En el presente caso de estudio, que es un dominio asociado a la medicina, la capacidad de lectura y adquisición por parte del ordenador podría ser útil en el estudio de casos similares a los de pacientes previos para generar un diagnóstico preciso, o bien podría emplearse para realizar análisis estadísticos de forma eficiente para investigaciones científicas.

También es posible utilizar los resultados adquiridos durante el análisis de patrones con fines didácticos. Al estudiar los patrones lingüísticos en los documentos de un dominio específico, se puede evaluar el estilo de redacción esperado en ellos. De esa manera, sería posible realizar manuales de estilo adaptados a la temática estudiada.

Por tanto, se puede apreciar la utilidad que tendría esta herramienta en empresas y organizaciones de distinta índole, por lo que invertir en esta clase de herramientas puede resultar considerablemente rentable.

4.2. Presupuesto

4.2.1 Sueldo

Se han identificado los roles del personal asociado a este proyecto. En total, se dispone de **1 ingeniero** y **2 tutores**. Por cada rol se ha calculado su coste por hora, así como su coste anual, asumiendo una jornada de **dos horas y media** diarias y **22 días laborables por mes**.

Puesto	€/hora	Coste anual (€)	#
Ingeniero	18	11.880,00	1
Tutor	45	29.700,00	2
	Coste total anual (€)	**71.280,00**	**3**

Figura 6. Presupuesto. Gastos por sueldo

4.2.2 Sueldo por etapas

A continuación, se especifica el tiempo total empleado por cada integrante del personal asignado a este proyecto. A partir del número de horas trabajadas por cada etapa se obtiene el coste total de cada uno.

Puesto Etapa	Tutor 1 (Horas)	Tutor 2 (Horas)	Ingeniero (Horas)
Etapa 0	15	5	2,5
Etapa 1	0	0	65
Etapa 2	30	0	62,5
Etapa 3	2,5	12,5	67,5
Etapa 4	0	0	45
Etapa 5	2,5	2,5	42,5
Etapa 6	0	0	40
TOTAL HORAS	**50**	**20**	**325**
COSTE	**2.250,00**	**900,00**	**5.850,00**
		COSTE TOTAL DEL PERSONAL (€)	**9.000,00**

Figura 7. Presupuesto. Sueldo divido por etapas

4.2.3 Hardware

En este proyecto se ha empleado un ordenador ASUS Desktop PC CG8250 Series con procesador Core i7-2600, 1TB de memoria en disco y 8GB de memoria RAM.

Activo	COSTE (€)
Ordenador de trabajo	902,00
	902,00

Figura 8. Presupuesto. Gastos por hardware

4.2.4 Software

Se ha adquirido y empleado las siguientes herramientas software en el ordenador de trabajo:

Software	COSTE (€)
Microsoft Office 2013*	0,00
Eclipse	0,00
Windows 10	0,00
BoilerPlates**	0,00
Pdf2txt**	0,00
	0,00

Figura 9. Presupuesto. Gastos por software

*Este producto fue adquirido gratuitamente gracias a un acuerdo entre Microsoft y la Universidad Carlos III de Madrid.

**Estos productos fueron proporcionados por la Universidad Carlos III de Madrid para la elaboración de este proyecto.

4.2.5 Otros

En la siguiente tabla se adjuntan otros costes ajenos al personal y el ordenador de trabajo.

Concepto	COSTE (€)
Electricidad (8 meses)	452,00
Conexión a Internet (8 meses)	320,00
CD de memoria	5,00
Paquete de 500 folios DINA4	3,75
Caja de bolígrafos Pilot (12 ud.)	13,00
	793,75

Figura 10. Presupuesto. Otros gastos

4.2.6 Total

Por último, se adjunta el presupuesto final asignado a este proyecto, teniendo en consideración los siguientes costes:

- Costes directos
  - Personal
  - Material (hardware, software y otros gastos)

- Costes indirectos
  - Margen de riesgo del proyecto (establecido al 8%)
  - Beneficio deseado (establecido al 12,5%)
  - Impuesto sobre el Valor Añadido (IVA), actualmente 21%

Concepto	Coste individual (€)	Coste acumulado (€)
Costes directos (personal y material)	10.693,75	10.693,75
Margen de riesgo (8%)	855,50	11.549,25
Beneficio (12,5%)	1.336,72	12.885,97
IVA (21%)	2.245,69	15.131,66
	COSTE FINAL DEL PROYECTO (€)	**15.131,66**

Figura 11. Presupuesto. Gastos totales

5. Planificación del proyecto

El proyecto se ha dividido en las siguientes etapas:

- **Etapa 0: Introducción del proyecto**. Esta etapa corresponde a las reuniones iniciales para introducir el proyecto y establecer los objetivos que se perseguirán, así como instalar los recursos necesarios en el ordenador de trabajo.

- **Etapa 1: Adquisición de la terminología**. Esta etapa corresponde a los procedimientos realizados para encontrar y extraer la terminología del dominio de estudio en un fichero para su posterior integración en la herramienta BoilerPlates.

- **Etapa 2: Agrupación de conceptos en semánticas**. Esta etapa corresponde a la implementación y revisión del proyecto software que permite la agrupación de los conceptos extraídos en la etapa anterior en un número variable de semánticas.

- **Etapa 3: Preparación de la herramienta**. Esta etapa corresponde a la inserción de los recursos necesarios para realizar el análisis correspondiente en BoilerPlates, que son los conceptos agrupados en semánticas y la documentación del dominio previamente procesada.

- **Etapa 4: Uso de la herramienta**. Esta etapa engloba el proceso de ejecución y revisión del funcionamiento de la herramienta BoilerPlates.

- **Etapa 5: Extracción de resultados**. Esta etapa corresponde a la adaptación de los ficheros generados en BoilerPlates para su posterior evaluación y conclusiones.

- **Etapa 6: Documentación**. Esta etapa corresponde a la redacción del presente documento.

La reunión inicial fue convocada el día 14 de julio de 2016, mientras que el comienzo oficial del proyecto equivale al inicio de la primera tarea de la etapa 1, el día 16 de agosto de 2016. El proyecto se dio por concluido el día 24 de enero de 2017, a falta de la redacción de la memoria.

La jornada de trabajo fue de dos horas y media diarias, de lunes a viernes. Al corresponder a un día festivo, los siguientes días no se consideraron en la planificación:

- 12 de octubre
- 1 de noviembre
- 6, 8 y 25 de diciembre
- 2 y 6 de enero

A continuación, se adjunta la planificación del proyecto en forma tabular:

Id	Descripción	Duración	Comienzo	Final	Pred.	Responsable(s)
0	**ETAPA 0: INTRODUCCIÓN DEL PROYECTO**					
1	Presentación y definición inicial del proyecto	1 día	Jue 07/14/2016	Jue 07/14/2016	--	Ingeniero, Tutores
2	Adquisición de documentación del dominio	7 días	Vie 07/15/2016	Lun 07/25/2016	1	Tutores
3	Instalación de recursos software en el ordenador de trabajo	1 día	Vie 07/15/2016	Vie 07/15/2016	1	Ingeniero, Tutores
4	**ETAPA 1: ADQUISICIÓN DE LA TERMINOLOGÍA**					
5	Búsqueda de tesauros con terminología del dominio	4 días	Mar 08/16/2016	Vie 08/19/2016	--	Ingeniero
6	Estudio de metodología de extracción de conceptos	5 días	Lun 08/22/2016	Vie 08/26/2016	3	Ingeniero
7	Adquisición del fichero de conceptos (desc2016.xml)	1 día	Lun 08/29/2016	Lun 08/29/2016	3	Ingeniero
8	Implementación del programa de extracción de conceptos	15 días	Mar 08/30/2016	Lun 09/19/2016	5	Ingeniero
9	Adquisición de los ficheros de conceptos	1 día	Mar 09/20/2016	Mar 09/20/2016	6	Ingeniero
10	**ETAPA 2: AGRUPACIÓN DE CONCEPTOS EN SEMÁNTICAS**					
11	Estudio de metodología de agrupación de conceptos	3 días	Mie 09/21/2016	Vie 09/23/2016	7	Ingeniero, Tutores
12	Implementación inicial del programa de agrupación de conceptos	12 días	Lun 09/26/2016	Mar 10/11/2016	8	Ingeniero
13	Adaptación del programa de agrupación a las especificaciones de la herramienta BoilerPlates	9 días	Jue 10/13/2016	Jue 10/20/2016	9	Ingeniero, Tutores
14	Generación final de los ficheros de semánticas	1 día	Vie 10/21/2016	Vie 10/21/2016	10	Ingeniero
15	**ETAPA 3: PREPARACIÓN DE LA HERRAMIENTA**					
16	Inserción de los conceptos en la herramienta BoilerPlates	12 días	Lun 10/24/2016	Mie 11/09/2016	11	Ingeniero
17	Conversión de los documentos del dominio a formato .txt	4 días	Jue 11/10/2016	Mar 11/15/2016	2	Ingeniero
18	Separación de los	2 días	Mie 11/16/2016	Jue 11/17/2016	13	Ingeniero

	documentos en lotes					
19	Identificación de los escenarios de estudio e información a extraer	5 días	Vie 11/18/2016	Jue 11/24/2016	--	Ingeniero, Tutores
20	Elaboración de un proceso de extracción de prueba	4 días	Vie 11/25/2016	Mie 11/30/2016	12, 14	Ingeniero
21	**ETAPA 4: USO DE LA HERRAMIENTA**					
22	Extracción de patrones en los ocho escenarios de estudio	18 días	Jue 12/01/2016	Mie 12/28/2016	16	Ingeniero
23	**ETAPA 5: EXTRACCIÓN DE RESULTADOS**					
24	Estudio de metodología de agrupación y extracción de resultados	1 día	Jue 12/29/2016	Jue 12/29/2016	17	Ingeniero, Tutores
25	Implementación de los programas de adaptación y agrupación de resultados	4 días	Vie 12/30/2016	Jue 01/05/2016	18	Ingeniero
26	Procedimiento de extracción de los resultados de cada escenario de estudio	8 días	Lun 01/09/2016	Mie 01/18/2016	19	Ingeniero
27	Análisis de los resultados obtenidos	4 días	Jue 01/19/2016	Mar 01/24/2016	20	Ingeniero
28	**ETAPA 6: DOCUMENTACIÓN**					
29	Redacción de la memoria	16 días	Mie 01/25/2016	Mie 02/15/2016	--	Ingeniero

Figura 12. Planificación del proyecto en forma tabular

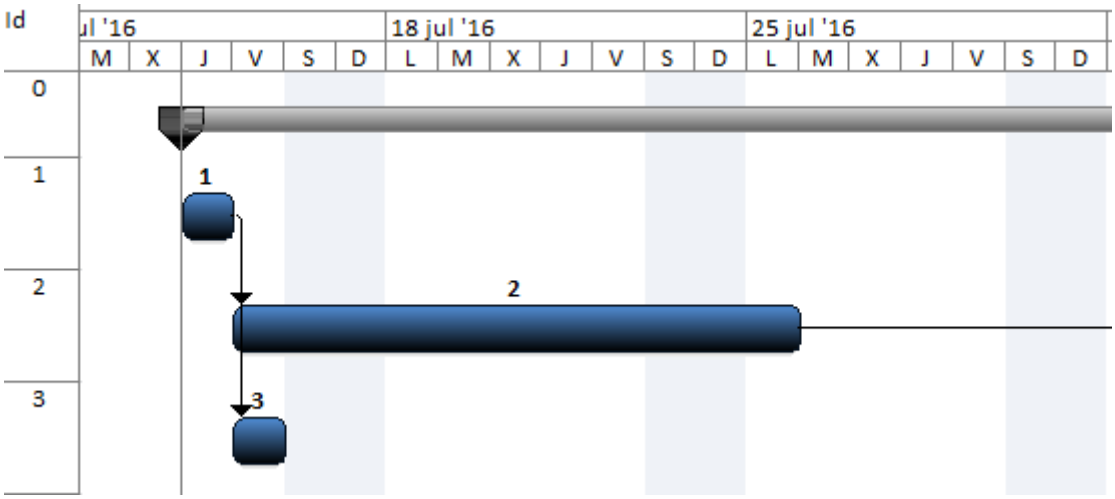

Figura 13. Diagrama de Gantt. Etapa 0

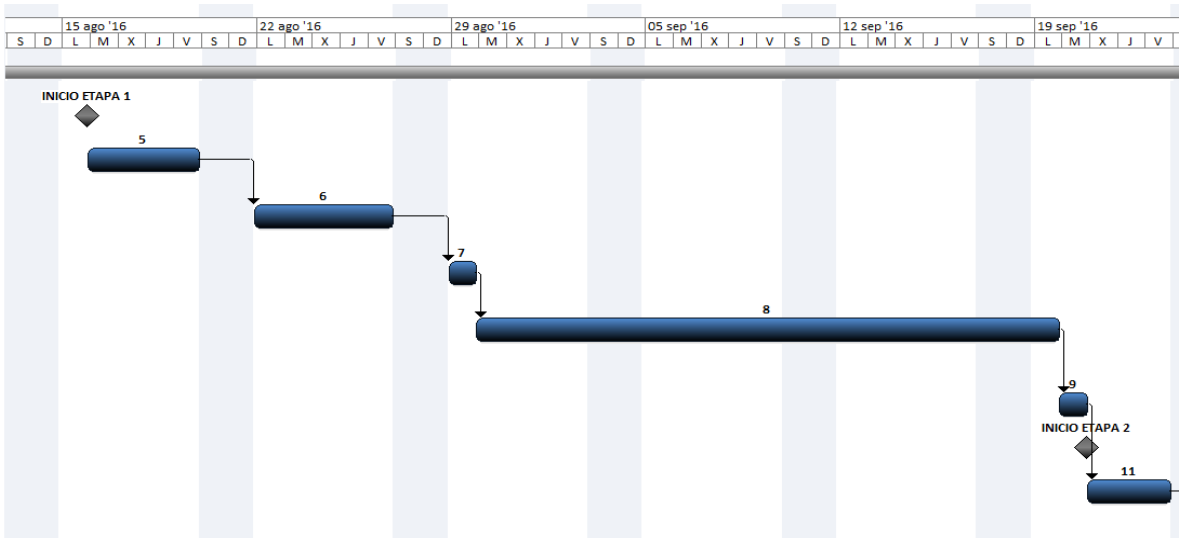

Figura 14. Diagrama de Gantt. Etapa 1

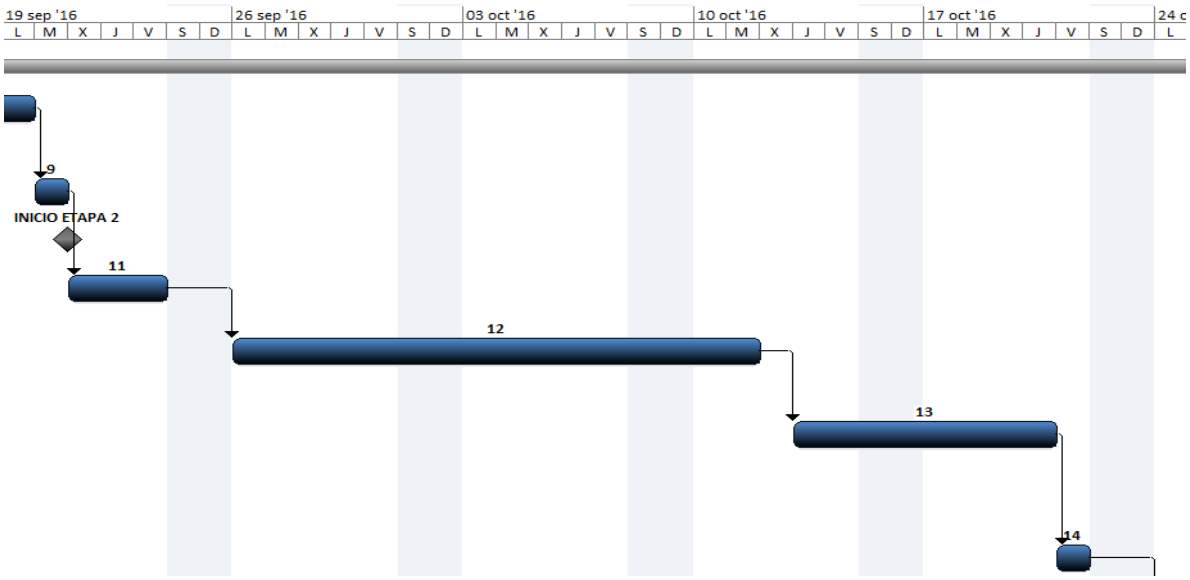

Figura 15. Diagrama de Gantt. Etapa 2

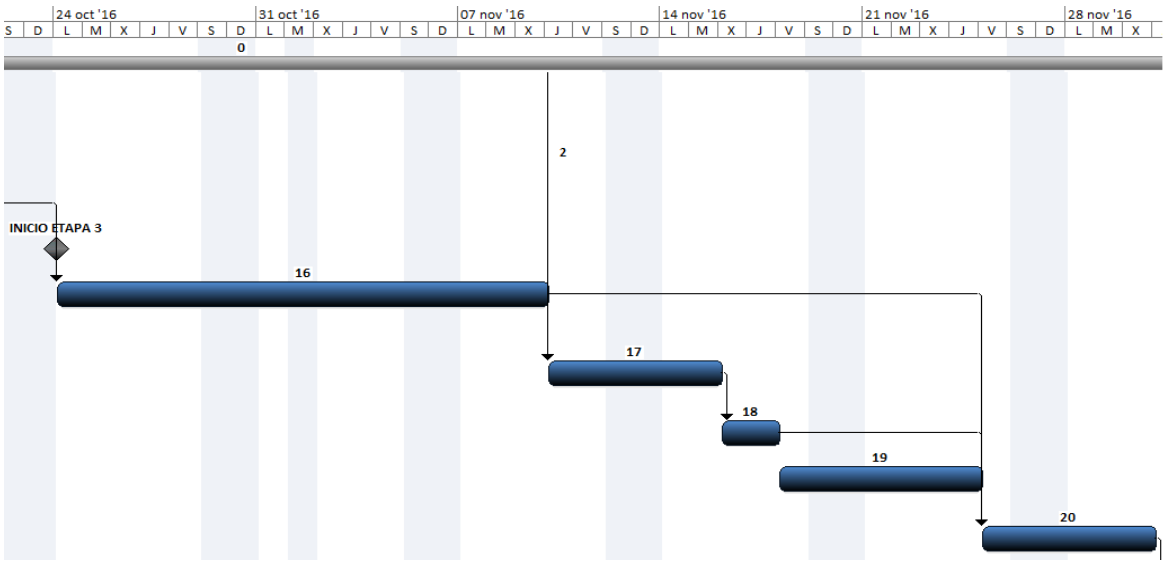

Figura 16. Diagrama de Gantt. Etapa 3

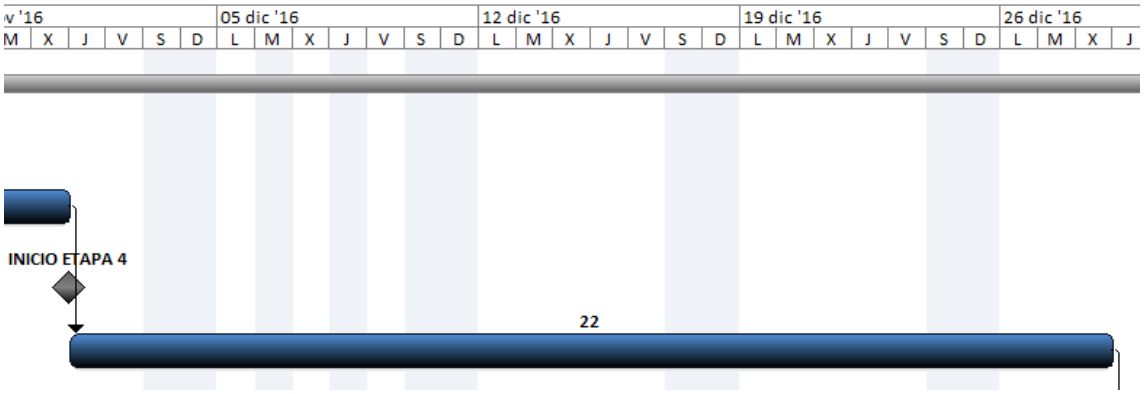

Figura 17. Diagrama de Gantt. Etapa 4

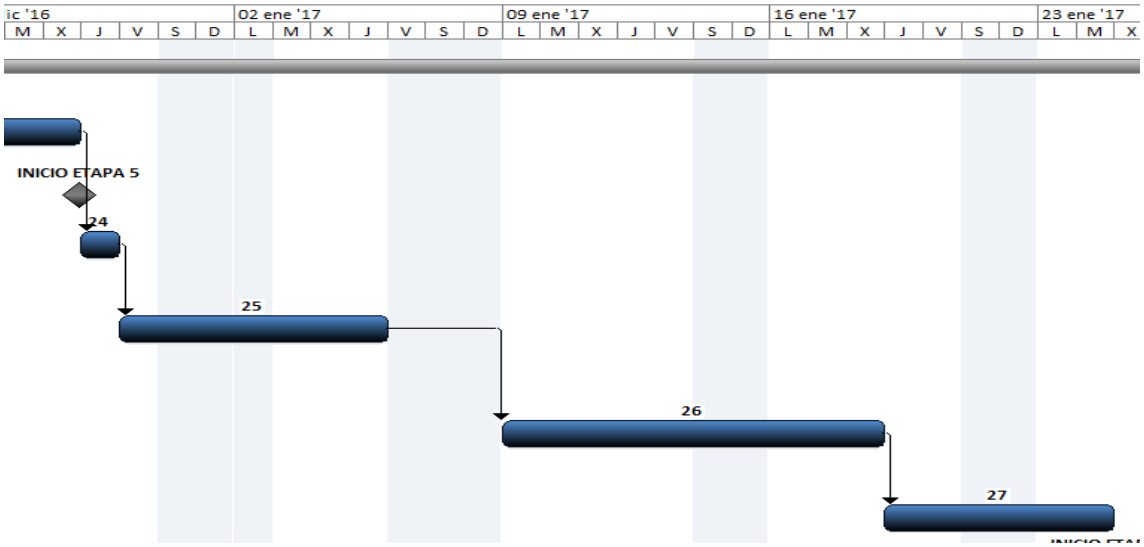

Figura 18. Diagrama de Gantt. Etapa 5

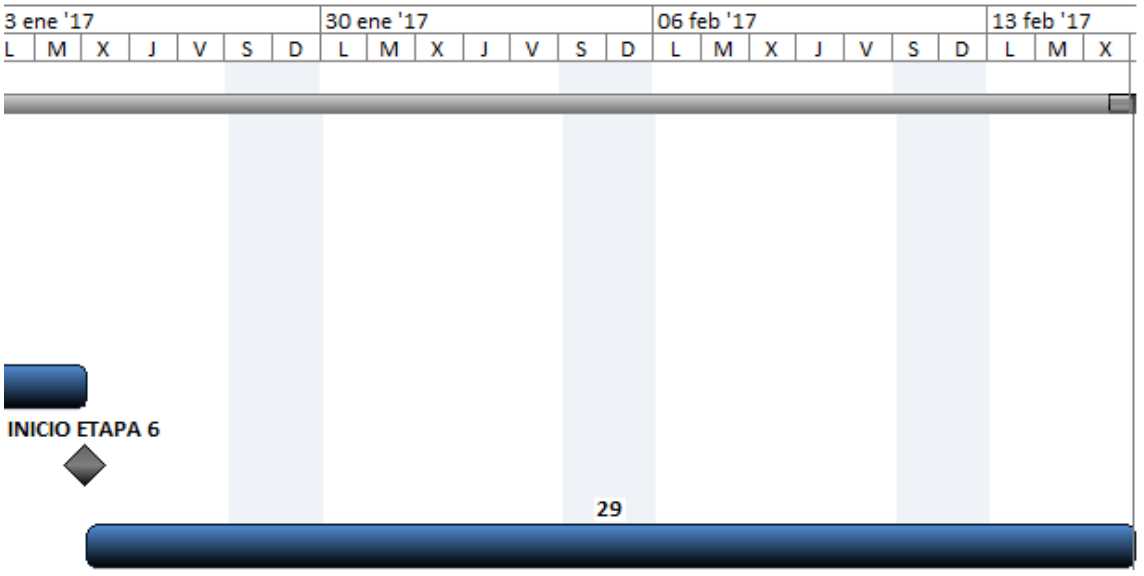

Figura 19. Diagrama de Gantt. Etapa 6

6. Arquitectura del proyecto

En esta sección se definirán los agentes y componentes que componen la arquitectura del proyecto, definiendo su funcionalidad y el paso de información entre sí.

6.1. Esquema general del proyecto

El esquema gráfico del proyecto es el siguiente:

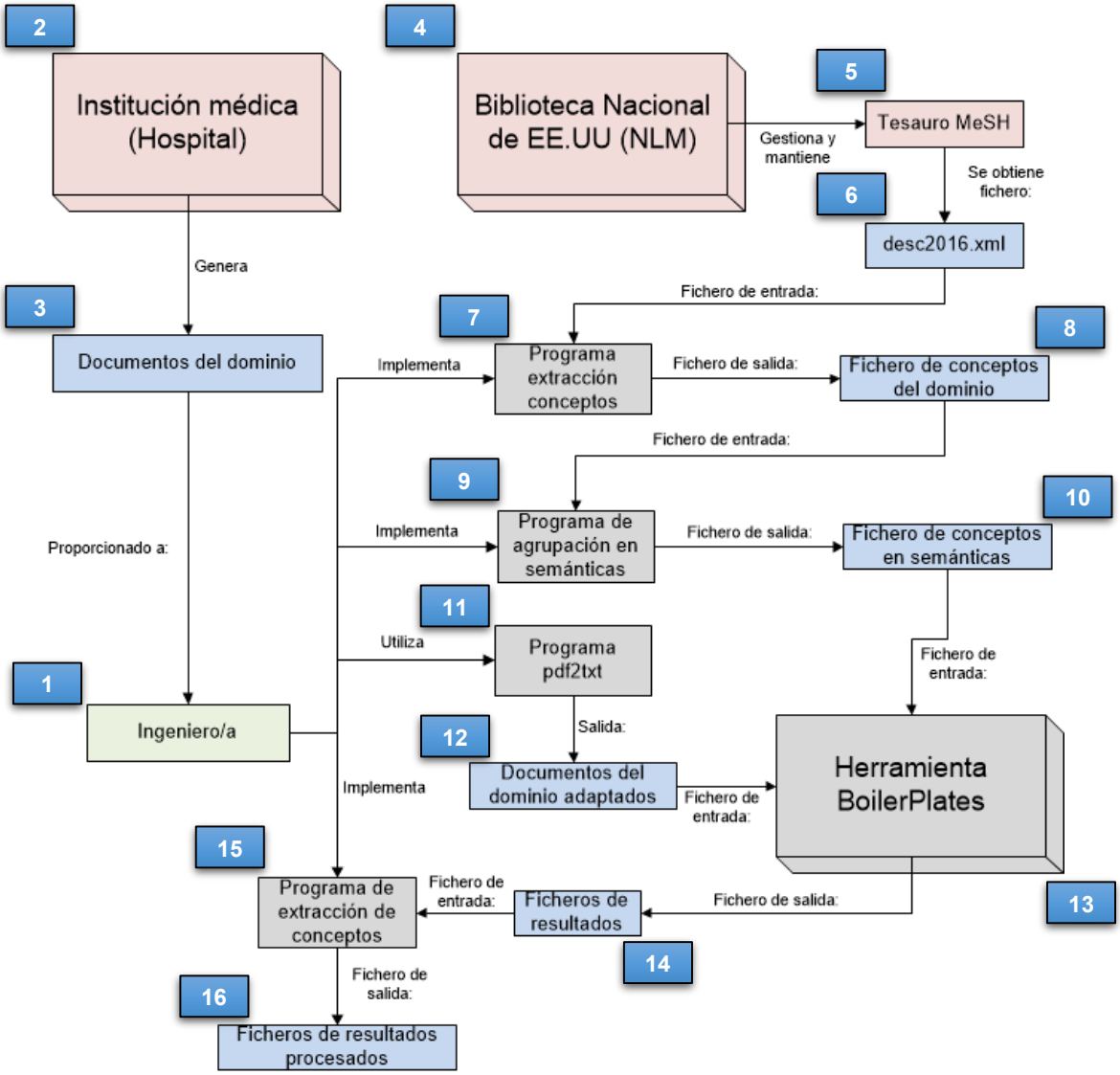

Figura 20. Esquema general del proyecto

En el siguiente apartado se definirá cada componente de forma individual.

6.2. Descripción detallada de los componentes

Id: **1**	Nombre del componente/agente: **Ingeniero/a**	Tipo: **Agente interno**
Descripción:	Este agente es el desarrollador principal del proyecto.	
Objetivo(s):	Controlar el proceso completo de extracción de patrones lingüísticos, desde la adquisición del vocabulario del dominio de estudio hasta el análisis final de los resultados. Para ello, realizará la implementación de los programas necesarios para cada fase del proyecto.	
Relaciones (entrada):	• Adquisición de los documentos del dominio (**2**)	
Relaciones (salida):	• Implementación del programa de extracción de terminología (**5**) • Implementación del programa de agrupamiento en semánticas (**7**) • Implementación de los programas de extracción de resultados de BoilerPlates (**13**)	
Referencia:	Secciones 7.2, 7.3, 7.8	

Figura 21. Descripción del agente: Ingeniero/a

Id: **2**	Nombre del componente/agente: **Institución médica (Hospital)**	Tipo: **Agente externo**
Descripción:	Este agente representa un hospital del cual se obtendrán los documentos relacionados con el dominio de estudio.	
Objetivo(s):	Proporcionar un lote de documentos oficiales que correspondan al dominio de estudio.	
Relaciones (entrada):	• Ninguna	
Relaciones (salida):	• Proporciona los documentos del dominio (**3**)	
Referencia:	Sección 7.1	

Figura 22. Descripción del agente: Institución médica (Hospital)

Id: **3**	Nombre del componente/agente: **Documentos del dominio**	Tipo: **Fichero de datos**
Descripción:	Estos ficheros corresponden al lote de documentos del dominio.	
Objetivo(s):	Representar el contenido y estilo de escritura esperados para el dominio de estudio.	
Relaciones (entrada):	• Obtenidos a partir de una institución médica (**2**)	
Relaciones (salida):	• Son empleados por el/la agente “Ingeniero/a” (**1**).	
Referencia:	Sección 7.1	

Figura 23. Descripción del componente: Documentos del dominio

Id: **4**	Nombre del componente/agente: **Biblioteca Nacional de Medicina de los EE.UU. (NLM)**	Tipo: **Agente externo**
Descripción:	Este agente externo representa la institución consultada para realizar la adquisición de la terminología.	
Objetivo(s):	Mostrar los conceptos de medicina de forma jerárquica y de fácil acceso.	
Relaciones (entrada):	• Ninguna	
Relaciones (salida):	• Desarrollo y mantenimiento del tesauro MeSH (**5**)	
Referencia:	Sección 7.2	

Figura 24. Descripción del agente: NLM

Id: **5**	Nombre del componente/agente: **MeSH**	Tipo: **Componente externo**
Descripción:	Este componente es el tesauro desarrollado por la NLM para el acceso gratuito a los conceptos relacionados con la medicina en general.	
Objetivo(s):	Generar un fichero en formato XML para la posterior extracción de términos del dominio a un formato de texto plano.	
Relaciones (entrada):	• Creación y mantenimiento realizado por el NLM (**3**)	
Relaciones (salida):	• Fichero de conceptos de medicina (desc2016.xml) (**6**)	
Referencia:	Sección 7.2	

Figura 25. Descripción del componente: MeSH

Id: **6**	Nombre del componente/agente: **desc2016.xml**	Tipo: **Fichero de datos**
Descripción:	Este fichero representa el contenido completo del tesauro MeSH escrito en lenguaje de marcas.	
Objetivo(s):	Mostrar todos los conceptos que componen el tesauro, así como los sinónimos y terminología relacionada de cada uno.	
Relaciones (entrada):	• Fichero obtenido a partir de MeSH (**5**)	
Relaciones (salida):	• Este fichero se emplea en la extracción de los conceptos específicos del dominio de estudio (**7**)	
Referencia:	Sección 7.2	

Figura 26. Descripción del componente: desc2016.xml

Id: **7**	Nombre del componente/agente: **Algoritmo de extracción de conceptos**	Tipo: **Implementación interna**
Descripción:	Programa empleado para, a partir de un fichero de conceptos, extraer sólo aquellos que correspondan al dominio de estudio.	
Objetivo(s):	Adquirir los conceptos del dominio de estudio, con sus sinónimos y terminología asociada, y exportarlos a un fichero de texto plano.	
Relaciones (entrada):	• Recibe como entrada el fichero desc2016.xml (**6**)	
Relaciones (salida):	• Genera como salida un fichero plano con los conceptos específicos del dominio (**8**)	
Referencia:	Sección 7.2	

Figura 27. Descripción del componente: Algoritmo de extracción de conceptos

Id: **8**	Nombre del componente/agente: **Fichero de conceptos del dominio**	Tipo: **Fichero de datos**
Descripción:	Este fichero representa el vocabulario específico del dominio de estudio en texto plano.	
Objetivo(s):	Representar los conceptos del dominio y su descripción correspondiente de forma que sea fácil de leer y procesar.	
Relaciones (entrada):	• Es un fichero generado a partir del algoritmo de extracción de conceptos (**7**)	
Relaciones (salida):	• Se emplea como entrada al algoritmo de agrupamiento en semánticas (**9**)	
Referencia:	Secciones 7.2, 7.3	

Figura 28. Descripción del componente: Fichero de conceptos del dominio

Id: **9**	Nombre del componente/agente: **Algoritmo de agrupamiento en semánticas**	Tipo: **Implementación interna**
Descripción:	Este programa representa un conjunto de clases que permiten agrupar los conceptos del dominio en semánticas.	
Objetivo(s):	Agrupar los conceptos del dominio en un número variable de semánticas, de forma en que los conceptos de una temática específica se encuentren en el mismo grupo.	
Relaciones (entrada):	• Recibe como entrada el fichero de conceptos del dominio (**8**)	
Relaciones (salida):	• Genera como salida el fichero de semánticas del dominio (**10**)	
Referencia:	Sección 7.3	

Figura 29. Descripción del componente: Algoritmo de agrupamiento en semánticas

Id: **10**	Nombre del componente/agente: **Fichero de semánticas**	Tipo: **Fichero de datos**
Descripción:	Este fichero representa los conceptos del dominio agrupados en semánticas de diversas temáticas.	
Objetivo(s):	Mostrar las semánticas generadas y los conceptos insertados en cada una de forma que sea fácil de leer y procesar.	
Relaciones (entrada):	• Es un fichero generado a partir del algoritmo de agrupamiento en semánticas (**9**)	
Relaciones (salida):	• Se emplea como entrada para la herramienta BoilerPlates (**13**)	
Referencia:	Secciones 7.3, 7.4	

Figura 30. Descripción del componente: Fichero de semánticas

Id: **11**	Nombre del componente/agente: **Programa pdf2txt**	Tipo: **Componente interno**
Descripción:	Este componente representa el programa empleado para convertir los documentos del dominio a un formato aceptado por BoilerPlates.	
Objetivo(s):	Realizar la conversión al formato .txt del mayor número de documentos posibles.	
Relaciones (entrada):	• Es un programa proporcionado al ingeniero (**1**)	
Relaciones (salida):	• Genera los ficheros del dominio adaptados (**12**)	
Referencia:	Sección 7.1	

Figura 31. Descripción del componente: Programa pdf2txt

Id: **12**	Nombre del componente/agente: **Ficheros del dominio adaptados**	Tipo: **Fichero de datos**
Descripción:	Este conjunto de ficheros es obtenido después de utilizar la herramienta BoilerPlates.	
Objetivo(s):	Representar los textos de los documentos del dominio en un formato legible por BoilerPlates.	
Relaciones (entrada):	• Es el lote de ficheros generado por pdf2txt (**11**)	
Relaciones (salida):	• Se emplea como entrada para la herramienta BoilerPlates (**13**)	
Referencia:	Sección 7.1	

Figura 32. Descripción del componente: Ficheros del dominio adaptados

Id: **13**	Nombre del componente/agente: **Herramienta BoilerPlates**	Tipo: **Componente interno**
Descripción:	Es una herramienta empleada para la extracción de patrones a partir de la terminología insertada en su base de datos.	
Objetivo(s):	Realizar la extracción de patrones a partir de los documentos recibidos y la terminología del dominio.	
Relaciones (entrada):	• Recibe como entrada el fichero de semánticas (**10**) • Recibe como entrada los documentos del dominio de estudio adaptados (**12**)	
Relaciones (salida):	• Genera los ficheros de patrones descubiertos a lo largo del proceso de extracción (**14**)	
Referencia:	Secciones 7.4, 7.5, 7.6, 7.7	

Figura 33. Descripción del componente: Herramienta BoilerPlates

Id: **14**	Nombre del componente/agente: **Ficheros de patrones descubiertos**	Tipo: **Fichero de datos**
Descripción:	Este conjunto de ficheros representa los patrones descubiertos tras utilizar la herramienta BoilerPlates.	
Objetivo(s):	Representar los resultados obtenidos por la herramienta BoilerPlates para su posterior procesamiento y análisis.	
Relaciones (entrada):	• Los ficheros son generados por la herramienta BoilerPlates (**13**)	
Relaciones (salida):	• Los ficheros deben ser adaptados (**16**) antes de realizar su análisis. Para ello se emplean programas de adaptación (**15**)	
Referencia:	Sección 7.8	

Figura 34. Descripción del componente: Fichero de datos

Id: **15**	Nombre del componente/agente: **Programas de adaptación**	Tipo: **Implementación interna**
Descripción:	Son implementaciones realizadas para la extracción de los resultados de BoilerPlates.	
Objetivo(s):	Procesar los resultados obtenidos por BoilerPlates para realizar el análisis de los patrones descubiertos.	
Relaciones (entrada):	• Recibe como entrada los ficheros generados por BoilerPlates (**12**)	
Relaciones (salida):	• Los ficheros quedarán adaptados para extraer la información que se empleará en el análisis de resultados (**16**)	
Referencia:	Sección 7.8	

Figura 35. Descripción del componente: Programas de adaptación

Id: **16**	Nombre del componente/agente: **Ficheros de resultados adaptados**	Tipo**:** **Fichero de datos**
Descripción**:**	Este conjunto de ficheros representa los resultados adaptados para el análisis.	
Objetivo(s):	Representar los patrones descubiertos por BoilerPlates de forma que se pueda extraer la información necesaria para el análisis.	
Relaciones (entrada):	• El fichero fue generado a partir de los programas de adaptación (**15**)	
Relaciones (salida):	• Ninguna	
Referencia:	Secciones 7.8, 8	

Figura 36. Descripción del componente: Ficheros de resultados adaptados

7. Desarrollo del proyecto

En esta sección, se explicarán todos los procesos realizados desde la situación inicial del proyecto hasta la extracción de los resultados.

7.1. Búsqueda y adaptación de documentos del dominio

Para realizar el proceso de extracción de patrones sobre la sordera genética, es necesario adquirir previamente documentos que correspondan al dominio de estudio.

Para ello, se solicitó al Servicio de Genética del Hospital Ramón y Cajal una serie de publicaciones científicas pertenecientes a ámbito de la sordera genética. La Unidad 728 del CIBERER de enfermedades raras realizó la compilación de 867 documentos digitales en formato pdf.

7.1.1 Conversión a txt

Una vez realizada la adquisición de los documentos del dominio, es necesario convertirlos a un formato que la herramienta de extracción de patrones sea capaz de reconocer. En este caso, es necesario volcar la información de los documentos al formato .txt.

Con este propósito, se empleó la herramienta **pdf2txt**. A partir de los ficheros en formato pdf en la carpeta de entrada, el programa genera ficheros en formato txt en la carpeta de salida. Consulte el Anexo D para obtener más información acerca de pdf2txt.

Para nuestro proyecto, se insertó todos los documentos del dominio en la carpeta de entrada y se ejecutó la herramienta. Durante el proceso, se pudo observar que algunos documentos no fueron procesados debido a que estaban cifrados.

En total, se pudo convertir un total de 663 documentos a txt mediante el uso de la herramienta.

7.1.2 Lotes de documentos

Una vez generados los ficheros txt, el siguiente paso era agruparlos en uno o varios ficheros para que la herramienta BoilerPlates los procese.

La idea original era almacenar los 663 documentos procesados en un solo fichero txt, con el objetivo de reducir el coste de realizar la extracción de

patrones. Sin embargo, el archivo que se generaría de ese modo es demasiado grande: la herramienta BoilerPlates puede procesar textos de una longitud máxima aproximada de dos millones de caracteres, mientras que la compilación de todos nuestros ficheros tenía más de cuatro millones de caracteres.

En su lugar, se decidió realizar lotes de documentos. Cada lote contendría un número fijo de documentos procesados, de forma que la herramienta pudiese procesarlos. Se estableció que cada lote contendría 50 ficheros, el último teniendo menos al no ser una división exacta. El inconveniente de esta alternativa es que aumentará el coste temporal de realizar una extracción de patrones para todos los escenarios posibles.

Para realizar este proceso, se realizó un programa Java que realizase este proceso. Su funcionalidad es sencilla:

1. Se insertan los ficheros en una carpeta de entrada.
2. Se lee cada fichero, volcando su contenido en un fichero de salida.
3. Al llegar a 50 lecturas, se crea un nuevo fichero de salida y se sigue introduciendo los datos en él.
4. Se repiten los pasos 2 y 3 hasta que se hayan leído todos los ficheros.

A continuación, se adjunta una tabla con los resultados obtenidos con el programa:

Nombre del lote	Nº de documentos	Caracteres
merged1.txt	50	5.677.056
merged2.txt	50	1.843.200
merged3.txt	50	1.572.864
merged4.txt	50	1.622.016
merged5.txt	50	1.941.504
merged6.txt	50	1.884.160
merged7.txt	50	1.540.096
merged8.txt	50	1.695.744
merged9.txt	50	1.835.008
merged10.txt	50	1.695.744
merged11.txt	50	2.211.840
merged12.txt	50	1.736.704
merged13.txt	50	1.810.432
merged14.txt	13	411.026

Figura 37. Longitud de los lotes de documentos generados

Por lo general, la varianza de caracteres entre lotes es pequeña, exceptuando por los lotes 1 y 14. El lote 14 es de tamaño muy reducido en comparación por el resto, si bien es justificable ya que sólo representa 13 documentos.

En cuanto al lote 1, el tamaño total es muy superior al resto, más de un 150% más largo que el segundo lote más extenso. Es posible que esto sea debido a que los primeros ficheros en procesarse eran también los más largos.

7.2. Extracción de los términos del MeSH

7.2.1 Localización de la terminología

El primer paso para la elaboración del estudio es extraer la terminología necesaria para generar la base de datos de BoilerPlates. Para ello, se empleará la información ofrecida por MeSH (explicado más atrás) para realizar las siguientes tareas:

- Identificar la posición de los descriptores relacionados con nuestro dominio.
- Extraer el nombre de los descriptores seleccionados, así como sus términos y sinónimos potenciales.

Ya que el dominio específico de la medicina que trata este proyecto es sobre la *sordera genética*, se hizo una consulta directa a la página web de MeSH sobre los términos “sordera” (“deafness” o “hearing loss”) y “genética” (“genetics”). Al final, se concluyó que los subárboles que contienen los descriptores necesarios tienen como nodo raíz “C09.218.458.341” para la sordera y “G05” para la genética.

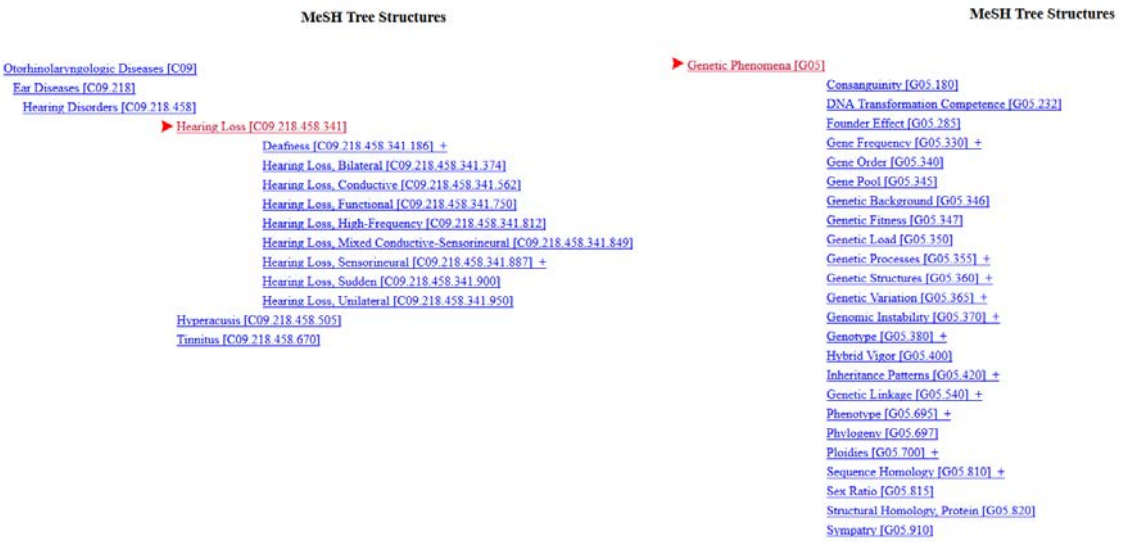

Figura 38: Árboles jerárquicos de los descriptores a extraer [44, 45]

Una vez obtenidas localizaciones de los descriptores, el siguiente paso sería extraer los conceptos y sus términos para su posterior inserción en la base de datos.

Sin embargo, aquí surge un problema: si bien el número de descriptores en total para la sordera genética es lo suficientemente bajo para realizar la extracción de forma manual (17 descriptores) mediante el uso de un procesador de texto, el árbol correspondiente a la genética es considerablemente grande (294 descriptores).

Debido a la gran cantidad de conceptos a procesar, se consideró el uso de alguna herramienta obtenida de Internet o el uso de una funcionalidad implementada en la página web que permitiese realizar la extracción de forma manual. Sin embargo, no se encontró ningún programa pensado para MeSH y la página web de MeSH no ofrece dicha posibilidad.

7.2.2 Extracción de la terminología

Ante esta situación, se decidió que sería necesario codificar un programa que permitiese la exportación de los descriptores. Se optó por el lenguaje de programación **Java** ya que se disponía de mayor experiencia programando con dicho lenguaje.

El primer paso fue descargar la base de datos MeSH en formato XML a partir de la página web. El fichero, **desc2016.xml**, tiene un tamaño de aproximadamente 280MB descomprimido y su extensión es de más de nueve millones de líneas de texto en lenguaje de marcas.

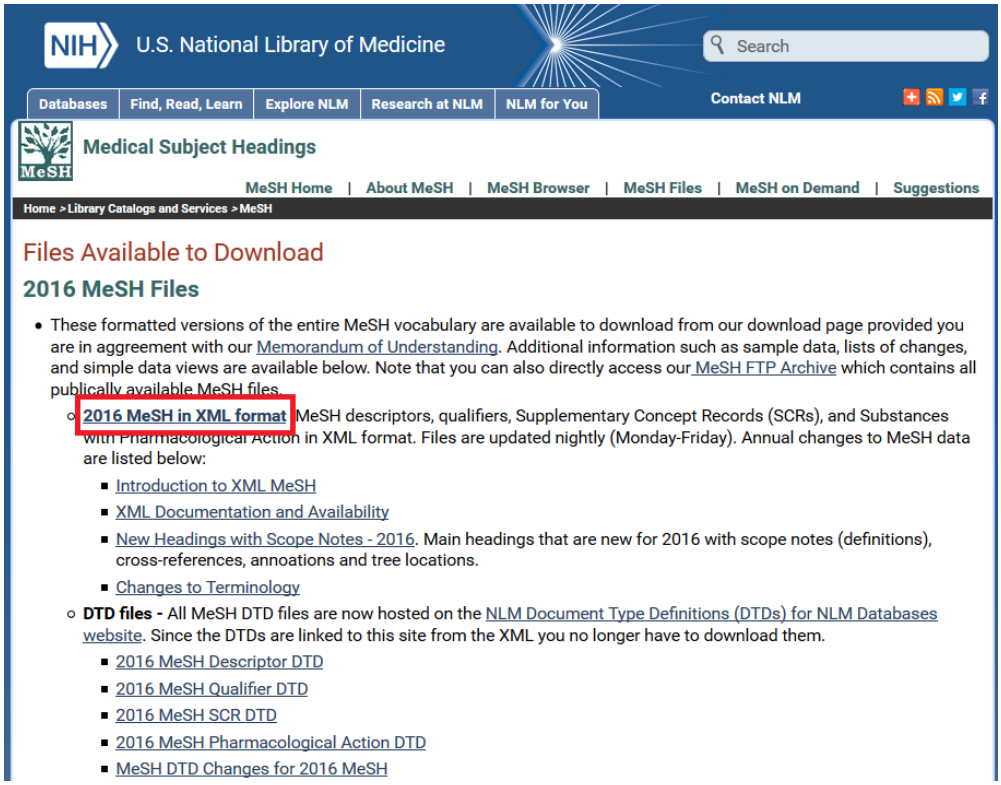

Figura 39: Enlace de descarga del fichero desc2016.xml [46]

Al estudiar el texto contenido en el fichero, se identificaron los “tags” que contienen la información necesaria para nuestro proyecto, que son los siguientes:

- **<DescriptorRecord>(…)</DescriptorRecord>**: define el inicio y fin de un descriptor.

- **<ConceptName>(…)<ConceptName>**: nombre del concepto. Es posible que haya más de un nombre de concepto. El primero se establecerá como el nombre del descriptor y el resto se considerarán sinónimos.

- **<TreeNumber>(…)</TreeNumber>**: identificador del descriptor. Es posible que haya varios identificadores. El primero de ellos es el que define su posición en el árbol jerárquico correspondiente.

- **Abbreviation>(…) </Abbreviation>**: calificador(es) del descriptor.

- **<ScopeNote>(…)</ScopeNote>**: información del descriptor.

- **<Annotation>(…)</Annotation>**: anotaciones del descriptor.

- **<TermUI>(…)</TermUI>**: términos asociados al descriptor.

Durante el estudio del fichero desc2016.xml se descubrió la existencia de un problema adicional: los descriptores no están ordenados por orden jerárquico, por lo que no será posible reducir el cómputo del programa estudiando sólo secciones determinadas del fichero. En su lugar, será necesario estudiar todos los conceptos contenidos en él para asegurarse de que no falta ningún descriptor sin extraer tras el procedimiento.

7.2.3 Implementación del programa de extracción de descriptores

Para todas las implementaciones realizadas en el curso de este proyecto se ha creado un proyecto en Eclipse conocido como “MeSHParser”. Cada algoritmo implementado se ha introducido en un paquete independiente para acceder a su contenido con facilidad.

El primer programa implementado, “customParser”, permite estudiar el fichero desc2016 (insertado en el proyecto Eclipse) para exportar la información de todos los descriptores cuyo identificador sea o comience por una constante string (*subCategory*) establecido previamente, con un coste computacional relativamente bajo (aproximadamente 10 segundos).

Por ejemplo, si se establece el valor de *subCategory* a “C09.218.458.341”, el algoritmo extraerá el descriptor asociado a dicho identificador y al resto de nodos contenidos en el subárbol. Los resultados serán exportados a un fichero diferente en formato txt.

El algoritmo sigue los siguientes pasos:

Apertura de ficheros: se abren los ficheros desc2016.xml y un fichero de salida (output). Además, se crea un buffer de lectura para desc2016 y un buffer de escritura para el fichero output.

Declaración de estructuras de datos: para almacenar temporalmente la información de los descriptores, se emplearán variables de tipo ArrayList<string> y string. Dichas variables son las siguientes:

- *conceptName*: variable donde se almacena el nombre del descriptor.
- *description:* variable donde se almacena la descripción.
- *notes:* variable donde se almacena las notas asociadas al descriptor.
- *nodes*: variable donde se almacenan los identificadores del descriptor.

- *qualifiers*: variable donde se almacenan los calificadores del descriptor.

- *terms*: variable donde se almacenan los términos del descriptor.

- *synonyms:* variable donde se almacenan los posibles sinónimos del descriptor.

Procedimiento: se empieza a leer desc2016 y se realizará una acción según la línea procesada. Mientras no se llegue al final del fichero:

- Si se lee <TreeNumber>, se introduce el nodo contenido en el "tag" en la variable *nodes*. Se comprueba si el descriptor está contenido en el árbol jerárquico que se desea extraer: si es el primer nodo leído de este descriptor corresponde al árbol jerárquico establecido con la constante *subCategory*, se registra este hecho con una variable booleana.

- Si se lee <Abbreviation>, se introduce el calificador contenido en el "tag" en la variable *qualifiers*.

- Si se lee <ScopeNote>, se introduce la descripción contenida en el "tag" en la variable *qualifiers*.

- Si se lee <Annotation>, se introduce las notas contenidas en el "tag" en la variable *notes*.

- Si se lee <ConceptName>, se introduce el nombre del descriptor en el "tag" en la variable *conceptName*. Si no es el primer nombre encontrado en el descriptor, es decir, se ha leído otro "tag" <ConceptName> en el mismo descriptor, se añade en la variable *synonyms*.

- Si se lee <TermUI>, se introduce el término contenido en el "tag" en la variable *terms*.

- Al leer </DescriptorRecord> se habrá llegado al final de la entrada para el descriptor. Entonces se realizará lo siguiente:

    - Si el descriptor está contenido en el árbol jerárquico correcto, entonces se vuelca la información contenida en las variables del sistema en el fichero de salida, siguiendo el formato como aparece en la *Figura 40: Formato de los descriptores tras su extracción*. Después, se reinician las variables de datos.

    - Si el descriptor no está contenido en el árbol jerárquico, se reinician las variables de datos sin volcar previamente los datos en el fichero de salida.

```
CONCEPT: Deafness
    Node(s): C09.218.458.341.186  C10.597.751.418.341.186  C23.888.592.763.393.341.186
    Description: A general term for the complete loss of the ability to hear from both ears.
    Notes: differentiate from HEARING LOSS, BILATERAL; see various specific terms under HEARING LOSS
    Term(s): Deafness ; Bilateral Deafness ; Deafness, Bilateral ; Hearing Loss, Complete ; Complete ; Complete Hearing L
    Qualifiers: PS  IM  MI  DI  PP  RA  RI  RH  EN  ET  MO  EP  PC  BL  CF  CI  CN  DH  DT  TH  PX  UR  VI  US  VE  HI  C
    Possible synonym(s): Prelingual Deafness ; Deafness, Acquired ; Hearing Loss, Extreme

---
```

Figura 40: Formato de los descriptores tras su extracción

Al terminar el algoritmo, el fichero output contendrá todos los descriptores solicitados. Ya que los conceptos “sordera” y “genética” están en distintos árboles jerárquicos, se ha realizado el procedimiento dos veces, obteniendo dos ficheros de salida con los descriptores de cada uno.

7.3. Agrupación de los conceptos en semánticas

Una vez realizado el proceso explicado en el apartado anterior, se disponen de dos ficheros de conceptos relativos a la sordera y a la genética, respectivamente, y siguen el formato mostrado en la *Figura 40: Formato de los descriptores tras su extracción*.

Antes de comenzar la inserción de los conceptos en la base de datos de BoilerPlates, es necesario realizar un paso intermedio, que consiste en la agrupación de los conceptos en un número indefinido de semánticas.

Para realizar el correspondiente análisis semántico durante la experimentación, los conceptos deben tener relaciones semánticas entre ellos. Realizar la agrupación de forma manual sería una labor demasiado laboriosa, y exigiría disponer de un mayor conocimiento del dominio de estudio. Por tanto, se planteó cómo generar las semánticas de forma automatizada y eficiente, utilizando sólo la información adquirida hasta este punto del proyecto.

Ya que los conceptos extraídos del MeSH disponen de un código identificador, que indica su posición en el árbol jerárquico correspondiente, es posible realizar una agrupación eficaz a partir de ellos. El algoritmo planteado para resolver este problema sería el siguiente, escrito en lenguaje informal:

1. Escoger un concepto y obtener su código identificador, por ejemplo: “G05.365.590”.

2. Verificar cuántos conceptos están contenidos en el código especificado. Por ejemplo, si el código es “G05.365.590”, se buscarán todos los conceptos cuyo código sea “G05.365.590.XXX…”

3. Se calcula cuántos conceptos se han obtenido:

a. Si se han encontrado al menos un número específico (X) de conceptos, se crea una semántica con ellos. Los conceptos añadidos ya no podrán ser utilizados en otras semánticas.

b. Si no, se debe repetir el paso 2, reduciendo el nivel de profundidad mínima de búsqueda. Si el código de búsqueda era "G05.365.590", ahora se convertirá en "G05.365".

4. Este proceso se repite hasta que no queden conceptos sin insertar o no se pueda completar el proceso para los conceptos restantes.

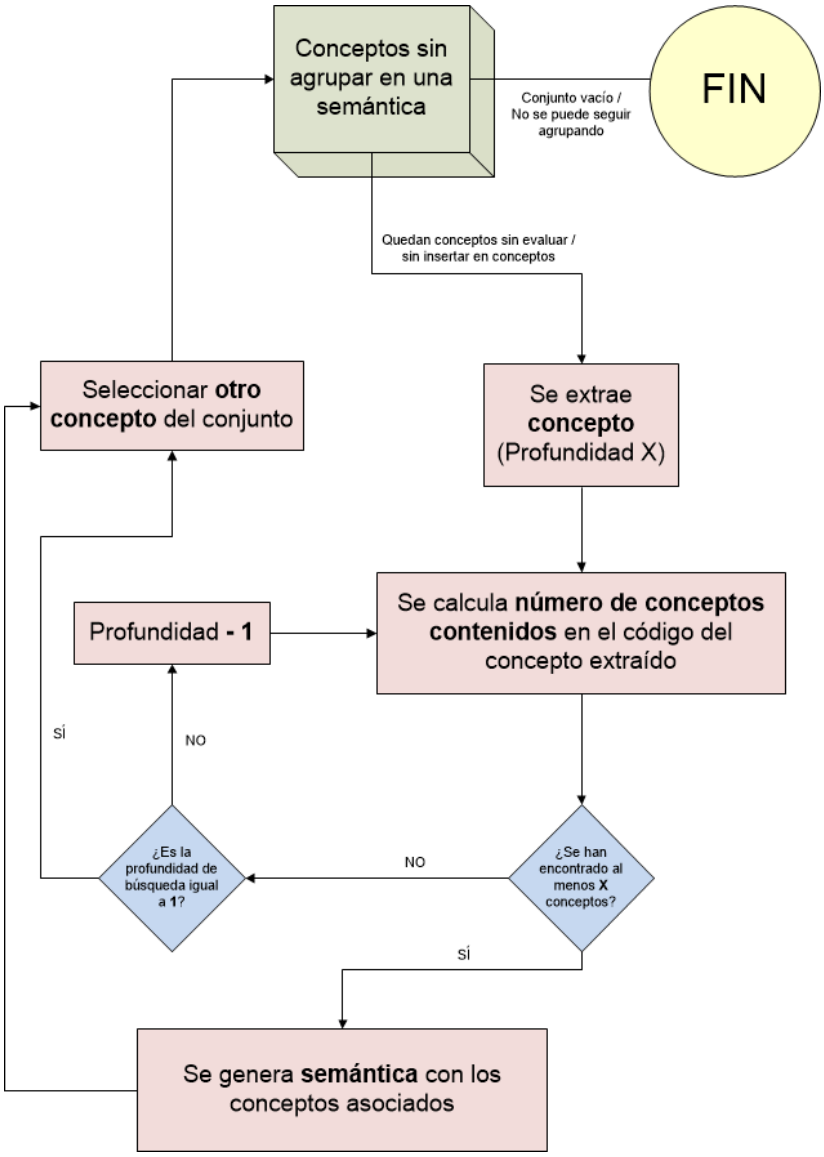

Figura 41. Diagrama de flujo del algoritmo de agrupación en semánticas

Para este proyecto, se decidió que el mínimo número de conceptos (la variable X definida anteriormente) para formar una semántica será **5**. Asimismo, con el objetivo de minimizar la posibilidad de tener conceptos sin

agrupar, el procedimiento se iterará varias veces, empezando con los conceptos más específicos hasta los más generalistas.

Adicionalmente, se detectaron algunos problemas adicionales que el programa debía resolver:

- Los conceptos contienen una lista de términos asociados. Se detectó que en ocasiones se repetían algunos términos en la lista, lo que podría dificultar el proceso de extracción de patrones mediante BoilerPlates. Por tanto, es necesario asegurarse que no haya términos repetidos en un concepto.

- Se decidió que los sinónimos asociados a cada concepto también se considerarán términos. Por tanto, deberán ser insertados en la lista de términos correspondiente.

La salida deseada sería un fichero de texto (txt) que contenga las semánticas reconocidas, cada una con la siguiente información:

- Identificador y nombre genérico de la semántica. Cada semántica contiene los conceptos agrupados en ella.

- Por cada concepto, se mostrarán sus términos asociados.

El programa para la agrupación en semánticas se ha implementado en un nuevo paquete del proyecto, denominado “groups”. En las siguientes subsecciones se explicarán las tres clases que componen el programa y su funcionalidad.

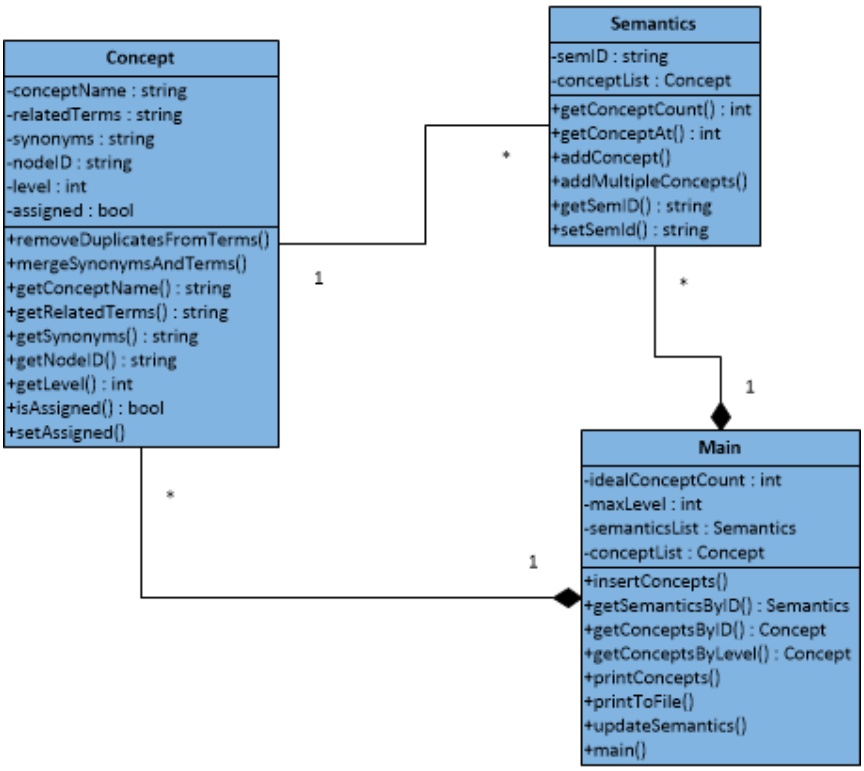

Figura 42. Esquema de las clases del algoritmo de agrupación

7.3.1 Programa de agrupación en semánticas: la clase “Concept”

Esta clase define las estructuras de datos necesarias para almacenar un concepto y su estado en el procedimiento.

La clase “Concept” está formada por los siguientes atributos:

- **string conceptName**: Nombre del concepto.
- **string relatedTerms**: Términos asociados al concepto. Cada término está separado por “;”.
- **string synonyms**: Sinónimos asociados al concepto.
- **string nodeID**: Identificador del concepto en el árbol jerárquico del MeSH.
- **int level**: Nivel de profundidad del concepto en el árbol. Este valor se obtiene a partir del código identificador. Por ejemplo, el código “G05.365.590” indica que la profundidad tiene el valor 3.

- **boolean assigned**: Determina si el concepto ha sido asociado a una semántica (*true*) o no (*false*).

La clase “Concept” contiene los siguientes métodos:

- **public Concept(string cN, string rT, string s, string nID, int l)** : Constructor de la clase**.**
- **public void removeDuplicatesFromTerms()** : Este método permite eliminar términos duplicados en el concepto. La variable *relatedTerms* quedará actualizada, sin repeticiones en sus términos.
- **public void mergeSynonymsAndTerms()** : Este método inserta los sinónimos en la lista de términos del concepto. La variable *relatedTerms* quedará actualizada, añadiendo el valor de la variable *synonyms*.
- **getters y setters**: Métodos empleados para obtener y actualizar los valores de los atributos.

7.3.2 Programa de agrupación en semánticas: la clase “Semantics”

Esta clase representa las semánticas, que son las agrupaciones lógicas de los conceptos.

La clase contiene los siguientes atributos:

- **string semID**: identificador de la semántica. Este valor será el código de búsqueda con el que se hayan encontrado al menos el número mínimo de conceptos.
- **ArrayList<Concept> conceptList**: es una estructura de datos en donde se insertarán los conceptos de la semántica.

Los métodos encontrados en esta clase son los siguientes:

- **public Semantics(String sID)** : Constructor de la clase.
- **public int getConceptCount()** : Devuelve el número de conceptos asociados a la semántica.

- **public Concept getConceptAt(int index)** : Devuelve el concepto situado en una posición (index) dada de la lista de conceptos. Devuelve *null* si el valor de index no es válido.

- **public void addConcept(Concept c)** : Añade un concepto a la semántica. Para ello, se establece el atributo *assigned* del concepto a *true* y se inserta en *conceptList*.

- **public void addMultipleConcepts(ArrayList<Concept> c)** : Este método permite insertar varios conceptos al mismo tiempo en la semántica. El procedimiento es similar al explicado en *addConcept()*.

- **getters y setters**: Métodos empleados para obtener y actualizar los valores de los atributos.

7.3.3 Programa de agrupación en semánticas: la clase "Main"

Esta clase representa el algoritmo de agrupación de conceptos en semánticas. Para ello, emplea las clases *Semantics* y *Concept*, definidas anteriormente.

En primer lugar, se deberá obtener el fichero de conceptos del dominio a partir del fichero desc2016.xml. Este proceso fue definido en el apartado "Extracción de la terminología".

El procedimiento se divide en los siguientes pasos:

Declaración de variables y estructuras, Se declaran dos variables de tipo ArrayList: una representará los conceptos leídos del fichero de conceptos (conceptList) y la otra representará las semánticas reconocidas durante la agrupación (semanticsList). Asimismo, se definen los ficheros de entrada y salida y los buffers necesarios para su procesamiento.

Lectura del fichero de entrada y generación de la lista de conceptos. Se abre el fichero de términos extraído de MeSH y se realiza una de las siguientes acciones según la línea leída:

- Si se lee "CONCEPT: …" se habrá leído el nombre del nuevo concepto. Se registra el nombre contenido en esta línea.

- Si se lee “Node(s): …” se habrá leído los identificadores en el árbol jerárquico del concepto. Para ello, se extrae el primer identificador leído en la línea (el resto quedan ignorados).

  A continuación, se establece el nivel de profundidad del concepto. Para ello, se lee el número de puntos (“.”) en el código de identificación. Por ejemplo, “G05.365.590” indica que la profundidad tiene el valor 3. A lo largo del proceso de lectura de conceptos se registrará el nivel de profundidad más alto de todos.

- Si se lee “Term(s): …” se habrán leído los términos del concepto. Se registran los términos leídos.

- Si se lee “Possible synonym(s): …” se habrán leído los sinónimos del concepto. Se registran los sinónimos leídos.

- Una vez que se llega al final de un concepto (marcado como “---“), se crea un objeto de tipo *Concept* con la información extraída y se almacena en la lista de conceptos.

Este proceso se repetirá hasta llegar al final del fichero. Una vez finalizada la lectura, todos los conceptos adquiridos de MeSH habrán sido insertados en la lista.

Proceso de agrupación en semánticas. Una vez obtenidos los conceptos a partir del fichero de entrada, se procederá a realizar el proceso de generación de semánticas. Para ello, se realizarán los siguientes pasos:

- En el proceso de lectura de conceptos, se detectó el nivel de profundidad más alto que la lista. En primer lugar, se obtendrá un subconjunto de todos los conceptos que estén a dicho nivel.
- A continuación, por cada concepto del subconjunto:

  Si todavía no ha sido asignado a una semántica:

  - Se recupera su código identificador.

  - Se obtienen todos los conceptos (del conjunto total) que contengan el identificador estudiado.

  - Si la suma total de todos los conceptos es mayor o igual a 5, se creará un nuevo objeto *Semantics* con los conceptos descubiertos y se añadirá a la lista de semánticas. Los conceptos insertados

tendrán su variable *assigned* a *true*. En este momento se pasa al siguiente concepto del subconjunto.

- Si la suma total es menor que 5, se reduce en un nivel el código de identificación y se reinicia la búsqueda de conceptos.

- Si se llega al nivel 0 (es decir, no se ha logrado asociar el código a ninguna semántica), se pasa al siguiente concepto, y este hecho se notificará al final del programa.

Una vez estudiados todos los conceptos del subconjunto, se repite el proceso, reduciendo el nivel de profundidad de los conceptos del subconjunto en 1. Los conceptos ya hayan sido seleccionados para ser registrados en una semántica no podrán aparecer en el subconjunto. El proceso termina cuando se haya llegado al nivel 0.

A continuación, se mostrará un ejemplo del procedimiento. Asuma que los conceptos a agrupar son:

- 1) G05.001.002.003
- 2) G05.001.002.004
- 3) G05.001.002.005
- 4) G05.001.002.006
- 5) G05.001.002.007
- 6) G05.001.003

En este caso, se asumirá que el nivel más profundo de todos los conceptos es el mostrado en el primer concepto, “G05.001.002.003” (nivel 4). Se empieza a realizar la agrupación por dicho concepto.

En primer lugar, se identificarán todos los conceptos que contengan el código “G05.001.002.003”. Sólo hay un concepto que cumple la condición:

- 1) **G05.001.002.003** – Válido
- 2) **G05.001.002.004** – No válido
- 3) **G05.001.002.005** – No válido
- 4) **G05.001.002.006** – No válido
- 5) **G05.001.002.007** – No válido
- 6) **G05.001.003** – No válido

Al no haber suficientes conceptos para generar una semántica, se reduce el código de identificación a “G05.001.002”. Se vuelve a realizar la búsqueda de conceptos que contengan el código. En este caso, los conceptos 1, 2, 3, 4 y 5 cumplen la condición:

- 1) **G05.001.002**.003 – Válido
- 2) **G05.001.002**.004 – Válido
- 3) **G05.001.002**.005 – Válido
- 4) **G05.001.002**.006 – Válido
- 5) **G05.001.002**.007 – Válido
- 6) **G05.001.003** – No válido

Al ser cinco conceptos, se genera una semántica a partir de ellos. Estos conceptos ya no podrán ser seleccionados en futuros procesos de creaciones de semánticas.

Escritura del fichero de salida. Una vez completado el procedimiento de agrupación en semánticas, se volcará la información de las semánticas en un nuevo fichero.

```
SEMANTICS --> C09.218.458.341
SEM.ID = 1183
SEM.NAME = Semantic1
    CONCEPT: Deafness
        SYNONYMS: Prelingual Deafness ; Deafness, Acquired ; Hearing Loss, Extreme
        TERMS: Acquired Deafness ; Deafness ; Bilateral Deafness ; Deaf Mutism ; Extreme Hearing Loss ;

    CONCEPT: Hearing Loss, Sudden
        SYNONYMS: Sudden ; Deafness, Sudden
        TERMS: Sudden Deafness ; Sudden ; Deafness, Sudden ; Sudden Hearing Loss ; Hearing Loss, Sudden

    CONCEPT: Hearing Loss, Bilateral
        SYNONYMS: Bilateral
        TERMS: Hearing Loss, Bilateral ; Loss, Bilateral Hearing ; Bilateral Hearing Losses ; Bilateral

    CONCEPT: Hearing Loss, Central
        SYNONYMS: Central ; Cortical Deafness
        TERMS: Central ; Central Hearing Loss ; Cortical Deafness ; Deafness, Cortical ; Hearing Loss,

    CONCEPT: Hearing Loss, Conductive
        SYNONYMS: Conductive
        TERMS: Conductive ; Hearing Loss, Conductive ; Conductive Hearing Loss
```

Figura 43. Formato de una semántica y sus conceptos tras su agrupación

El fichero de salida resultante tiene la siguiente estructura:

- **SEMANTICS --> …** : El identificador del concepto por el cual se generó la semántica.

- **SEM.ID = …** : Es el identificador que tendrá la semántica al ser insertada en la base de datos BoilerPlates (consulte la sección 7.4.2).

- **SEM. NAME = …** : Por defecto, el programa inserta un nombre genérico a cada semántica. Para este proyecto, se introdujo un nombre relacionado con los conceptos agrupados de forma manual, utilizando un procesador de texto.

- **CONCEPT: …** : Define un concepto, y está formado por:

    - **SYNONYMS: …** : Sinónimos del concepto.

    - **TERMS: …** : Términos del concepto. Se puede observar que los sinónimos también aparecen en este campo.

El resto de conceptos repiten la misma estructura. Para diferenciar una semántica de otra, se emplea un símbolo de separación genérico ("---").

Al disponer de dos ficheros de entrada, uno correspondiente a la sordera y otro correspondiente a la genética, se ha realizado el procedimiento dos veces, variando el fichero de entrada, y se han obtenido dos ficheros de semánticas.

Una vez generados los ficheros de semánticas, el siguiente paso es volcar toda la información obtenida del MeSH en la herramienta BoilerPlates. En la siguiente sección se explicará el funcionamiento de la herramienta, así como el proceso de inserción de la terminología.

7.4. Herramienta BoilerPlates

BoilerPlates es una herramienta que permite realizar el estudio, oración por oración, de uno o varios documentos de texto para su correspondiente análisis léxico, sintáctico y semántico. Puede encontrar más información sobre el desarrollo y evolución de la herramienta en la sección 3.3.1.

La herramienta asigna cada palabra de las oraciones, definida como "token" en la herramienta, su categoría gramatical y semántica. Asimismo, genera patrones cuando aparecen con un mínimo de frecuencia al menos dos tokens consecutivos en el texto analizado.

BoilerPlates necesita emplear dos bases de datos para funcionar: **"Rqa Quality Analyzer"** y **"RequirementsClassification"**. En los siguientes apartados se explicarán su propósito y estructura.

7.4.1 Base de datos Rqa Quality Analyzer v4.1

Esta base de datos contiene la información de los términos, y categorías sintácticas y semánticas que emplea la herramienta para estudiar los tokens leídos y extraer patrones a partir de ellos. La base de datos está contenida por varias tablas. Este estudio se centrará en las tablas **Vocabulary**, **Rules_Families** y **Grammatical**:

- **Vocabulary**: esta tabla contiene los términos que forman el vocabulario del dominio. Para este proyecto, en esta tabla se introducirá la lista de términos que forman cada concepto extraído del MeSH.

- **Rules_Families**: esta tabla contiene las categorías sintácticas de los elementos del vocabulario. En nuestro caso, se incluirán los conceptos encontrados en el MeSH.

- **Grammatical**: esta tabla contiene las categorías semánticas de los elementos del vocabulario. En esta tabla se introducirán las semánticas generadas durante el proceso de agrupación de conceptos.

Las tablas están relacionadas de la siguiente manera:

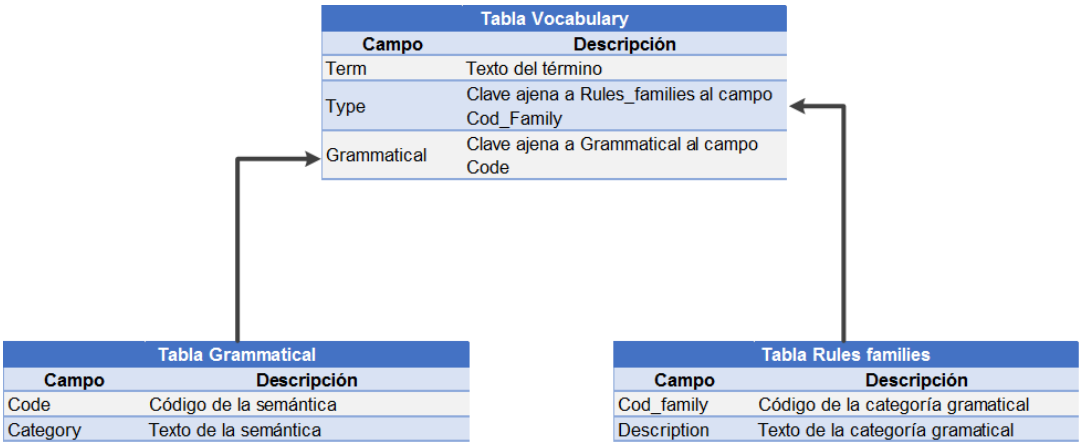

Figura 44. Relación de las tablas de Rqa Quality Analyzer

A continuación, se mostrará un ejemplo de las relaciones entre las tablas con una semántica de nuestro dominio. En particular, se estudiará la semántica “deafness”, que tiene el identificador “1183”. En la tabla Grammatical aparecerá la siguiente entrada:

Tabla Grammatical	
Campo	**Valor**
Code	1183
Category	deafness

Figura 45. Ejemplo de inserción para la tabla Grammatical

La semántica “deafness” contiene 16 conceptos, que se insertarán en la tabla Rules_Families. Los primeros cuatro conceptos de esta semántica aparecerán de la siguiente manera:

Tabla Rules families	
Campo	**Valor**
Cod_family	1553
Description	Deafness

Figura 46. Primer ejemplo de inserción en la tabla Rules Families

Tabla Rules families	
Campo	**Valor**
Cod_family	1554
Description	Hearing Loss, Sudden

Figura 47. Segundo ejemplo de inserción en la tabla Rules Families

Tabla Rules families	
Campo	**Valor**
Cod_family	1555
Description	Hearing Loss, Bilateral

Figura 48. Tercer ejemplo de inserción en la tabla Rules Families

Tabla Rules families	
Campo	**Valor**
Cod_family	1556
Description	Hearing Loss, Central

Figura 49. Cuarto ejemplo de inserción en la tabla Rules Families

Por último, cada concepto tiene una lista de términos asociados. En el caso de “Hearing Loss, Sudden” (1154) hay disponibles cinco términos asociados. En la tabla Vocabulary aparecerán de la siguiente manera:

Tabla Vocabulary	
Campo	**Valor**
Term	Sudden Deafness
Type	1154
Grammatical	1183

Figura 50. Primer ejemplo de inserción en la tabla Vocabulary

Tabla Vocabulary	
Campo	**Valor**
Term	Sudden
Type	1154
Grammatical	1183

Figura 51. Segundo ejemplo de inserción en la tabla Vocabulary

Tabla Vocabulary	
Campo	**Valor**
Term	Deafness, Sudden
Type	1154
Grammatical	1183

Figura 52. Tercer ejemplo de inserción en la tabla Vocabulary

Tabla Vocabulary	
Campo	**Valor**
Term	Sudden Hearing Loss
Type	1154
Grammatical	1183

Figura 53. Cuarto ejemplo de inserción en la tabla Vocabulary

Tabla Vocabulary	
Campo	**Valor**
Term	Hearing Loss, Sudden
Type	1154
Grammatical	1183

Figura 54. Quinto ejemplo de inserción en la tabla Vocabulary

Se puede apreciar que la relación entre las tablas Rules_Families y Grammatical se realiza a través de la tabla Vocabulary. Mientras que las semánticas y los conceptos se insertan por separado, los términos asociados deben mantener una relación con la semántica (campo “Grammatical”) y con el concepto (campo “Type”) al que está asociado. Esto se consigue por medio de una clave ajena al identificador de cada uno.

Antes de empezar el proyecto, las tablas contienen terminología genérica para todos los dominios (es decir, preposiciones, verbos comunes…). Es necesario volcar la información de nuestro dominio sin provocar un conflicto con los datos que ya están en la base de datos. Este proceso se explicará a continuación.

7.4.2 Inserción de la terminología en Rqa Quality Analyzer v4.1

Para trabajar con la terminología extraída del MeSH, es necesario insertar las semánticas, conceptos y términos en sus respectivas tablas (Grammatical, Rules_Families y Vocabulary, respectivamente). Para ello, se implementó un algoritmo sencillo por cada elemento, que permita adaptar la información para su inserción como entradas de la base de datos.

La información que se extraerá por cada elemento se muestra en la siguiente imagen:

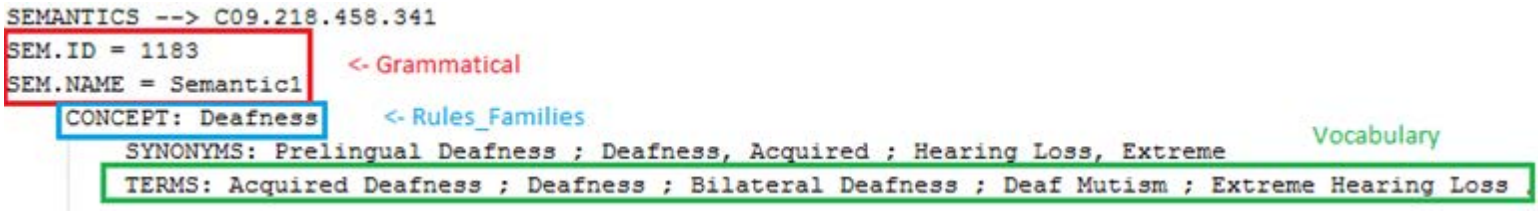

Figura 55. Información a extraer para cada tabla

En la tabla Grammatical, las semánticas fueron insertadas mediante la siguiente estructura:

CODE	CATEGORY	NORMALISED_RSHP	ROTATE_CONCEPT	…
<SEM. ID>	<SEM. NAME>		FALSO	…

Figura 56. Estructuración de las semánticas para su inserción en BoilerPlates

En donde:

- **SEM. ID** es el identificador de la semántica en la base de datos, extraído directamente del fichero de agrupaciones. Un estudio previo de Rqa Quality Analyzer indicó que, de todas las semánticas genéricas insertadas previamente, el identificador más alto era **1165**. Por tanto, las semánticas de nuestro dominio tendrán los identificadores 1166 en adelante. Consulte el Anexo A para conocer las semánticas que aparecen en la base de datos.

- **<SEM. NAME>** es el nombre de la semántica, extraído directamente del fichero de agrupaciones. Como ya se explicó anteriormente, se introdujo un nombre específico para cada semántica mediante un procesador de texto.

- El resto de campos no son relevantes para nuestro estudio, por lo que se introdujo un valor genérico.

En la tabla Rules_Families, los conceptos fueron insertados mediante la siguiente estructura:

COD_FAMILY	DESCRIPTION	Is_Software	GENERIC_FAMILY	TYPE	…
<CON. ID>	<CON. NAME>	FALSO	1119	1	…

Figura 57. Estructuración de los conceptos para su inserción en BoilerPlates

En donde:

- **CON. ID** es el identificador del concepto en la base de datos. Se introdujo un identificador nuevo al procesar cada concepto. Un estudio previo de la base de datos indicó que el identificador más alto de todos los conceptos genéricos era **1258**. Por tanto, todos los conceptos del dominio tendrán el identificador 1259 en adelante. Consulte el Anexo B para conocer los conceptos que aparecen en la base de datos.

- **CON. NAME** es el nombre del concepto, extraído directamente del fichero de agrupaciones.

- El resto de campos no son relevantes para nuestro estudio, por lo que se introdujo un valor genérico.

En la tabla Vocabulary, los términos fueron insertados mediante la siguiente estructura:

COD_T	LANGUAGE	TERM	TYPE	GRAMMATICAL	…
<TER. ID>	22	<TER. NAME>	<CON. ID>	<SEM. ID>	…

Figura 58. Estructuración de los términos para su inserción en BoilerPlates

En donde:

- **TER. ID** es el identificador del término en la base de datos. Se introdujo un identificador nuevo al procesar cada término. Un estudio previo de la base de datos indicó que el identificador más alto de todos los términos genéricos era **77829**. Por tanto, todos los conceptos del dominio tendrán el identificador 77830 en adelante.

- **TER. NAME** es el nombre del término, extraído directamente del fichero de agrupaciones.

- **CON. ID** es el identificador del concepto al cual el término pertenece.

- **SEM. ID** es el identificador de la semántica al cual el término pertenece.

Una vez realizados todos los procesos de inserción, la base de datos Rqa Quality Analyzer contendrá tanto la terminología genérica como la terminología extraída del MeSH, por lo que estará preparada para realizar el proceso de extracción de patrones.

7.4.3 Base de datos RequirementsClassification

Esta base de datos contiene la información asociada a los patrones base, así como los patrones finales obtenidos tras completar un proceso de análisis de uno o varios documentos. Utiliza Rqa Quality Analyzer como referencia para llevar a cabo sus funcionalidades, pero no la modifica en ningún momento.

RequirementsClassification está formado por cinco tablas, que se definirán a continuación:

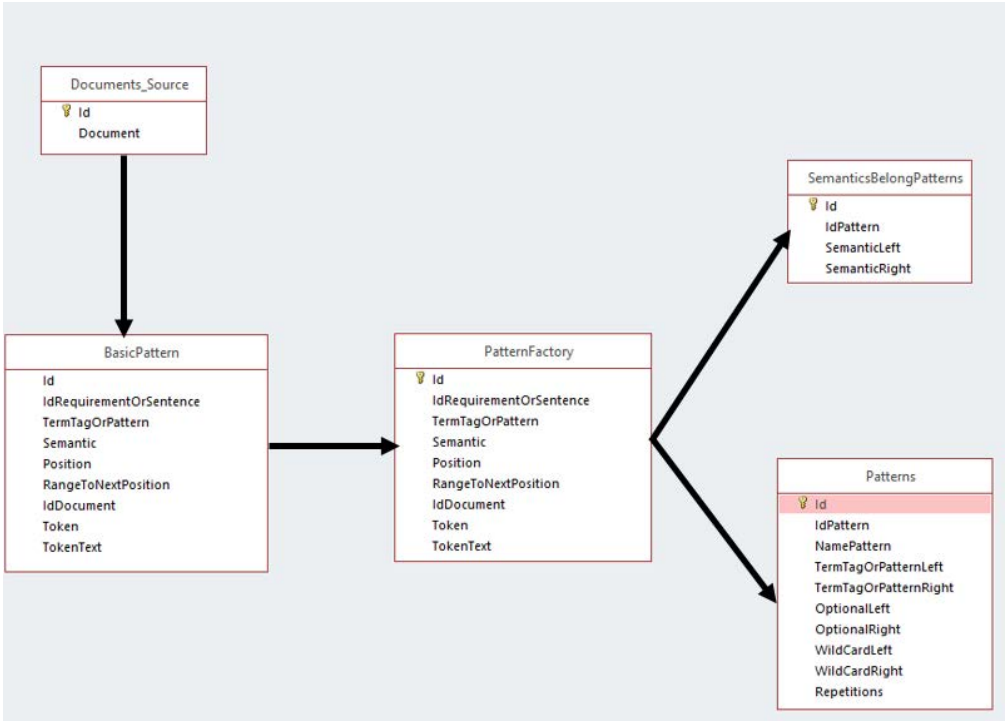

Figura 59. Relación de las tablas de RequirementsClassification

- **BasicPattern**: esta tabla contiene todos los tokens leídos en los documentos analizados. Por cada token, se puede verificar su documento de precedencia, así como el identificador de su concepto asociado.

- **Documents_Source**: en esta tabla se recogen todos los ficheros de texto almacenados en la herramienta. Se introduce una nueva entrada cada vez que se analiza un documento.

- **PatternFactory**: esta tabla se emplea para generar los patrones. En la primera iteración del proceso, la tabla es una copia de BasicPattern. En los pasos siguientes, se sustituyen términos que aparecen con frecuencia por un patrón. El proceso continúa hasta llegar a una condición de parada, que indica la frecuencia mínima de aparición de pares de términos.

- **Patterns**: esta tabla contiene los patrones que se han generado en el proceso. Por cada uno se puede conocer su identificación (un valor negativo) y su composición.

  Cada patrón contiene dos elementos, situados en la “parte izquierda” y la “parte derecha”. En el caso más sencillo, estos elementos referencian semánticas registradas en Rqa Quality Analyzer. Un ejemplo básico de patrón muy frecuente en el inglés sería por

ejemplo “Adjective” (izquierda) + “Noun” (derecha), siendo “Adjective” y “Noun” dos semánticas.

Sin embargo, es posible que un elemento del patrón referencie a otro patrón que haya sido descubierto previamente. Por ejemplo, en es posible encontrarse con el patrón “-1” (izquierda) + “Noun” (derecha). Esto significa que el elemento izquierdo del patrón referencia a otro patrón, que a su vez contendrá sus respectivos elementos en su izquierda y en su derecha. Esto implica que **un patrón puede tener longitud variable**, siendo el mínimo 2 en el caso más sencillo.

- **SemanticsBelongPatters**: esta tabla contiene la asociación de las semánticas a los elementos de los patrones.

7.5. Uso de la herramienta BoilerPlates

A continuación, se procederá a explicar el uso de la herramienta a partir de la interfaz de usuario ofrecida en ella. Una vez adquirido el fichero en donde se encuentra todos los recursos del programa, se hace doble clic en el ejecutable (BoilerPlates.exe).

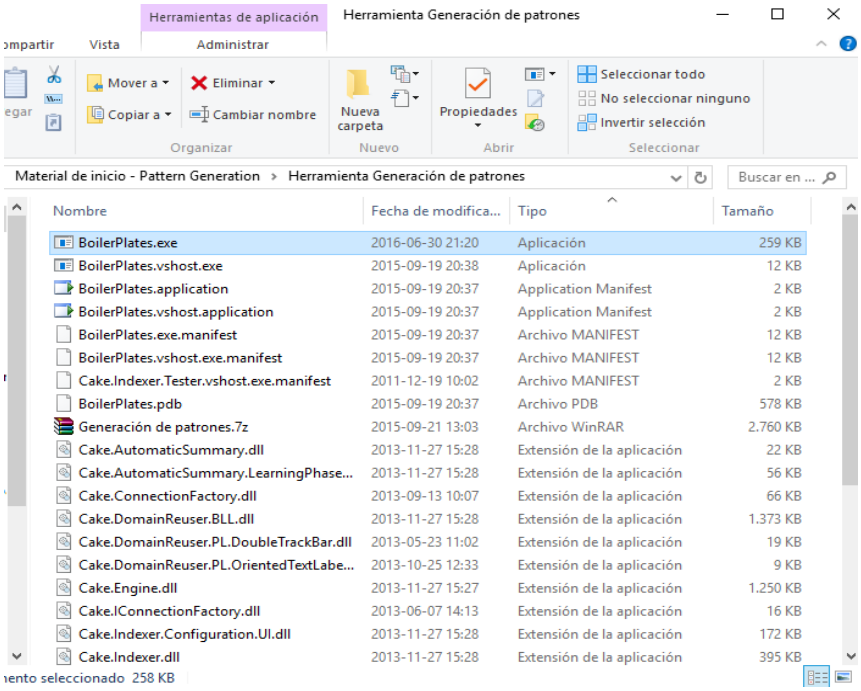

Figura 60: Contenido de la herramienta BoilerPlates

7.5.1 Conexión a la base de datos

Al ejecutar la herramienta aparecerá la interfaz de usuario, en donde se podrán realizar las acciones necesarias para realizar el proceso de extracción de patrones. La ventana contiene una serie de botones que representan las funcionalidades ofrecidas por la herramienta.

El primer paso es conectarse a la base de datos. Para ello, se debe establecer el tipo y parámetros de la conexión desde la sección “Conexión a la base de datos”.

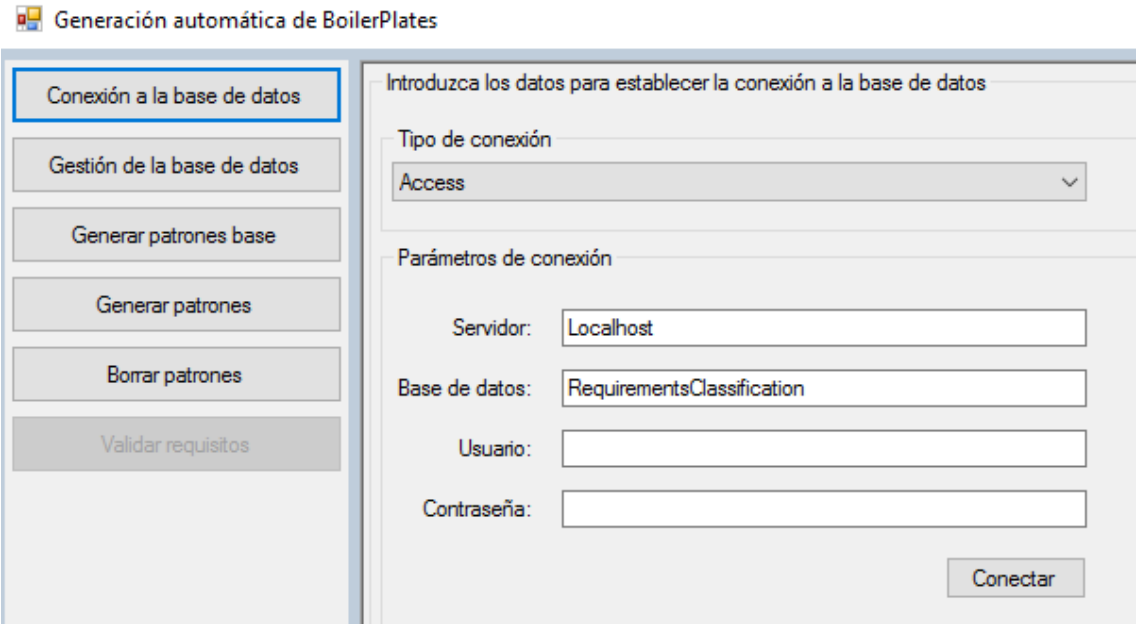

Figura 61. BoilerPlates - Conexión a la base de datos

La configuración ofrecida por defecto permite realizar la conexión. Por tanto, en esta ventana se pulsa el botón Conectar. Si la conexión ha sido establecida correctamente, recibirá el siguiente mensaje de confirmación:

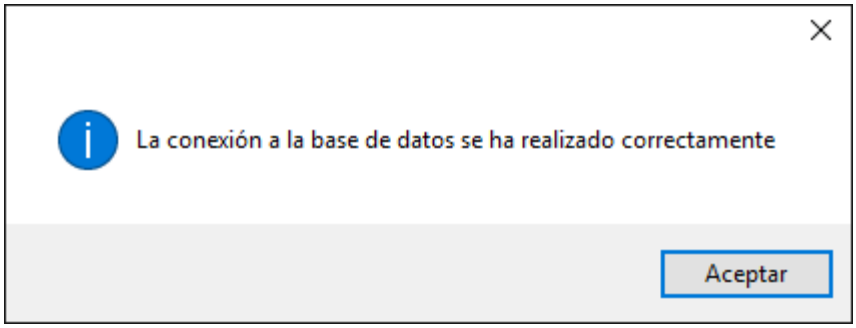

Figura 62. Mensaje recibido al realizar la conexión

7.5.2 Gestión de la base de datos

Esta sección permite realizar el borrado de información de una o varias tablas de la base de datos. Esta funcionalidad se emplea para realizar nuevos procesos de extracción de patrones.

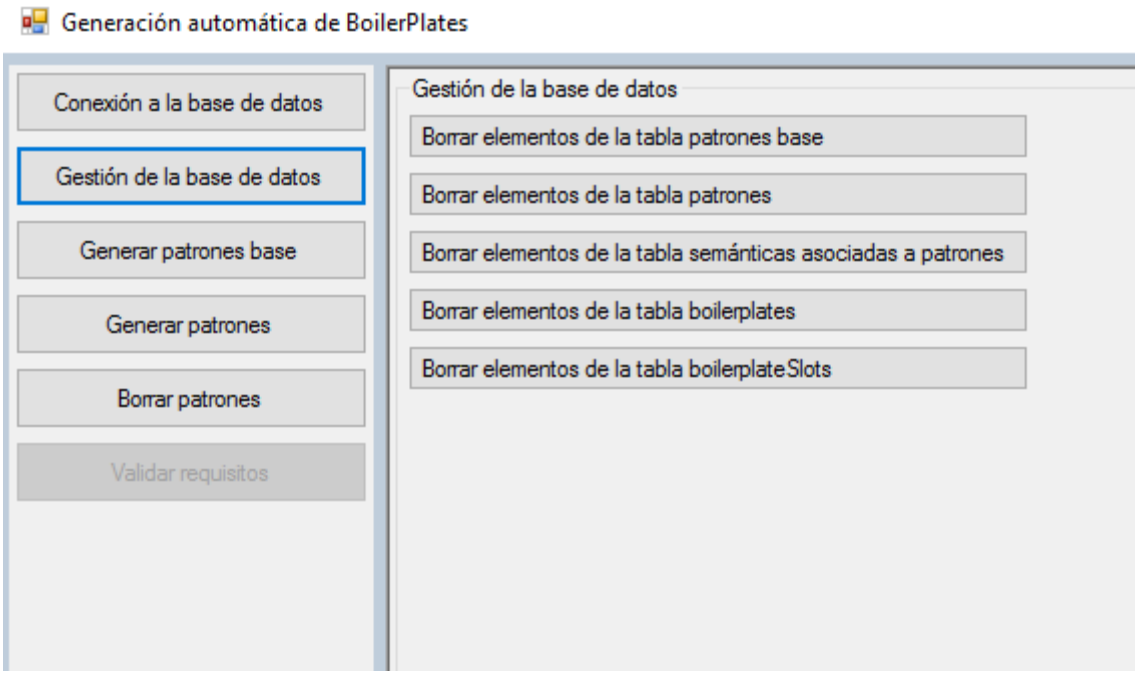

Figura 63. BoilerPlates - Gestión de la base de datos

Para este proyecto, se han empleado los siguientes mecanismos de borrado:

- **Borrar elementos de la tabla patrones base**. Los patrones base reconocidos en el proceso inicial fueron borrados cada vez que se realizaba el estudio sobre un nuevo fichero.

- **Borrar elementos de la tabla patrones** y **Borrar elementos de la tabla semánticas asociadas a patrones**. Tras almacenar en un fichero aparte los resultados de cada prueba, se borran las tablas de patrones reconocidos para realizar una nueva prueba sobre el mismo fichero, pero con distinta configuración.

7.5.3 Generar patrones base

Esta sección permite generar los patrones base de los documentos seleccionados para el estudio.

Un patrón base es un patrón binario generado al realizar el procesamiento de los documentos de entrada. A partir de estos, se generan los patrones sintáctico-semánticos finales. Por lo general, los patrones base se adquieren una sola vez para luego obtener diferentes conjuntos de resultados variando los parámetros de la herramienta. Para la experimentación realizada en este proyecto, se obtuvieron los patrones base por cada lote de documentos empleado.

Mientras se extraen los patrones base, la herramienta identifica diferentes frases cuando están separadas por un punto (“.”). La sección se divide en dos pestañas:

- **“Generar patrones desde un documento”**. Desde aquí la herramienta permite seleccionar un fichero que se encuentre en el ordenador. para ello, se especifica su ruta mediante el botón Selecionar documento y, a continuación, se hace clic en el botón Iniciar proceso de generación de patrones.

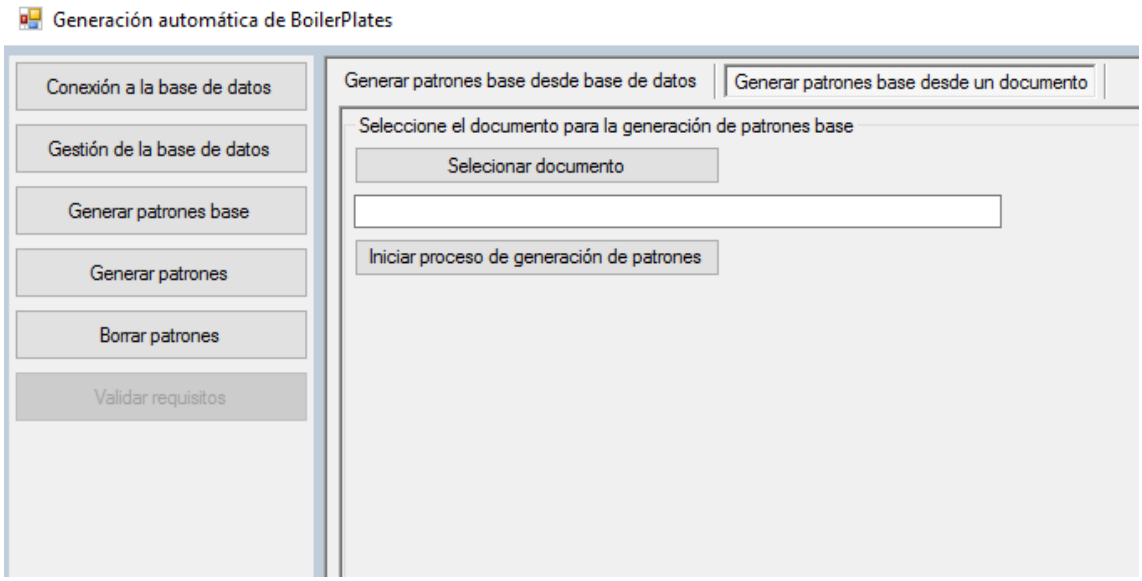

Figura 64. BoilerPlates. Generar patrones base desde un documento

- **“Generar patrones base desde base de datos”**. En esta pestaña se muestran los documentos que han sido empleados para el proceso de generación de los patrones base.

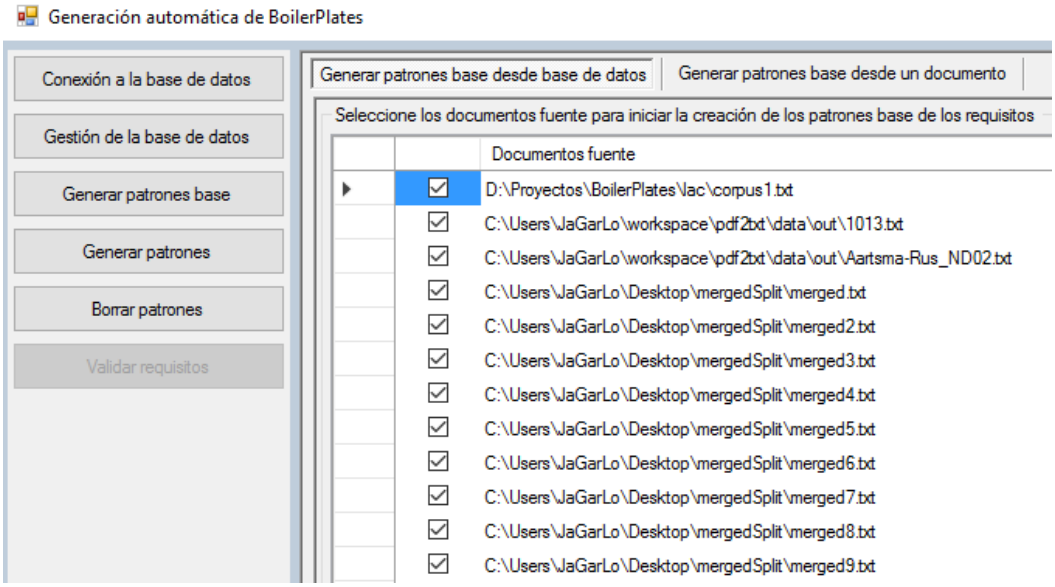

Figura 65. BoilerPlates. Generar patrones base desde la base de datos

7.5.4 Generar patrones

Esta sección permite empezar el proceso de extracción de patrones, una vez determinados los patrones base.

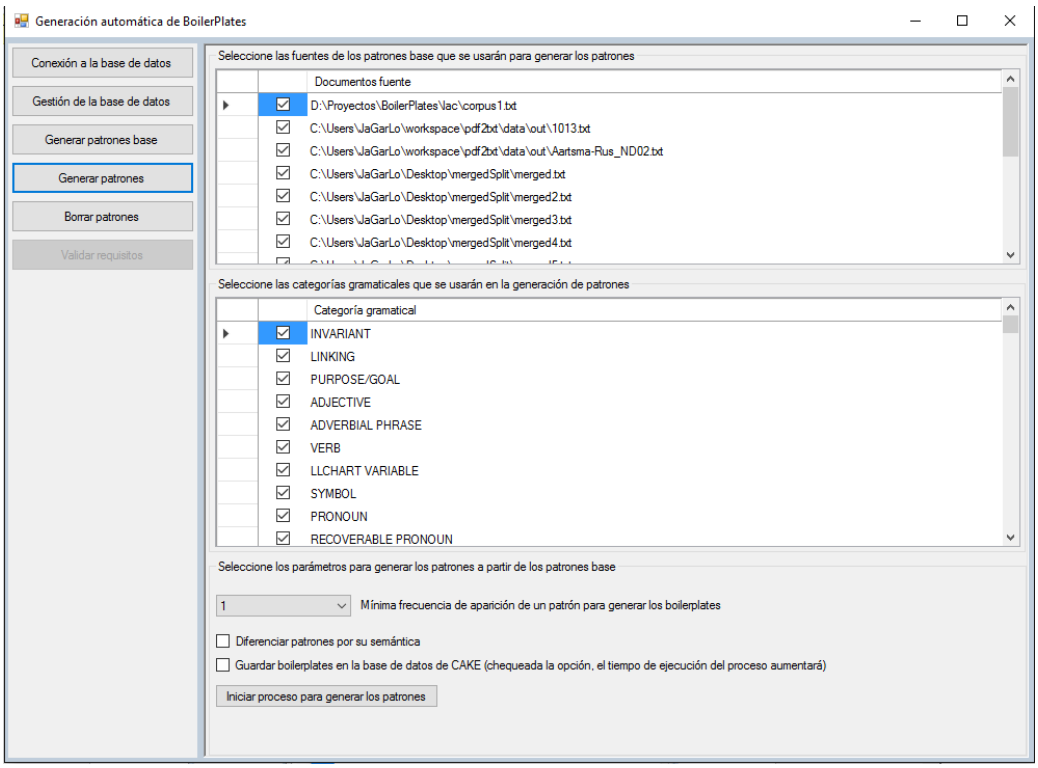

Figura 66. BoilerPlates - Generar patrones

En la ventana “Documentos fuente” se pueden seleccionar los documentos por los cuales se realizará la extracción de datos. Para ello, se marcan los ficheros que se deseen analizar.

En la ventana “Seleccione las categorías gramaticales…” se puede seleccionar qué categorías de términos se considerarán en el estudio. Por defecto, todas las categorías que estén almacenadas en la base de datos serán empleadas en el estudio.

A continuación, se establecen los siguientes parámetros:

- **Mínima frecuencia de patrón**: determina el número mínimo de apariciones de un patrón necesario para registrarlo. En este estudio se utilizarán los valores 1, 5, 10 y 20.

- **Diferenciar patrones por su semántica**: determina si se considerará la semántica durante el proceso de extracción de patrones. En este estudio se estudiarán las diferencias entre ignorar y considerar la semántica.

Una vez terminado de configurar el análisis, se hace clic en el botón 

.

7.5.5 Borrar patrones

Una vez terminado el proceso de extracción de patrones, esta sección permite consultar los resultados obtenidos, pudiendo realizar modificaciones sobre los patrones si fuese necesario.

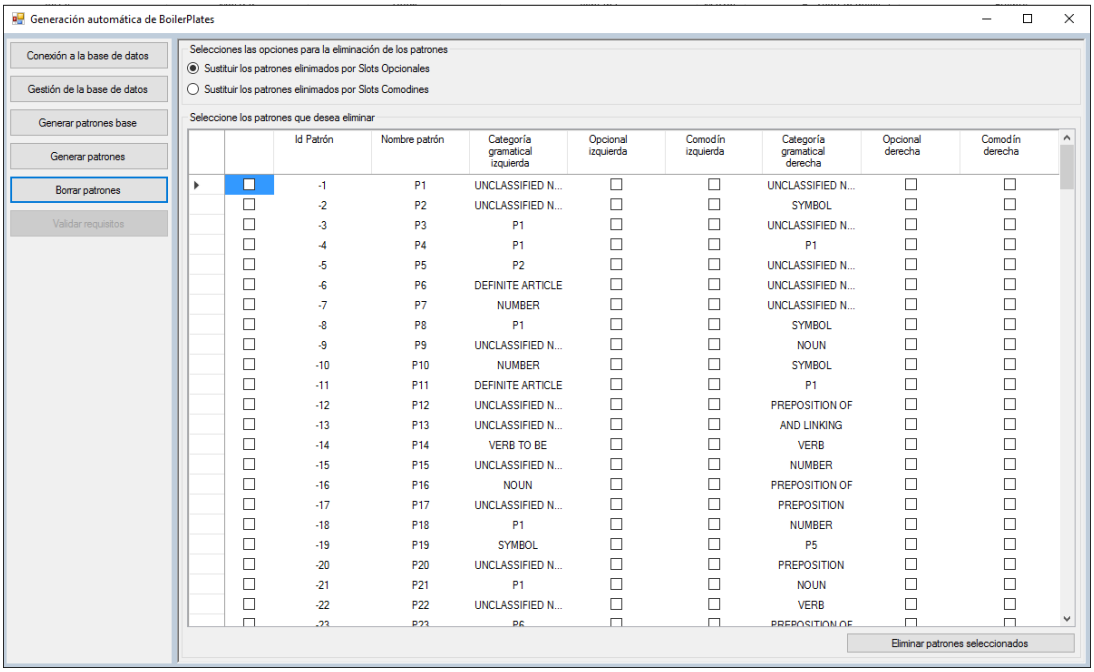

Figura 67. BoilerPlates - Borrar patrones

Si se desea borrar uno o varios patrones específicos, se marcan las casillas de verificación correspondientes en la primera columna de la tabla y se hace clic en el botón 

.

Además, es posible reemplazar las categorías gramaticales situadas a la izquierda y a la derecha de cada patrón por slots “comodín” o slots “opcionales”.

En este estudio no se han empleado ninguna de las funcionalidades ofrecidas en esta sección. Los patrones se han estudiado “tal cual” y no han sido modificados ni borrados.

7.6. Funcionamiento de la herramienta

Como ya se ha explicado anteriormente, BoilerPlates divide el texto a analizar en varias oraciones y determina los patrones lingüísticos a partir de ellos.

Antes de proceder al análisis de patrones, la herramienta realiza dos acciones previas:

- **Tokenización**. En primer lugar, la herramienta divide el texto recibido en palabras y frases. Cada palabra se podrá identificar debido a que hay un espacio, coma o punto que lo separa de otra, mientras que una frase se reconocerá con una separación de un punto (“.”).

- **Normalización**. Con el objetivo de mejorar la eficacia de la extracción de los patrones, la herramienta estandariza las palabras de los documentos. Algunos procesos de la normalización son, por ejemplo, el paso del plural al singular o el cambio de los verbos a infinitivo.

7.6.1 Funcionamiento: obtención de tokens

Tras completar los procesos previos (tokenización y normalización), la herramienta comienza a analizar el texto de análisis para crear los patrones base. Para ello, BoilerPlates emplea la información contenida en la base de datos Rqa Quality Analyzer (las tablas Rules_Families, Vocabulary y Grammatical) para detectar tanto los términos de carácter generalista que venían incluidos antes de comenzar el proyecto como los términos que hemos introducido relativos a la sordera genética.

Una vez procesado el texto, la tabla BasicPattern de la base de datos RequirementsClassification contendrá todos los tokens que haya logrado reconocer, así como la categoría gramatical asociada a cada uno. En el caso de que el token no haya sido categorizado, su categoría gramatical se establecerá a "**Unclassified Noun**" (sustantivo sin clasificar), cuyo identificador en Rqa Quality Analyzer en la tabla Rules_Families es **1144**.

En el resto del procedimiento, los contenidos de la tabla BasicPattern no serán modificados.

7.6.2 Funcionamiento: obtención de patrones básicos

A continuación, la herramienta buscará secuencias de semánticas de longitud 2 que aparezcan con un mínimo de frecuencia. Los resultados de esta operación dependerán fundamentalmente del contenido del texto y las semánticas que la herramienta utiliza para trabajar. A continuación, se muestran ejemplos de patrones básicos:

Nombre del Patrón	Sem. Izquierda	Sem. Derecha
P1	Adjective	Noun
P2	Noun	Verb
P3	Adjective	Hearing Loss
P4	Selection, Genetic	Verb

Figura 68. Ejemplos de patrones básicos

En el ejemplo, se puede observar que los patrones P1 y P2 emplean terminología ajena al dominio de estudio. Sin embargo, la herramienta ha empleado los conceptos “Hearing Loss” y “Selection, Genetic” para formar los patrones P3 y P4.

En el caso de que se desee realizar una distinción por semántica, la herramienta también considerará tanto las categorías gramaticales definidas en Rules_Families como las semánticas añadidas en Grammatical.

Por tanto, con el patrón P1 del ejemplo de antes (Adjective + Noun) y considerando las semánticas S1, S2 y S3, es posible que se obtenga los siguientes resultados si se realiza distinción por semántica:

Nombre del Patrón	Sem. Izquierda	Sem. Derecha
P1	Adjective S1	Noun
P2	Adjective S2	Noun
P3	Adjective	Noun S3

Figura 69. Ejemplos de patrones básicos diferenciando por semántica

Como se puede apreciar, se han creado tres patrones que tienen las mismas categorías gramaticales (Adjective + Noun) pero son diferentes ya que están contenidas en distintas semánticas.

7.6.3 Funcionamiento: generación de los patrones de frecuencia

Tras generar los patrones básicos por medio del procedimiento explicado anteriormente, la herramienta empieza a estudiar qué pares de tokens son más frecuentes en la tabla BasicPattern. Dichos pares son sustituidos por un patrón, cuyo identificador será un **número negativo**.

Todo el proceso de emparejamiento se contempla en la tabla PatternFactory. Como se explicó anteriormente, en la primera iteración el contenido de esta tabla es igual a la de BasicPattern. Al final del proceso, se podrá ver por cada frase las parejas de elementos que hayan sido sustituidas por patrones.

Los patrones de frecuencia superior a 2 recibirán el nombre de “patrones compuestos”, ya que uno o los dos elementos que los componen son patrones y, por tanto, su longitud es mayor que 2. Véase el siguiente ejemplo:

Nombre del Patrón	Sem. Izquierda	Sem. Derecha
P1	Adjective	Noun
P2	-1	Verb
P3	-2	Preposition

Figura 70. Ejemplo de patrones compuestos

P1 es un patrón simple: está formado por la pareja de categorías gramaticales Adjetive + Noun. P2 tiene en su parte izquierda el patrón P1, por lo que su secuencia será el contenido del patrón P1 y un verbo (Verb). P3 tiene en su parte izquierda el patrón P2, por lo que su secuencia será el patrón P2 y una preposición (Preposition).

Por tanto, los patrones del ejemplo tienen las siguientes características:

Patrón	Secuencia	Long.
P1	Adjective + Noun	2
P2	Adjective + Noun + Verb	3
P3	Adjective + Noun + Verb + Preposition	4

Figura 71. Resultado del análisis de los patrones complejos de la tabla anterior

El resultado final de la extracción de patrones se podrá contemplar en la tabla Patterns. En ella, aparecerán tanto los patrones básicos como los patrones compuestos. El análisis de los resultados se realizará mediante el uso de los contenidos de esta tabla.

7.7. Ejemplo de extracción de patrones

En esta sección, se empleará un texto de ejemplo para observar el proceso de extracción de patrones con distintas configuraciones.

Asuma que la herramienta debe procesar lo siguiente:

“Javier eats apples at the park. Laura works at the office. José waits.”

En primer lugar, se divide el texto en frases:

Frase 1 (F1): “Javier eats apples at the park.”

Frase 2 (F2): “Laura works at the office.”

Frase 3 (F3): “José waits.”

La herramienta ha detectado tres frases gracias al punto (“.”) que los separa. A continuación, se procede a estudiar cada token:

Id. Frase	Pos. token	Token leído	Cat. gramat.	Token norm.
F1	0	“Javier”	NOUN	Javier
F1	1	“eats”	VERB	to eat
F1	2	“apples”	NOUN	apple
F1	3	“at”	PREP	at
F1	4	“the”	DT	the
F1	5	“park”	NOUN	park

Figura 72. Estudio de tokens de la primera frase

Id. Frase	Pos. token	Token leído	Cat. gramat.	Token norm.
F2	0	Laura	NOUN	Laura
F2	1	works	VERB	to work
F2	2	at	PREP	at
F2	3	the	DT	the
F2	4	office	NOUN	office

Figura 73. Estudio de tokens de la segunda frase

Id. Frase	Pos. token	Token leído	Cat. gramat.	Token norm.
F3	0	José	NOUN	José
F3	1	waits	VERB	to wait

Figura 74. Estudio de tokens de la tercera frase

Por tanto, las categorías gramaticales que aparecen en las frases son:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	NOUN	VERB	NOUN	PREP	DT	NOUN
F2	NOUN	VERB	PREP	DT	NOUN	--
F3	NOUN	VERB				

Figura 75. Análisis sintáctico de las frases

7.7.1 Obtención de patrones con frecuencia mínima 2

La herramienta ha realizado los pasos necesarios para obtener la información contemplada en la *Figura 75. Análisis sintáctico de las frases*. A continuación, se procederá a extraer los patrones que pueda haber en el texto recibido. Se asume que el mínimo de apariciones de un patrón para ser reconocido es 2 y no se realiza distinción por semántica.

Se puede apreciar que la secuencia “NOUN + VERB” es la más frecuente, con tres repeticiones. Por tanto, se convierte en el patrón P1:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	P1	NOUN	PREP	DT	NOUN	--
F2	P1	PREP	DT	NOUN	--	--
F3	P1	--	--	--	--	--

Figura 76. Creación del patrón P1 (Noun + Verb)

La secuencia que más se repite ahora es “DT + NOUN” se repite dos veces. A partir de ella se genera el patrón P2:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	P1	NOUN	PREP	P2	--	--
F2	P1	PREP	P2	--	--	--
F3	P1	--	--	--	--	--

Figura 77. Creación del patrón P2 (DT + NOUN)

Además, a partir de P2, se ha generado la secuencia “PREP + P2”, que se repite dos veces. La herramienta genera el patrón P3:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	P1	NOUN	P3	--	--	--
F2	P1	P3	--	--	--	--
F3	P1	--	--	--	--	--

Figura 78. Creación del patrón P3 (PREP + P2)

Ya no se pueden extraer más patrones a partir de las frases, por lo que el proceso se considera terminado. La tabla Patterns tendrá el siguiente contenido:

Patrón	Id. Patrón	Elem. izquierda	Elem. derecha
P1	-1	NOUN	VERB
P2	-2	DT	NOUN
P3	-3	PREP	-2

Figura 79. Estado de la tabla Patterns tras completar el proceso

7.7.2 Obtención de patrones con frecuencia mínima 3

Ahora, se repetirá el proceso de análisis sobre el contenido mostrado en la *Figura 75. Análisis sintáctico de las frases*, pero esta vez asignando una frecuencia mínima de 3 para asignar un patrón. No se considera semántica.

De nuevo, se observa que la secuencia “NOUN + VERB” se repite tres veces. Por tanto, a partir de ella se obtiene el patrón P1:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	P1	NOUN	PREP	DT	NOUN	--
F2	P1	PREP	DT	NOUN	--	--
F3	P1	--	--	--	--	--

Figura 80. Creación del patrón P1 (Noun + Verb)

La siguiente secuencia que se repite es “DT + NOUN”. Sin embargo, ya que sólo se repite dos veces, no llega al mínimo exigido y, por tanto, no se genera un patrón a partir de ella. Al no poder seguir avanzando, el proceso se termina. La tabla Patterns tendrá la siguiente apariencia:

Patrón	Id. Patrón	Elem. izquierda	Elem. derecha
P1	-1	NOUN	VERB

Figura 81. Estado de la tabla Patterns tras completar el proceso

7.7.3 Obtención de patrones con frecuencia mínima 2, con distinción de semántica

En esta prueba, se realizará el análisis sobre el contenido mostrado en la *Figura 75. Análisis sintáctico de las frases*, esta vez considerando la semántica. El mínimo de frecuencia para generar un patrón será 2.

Asuma que la tabla Grammatical de Rqa Quality Analyzer contiene las siguientes entradas:

Tabla Grammatical		
Campo	Descripción	Valor
Code	Código de la semántica	1
Category	Texto de la semántica	Location Outdoors

Figura 82. Entrada para la semántica "Location Outdoors"

Tabla Grammatical		
Campo	Descripción	Valor
Code	Código de la semántica	2
Category	Texto de la semántica	Location Indoors

Figura 83. Entrada para la semántica "Location Indoors"

Tabla Grammatical		
Campo	Descripción	Valor
Code	Código de la semántica	3
Category	Texto de la semántica	Action Food

Figura 84. Entrada para la semántica "Action Food"

Tabla Grammatical		
Campo	Descripción	Valor
Code	Código de la semántica	4
Category	Texto de la semántica	Action Other

Figura 85. Entrada para la semántica "Action Other"

Además, en la tabla Vocabulary se han definido los siguientes términos:

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

Figura 86. Entrada para el término "park"

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

Figura 87. Entrada para el término "office"

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

Figura 88. Entrada para el término "to eat"

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	to work
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	4

Figura 89. Entrada para el término "to work"

Tabla Vocabulary		
Campo	Descripción	Ejemplo
Term	Texto del término	to wait
Type	Clave ajena a Rules_families al campo Cod_Family	1119
Grammatical	Clave ajena a Grammatical al campo Code	4

Figura 90. Entrada para el término "to wait"

Al diferenciar por semántica, la herramienta tendrá en consideración las semánticas de las categorías gramaticales que ha reconocido en el texto. Por tanto, la situación inicial es la siguiente:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	NOUN	VERB (S3)	NOUN	PREP	DT	NOUN (S1)
F2	NOUN	VERB (S4)	PREP	DT	NOUN (S2)	--
F3	NOUN	VERB (S4)	--	--	--	--

Figura 91. Análisis sintáctico de las frases, incluyendo semántica

Se puede apreciar que la secuencia más frecuente es “NOUN + VERB (S4)”, que se repite dos veces. A partir de ella se genera el patrón P1:

Id. Frase	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
F1	NOUN	VERB (S3)	NOUN	PREP	DT	NOUN (S1)
F2	P1	PREP	DT	NOUN (S2)	--	--
F3	P1	--	--	--	--	--

Figura 92. Creación del patrón P1 (NOUN + VERB S4)

Ante esta situación, ya no se puede seguir extrayendo patrones: la única secuencia que tiene repeticiones es “DT + NOUN”. Sin embargo, ya que el término categorizado como “NOUN” al final de las frases F1 y F2 no

corresponden a la misma semántica, no se puede generar un patrón a partir de ella.

Por tanto, la tabla Patterns tendrá la siguiente apariencia:

Patrón	Id. Patrón	Elem. izquierda	Elem. derecha
P1	-1	NOUN	VERB (S4)

Figura 93. Estado de la tabla Patterns tras completar el proceso

7.8. Realización del análisis de los lotes de documentos

Una vez insertada la terminología extraída del MeSH y procesados los documentos de entrada, estamos listos para empezar el proceso de extracción de patrones.

Los objetivos que deberá cumplir la herramienta BoilerPlates son los siguientes:

- Leer el contenido de los ficheros que reciba.
- Procesar el texto y extraer los patrones que pueda identificar siguiendo la metodología estudiada en los apartados anteriores.
- Introducir en la tabla Patterns (en la base de datos RequirementsClassification) la información de todos los patrones detectados en el sistema.

Los escenarios que se contemplarán en este estudio son los siguientes:

1. Mínima frecuencia **1**, **sin** distinción de semántica.
2. Mínima frecuencia **5**, **sin** distinción de semántica.
3. Mínima frecuencia **10**, **sin** distinción de semántica.
4. Mínima frecuencia **20**, **sin** distinción de semántica.
5. Mínima frecuencia **1**, **con** distinción de semántica.
6. Mínima frecuencia **5**, **con** distinción de semántica.
7. Mínima frecuencia **10**, **con** distinción de semántica.
8. Mínima frecuencia **20**, **con** distinción de semántica.

Ya que disponemos de 14 lotes de ficheros, por cada escenario será necesario iterar el proceso 14 veces, uno por cada lote obtenido. Por tanto, la

herramienta BoilerPlates deberá ser utilizada **112** veces para estudiar los ocho escenarios.

A continuación, se adjunta una tabla con el coste temporal aproximado de realizar el análisis:

Configuración	T. ejec. (general)	t. ejec. (merged1)	t. ejec. (merged14)
Mínima frec. 1	18 horas	26 horas	3 horas
Mínima frec. 5	10 horas	18 horas	1 hora
Mínima frec. 10	4 horas	10 horas	< 1 hora
Mínima frec. 20	2 horas	4 horas	< 1 hora

Figura 94. Tiempos de ejecución aproximados

Como puede apreciarse, el coste temporal de realizar el proceso por cada lote es inversamente proporcional a la frecuencia mínima seleccionada. Esto es debido a que una condición menos rígida de frecuencia permite generar más patrones, aumentando la carga de trabajo.

Como era de esperar, el coste temporal es directamente proporcional al tamaño del fichero procesado. La herramienta tardó más en procesar el fichero merged1.txt (el más grande de todos) con una diferencia considerable. El fichero merged14.txt, el más pequeño, fue procesado mucho más deprisa que el resto. Los resultados sugieren que el coste computacional de la herramienta es exponencial según la cantidad de texto a procesar.

Por último, cabe destacar que la diferenciación por semántica influye poco en el coste del análisis, por lo que se ha considerado despreciable.

En total, se ha tardado un total de **18 días** para obtener toda la información necesaria para los ocho escenarios identificados.

7.8.1 Agrupación de los experimentos

En el caso de que hubiésemos empleado un solo fichero de texto para el estudio, a partir de este momento estaríamos listos para extraer la información de RequirementsClassification y realizar el análisis estadístico.

Sin embargo, al haber dividido el contenido de los documentos en lotes, ahora disponemos de 14 listas de resultados en cada escenario: una por cada lote que se ha procesado en la herramienta. Por tanto, el último paso antes de empezar con el análisis de resultados es **agrupar los experimentos en una sola lista de resultados por escenario.**

Sin embargo, el problema no es de carácter trivial: la solución no consiste en unir directamente las tablas Pattern obtenidas en una. Esto es

debido a que se los experimentos por cada fichero se realizaron por separado, borrando el contenido de las tablas obtenidas tras completar el análisis. Por tanto, es muy probable que tengamos patrones con el mismo nombre pero con diferentes parámetros si se realiza la unión directa. Por ejemplo, observe las siguientes tablas:

Tabla Patterns de merged1.txt		
Nombre del Patrón	Sem. Izquierda	Sem. Derecha
P1	Adjective	Noun
P2	Noun	Verb
P3	Adjective	Hearing Loss

Figura 95. Primera tabla del problema de la unión

Tabla Patterns de merged2.txt		
Nombre del Patrón	Sem. Izquierda	Sem. Derecha
P1	Adjective	Noun
P2	Prep	Verb
P3	Noun	Genetics

Figura 96. Segunda tabla del problema de la unión

Como se puede observar, el patrón P1 es el mismo para los dos ficheros de entrada, mientras que los patrones P2 y P3 son diferentes, pero se llaman igual. Al unirlos se perderá información fundamental en el análisis. Por tanto, es necesario encontrar un algoritmo que permita unir los catorce ficheros de resultados respetando las repeticiones y renombrando los patrones que sean nuevos.

Durante el estudio de este problema, se propuso una mejora adicional a los resultados obtenidos en las tablas Patterns. Ya que es posible que un patrón sea compuesto (es decir, que uno de sus elementos sea un patrón), se decidió obtener la longitud por cada patrón. La longitud es el número de elementos que forman la cadena completa del patrón. Por ejemplo:

Patrón	Id. Patrón	Elem. izquierda	Elem. derecha	Longitud
P1	-1	NOUN	VERB	2
P2	-2	DT	NOUN	2
P3	-3	PREP	-2	3

Figura 97. Ejemplo de longitud de patrones

Los patrones P1 y P2 sólo contienen categorías gramaticales, por lo que tienen longitud 2. Sin embargo, el patrón 3 referencia a P2, por lo que su secuencia sería “PREP + DT + NOUN” y, por tanto, su longitud es 3. Se puede implementar un algoritmo que permita la obtención de la longitud por cada patrón de forma directa.

En resumen, antes del proceso de análisis de resultados se realizarán dos procesos más:

- Obtención de la longitud de cada patrón.

- Renombrado y agrupamiento de los patrones en una sola tabla de resultados.

Ambos procesos se llevarán a cabo con un programa escrito en Java por separado.

7.8.2 Obtención de las longitudes de cada patrón

El programa implementado recibe por parámetros el fichero de entrada con todos los patrones obtenidos durante el proceso de análisis. Para ello, se han unido las tablas correspondientes a cada lote de documentos, con los patrones de cada uno ordenados por su orden de descubrimiento (reconocido por su identificador numérico). Además, cada patrón tendrá un valor numérico que representa la tabla de origen (un número del 1 al 14).

El algoritmo es el siguiente:

1. Se declaran las variables y recursos necesarios para leer un fichero, así como una estructura de datos (patternLength) en donde se almacenarán las longitudes reconocidas de cada patrón.

2. Se lee una línea del fichero de entrada, que representa un patrón.

3. Se evalúa el contenido de la parte izquierda del patrón:
    a. Si el valor es positivo, es una categoría gramatical. Se suma 1 a la longitud total.

    b. Si el valor es negativo, es un patrón, por lo que se suma a la longitud total la longitud del patrón. Este valor se obtiene a partir de la variable patternLength.

4. Se evalúa el contenido de la parte derecha del patrón de una forma similar al paso 3. Las longitudes de la parte izquierda y derecha se suman.

5. Se añade a la variable patternLength una dupla que consiste en el identificador del patrón estudiado y la longitud calculada.

6. Se escribe en el fichero de salida el contenido del patrón, con una columna adicional que indica su longitud.

7. Se repiten los pasos 1 a 6 hasta que se lea un patrón perteneciente a una nueva tabla. Cuando esto sucede, se borra todo el contenido de patternLength para evitar que se estudie la longitud con patrones ajenos a la nueva tabla.

8. Se repiten todos los pasos hasta que no haya más patrones por estudiar.

La salida del algoritmo es un fichero que contiene los patrones en el mismo orden con una columna más que determina la longitud del patrón.

7.8.3 Proceso de agrupación de patrones

Este programa recibe como parámetro de entrada el fichero procesado en el apartado anterior. El algoritmo funciona de la siguiente manera:

1. Se declaran las variables y recursos necesarios para leer un fichero. Además, se declararán las estructuras de datos *patternList* (almacena los patrones procesados) y *translator* (almacena los nombres de los patrones leídos en cada tabla).

2. Se lee una línea del fichero de entrada, que representa un patrón.

3. Se obtienen los valores de los elementos a la izquierda y a la derecha del patrón.

    a. En el caso de que el valor sea negativo, estamos ante un patrón, por lo que se reemplaza el valor leído por el valor que está asociado al identificador del patrón en la variable *translator*.

4. Una vez procesados las partes izquierda y derecha del patrón, se estudia si ya existe un patrón con el mismo contenido.

    a. Si en *patternList* ya existe un patrón igual, se actualiza el patrón de *patternList* sumando el número de repeticiones de cada uno.

   b. Si no es así, estamos ante un nuevo patrón. Se genera un identificador único al patrón (“p” seguido de un número) y se introduce en *patternList*.

5. Se añade una nueva entrada al traductor, que consiste en una dupla con el nombre original del patrón y su nombre durante el proceso.

6. Se repiten los pasos 1 a 6 hasta que se lea un patrón perteneciente a una nueva tabla. Cuando esto sucede, se borra todo el contenido de *translator* para evitar que se asocien patrones distintos pero con identificador igual en diferentes tablas.

7. Se repiten todos los pasos hasta que no haya más patrones por estudiar.

El resultado del algoritmo es un fichero que contiene la siguiente información de cada patrón:

- **Identificador**: identificador único (“p” seguido de un número).

- **Parte izquierda**: elemento en la parte izquierda de un patrón (otro patrón o una categoría gramatical).

- **Parte derecha**: elemento en la parte derecha de un patrón (otro patrón o una categoría gramatical).

- **Longitud**: longitud del patrón, cuyo valor fue obtenido con el algoritmo del apartado anterior.

- **Repeticiones**: número de repeticiones del patrón en los textos estudiados. Es la suma de todos los patrones que el algoritmo haya considerado iguales.

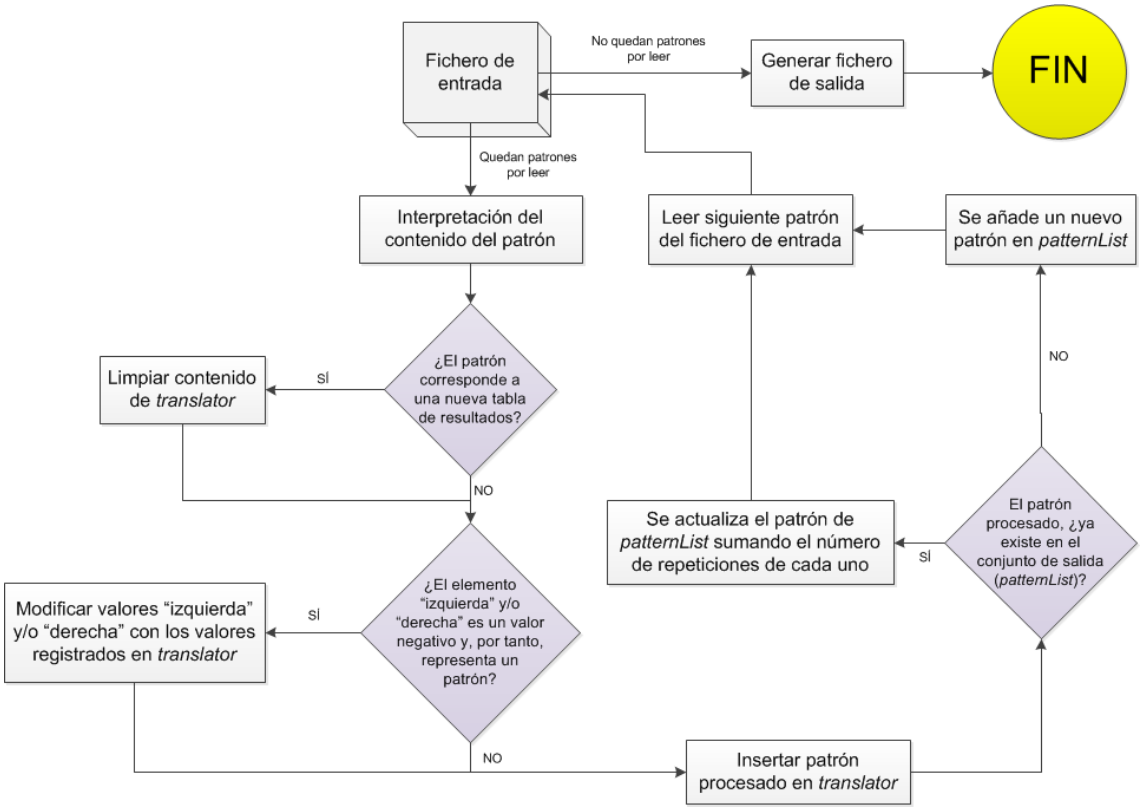

Figura 98. Diagrama de flujo para el algoritmo de unión de resultados

8. Experimentación

En esta sección se mostrarán los resultados obtenidos en los procesos de extracción de patrones en los ocho escenarios identificados (véase sección 7.8).

Por cada escenario, se ha extraído la siguiente información:

- **Estudio de patrones generados**. Se calcula el número total de patrones obtenidos. Además, se muestran los 100 patrones que más se han repetido en el proceso, así como una gráfica que muestra la proporción de repeticiones de los patrones cuya frecuencia supera el 1% de la suma total.

- **Estudio de los patrones del dominio**. Se calcula el número total de patrones, tanto simples como compuestos, que contengan un elemento asociado al dominio de estudio, y se muestran los 100 patrones más repetidos en dicho subconjunto. Asimismo, se calcula la proporción entre patrones que contienen términos del dominio como los que no.

- **Estudio de categorías gramaticales**. Se muestran las categorías gramaticales que más se repiten en los patrones, filtrando entre los conceptos asociados a "Genética" y "Sordera". Además, se muestra la proporción de repeticiones entre categorías correspondientes al dominio como los que no.

- **Estudio de patrones aplicando una fórmula de ponderación**. Se emplea una fórmula que determina el valor útil de cada patrón. Este valor es directamente proporcional a su longitud y a su número de repeticiones en los documentos de estudio, e inversamente proporcional al número de conceptos cuya categoría no está clasificada.

  La fórmula empleada es la siguiente:

$$\boldsymbol{Ponderación} = (5 * longitud) * (0{,}5 * repeticiones) * multDefinido$$

Donde *"multDefinido"* puede adoptar los siguientes valores:

$$\text{multDefinido} = \begin{cases} 1 \text{ ; si todos los términos están clasificados} \\ 0{,}1 * \%\text{términosConocidos ; en otro caso} \end{cases}$$

Donde *"longitud"* acepta los siguientes valores:

$$\text{longitud} = \begin{cases} \text{longitud}*0{,}2; \text{ si el patrón tiene longitud 2.} \\ \text{longitud ; en otro caso} \end{cases}$$

Por ejemplo, sea p1 uno de los patrones descubiertos durante el proceso de análisis:

Patrón	Secuencia	Repeticiones
p1	1144 + 1208 + 1236	10.000

Figura 99. Ejemplo de patrón descubierto en el análisis

El patrón p1 tiene dos términos que la herramienta conoce (1208, "QUANTIFIER DETERMINER" y 1236, "PREPOSITION BY"), mientras que desconoce un término (1144, "UNCLASSIFIED NOUN"). Al aplicar estos valores a la fórmula, el valor de ponderación es:

$$Ponderación = (5 * 3) * (0{,}5 * 10000) * 0{,}67 * 0{,}1 = \mathbf{5.025\ ud}$$

Toda la información correspondiente a los cuatro diferentes enfoques de estudio fue extraída después de realizar el proceso de extracción de patrones y realizar la agrupación de los experimentos correspondiente a cada escenario.

Los elementos contenidos en los patrones se exponen mediante el uso de su identificador en la base de datos Rqa Quality Analyzer. Consulte el Anexo A y el Anexo B para obtener una lista de identificadores de semánticas y categorías con su nombre correspondiente.

8.1. Experimento 1: frecuencia = 1, sin semántica

8.1.1 Estudio de patrones generados

Se han generado un total de 140.909 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.571
7	p8	1144	1119	2	53.455
8	p7	p2	1144	3	45.762
9	p9	1123	1144	2	43.159
10	p10	1224	p1	3	38.738
11	p11	p4	p4	6	32.092
12	p12	1144	1237	2	29.788
13	p13	1144	1248	2	28.730
14	p14	1230	1108	2	27.125
15	p52	1110	1144	2	26.481
16	p15	1144	1213	2	24.086
17	p17	1119	1237	2	21.721
18	p21	p6	1237	3	19.231
19	p20	1144	1108	2	18.493
20	p26	p1	1119	3	17.675
21	p18	1123	1110	2	17.386
22	p19	1144	1230	2	16.448
23	p27	1144	1151	2	15.980
24	p22	p1	1237	3	15.380
25	p25	p1	1248	3	14.989
26	p28	1144	1123	2	14.883
27	p24	1110	p7	4	14.726
28	p16	1123	1123	2	14.007
29	p29	1144	1229	2	12.439
30	p31	1151	1110	2	11.938
31	p64	1110	p1	3	11.818
32	p36	p1	1213	3	11.450
33	p23	p1	p3	5	11.411
34	p33	1224	p3	4	11.407
35	p30	p1	p2	4	10.147

36	p37	p8	1110	3	9.207
37	p34	p1	1108	3	9.115
38	p47	1224	1119	2	8.426
39	p38	p2	1248	3	8.380
40	p40	1144	1228	2	8.359
41	p56	1230	1151	2	8.314
42	p42	p1	1123	3	8.257
43	p35	p10	1237	4	8.072
44	p48	1230	1103	2	7.923
45	p32	p1	1230	3	7.918
46	p39	1151	1108	2	7.359
47	p43	1224	p17	3	7.076
48	p44	1144	1197	2	7.064
49	p76	1110	1248	2	7.051
50	p60	p1	1151	3	6.779
51	p45	1224	1103	2	6.755
52	p67	p3	1119	4	6.570
53	p54	p11	p11	12	6.517
54	p68	1123	p1	3	6.460
55	p66	1144	1158	2	6.458
56	p49	1158	1237	2	6.420
57	p46	1144	1286	2	6.337
58	p57	1224	p5	5	6.258
59	p81	p4	1144	4	6.132
60	p53	p2	p1	4	6.117
61	p51	p6	1119	3	5.863
62	p50	1144	1236	2	5.842
63	p63	1230	p15	3	5.718
64	p62	1144	1103	2	5.475
65	p55	1213	1223	2	5.442
66	p70	p3	1123	4	5.382
67	p95	1119	1144	2	5.343
68	p61	p2	1119	3	5.322
69	p73	p3	1248	4	5.172
70	p69	1144	p9	3	5.168
71	p71	p8	1237	3	5.120
72	p59	p2	p2	4	5.107
73	p78	1108	1213	2	4.932
74	p72	1108	1229	2	4.930
75	p77	p5	p1	6	4.914
76	p75	p2	1286	3	4.801
77	p91189	1144	p52	3	4.741
78	p90	1151	1144	2	4.679
79	p110	p4	p1	5	4.447
80	p146	1230	1144	2	4.233

81	p84	p6	1248	3	4.195
82	p83	1119	1110	2	4.193
83	p101	p5	p5	8	4.175
84	p41	p16	p16	4	4.143
85	p102	p3	1213	4	4.043
86	p87	1223	p1	3	3.924
87	p82	p3	1237	4	3.914
88	p105	p14	1213	3	3.776
89	p114	1108	1221	2	3.748
90	p91190	p52	p52	4	3.732
91	p86	1213	1144	2	3.667
92	p93	1223	1103	2	3.562
93	p88	p4	1248	4	3.536
94	p103	p1	1229	3	3.459
95	p107	p3	1151	4	3.420
96	p111	1103	1213	2	3.414
97	p115	p14	1236	3	3.390
98	p80	1221	1144	2	3.388
99	p122	1223	1144	2	3.375
100	p116	p14	1229	3	3.334

Figura 100. Patrones más repetidos del experimento 1

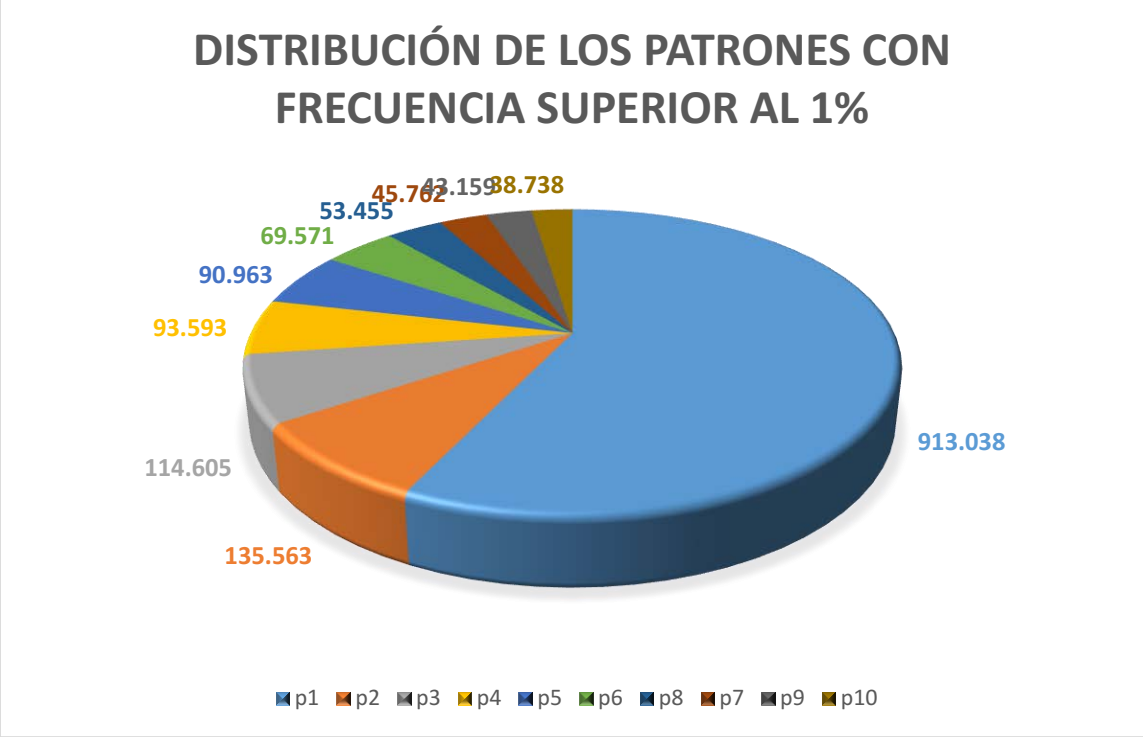

Figura 101. Gráfico de los patrones más frecuentes del experimento 1

8.1.2 Estudio de los patrones del dominio

De los 140.909 patrones generados en este escenario, 20.666 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 14,66% de los patrones pertenecen al dominio de estudio.

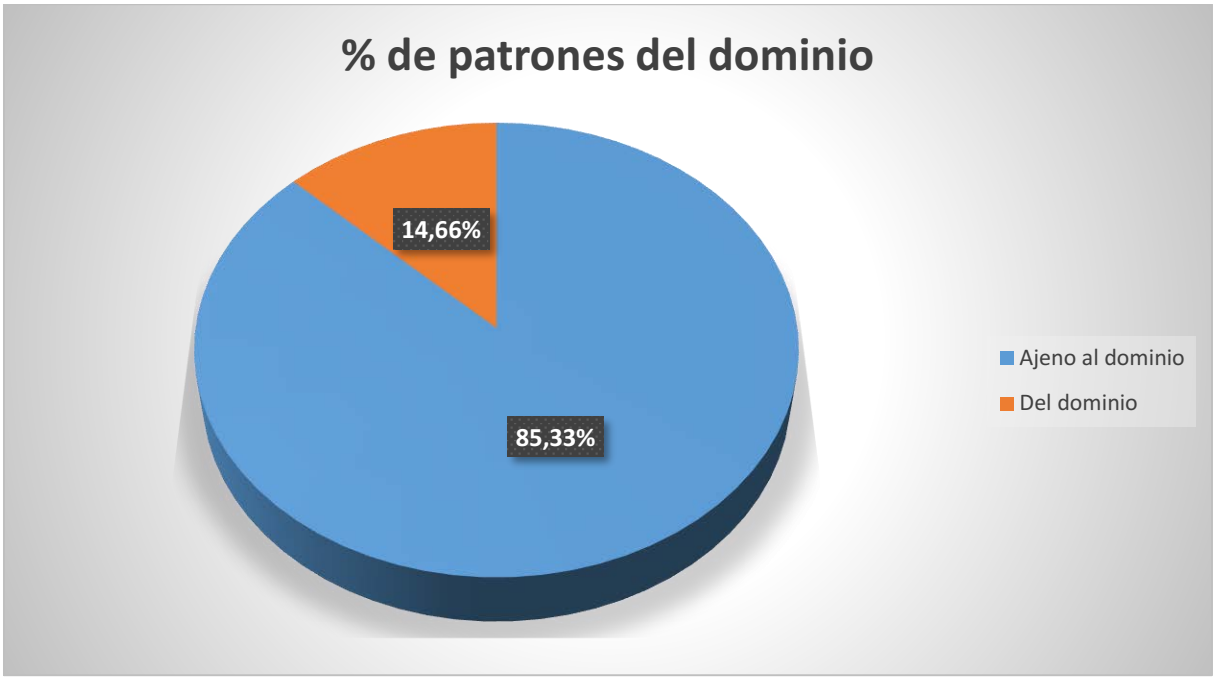

Figura 102. Proporción de patrones en el experimento 1

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p46	1144	1286	2	6.337
2	p75	p2	1286	3	4.801
3	p148	1300	1123	2	3.206
4	p98	1411	1213	2	3.181
5	p126	p6	1309	3	2.582
6	p219	1144	1411	2	2.424
7	p294	1144	1309	2	2.169
8	p109	p46	1110	3	2.044
9	p70197	1224	1293	2	1.922
10	p184	1286	1110	2	1.793
11	p100	1291	1144	2	1.132
12	p438	p1	1286	3	1.098
13	p251	1110	p75	4	1.097
14	p460	p3	1286	4	1.024

15	p401	p4	1286	4	935
16	p478	1110	1286	2	881
17	p339	1224	1411	2	878
18	p324	1224	1309	2	808
19	p705	1286	1151	2	791
20	p5249	1144	1288	2	769
21	p112	p35	p100	6	731
22	p420	1103	1286	2	671
23	p684	1300	p9	3	657
24	p70223	1293	1144	2	647
25	p509	p6	1411	3	644
26	p881	1309	1144	2	583
27	p427	1411	1237	2	522
28	p70206	p81	p184	6	516
29	p369	1286	1144	2	506
30	p596	p31	1286	3	495
31	p311	1144	1564	2	488
32	p70233	p15	1293	3	455
33	p1076	p1	1309	3	436
34	p613	1221	1411	2	419
35	p2157	1288	1144	2	377
36	p880	1300	1144	2	368
37	p373	p98	p126	5	359
38	p683	1300	p18	3	352
39	p70316	1144	1293	2	340
40	p91208	p91196	1286	4	337
41	p70230	1293	1411	2	335
42	p70478	p52	1286	3	334
43	p1681	1123	1309	2	333
44	p830	1224	1300	2	327
45	p1551	1363	1123	2	326
46	p691	1309	1110	2	326
47	p794	1345	1228	2	314
48	p954	1224	1291	2	313
49	p1890	1224	1294	2	302
50	p1315	1291	p1	3	284
51	p100715	p5221	1286	4	281
52	p70242	p1	1411	3	274
53	p956	1224	1558	2	272
54	p561	p30	1286	5	269
55	p1394	p5	1286	5	266
56	p1931	1286	p1	3	266
57	p840	p13	1286	3	265
58	p81470	p9	1293	3	257
59	p822	p6	1526	3	255

60	p1659	p6	1308	3	254
61	p1215	1248	1286	2	243
62	p72001	1224	1269	2	232
63	p784	1144	1283	2	227
64	p2184	1411	1248	2	226
65	p4531	p148	p3	5	225
66	p955	1224	1526	2	225
67	p1106	p25	1286	4	223
68	p2959	1123	1286	2	221
69	p3672	p10	1309	4	218
70	p968	p48	1286	3	218
71	p1242	p148	1237	3	217
72	p70227	p148	1248	3	216
73	p1739	1309	1213	2	207
74	p1713	1411	1144	2	205
75	p1013	1224	1564	2	204
76	p1023	p46	1144	3	200
77	p519	1564	1213	2	199
78	p68560	1221	1309	2	198
79	p70302	1293	p1	3	198
80	p1791	p76	p75	5	197
81	p757	1144	1553	2	195
82	p4708	p77	1286	7	192
83	p70367	p81	1286	5	192
84	p5445	1288	p1	3	188
85	p1858	p3	1309	4	186
86	p1539	1224	1553	2	185
87	p3050	1411	1110	2	184
88	p706	1300	p99	3	180
89	p71249	1144	1300	2	180
90	p1538	1224	1459	2	177
91	p100741	p219	1213	3	175
92	p1321	p122	1309	3	175
93	p70479	p12	1293	3	174
94	p70580	1293	1248	2	174
95	p536	1144	1312	2	169
96	p70930	p6	1293	3	169
97	p1017	1411	p86	3	166
98	p70396	p1	1288	3	166
99	p70249	p70187	1286	6	165
100	p1428	1248	p75	4	159

Figura 103. Patrones del dominio más repetidos del experimento 1

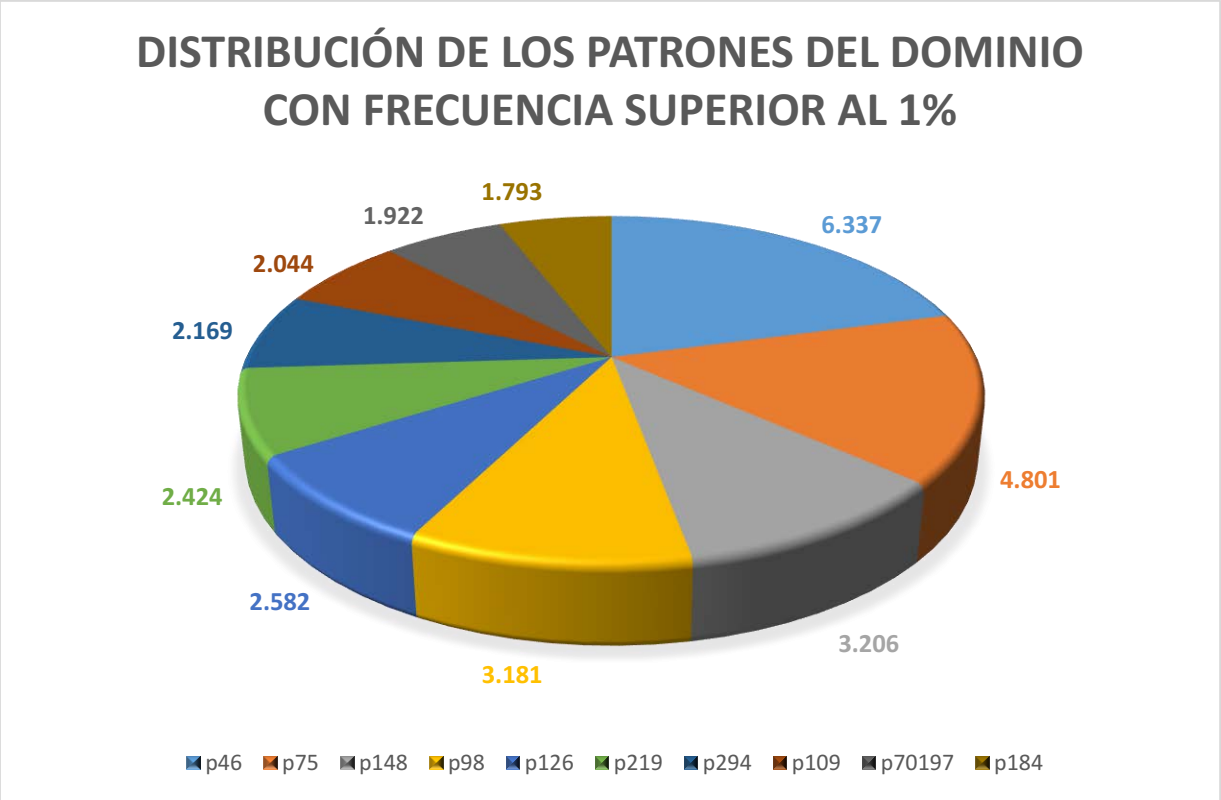

Figura 104. Gráfico de patrones del dominio en el experimento 1

8.1.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 221 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1218	QUANTIFIER DETERMINER	1.377
2	1213	PREPOSITION	1.222
3	1151	ADVERB	1.193
4	1119	NOUN	1.164
5	1123	NUMBER	1.138
6	1108	VERB	1.102
7	1144	UNCLASSIFIED NOUN	1.066
8	1103	ADJECTIVE	922
9	1110	SYMBOL	740
10	1248	AND LINKING	737

Figura 105. Categorías más frecuentes en el experimento 1

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	491
2	1411	Mutation	466
3	1300	Exons	424
4	1309	Genes	413
5	1293	Genes, Neurofibromatosis 1	382
6	1288	Genes, Tumor Suppressor	322
7	1291	Genes, Wilms Tumor	256
8	1526	Phenotype	181
9	1308	Alleles	180
10	1294	Genes, Neurofibromatosis 2	178

Figura 106. Categorías de genética más frecuentes en el experimento 1

Limitando los resultados a sólo los conceptos asociados a la sordera, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	248
2	1553	Deafness	184
3	1561	Hearing Loss, Sensorineural	180
4	1558	Hearing Loss, Functional	126
5	1563	Wolfram Syndrome	57
6	1559	Hearing Loss, High-Frequency	56
7	1555	Hearing Loss, Bilateral	46
8	1556	Hearing Loss, Central	34
9	1562	Presbycusis	17
10	1565	Hearing Loss, Unilateral	16

Figura 107. Categorías de sordera más frecuentes en el experimento 1

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

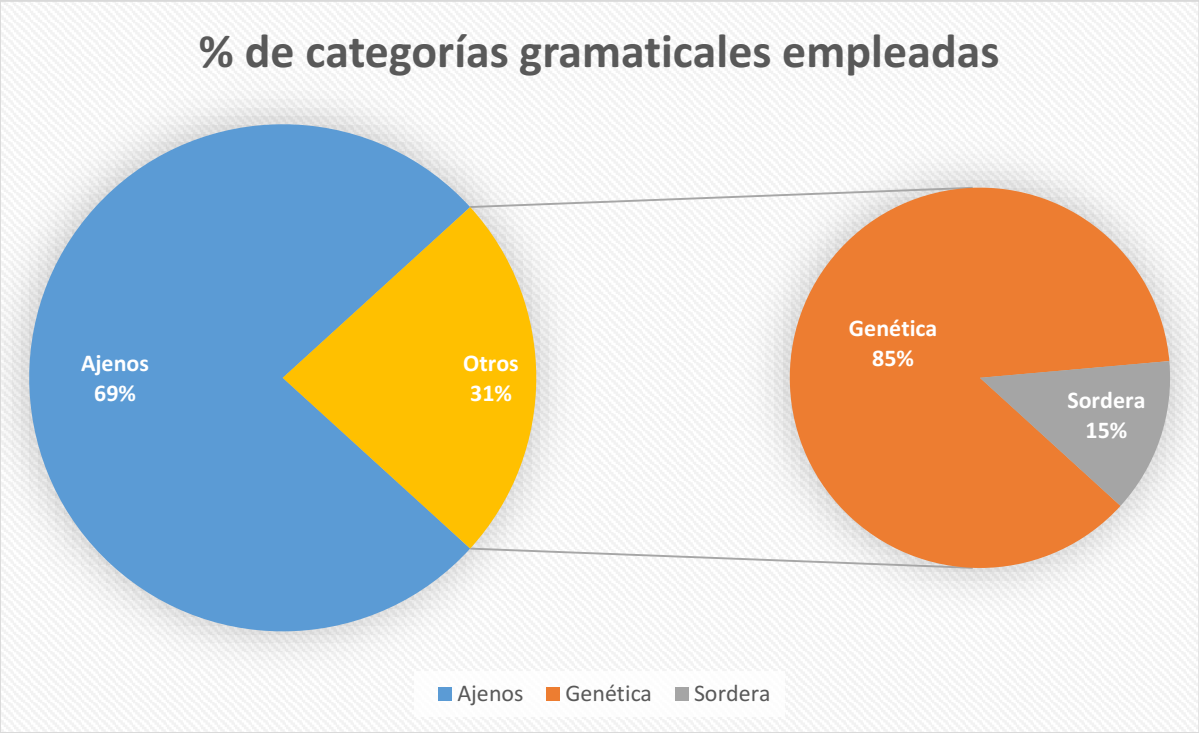

Figura 108. Proporción de categorías gramaticales en el experimento 1

Aproximadamente el 69% de elementos que forman los patrones son ajenos a nuestro dominio. El 31% forman parte del dominio, de los cuales el 85% están asociados a la "Genética" y el 15% restante a la "Sordera".

8.1.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p43	1224	p17	3	7.076	**53.070**
2	p41	p16	p16	4	4.143	**41.430**
3	p104	p41	p41	8	1.619	**32.380**
4	p105	p14	1213	3	3.776	**28.320**
5	p115	p14	1236	3	3.390	**25.425**
6	p116	p14	1229	3	3.334	**25.005**
7	p4	p1	1110	3	93.593	**23.866**
8	p287	p104	p104	16	552	**22.080**
9	p160	p18	1123	3	2.215	**16.612**
10	p11	p4	p4	6	32.092	**16.366**
11	p14	1230	1108	2	27.125	**13.562**

12	p164	1224	p49	3	1.672	**12.540**
13	p191	p18	p18	4	1.240	**12.400**
14	p7	p2	1144	3	45.762	**11.669**
15	p17	1119	1237	2	21.721	**10.860**
16	p247	p48	1213	3	1.330	**9.975**
17	p270	p48	1229	3	1.320	**9.900**
18	p10	1224	p1	3	38.738	**9.878**
19	p21	p6	1237	3	19.231	**9.663**
20	p18	1123	1110	2	17.386	**8.693**
21	p295	p14	p55	4	860	**8.600**
22	p86909	p287	p287	32	105	**8.400**
23	p379	p48	1228	3	1.062	**7.965**
24	p24	1110	p7	4	14.726	**7.363**
25	p348	p287	1123	17	173	**7.352**
26	p16	1123	1123	2	14.007	**7.003**
27	p54	p11	p11	12	6.517	**6.647**
28	p165	1230	p39	3	852	**6.390**
29	p31	1151	1110	2	11.938	**5.969**
30	p331	1233	p14	3	775	**5.812**
31	p125395	p125394	p125392	60	36	**5.400**
32	p304	p31	1203	3	718	**5.385**
33	p492	1223	p17	3	672	**5.040**
34	p415	1153	p56	3	662	**4.965**
35	p107471	p3753	p3753	4	492	**4.920**
36	p435	1203	p56	3	644	**4.830**
37	p529	p14	1228	3	629	**4.717**
38	p125394	p86909	p287	48	39	**4.680**
39	p37	p8	1110	3	9.207	**4.626**
40	p791	1233	p105	4	455	**4.550**
41	p26	p1	1119	3	17.675	**4.507**
42	p315	1108	p55	3	593	**4.447**
43	p548	1233	p116	4	438	**4.380**
44	p47	1224	1119	2	8.426	**4.213**
45	p38	p2	1248	3	8.380	**4.210**
46	p56	1230	1151	2	8.314	**4.157**
47	p659	1110	p31	3	543	**4.072**
48	p35	p10	1237	4	8.072	**4.036**
49	p48	1230	1103	2	7.923	**3.961**
50	p22	p1	1237	3	15.380	**3.921**
51	p25	p1	1248	3	14.989	**3.822**
52	p518	p14	1166	3	508	**3.810**
53	p107473	p107471	p107471	8	190	**3.800**
54	p596	p31	1286	3	495	**3.712**
55	p39	1151	1108	2	7.359	**3.679**
56	p965	1103	p17	3	483	**3.622**

57	p76	1110	1248	2	7.051	**3.525**
58	p70216	p56	p78	4	348	**3.480**
59	p720	p14	1151	3	463	**3.472**
60	p511	1203	p39	3	461	**3.457**
61	p307	p99	1123	3	454	**3.405**
62	p2	1144	1110	2	135.563	**3.389**
63	p45	1224	1103	2	6.755	**3.377**
64	p70188	p47	1237	3	450	**3.375**
65	p672	p45	p17	4	334	**3.340**
66	p502	p48	1236	3	430	**3.225**
67	p49	1158	1237	2	6.420	**3.210**
68	p798	p274	p115	5	245	**3.062**
69	p832	1228	p246	3	408	**3.060**
70	p64	1110	p1	3	11.818	**3.013**
71	p630	p47	1248	3	400	**3.000**
72	p190	1221	p79	3	396	**2.970**
73	p1136	1240	p105	4	297	**2.970**
74	p51	p6	1119	3	5.863	**2.946**
75	p508	p14	1221	3	392	**2.940**
76	p36	p1	1213	3	11.450	**2.919**
77	p70248	1203	p114	3	387	**2.902**
78	p63	1230	p15	3	5.718	**2.873**
79	p495	p39	1229	3	382	**2.865**
80	p33	1224	p3	4	11.407	**2.851**
81	p453	p39	1213	3	373	**2.797**
82	p55	1213	1223	2	5.442	**2.721**
83	p61	p2	1119	3	5.322	**2.674**
84	p625	p48	p140	4	267	**2.670**
85	p683	1300	p18	3	352	**2.640**
86	p107487	p107473	p107473	16	65	**2.600**
87	p71	p8	1237	3	5.120	**2.572**
88	p544	p14	1248	3	342	**2.565**
89	p59	p2	p2	4	5.107	**2.553**
90	p1163	p56	p111	4	255	**2.550**
91	p30	p1	p2	4	10.147	**2.536**
92	p409	p18	1248	3	333	**2.497**
93	p854	p55	1103	3	329	**2.467**
94	p78	1108	1213	2	4.932	**2.466**
95	p72	1108	1229	2	4.930	**2.465**
96	p75	p2	1286	3	4.801	**2.412**
97	p10358	p16	1123	3	321	**2.407**
98	p765	p194	1236	3	320	**2.400**
99	p507	p45	1119	3	320	**2.400**
100	p34	p1	1108	3	9.115	**2.324**

Figura 109. Patrones con mayor ponderación del experimento 1

El patrón con la mayor ponderación es p43, cuya secuencia es la siguiente:

- **p43**: “1224 + p17” = “1224 + 1119 + 1237” = “DEFINITE ARTICLE + NOUN + PREPOSITION OF”.

El patrón tiene longitud 3 y se repitió 7.076 veces en el texto, por lo que ocupa la posición 47º en la lista de patrones más frecuentes en este escenario de estudio.

Algunos otros patrones ponderados de interés encontrados en la lista son:

- **p164**: “1224 + p49” = “1224 + 1158 + 1237” = “DEFINITE ARTICLE + NOT GROUPING NOUN + PREPOSITION OF”.
- **p295**: “p14 + p55” = “1230 + 1108 + p55” = “1230 +1108 + 1213 + 1223” = “VERB TO BE + VERB + PREPOSITION + INDEFINITE ARTICLE”.
- **p70216**: “p56 + p78” = “1230 + 1151 + p78” = “1230 + 1151 + 1108 + 1213” = “VERB TO BE + ADVERB + VERB + PREPOSITION

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p596	p31	1286	3	495	**3.712**
2	p683	1300	p18	3	352	**2.640**
3	p75	p2	1286	3	4.801	**2.412**
4	p968	p48	1286	3	218	**1.635**
5	p1242	p148	1237	3	217	**1.627**
6	p70227	p148	1248	3	216	**1.620**
7	p148	1300	1123	2	3.206	**1.603**
8	p98	1411	1213	2	3.181	**1.590**
9	p706	1300	p99	3	180	**1.350**
10	p126	p6	1309	3	2.582	**1.297**
11	p70251	p148	p262	4	111	**1.110**
12	p70616	p111	1293	3	139	**1.042**
13	p109	p46	1110	3	2.044	**1.027**
14	p1046	p18	1286	3	133	**997**

15	p70197	1224	1293	2	1.922	**961**
16	p100816	1411	p787	3	127	**952**
17	p215	p214	p112	18	327	**912**
18	p184	1286	1110	2	1.793	**896**
19	p70556	p98	p70197	4	86	**860**
20	p1994	p83	1286	3	111	**832**
21	p251	1110	p75	4	1.097	**822**
22	p1127	p273	1286	3	108	**810**
23	p70498	1237	p70197	3	107	**802**
24	p1999	p76	1286	3	101	**757**
25	p1217	1300	p307	4	73	**730**
26	p1888	1224	p98	3	91	**682**
27	p1216	1291	p436	3	89	**667**
28	p70899	p17	p70197	4	66	**660**
29	p1209	1223	p98	3	79	**592**
30	p2506	1363	p99	3	78	**585**

Figura 110. Patrones del dominio con mejor ponderación del experimento 1

El patrón del dominio con la mayor ponderación es p596, cuya secuencia es la siguiente:

- **p596**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 495 veces en el texto, por lo que ocupa la posición 30º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.2. Experimento 2: frecuencia = 5, sin semántica

8.2.1 Estudio de patrones generados

Se han generado un total de 14.891 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.571
7	p8	1144	1119	2	53.455
8	p7	p2	1144	3	45.762
9	p9	1123	1144	2	43.159
10	p10	1224	p1	3	38.738
11	p11	p4	p4	6	32.092
12	p12	1144	1237	2	29.788
13	p13	1144	1248	2	28.730
14	p14	1230	1108	2	27.125
15	p52	1110	1144	2	26.481
16	p15	1144	1213	2	24.086
17	p17	1119	1237	2	21.721
18	p21	p6	1237	3	19.231
19	p20	1144	1108	2	18.493
20	p26	p1	1119	3	17.675
21	p18	1123	1110	2	17.386
22	p19	1144	1230	2	16.448
23	p27	1144	1151	2	15.980
24	p22	p1	1237	3	15.380
25	p25	p1	1248	3	14.989
26	p28	1144	1123	2	14.883
27	p24	1110	p7	4	14.726
28	p16	1123	1123	2	14.007
29	p29	1144	1229	2	12.439
30	p31	1151	1110	2	11.938
31	p64	1110	p1	3	11.818
32	p36	p1	1213	3	11.450
33	p23	p1	p3	5	11.411
34	p33	1224	p3	4	11.407
35	p30	p1	p2	4	10.147

36	p37	p8	1110	3	9.207
37	p34	p1	1108	3	9.115
38	p47	1224	1119	2	8.426
39	p38	p2	1248	3	8.380
40	p40	1144	1228	2	8.359
41	p56	1230	1151	2	8.314
42	p42	p1	1123	3	8.257
43	p35	p10	1237	4	8.072
44	p48	1230	1103	2	7.923
45	p32	p1	1230	3	7.918
46	p39	1151	1108	2	7.359
47	p43	1224	p17	3	7.076
48	p44	1144	1197	2	7.064
49	p76	1110	1248	2	7.051
50	p60	p1	1151	3	6.779
51	p45	1224	1103	2	6.755
52	p67	p3	1119	4	6.570
53	p54	p11	p11	12	6.517
54	p68	1123	p1	3	6.460
55	p66	1144	1158	2	6.458
56	p49	1158	1237	2	6.420
57	p46	1144	1286	2	6.337
58	p57	1224	p5	5	6.258
59	p81	p4	1144	4	6.132
60	p53	p2	p1	4	6.117
61	p51	p6	1119	3	5.863
62	p50	1144	1236	2	5.842
63	p63	1230	p15	3	5.718
64	p62	1144	1103	2	5.475
65	p55	1213	1223	2	5.442
66	p70	p3	1123	4	5.382
67	p95	1119	1144	2	5.343
68	p61	p2	1119	3	5.322
69	p73	p3	1248	4	5.172
70	p69	1144	p9	3	5.168
71	p71	p8	1237	3	5.120
72	p59	p2	p2	4	5.107
73	p78	1108	1213	2	4.932
74	p72	1108	1229	2	4.930
75	p77	p5	p1	6	4.914
76	p75	p2	1286	3	4.801
77	p11737	1144	p52	3	4.741
78	p90	1151	1144	2	4.679
79	p110	p4	p1	5	4.447
80	p146	1230	1144	2	4.233

81	p84	p6	1248	3	4.195
82	p83	1119	1110	2	4.193
83	p101	p5	p5	8	4.175
84	p41	p16	p16	4	4.143
85	p102	p3	1213	4	4.043
86	p87	1223	p1	3	3.924
87	p82	p3	1237	4	3.914
88	p105	p14	1213	3	3.776
89	p114	1108	1221	2	3.748
90	p11738	p52	p52	4	3.732
91	p86	1213	1144	2	3.667
92	p93	1223	1103	2	3.562
93	p88	p4	1248	4	3.536
94	p103	p1	1229	3	3.459
95	p107	p3	1151	4	3.420
96	p111	1103	1213	2	3.414
97	p115	p14	1236	3	3.390
98	p80	1221	1144	2	3.388
99	p122	1223	1144	2	3.375
100	p116	p14	1229	3	3.334

Figura 111. Patrones más repetidos del experimento 2

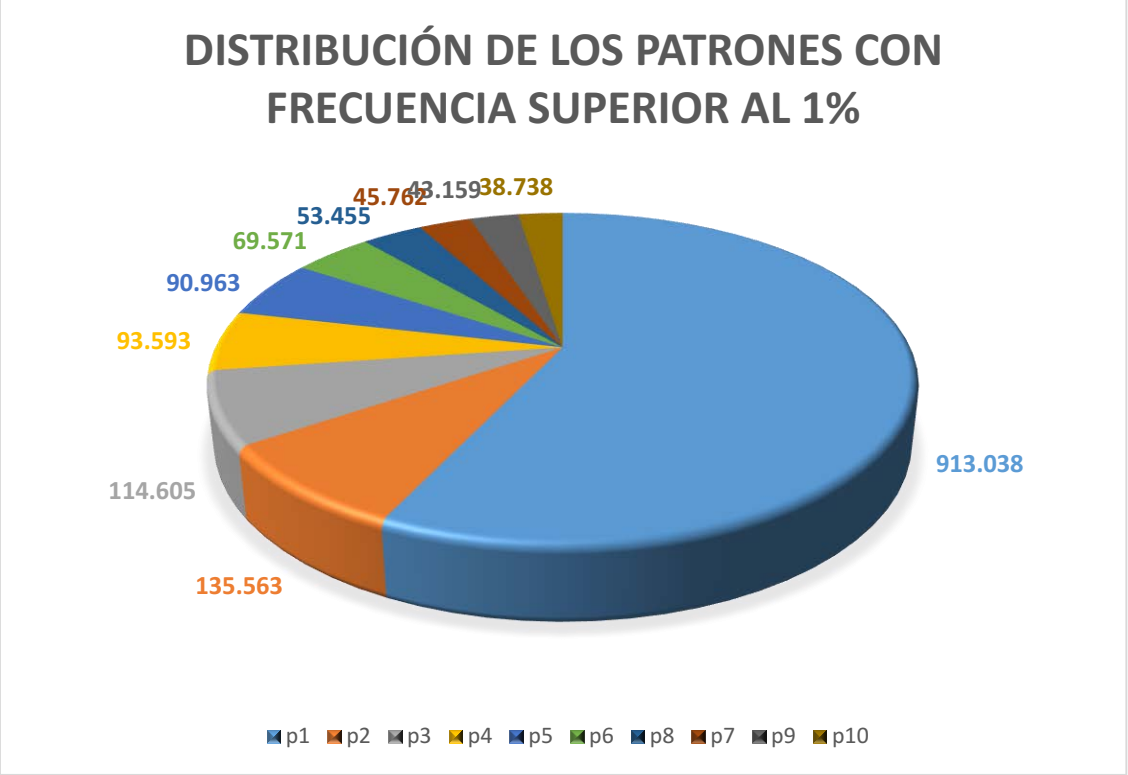

Figura 112. Gráfico de los patrones más frecuentes del experimento 2

8.2.2 Estudio de los patrones del dominio

De los 14.891 patrones generados en este escenario, 2.121 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 14,66% de los patrones pertenecen al dominio de estudio.

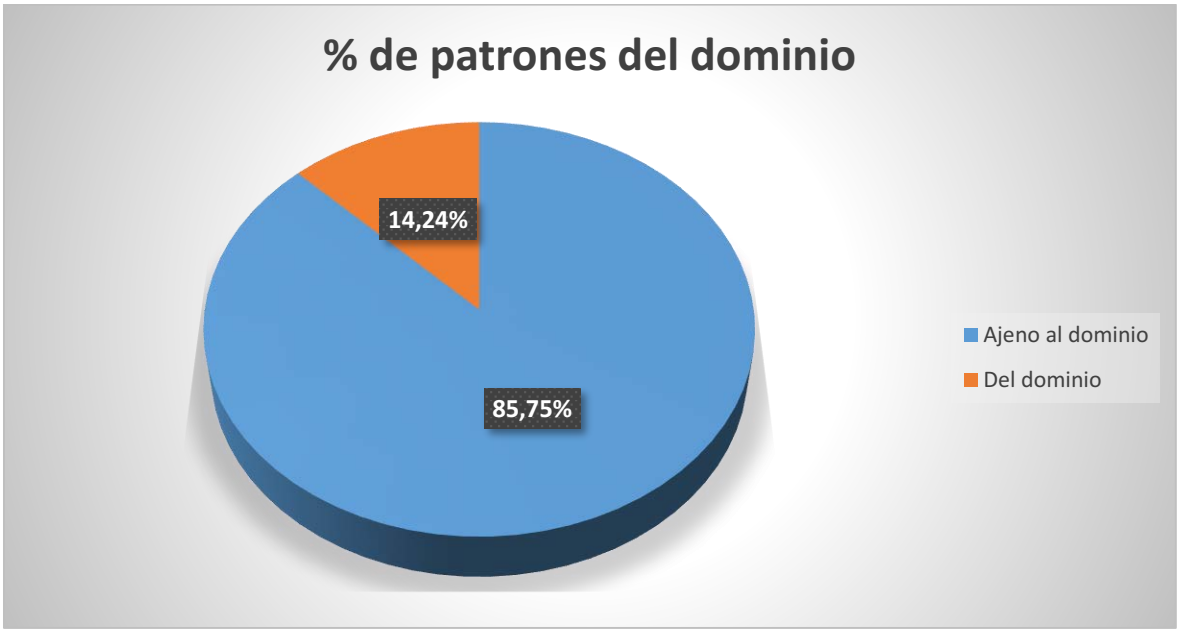

Figura 113. Proporción de patrones en el experimento 2

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Id	PatternName	LeftSide	RightSide	Length	Repeats
1	p46	1144	1286	2	6.337
2	p75	p2	1286	3	4.801
3	p148	1300	1123	2	3.206
4	p98	1411	1213	2	3.181
5	p126	p6	1309	3	2.582
6	p219	1144	1411	2	2.424
7	p294	1144	1309	2	2.169
8	p109	p46	1110	3	2.044
9	p10155	1224	1293	2	1.922
10	p184	1286	1110	2	1.793
11	p100	1291	1144	2	1.130
12	p438	p1	1286	3	1.098
13	p251	1110	p75	4	1.097

14	p460	p3	1286	4	1.024
15	p401	p4	1286	4	935
16	p478	1110	1286	2	881
17	p339	1224	1411	2	878
18	p324	1224	1309	2	804
19	p705	1286	1151	2	791
20	p5249	1144	1288	2	769
21	p112	p35	p100	6	729
22	p420	1103	1286	2	671
23	p684	1300	p9	3	652
24	p10182	1293	1144	2	645
25	p509	p6	1411	3	644
26	p881	1309	1144	2	583
27	p10164	p81	p184	6	516
28	p427	1411	1237	2	515
29	p369	1286	1144	2	496
30	p596	p31	1286	3	492
31	p311	1144	1564	2	483
32	p10193	p15	1293	3	455
33	p1076	p1	1309	3	433
34	p613	1221	1411	2	419
35	p2157	1288	1144	2	377
36	p880	1300	1144	2	365
37	p373	p98	p126	5	359
38	p683	1300	p18	3	352
39	p11756	p11744	1286	4	337
40	p10276	1144	1293	2	336
41	p1681	1123	1309	2	333
42	p10189	1293	1411	2	332
43	p10442	p52	1286	3	332
44	p691	1309	1110	2	321
45	p830	1224	1300	2	317
46	p1551	1363	1123	2	312
47	p794	1345	1228	2	311
48	p954	1224	1291	2	310
49	p1890	1224	1294	2	298
50	p1315	1291	p1	3	281
51	p12567	p5221	1286	4	281
52	p561	p30	1286	5	269
53	p10202	p1	1411	3	269
54	p956	1224	1558	2	264
55	p840	p13	1286	3	260
56	p1931	1286	p1	3	259
57	p1394	p5	1286	5	256
58	p822	p6	1526	3	255

59	p11463	p9	1293	3	254
60	p1659	p6	1308	3	252
61	p1215	1248	1286	2	233
62	p11144	1224	1269	2	228
63	p784	1144	1283	2	227
64	p955	1224	1526	2	223
65	p2184	1411	1248	2	219
66	p2959	1123	1286	2	217
67	p1242	p148	1237	3	217
68	p1106	p25	1286	4	216
69	p10186	p148	1248	3	216
70	p968	p48	1286	3	214
71	p4531	p148	p3	5	213
72	p3672	p10	1309	4	207
73	p1739	1309	1213	2	200
74	p1023	p46	1144	3	200
75	p1713	1411	1144	2	199
76	p10313	1221	1309	2	192
77	p4708	p77	1286	7	192
78	p10328	p81	1286	5	192
79	p1013	1224	1564	2	191
80	p10262	1293	p1	3	189
81	p757	1144	1553	2	187
82	p1791	p76	p75	5	187
83	p519	1564	1213	2	186
84	p1858	p3	1309	4	184
85	p3050	1411	1110	2	181
86	p706	1300	p99	3	180
87	p1539	1224	1553	2	176
88	p1321	p122	1309	3	175
89	p12605	p219	1213	3	175
90	p1538	1224	1459	2	169
91	p10545	1293	1248	2	169
92	p10915	p6	1293	3	167
93	p1017	1411	p86	3	166
94	p10209	p10145	1286	6	165
95	p5445	1288	p1	3	165
96	p536	1144	1312	2	161
97	p10443	p12	1293	3	160
98	p1428	1248	p75	4	157
99	p10990	1144	1300	2	156
100	p1682	1144	p98	3	154

Figura 114. Patrones del dominio más repetidos del experimento 2

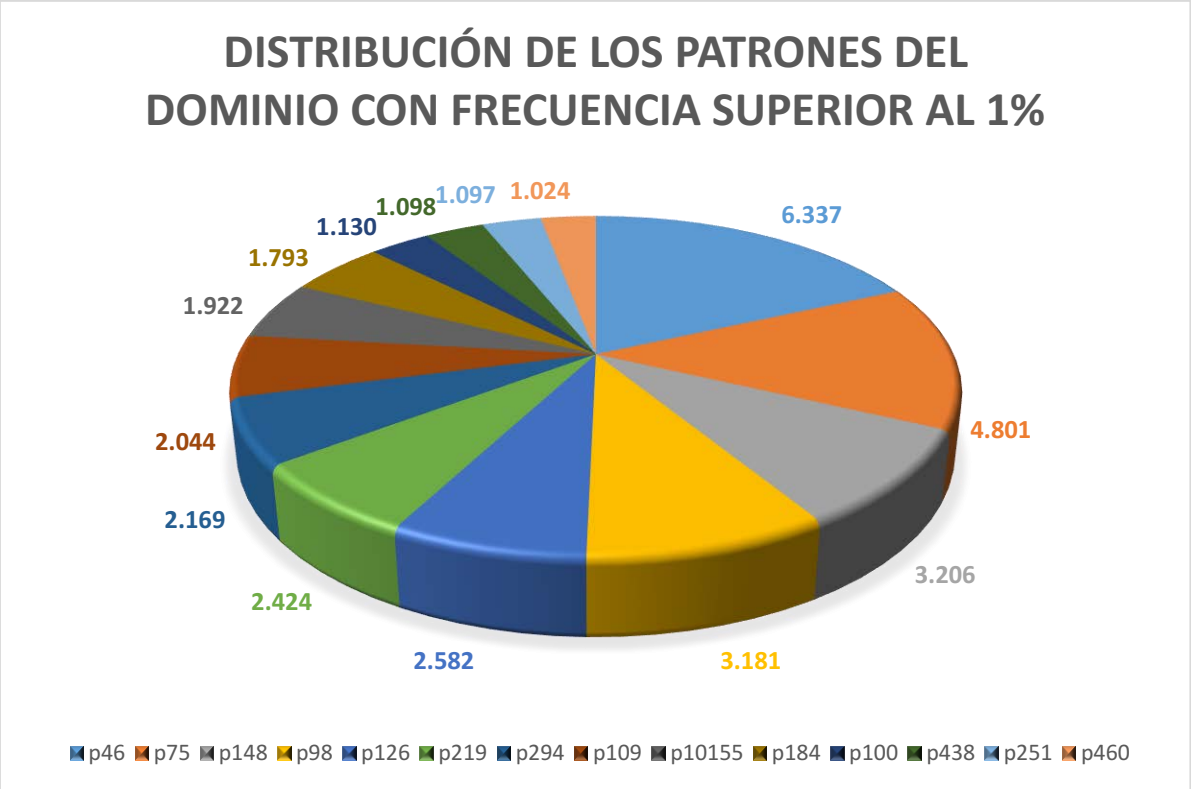

Figura 115. Gráfico de patrones del dominio en el experimento 2

8.2.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 221 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	446
2	1123	NUMBER	376
3	1119	NOUN	339
4	1151	ADVERB	336
5	1213	PREPOSITION	318
6	1108	VERB	307
7	1110	SYMBOL	241
8	1103	ADJECTIVE	231
9	1248	AND LINKING	222
10	1218	QUANTIFIER DETERMINER	221

Figura 116. Categorías más frecuentes en el experimento 2

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	160
2	1411	Mutation	104
3	1309	Genes	95
4	1300	Exons	84
5	1293	Genes, Neurofibromatosis 1	76
6	1288	Genes, Tumor Suppressor	43
7	1291	Genes, Wilms Tumor	41
8	1294	Genes, Neurofibromatosis 2	39
9	1363	Introns	37
10	1526	Phenotype	32

Figura 117. Categorías de genética más frecuentes en el experimento 2

Limitando los resultados a sólo los conceptos asociados a la sordera, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	56
2	1561	Hearing Loss, Sensorineural	33
3	1553	Deafness	33
4	1558	Hearing Loss, Functional	26
5	1563	Wolfram Syndrome	10
6	1555	Hearing Loss, Bilateral	7
7	1559	Hearing Loss, High-Frequency	6
8	1556	Hearing Loss, Central	6
9	1567	Usher Syndromes	4
10	1566	Hearing Loss, Mixed Conductive-Sensorineural	2

Figura 118. Categorías de sordera más frecuentes en el experimento 2

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

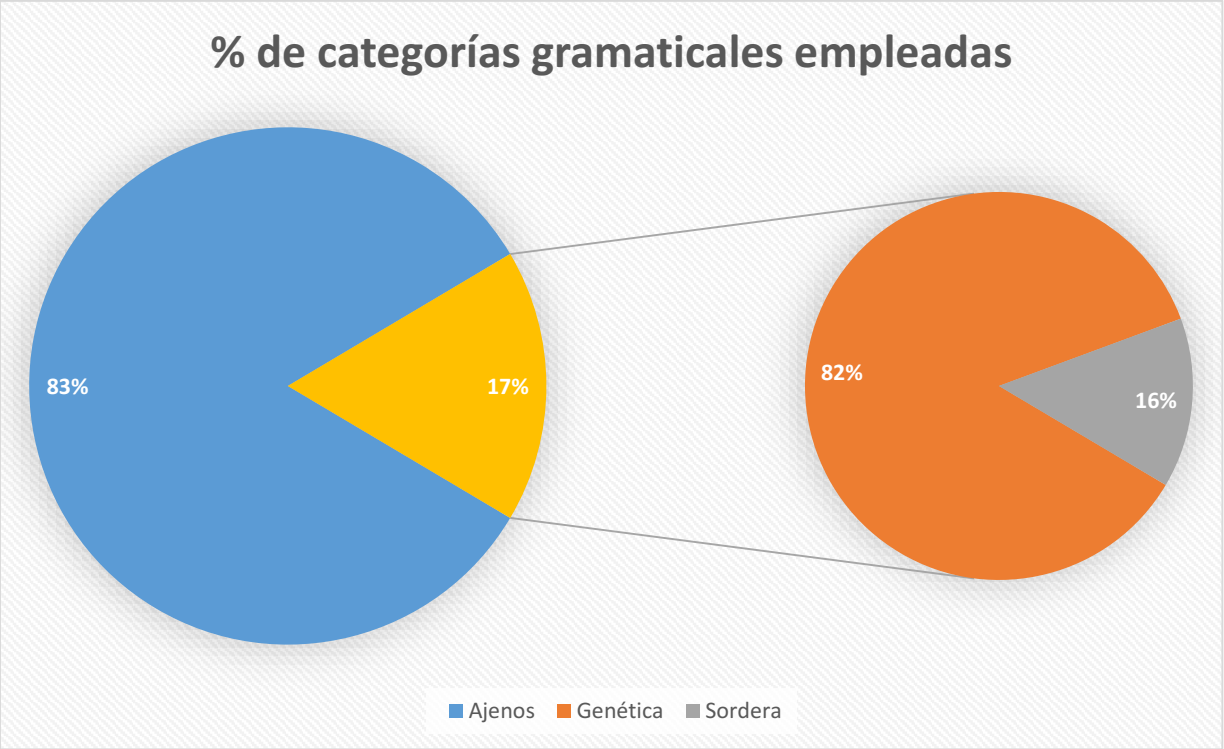

Figura 119. Proporción de categorías gramaticales en el experimento 2

Aproximadamente el 83% de elementos que forman los patrones son ajenos a nuestro dominio. El 17% forman parte del dominio, de los cuales el 82% están asociados a la "Genética" y el 16% restante a la "Sordera".

8.2.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p43	1224	p17	3	7.076	**53.070**
2	p41	p16	p16	4	4.143	**41.430**
3	p104	p41	p41	8	1.608	**32.160**
4	p105	p14	1213	3	3.776	**28.320**
5	p115	p14	1236	3	3.390	**25.425**
6	p116	p14	1229	3	3.334	**25.005**
7	p4	p1	1110	3	93.593	**23.866**
8	p287	p104	p104	16	552	**22.080**
9	p160	p18	1123	3	2.215	**16.612**
10	p11	p4	p4	6	32.092	**16.366**
11	p14	1230	1108	2	27.125	**13.562**

12	p164	1224	p49	3	1.672	**12.540**
13	p191	p18	p18	4	1.240	**12.400**
14	p7	p2	1144	3	45.762	**11.669**
15	p17	1119	1237	2	21.721	**10.860**
16	p247	p48	1213	3	1.330	**9.975**
17	p270	p48	1229	3	1.320	**9.900**
18	p10	1224	p1	3	38.738	**9.878**
19	p21	p6	1237	3	19.231	**9.663**
20	p18	1123	1110	2	17.386	**8.693**
21	p295	p14	p55	4	860	**8.600**
22	p379	p48	1228	3	1.062	**7.965**
23	p11456	p287	p287	32	99	**7.920**
24	p24	1110	p7	4	14.726	**7.363**
25	p348	p287	1123	17	169	**7.182**
26	p16	1123	1123	2	14.007	**7.003**
27	p54	p11	p11	12	6.517	**6.647**
28	p165	1230	p39	3	852	**6.390**
29	p31	1151	1110	2	11.938	**5.969**
30	p331	1233	p14	3	775	**5.812**
31	p14083	p14082	p14080	60	36	**5.400**
32	p304	p31	1203	3	718	**5.385**
33	p492	1223	p17	3	667	**5.002**
34	p415	1153	p56	3	662	**4.965**
35	p12916	p3753	p3753	4	488	**4.880**
36	p435	1203	p56	3	641	**4.807**
37	p529	p14	1228	3	629	**4.717**
38	p14082	p11456	p287	48	39	**4.680**
39	p37	p8	1110	3	9.207	**4.626**
40	p791	1233	p105	4	451	**4.510**
41	p26	p1	1119	3	17.675	**4.507**
42	p315	1108	p55	3	593	**4.447**
43	p548	1233	p116	4	438	**4.380**
44	p47	1224	1119	2	8.426	**4.213**
45	p38	p2	1248	3	8.380	**4.210**
46	p56	1230	1151	2	8.314	**4.157**
47	p659	1110	p31	3	543	**4.072**
48	p35	p10	1237	4	8.072	**4.036**
49	p48	1230	1103	2	7.923	**3.961**
50	p22	p1	1237	3	15.380	**3.921**
51	p25	p1	1248	3	14.989	**3.822**
52	p12918	p12916	p12916	8	190	**3.800**
53	p518	p14	1166	3	503	**3.772**
54	p596	p31	1286	3	492	**3.690**
55	p39	1151	1108	2	7.359	**3.679**
56	p965	1103	p17	3	483	**3.622**

57	p76	1110	1248	2	7.051	**3.525**
58	p10174	p56	p78	4	348	**3.480**
59	p511	1203	p39	3	461	**3.457**
60	p720	p14	1151	3	458	**3.435**
61	p307	p99	1123	3	454	**3.405**
62	p2	1144	1110	2	135.563	**3.389**
63	p45	1224	1103	2	6.755	**3.377**
64	p10146	p47	1237	3	450	**3.375**
65	p672	p45	p17	4	334	**3.340**
66	p49	1158	1237	2	6.420	**3.210**
67	p502	p48	1236	3	426	**3.195**
68	p832	1228	p246	3	404	**3.030**
69	p64	1110	p1	3	11.818	**3.013**
70	p798	p274	p115	5	241	**3.012**
71	p630	p47	1248	3	400	**3.000**
72	p51	p6	1119	3	5.863	**2.946**
73	p508	p14	1221	3	392	**2.940**
74	p1136	1240	p105	4	293	**2.930**
75	p36	p1	1213	3	11.450	**2.919**
76	p10208	1203	p114	3	387	**2.902**
77	p190	1221	p79	3	386	**2.895**
78	p63	1230	p15	3	5.718	**2.873**
79	p33	1224	p3	4	11.407	**2.851**
80	p495	p39	1229	3	378	**2.835**
81	p453	p39	1213	3	373	**2.797**
82	p55	1213	1223	2	5.442	**2.721**
83	p61	p2	1119	3	5.322	**2.674**
84	p625	p48	p140	4	267	**2.670**
85	p683	1300	p18	3	352	**2.640**
86	p12935	p12918	p12918	16	65	**2.600**
87	p71	p8	1237	3	5.120	**2.572**
88	p544	p14	1248	3	342	**2.565**
89	p59	p2	p2	4	5.107	**2.553**
90	p1163	p56	p111	4	255	**2.550**
91	p30	p1	p2	4	10.147	**2.536**
92	p409	p18	1248	3	331	**2.482**
93	p78	1108	1213	2	4.932	**2.466**
94	p72	1108	1229	2	4.930	**2.465**
95	p854	p55	1103	3	325	**2.437**
96	p75	p2	1286	3	4.801	**2.412**
97	p507	p45	1119	3	320	**2.400**
98	p765	p194	1236	3	314	**2.355**
99	p10177	p16	1123	3	313	**2.347**
100	p34	p1	1108	3	9.115	**2.324**

Figura 120. Patrones con mayor ponderación del experimento 2

El patrón con la mayor ponderación es p43, cuya secuencia es la siguiente:

- **p43**: “1224 + p17” = “1224 + 1119 + 1237” = “DEFINITE ARTICLE + NOUN + PREPOSITION OF”.

El patrón tiene longitud 3 y se repitió 7.076 veces en el texto, por lo que ocupa la posición 47º en la lista de patrones más frecuentes en este escenario de estudio.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p596	p31	1286	3	492	**3.690**
2	p683	1300	p18	3	352	**2.640**
3	p75	p2	1286	3	4.801	**2.412**
4	p1242	p148	1237	3	217	**1.627**
5	p10186	p148	1248	3	216	**1.620**
6	p968	p48	1286	3	214	**1.605**
7	p148	1300	1123	2	3.206	**1.603**
8	p98	1411	1213	2	3.181	**1.590**
9	p706	1300	p99	3	180	**1.350**
10	p126	p6	1309	3	2.582	**1.297**
11	p10211	p148	p262	4	111	**1.110**
12	p109	p46	1110	3	2.044	**1.027**
13	p10584	p111	1293	3	134	**1.005**
14	p10155	1224	1293	2	1.922	**961**
15	p12780	1411	p787	3	127	**952**
16	p1046	p18	1286	3	125	**937**
17	p215	p214	p112	18	327	**912**
18	p184	1286	1110	2	1.793	**896**
19	p251	1110	p75	4	1.097	**822**
20	p1127	p273	1286	3	105	**787**
21	p10462	1237	p10155	3	105	**787**
22	p10521	p98	p10155	4	74	**740**
23	p1217	1300	p307	4	73	**730**
24	p1994	p83	1286	3	90	**675**
25	p1999	p76	1286	3	85	**637**
26	p1888	1224	p98	3	81	**607**
27	p10883	p17	p10155	4	58	**580**

28	p112	p35	p100	6	729	**546**
29	p1216	1291	p436	3	68	**510**
30	p2506	1363	p99	3	64	**480**

Figura 121. Patrones del dominio con mayor ponderación del experimento 2

El patrón del dominio con la mayor ponderación es p653, cuya secuencia es la siguiente:

- **p653**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 492 veces en el texto, por lo que ocupa la posición 30º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.3. Experimento 3: frecuencia = 10, sin semántica

8.3.1 Estudio de patrones generados

Se han generado un total de 5.707 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.571
7	p8	1144	1119	2	53.455
8	p7	p2	1144	3	45.762
9	p9	1123	1144	2	43.159
10	p10	1224	p1	3	38.738
11	p11	p4	p4	6	32.092
12	p12	1144	1237	2	29.788
13	p13	1144	1248	2	28.730
14	p14	1230	1108	2	27.125
15	p52	1110	1144	2	26.481
16	p15	1144	1213	2	24.086
17	p17	1119	1237	2	21.721
18	p21	p6	1237	3	19.231
19	p20	1144	1108	2	18.493
20	p26	p1	1119	3	17.675
21	p18	1123	1110	2	17.386
22	p19	1144	1230	2	16.448
23	p27	1144	1151	2	15.980
24	p22	p1	1237	3	15.380
25	p25	p1	1248	3	14.989
26	p28	1144	1123	2	14.883
27	p24	1110	p7	4	14.726
28	p16	1123	1123	2	14.007
29	p29	1144	1229	2	12.439
30	p31	1151	1110	2	11.938
31	p64	1110	p1	3	11.818
32	p36	p1	1213	3	11.450
33	p23	p1	p3	5	11.411
34	p33	1224	p3	4	11.407
35	p30	p1	p2	4	10.147

36	p37	p8	1110	3	9.207
37	p34	p1	1108	3	9.115
38	p47	1224	1119	2	8.426
39	p38	p2	1248	3	8.380
40	p40	1144	1228	2	8.359
41	p56	1230	1151	2	8.314
42	p42	p1	1123	3	8.257
43	p35	p10	1237	4	8.072
44	p48	1230	1103	2	7.923
45	p32	p1	1230	3	7.918
46	p39	1151	1108	2	7.359
47	p43	1224	p17	3	7.076
48	p44	1144	1197	2	7.064
49	p76	1110	1248	2	7.051
50	p60	p1	1151	3	6.779
51	p45	1224	1103	2	6.755
52	p67	p3	1119	4	6.570
53	p54	p11	p11	12	6.517
54	p68	1123	p1	3	6.460
55	p66	1144	1158	2	6.458
56	p49	1158	1237	2	6.420
57	p46	1144	1286	2	6.337
58	p57	1224	p5	5	6.258
59	p81	p4	1144	4	6.132
60	p53	p2	p1	4	6.117
61	p51	p6	1119	3	5.863
62	p50	1144	1236	2	5.842
63	p63	1230	p15	3	5.718
64	p62	1144	1103	2	5.475
65	p55	1213	1223	2	5.442
66	p70	p3	1123	4	5.382
67	p95	1119	1144	2	5.343
68	p61	p2	1119	3	5.322
69	p73	p3	1248	4	5.172
70	p69	1144	p9	3	5.168
71	p71	p8	1237	3	5.120
72	p59	p2	p2	4	5.107
73	p78	1108	1213	2	4.932
74	p72	1108	1229	2	4.930
75	p77	p5	p1	6	4.914
76	p75	p2	1286	3	4.801
77	p4537	1144	p52	3	4.741
78	p90	1151	1144	2	4.679
79	p110	p4	p1	5	4.447
80	p146	1230	1144	2	4.233

81	p84	p6	1248	3	4.195
82	p83	1119	1110	2	4.193
83	p101	p5	p5	8	4.175
84	p41	p16	p16	4	4.143
85	p102	p3	1213	4	4.043
86	p87	1223	p1	3	3.918
87	p82	p3	1237	4	3.914
88	p105	p14	1213	3	3.776
89	p114	1108	1221	2	3.748
90	p4538	p52	p52	4	3.732
91	p86	1213	1144	2	3.667
92	p93	1223	1103	2	3.562
93	p88	p4	1248	4	3.536
94	p103	p1	1229	3	3.459
95	p107	p3	1151	4	3.420
96	p111	1103	1213	2	3.414
97	p115	p14	1236	3	3.390
98	p80	1221	1144	2	3.388
99	p122	1223	1144	2	3.375
100	p116	p14	1229	3	3.334

Figura 122. Patrones más repetidos del experimento 3

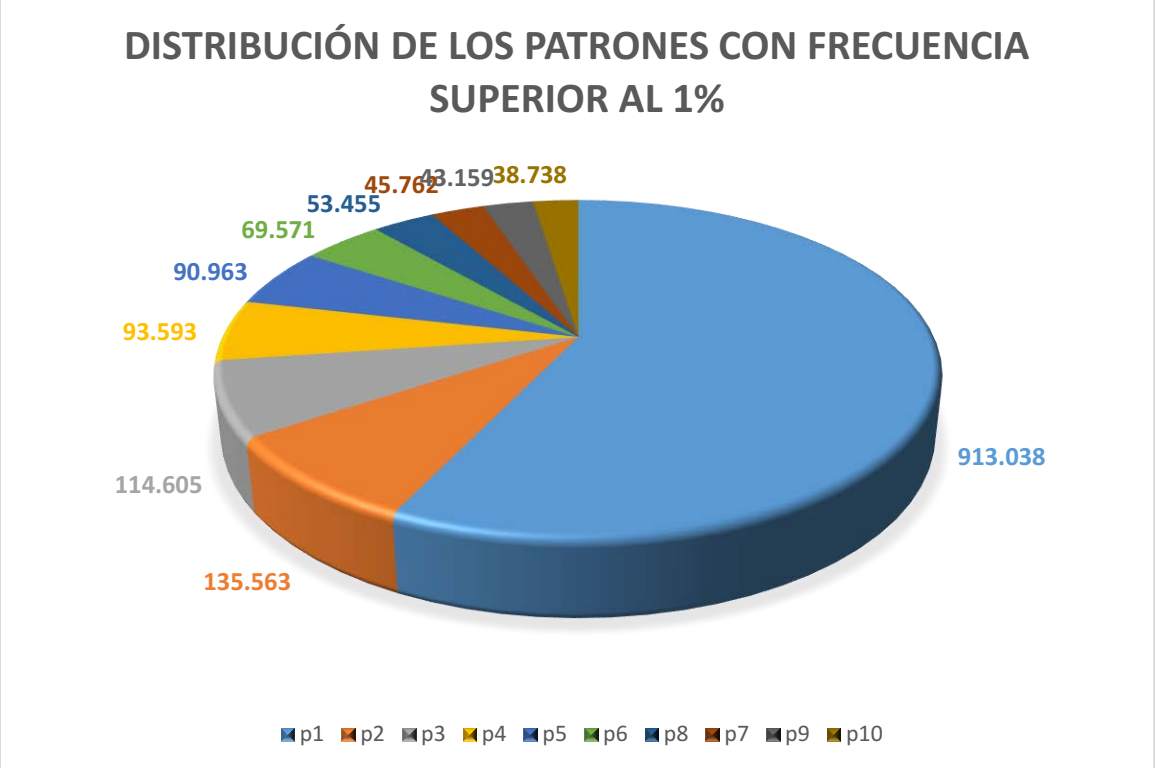

Figura 123. Gráfico de los patrones más frecuentes del experimento 3

8.3.2 Estudio de los patrones del dominio

De los 5.707 patrones generados en este escenario, 720 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 12,61% de los patrones pertenecen al dominio de estudio.

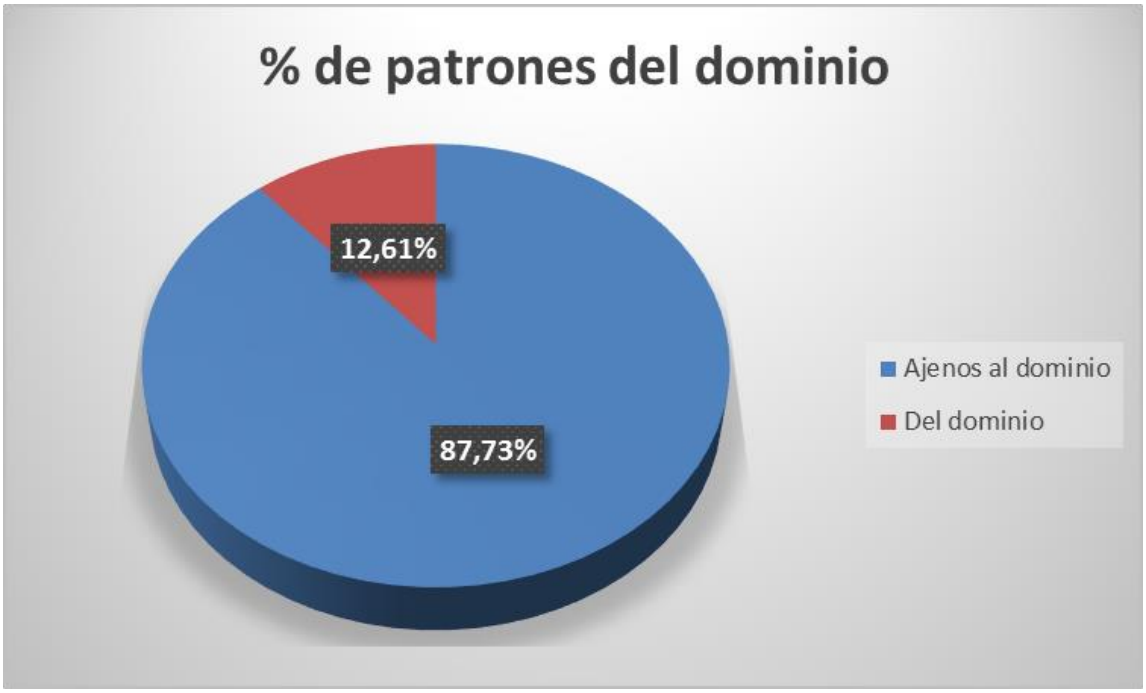

Figura 124. Proporción de patrones en el experimento 3

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p46	1144	1286	2	6.337
2	p75	p2	1286	3	4.801
3	p148	1300	1123	2	3.196
4	p98	1411	1213	2	3.181
5	p126	p6	1309	3	2.582
6	p219	1144	1411	2	2.424
7	p294	1144	1309	2	2.169
8	p109	p46	1110	3	2.044
9	p3936	1224	1293	2	1.922
10	p184	1286	1110	2	1.786
11	p100	1291	1144	2	1.130

12	p438	p1	1286	3	1.098
13	p251	1110	p75	4	1.097
14	p460	p3	1286	4	1.024
15	p401	p4	1286	4	925
16	p339	1224	1411	2	878
17	p478	1110	1286	2	873
18	p324	1224	1309	2	804
19	p705	1286	1151	2	791
20	p4306	1144	1288	2	741
21	p112	p35	p100	6	729
22	p420	1103	1286	2	661
23	p684	1300	p9	3	652
24	p3964	1293	1144	2	645
25	p509	p6	1411	3	644
26	p881	1309	1144	2	583
27	p3945	p81	p184	6	516
28	p427	1411	1237	2	515
29	p596	p31	1286	3	492
30	p369	1286	1144	2	477
31	p311	1144	1564	2	460
32	p3976	p15	1293	3	447
33	p1076	p1	1309	3	423
34	p613	1221	1411	2	395
35	p2157	1288	1144	2	368
36	p880	1300	1144	2	359
37	p4556	p4544	1286	4	337
38	p373	p98	p126	5	335
39	p1681	1123	1309	2	333
40	p683	1300	p18	3	332
41	p4074	1144	1293	2	330
42	p3971	1293	1411	2	318
43	p691	1309	1110	2	314
44	p4276	p52	1286	3	312
45	p1551	1363	1123	2	302
46	p794	1345	1228	2	302
47	p954	1224	1291	2	300
48	p1890	1224	1294	2	298
49	p830	1224	1300	2	288
50	p4896	p4046	1286	4	281
51	p1315	1291	p1	3	272
52	p3985	p1	1411	3	269
53	p561	p30	1286	5	262
54	p1931	1286	p1	3	251
55	p4509	p9	1293	3	248
56	p1394	p5	1286	5	241

57	p956	1224	1558	2	239
58	p822	p6	1526	3	230
59	p840	p13	1286	3	226
60	p1215	1248	1286	2	224
61	p3968	p148	1248	3	216
62	p1659	p6	1308	3	215
63	p4447	1224	1269	2	212
64	p4030	p148	p3	5	204
65	p2184	1411	1248	2	200
66	p968	p48	1286	3	196
67	p4132	p81	1286	5	192
68	p784	1144	1283	2	192
69	p2959	1123	1286	2	190
70	p1106	p25	1286	4	190
71	p955	1224	1526	2	187
72	p1242	p148	1237	3	187
73	p4115	1221	1309	2	183
74	p3672	p10	1309	4	182
75	p4059	1293	p1	3	181
76	p1023	p46	1144	3	177
77	p4061	p77	1286	7	176
78	p4973	p219	1213	3	169
79	p519	1564	1213	2	166
80	p3992	p3926	1286	6	165
81	p706	1300	p99	3	165
82	p1739	1309	1213	2	162
83	p757	1144	1553	2	160
84	p1321	p122	1309	3	160
85	p1791	p76	p75	5	160
86	p4311	1144	1300	2	156
87	p3050	1411	1110	2	154
88	p4278	p12	1293	3	154
89	p1013	1224	1564	2	152
90	p1539	1224	1553	2	148
91	p1713	1411	1144	2	147
92	p4177	1288	p1	3	146
93	p1017	1411	p86	3	146
94	p1682	1144	p98	3	140
95	p4761	p36	1293	4	135
96	p3041	1300	1248	2	134
97	p536	1144	1312	2	134
98	p4171	p1	1288	3	133
99	p1858	p3	1309	4	131
100	p4660	1293	1248	2	130

Figura 125. Patrones del dominio más repetidos del experimento 3

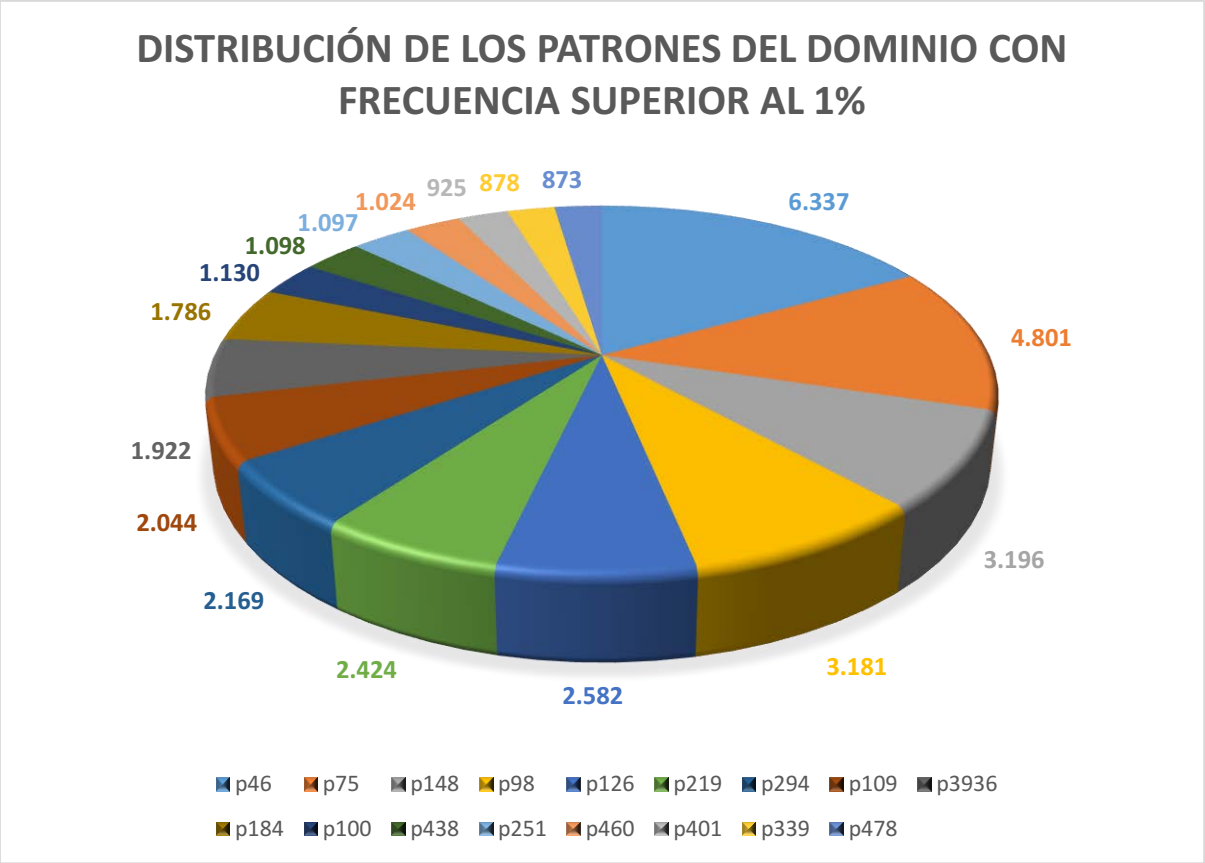

Figura 126. Gráfico de patrones del dominio en el experimento 3

8.3.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 111 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	299
2	1123	NUMBER	219
3	1151	ADVERB	196
4	1119	NOUN	193
5	1108	VERB	174
6	1213	PREPOSITION	174
7	1110	SYMBOL	141
8	1248	AND LINKING	133
9	1103	ADJECTIVE	123
10	1224	DEFINITE ARTICLE	111

Figura 127. Categorías más frecuentes en el experimento 3

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	83
2	1411	Mutation	46
3	1309	Genes	41
4	1300	Exons	38
5	1293	Genes, Neurofibromatosis 1	35
6	1288	Genes, Tumor Suppressor	22
7	1294	Genes, Neurofibromatosis 2	16
8	1291	Genes, Wilms Tumor	16
9	1363	Introns	12
10	1526	Phenotype	10

Figura 128. Categorías de genética más frecuentes en el experimento 3

Limitando los resultados a sólo los conceptos asociados a la sordera, las categorías empleadas han sido las siguientes, ordenadas por número de repeticiones:

Posición	Elemento	Nombre elemento	Repeticiones
1	1564	Hearing Loss	24
2	1561	Hearing Loss, Sensorineural	13
3	1553	Deafness	13
4	1558	Hearing Loss, Functional	9
5	1559	Hearing Loss, High-Frequency	4
6	1563	Wolfram Syndrome	2
7	1556	Hearing Loss, Central	2
8	1555	Hearing Loss, Bilateral	2

Figura 129. Categorías de sordera más frecuentes en el experimento 3

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

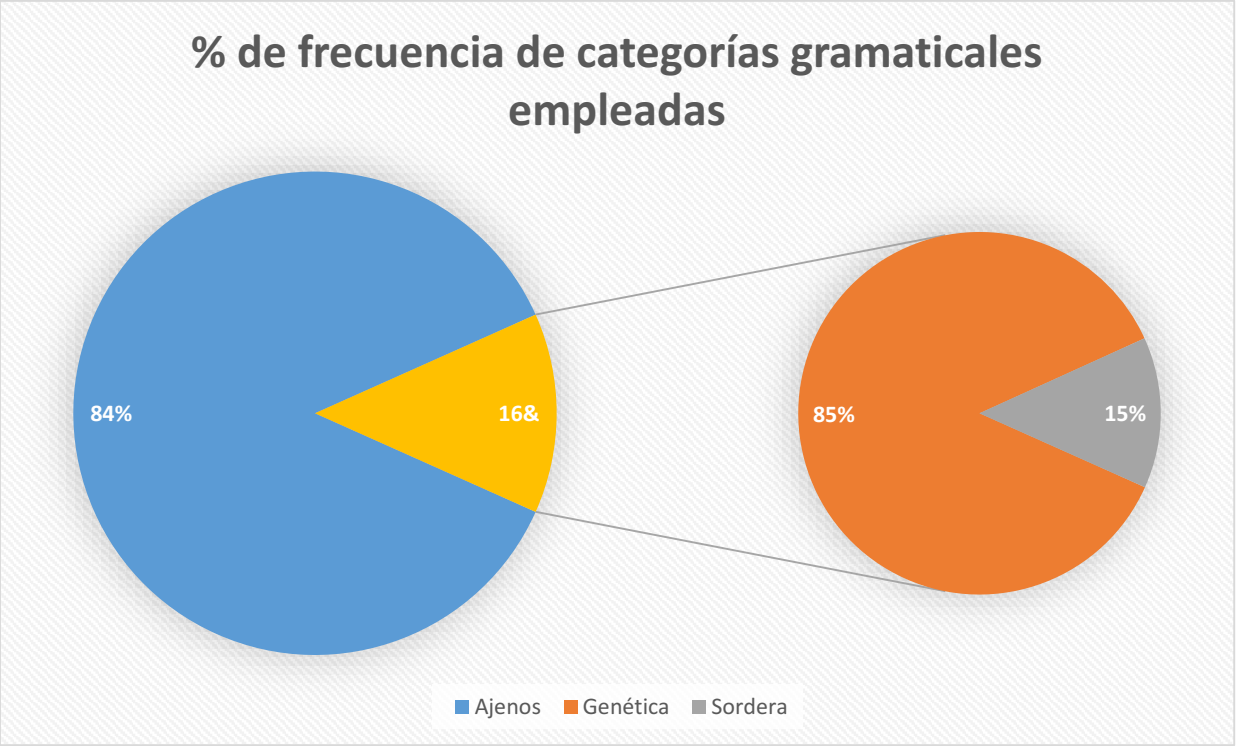

Figura 130. Proporción de categorías gramaticales en el experimento 3

Aproximadamente el 84% de elementos que forman los patrones son ajenos a nuestro dominio. El 16% forman parte del dominio, de los cuales el 85% están asociados a la "Genética" y el 15% restante a la "Sordera".

8.3.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p43	1224	p17	3	7.076	**53.070**
2	p41	p16	p16	4	4.143	**41.430**
3	p104	p41	p41	8	1.602	**32.040**
4	p105	p14	1213	3	3.776	**28.320**
5	p115	p14	1236	3	3.390	**25.425**
6	p116	p14	1229	3	3.334	**25.005**
7	p4	p1	1110	3	93.593	**23.866**
8	p287	p104	p104	16	545	**21.800**
9	p160	p18	1123	3	2.215	**16.612**
10	p11	p4	p4	6	32.092	**16.366**
11	p14	1230	1108	2	27.125	**13.562**

12	p164	1224	p49	3	1.672	**12.540**
13	p191	p18	p18	4	1.240	**12.400**
14	p7	p2	1144	3	45.762	**11.669**
15	p17	1119	1237	2	21.721	**10.860**
16	p247	p48	1213	3	1.330	**9.975**
17	p270	p48	1229	3	1.320	**9.900**
18	p10	1224	p1	3	38.738	**9.878**
19	p21	p6	1237	3	19.231	**9.663**
20	p18	1123	1110	2	17.386	**8.693**
21	p295	p14	p55	4	854	**8.540**
22	p379	p48	1228	3	1.062	**7.965**
23	p4495	p287	p287	32	99	**7.920**
24	p24	1110	p7	4	14.726	**7.363**
25	p348	p287	1123	17	169	**7.182**
26	p16	1123	1123	2	14.007	**7.003**
27	p54	p11	p11	12	6.517	**6.647**
28	p165	1230	p39	3	852	**6.390**
29	p31	1151	1110	2	11.938	**5.969**
30	p331	1233	p14	3	769	**5.767**
31	p5439	p5438	p5436	60	36	**5.400**
32	p304	p31	1203	3	699	**5.242**
33	p492	1223	p17	3	667	**5.002**
34	p415	1153	p56	3	656	**4.920**
35	p5016	p3753	p3753	4	488	**4.880**
36	p435	1203	p56	3	641	**4.807**
37	p529	p14	1228	3	629	**4.717**
38	p5438	p4495	p287	48	39	**4.680**
39	p37	p8	1110	3	9.207	**4.626**
40	p791	1233	p105	4	451	**4.510**
41	p26	p1	1119	3	17.675	**4.507**
42	p315	1108	p55	3	593	**4.447**
43	p548	1233	p116	4	431	**4.310**
44	p47	1224	1119	2	8.426	**4.213**
45	p38	p2	1248	3	8.380	**4.210**
46	p56	1230	1151	2	8.314	**4.157**
47	p35	p10	1237	4	8.072	**4.036**
48	p659	1110	p31	3	533	**3.997**
49	p48	1230	1103	2	7.923	**3.961**
50	p22	p1	1237	3	15.380	**3.921**
51	p25	p1	1248	3	14.989	**3.822**
52	p5019	p5016	p5016	8	190	**3.800**
53	p518	p14	1166	3	503	**3.772**
54	p596	p31	1286	3	492	**3.690**
55	p39	1151	1108	2	7.359	**3.679**
56	p965	1103	p17	3	473	**3.547**

57	p76	1110	1248	2	7.051	**3.525**
58	p3955	p56	p78	4	348	**3.480**
59	p720	p14	1151	3	458	**3.435**
60	p307	p99	1123	3	454	**3.405**
61	p2	1144	1110	2	135.563	**3.389**
62	p511	1203	p39	3	451	**3.382**
63	p45	1224	1103	2	6.755	**3.377**
64	p3927	p47	1237	3	450	**3.375**
65	p672	p45	p17	4	324	**3.240**
66	p49	1158	1237	2	6.420	**3.210**
67	p502	p48	1236	3	416	**3.120**
68	p832	1228	p246	3	404	**3.030**
69	p64	1110	p1	3	11.818	**3.013**
70	p630	p47	1248	3	400	**3.000**
71	p51	p6	1119	3	5.863	**2.946**
72	p508	p14	1221	3	392	**2.940**
73	p36	p1	1213	3	11.450	**2.919**
74	p798	p274	p115	5	232	**2.900**
75	p63	1230	p15	3	5.718	**2.873**
76	p33	1224	p3	4	11.407	**2.851**
77	p3991	1203	p114	3	380	**2.850**
78	p1136	1240	p105	4	285	**2.850**
79	p495	p39	1229	3	368	**2.760**
80	p55	1213	1223	2	5.442	**2.721**
81	p453	p39	1213	3	357	**2.677**
82	p61	p2	1119	3	5.322	**2.674**
83	p190	1221	p79	3	356	**2.670**
84	p5042	p5019	p5019	16	65	**2.600**
85	p71	p8	1237	3	5.120	**2.572**
86	p59	p2	p2	4	5.107	**2.553**
87	p30	p1	p2	4	10.147	**2.536**
88	p544	p14	1248	3	333	**2.497**
89	p683	1300	p18	3	332	**2.490**
90	p1163	p56	p111	4	248	**2.480**
91	p78	1108	1213	2	4.932	**2.466**
92	p72	1108	1229	2	4.930	**2.465**
93	p625	p48	p140	4	244	**2.440**
94	p409	p18	1248	3	324	**2.430**
95	p75	p2	1286	3	4.801	**2.412**
96	p854	p55	1103	3	316	**2.370**
97	p765	p194	1236	3	314	**2.355**
98	p34	p1	1108	3	9.115	**2.324**
99	p3958	p16	1123	3	306	**2.295**
100	p507	p45	1119	3	303	**2.272**

Figura 131. Patrones con mayor ponderación del experimento 3

El patrón con la mayor ponderación es p43, cuya secuencia es la siguiente:

- **p43**: “1224 + p17” = “1224 + 1119 + 1237” = “DEFINITE ARTICLE + NOUN + PREPOSITION OF”.

El patrón tiene longitud 3 y se repitió 7.076 veces en el texto, por lo que ocupa la posición 47º en la lista de patrones más frecuentes en este escenario de estudio.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p596	p31	1286	3	492	**3.690**
2	p683	1300	p18	3	332	**2.490**
3	p75	p2	1286	3	4.801	**2.412**
4	p3968	p148	1248	3	216	**1.620**
5	p148	1300	1123	2	3.196	**1.598**
6	p98	1411	1213	2	3.181	**1.590**
7	p968	p48	1286	3	196	**1.470**
8	p1242	p148	1237	3	187	**1.402**
9	p126	p6	1309	3	2.582	**1.297**
10	p706	1300	p99	3	165	**1.237**
11	p109	p46	1110	3	2.044	**1.027**
12	p3936	1224	1293	2	1.922	**961**
13	p3995	p148	p262	4	94	**940**
14	p215	p214	p112	18	327	**912**
15	p5056	1411	p787	3	120	**900**
16	p184	1286	1110	2	1.786	**893**
17	p251	1110	p75	4	1.097	**822**
18	p4342	p111	1293	3	104	**780**
19	p1046	p18	1286	3	101	**757**
20	p1217	1300	p307	4	73	**730**
21	p4303	1237	p3936	3	97	**727**
22	p1127	p273	1286	3	91	**682**
23	p1994	p83	1286	3	81	**607**
24	p112	p35	p100	6	729	**546**
25	p4715	p98	p3936	4	48	**480**
26	p401	p4	1286	4	925	**462**
27	p339	1224	1411	2	878	**439**
28	p478	1110	1286	2	873	**436**

29	p3953	p98	1300	3	58	**435**
30	p468	p11	p109	9	411	**416**

Figura 132. Patrones del dominio con mayor ponderación del experimento 3

El patrón del dominio con la mayor ponderación es p596, cuya secuencia es la siguiente:

- **p596**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 492 veces en el texto, por lo que ocupa la posición 29º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.4. Experimento 4: frecuencia = 20, sin semántica

8.4.1 Estudio de patrones generados

Se han generado un total de 2.847 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.571
7	p8	1144	1119	2	53.455
8	p7	p2	1144	3	45.762
9	p9	1123	1144	2	43.159
10	p10	1224	p1	3	38.738
11	p11	p4	p4	6	32.092
12	p12	1144	1237	2	29.788
13	p13	1144	1248	2	28.730
14	p14	1230	1108	2	27.125
15	p52	1110	1144	2	26.481
16	p15	1144	1213	2	24.086
17	p17	1119	1237	2	21.721
18	p21	p6	1237	3	19.231
19	p20	1144	1108	2	18.493
20	p26	p1	1119	3	17.675
21	p18	1123	1110	2	17.386
22	p19	1144	1230	2	16.448
23	p27	1144	1151	2	15.980
24	p22	p1	1237	3	15.380
25	p25	p1	1248	3	14.989
26	p28	1144	1123	2	14.883
27	p24	1110	p7	4	14.726
28	p16	1123	1123	2	14.007
29	p29	1144	1229	2	12.439
30	p31	1151	1110	2	11.938
31	p64	1110	p1	3	11.818
32	p36	p1	1213	3	11.450
33	p23	p1	p3	5	11.411
34	p33	1224	p3	4	11.407
35	p30	p1	p2	4	10.147

36	p37	p8	1110	3	9.207
37	p34	p1	1108	3	9.115
38	p47	1224	1119	2	8.426
39	p38	p2	1248	3	8.380
40	p40	1144	1228	2	8.359
41	p56	1230	1151	2	8.314
42	p42	p1	1123	3	8.257
43	p35	p10	1237	4	8.072
44	p48	1230	1103	2	7.923
45	p32	p1	1230	3	7.918
46	p39	1151	1108	2	7.359
47	p43	1224	p17	3	7.076
48	p44	1144	1197	2	7.064
49	p76	1110	1248	2	7.051
50	p60	p1	1151	3	6.779
51	p45	1224	1103	2	6.755
52	p67	p3	1119	4	6.570
53	p54	p11	p11	12	6.517
54	p68	1123	p1	3	6.460
55	p66	1144	1158	2	6.458
56	p49	1158	1237	2	6.420
57	p46	1144	1286	2	6.337
58	p57	1224	p5	5	6.258
59	p81	p4	1144	4	6.132
60	p53	p2	p1	4	6.117
61	p51	p6	1119	3	5.863
62	p50	1144	1236	2	5.842
63	p63	1230	p15	3	5.718
64	p62	1144	1103	2	5.475
65	p55	1213	1223	2	5.442
66	p70	p3	1123	4	5.382
67	p95	1119	1144	2	5.343
68	p61	p2	1119	3	5.322
69	p73	p3	1248	4	5.172
70	p69	1144	p9	3	5.168
71	p71	p8	1237	3	5.120
72	p59	p2	p2	4	5.107
73	p78	1108	1213	2	4.932
74	p72	1108	1229	2	4.930
75	p77	p5	p1	6	4.914
76	p75	p2	1286	3	4.801
77	p2390	1144	p52	3	4.741
78	p90	1151	1144	2	4.679
79	p110	p4	p1	5	4.447
80	p146	1230	1144	2	4.233

81	p84	p6	1248	3	4.195
82	p83	1119	1110	2	4.193
83	p101	p5	p5	8	4.175
84	p41	p16	p16	4	4.095
85	p102	p3	1213	4	4.043
86	p87	1223	p1	3	3.918
87	p82	p3	1237	4	3.914
88	p105	p14	1213	3	3.776
89	p114	1108	1221	2	3.748
90	p2391	p52	p52	4	3.732
91	p86	1213	1144	2	3.667
92	p93	1223	1103	2	3.562
93	p88	p4	1248	4	3.536
94	p103	p1	1229	3	3.459
95	p107	p3	1151	4	3.420
96	p111	1103	1213	2	3.414
97	p115	p14	1236	3	3.390
98	p80	1221	1144	2	3.370
99	p122	1223	1144	2	3.362
100	p116	p14	1229	3	3.334

Figura 133. Patrones más repetidos del experimento 4

Figura 134. Gráfico de los patrones más frecuentes del experimento 4

8.4.2 Estudio de los patrones del dominio

De los 2.847 patrones generados en este escenario, 300 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 10,53% de los patrones pertenecen al dominio de estudio.

Figura 135. Proporción de patrones en el experimento 4

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p46	1144	1286	2	6.337
2	p75	p2	1286	3	4.801
3	p148	1300	1123	2	3.196
4	p98	1411	1213	2	3.162
5	p126	p6	1309	3	2.582
6	p219	1144	1411	2	2.424
7	p294	1144	1309	2	2.169
8	p109	p46	1110	3	2.022
9	p2174	1224	1293	2	1.909
10	p184	1286	1110	2	1.786
11	p438	p1	1286	3	1.098
12	p100	1291	1144	2	1.092

13	p251	1110	p75	4	1.077
14	p460	p3	1286	4	1.012
15	p401	p4	1286	4	925
16	p478	1110	1286	2	873
17	p339	1224	1411	2	867
18	p324	1224	1309	2	804
19	p705	1286	1151	2	791
20	p2305	1144	1288	2	741
21	p112	p35	p100	6	729
22	p420	1103	1286	2	647
23	p509	p6	1411	3	644
24	p684	1300	p9	3	616
25	p2202	1293	1144	2	599
26	p881	1309	1144	2	554
27	p2183	p81	p184	6	516
28	p596	p31	1286	3	492
29	p369	1286	1144	2	461
30	p427	1411	1237	2	457
31	p311	1144	1564	2	447
32	p2216	p15	1293	3	428
33	p613	1221	1411	2	366
34	p1076	p1	1309	3	364
35	p2411	p2397	1286	4	337
36	p2157	1288	1144	2	326
37	p1681	1123	1309	2	307
38	p880	1300	1144	2	306
39	p1551	1363	1123	2	302
40	p683	1300	p18	3	297
41	p2210	1293	1411	2	293
42	p2571	p2298	1286	4	281
43	p830	1224	1300	2	272
44	p691	1309	1110	2	263
45	p2317	1144	1293	2	263
46	p2401	p52	1286	3	262
47	p1890	1224	1294	2	261
48	p794	1345	1228	2	241
49	p2485	p9	1293	3	236
50	p373	p98	p126	5	236
51	p1931	1286	p1	3	220
52	p1394	p5	1286	5	209
53	p2280	p148	p3	5	204
54	p2365	1224	1269	2	199
55	p2226	p1	1411	3	197
56	p956	1224	1558	2	187
57	p954	1224	1291	2	185

58	p2370	p81	1286	5	177
59	p1659	p6	1308	3	173
60	p822	p6	1526	3	166
61	p519	1564	1213	2	166
62	p561	p30	1286	5	166
63	p2235	p2164	1286	6	165
64	p2616	p219	1213	3	157
65	p1315	1291	p1	3	153
66	p1013	1224	1564	2	152
67	p2206	p148	1248	3	151
68	p840	p13	1286	3	143
69	p2556	1411	1248	2	140
70	p706	1300	p99	3	129
71	p757	1144	1553	2	127
72	p1215	1248	1286	2	126
73	p1682	1144	p98	3	126
74	p2241	p10	1309	4	125
75	p2320	1144	1300	2	124
76	p2740	1144	1328	2	124
77	p2741	p2740	p68	5	120
78	p2774	1123	1286	2	119
79	p1713	1411	1144	2	119
80	p1739	1309	1213	2	117
81	p1791	p76	p75	5	117
82	p1017	1411	p86	3	114
83	p2319	p77	1286	7	112
84	p1106	p25	1286	4	111
85	p1892	1224	1561	2	108
86	p2524	1293	1248	2	106
87	p2699	1411	p787	3	105
88	p1242	p148	1237	3	103
89	p1237	1309	1119	2	99
90	p536	1144	1312	2	99
91	p955	1224	1526	2	98
92	p2201	1294	1411	2	98
93	p2207	1294	1144	2	95
94	p753	p4	p46	5	95
95	p2609	p152	p1681	5	92
96	p2585	1144	1410	2	90
97	p2181	1144	p184	3	90
98	p2364	1288	p1	3	89
99	p2473	1221	1309	2	89
100	p2581	p12	1293	3	86

Figura 136. Patrones del dominio más repetidos del experimento 4

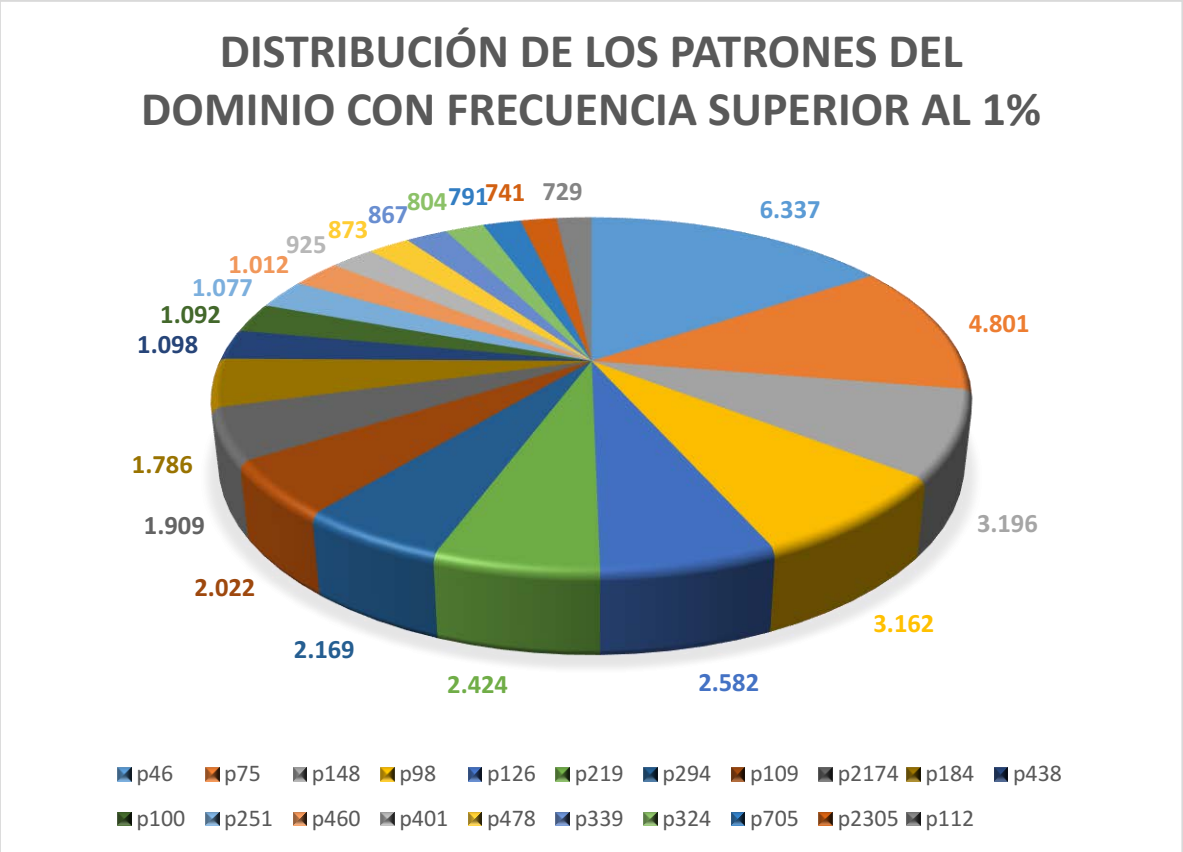

Figura 137. Gráfico de patrones del dominio en el experimento 4

8.4.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 85 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	213
2	1123	NUMBER	142
3	1151	ADVERB	121
4	1119	NOUN	120
5	1213	PREPOSITION	107
6	1108	VERB	106
7	1110	SYMBOL	99
8	1248	AND LINKING	85
9	1224	DEFINITE ARTICLE	78
10	1103	ADJECTIVE	67

Figura 138. Categorías más frecuentes en el experimento 4

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	49
2	1411	Mutation	23
3	1309	Genes	21
4	1300	Exons	20
5	1293	Genes, Neurofibromatosis 1	18
6	1288	Genes, Tumor Suppressor	8
7	1291	Genes, Wilms Tumor	8
8	1294	Genes, Neurofibromatosis 2	8
9	1312	Genes, Dominant	7
10	1526	Phenotype	6

Figura 139. Categorías de genética más frecuentes en el experimento 4

Limitando los resultados a sólo los conceptos asociados a la sordera, las categorías empleadas han sido las siguientes, ordenadas por número de repeticiones:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	12
2	1553	Deafness	8
3	1561	Hearing Loss, Sensorineural	7
4	1563	Wolfram Syndrome	2
5	1559	Hearing Loss, High-Frequency	2
6	1556	Hearing Loss, Central	2
7	1558	Hearing Loss, Functional	1

Figura 140. Categorías de sordera más frecuentes en el experimento 4

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

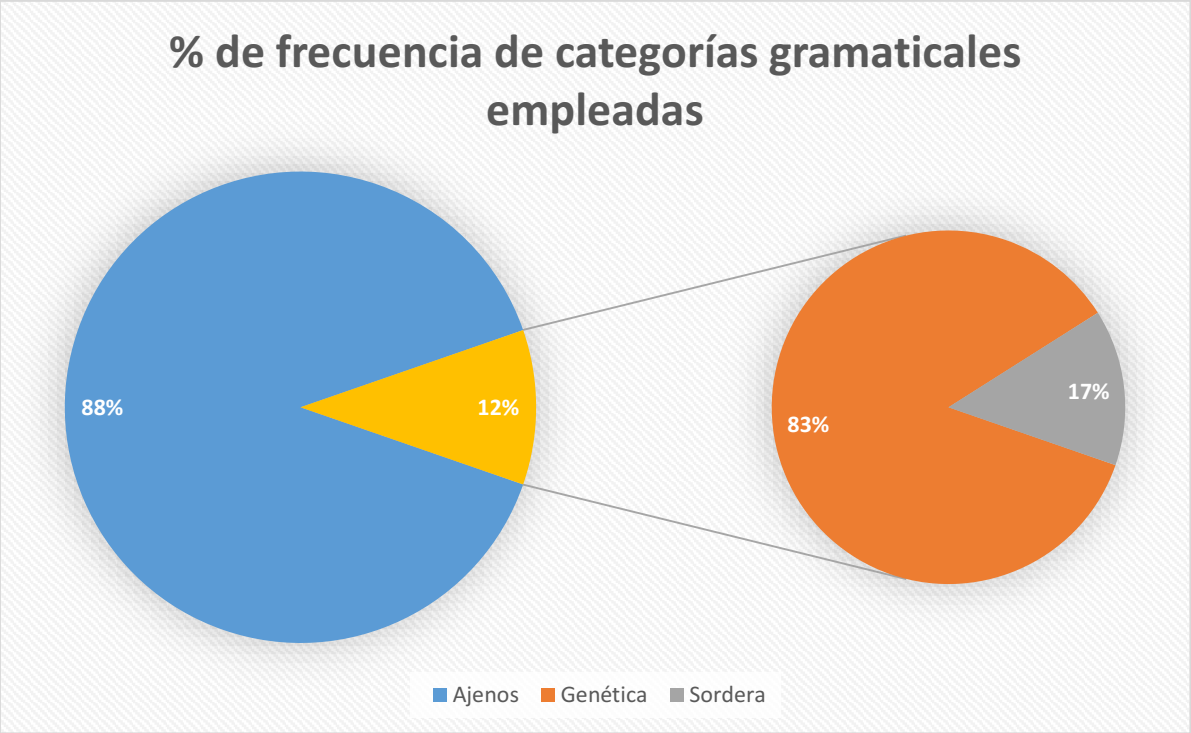

Figura 141. Proporción de categorías gramaticales en el experimento 4

Aproximadamente el 88% de elementos que forman los patrones son ajenos a nuestro dominio. El 12% forman parte del dominio, de los cuales el 83% están asociados a la "Genética" y el 17% restante a la "Sordera".

8.4.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p43	1224	p17	3	7.076	**53.070**
2	p41	p16	p16	4	4.095	**40.950**
3	p104	p41	p41	8	1.569	**31.380**
4	p105	p14	1213	3	3.776	**28.320**
5	p115	p14	1236	3	3.390	**25.425**
6	p116	p14	1229	3	3.334	**25.005**
7	p4	p1	1110	3	93.593	**23.866**
8	p287	p104	p104	16	531	**21.240**
9	p160	p18	1123	3	2.215	**16.612**
10	p11	p4	p4	6	32.092	**16.366**
11	p14	1230	1108	2	27.125	**13.562**

12	p164	1224	p49	3	1.653	**12.397**
13	p191	p18	p18	4	1.223	**12.230**
14	p7	p2	1144	3	45.762	**11.669**
15	p17	1119	1237	2	21.721	**10.860**
16	p10	1224	p1	3	38.738	**9.878**
17	p247	p48	1213	3	1.317	**9.877**
18	p270	p48	1229	3	1.307	**9.802**
19	p21	p6	1237	3	19.231	**9.663**
20	p18	1123	1110	2	17.386	**8.693**
21	p295	p14	p55	4	854	**8.540**
22	p379	p48	1228	3	1.049	**7.867**
23	p24	1110	p7	4	14.726	**7.363**
24	p348	p287	1123	17	169	**7.182**
25	p16	1123	1123	2	14.007	**7.003**
26	p2769	p287	p287	32	86	**6.880**
27	p54	p11	p11	12	6.517	**6.647**
28	p165	1230	p39	3	852	**6.390**
29	p31	1151	1110	2	11.938	**5.969**
30	p331	1233	p14	3	769	**5.767**
31	p2778	p2777	p2775	60	36	**5.400**
32	p304	p31	1203	3	685	**5.137**
33	p415	1153	p56	3	656	**4.920**
34	p2604	p2603	p2603	4	488	**4.880**
35	p492	1223	p17	3	650	**4.875**
36	p435	1203	p56	3	641	**4.807**
37	p529	p14	1228	3	629	**4.717**
38	p2777	p2769	p287	48	39	**4.680**
39	p37	p8	1110	3	9.207	**4.626**
40	p791	1233	p105	4	451	**4.510**
41	p26	p1	1119	3	17.675	**4.507**
42	p315	1108	p55	3	593	**4.447**
43	p47	1224	1119	2	8.426	**4.213**
44	p38	p2	1248	3	8.380	**4.210**
45	p56	1230	1151	2	8.314	**4.157**
46	p35	p10	1237	4	8.072	**4.036**
47	p48	1230	1103	2	7.923	**3.961**
48	p22	p1	1237	3	15.380	**3.921**
49	p659	1110	p31	3	513	**3.847**
50	p548	1233	p116	4	384	**3.840**
51	p25	p1	1248	3	14.989	**3.822**
52	p2607	p2604	p2604	8	190	**3.800**
53	p596	p31	1286	3	492	**3.690**
54	p39	1151	1108	2	7.359	**3.679**
55	p965	1103	p17	3	473	**3.547**
56	p76	1110	1248	2	7.051	**3.525**

57	p518	p14	1166	3	465	**3.487**
58	p2193	p56	p78	4	348	**3.480**
59	p2	1144	1110	2	135.563	**3.389**
60	p45	1224	1103	2	6.755	**3.377**
61	p2165	p47	1237	3	450	**3.375**
62	p511	1203	p39	3	439	**3.292**
63	p49	1158	1237	2	6.420	**3.210**
64	p307	p99	1123	3	402	**3.015**
65	p64	1110	p1	3	11.818	**3.013**
66	p502	p48	1236	3	401	**3.007**
67	p720	p14	1151	3	396	**2.970**
68	p51	p6	1119	3	5.863	**2.946**
69	p36	p1	1213	3	11.450	**2.919**
70	p63	1230	p15	3	5.718	**2.873**
71	p832	1228	p246	3	381	**2.857**
72	p33	1224	p3	4	11.407	**2.851**
73	p495	p39	1229	3	368	**2.760**
74	p2234	1203	p114	3	366	**2.745**
75	p55	1213	1223	2	5.442	**2.721**
76	p61	p2	1119	3	5.322	**2.674**
77	p190	1221	p79	3	356	**2.670**
78	p630	p47	1248	3	344	**2.580**
79	p71	p8	1237	3	5.120	**2.572**
80	p59	p2	p2	4	5.107	**2.553**
81	p30	p1	p2	4	10.147	**2.536**
82	p78	1108	1213	2	4.932	**2.466**
83	p72	1108	1229	2	4.930	**2.465**
84	p672	p45	p17	4	244	**2.440**
85	p508	p14	1221	3	323	**2.422**
86	p75	p2	1286	3	4.801	**2.412**
87	p1136	1240	p105	4	241	**2.410**
88	p34	p1	1108	3	9.115	**2.324**
89	p683	1300	p18	3	297	**2.227**
90	p453	p39	1213	3	297	**2.227**
91	p854	p55	1103	3	283	**2.122**
92	p84	p6	1248	3	4.195	**2.107**
93	p42	p1	1123	3	8.257	**2.105**
94	p83	1119	1110	2	4.193	**2.096**
95	p2196	p16	1123	3	279	**2.092**
96	p32	p1	1230	3	7.918	**2.019**
97	p1137	p287	p41	20	40	**2.000**
98	p765	p194	1236	3	265	**1.987**
99	p2826	p2607	p2607	16	48	**1.920**
100	p2275	p56	1108	3	252	**1.890**

Figura 142. Patrones con mayor ponderación del experimento 4

El patrón con la mayor ponderación es p43, cuya secuencia es la siguiente:

- **p43**: “1224 + p17” = “1224 + 1119 + 1237” = “DEFINITE ARTICLE + NOUN + PREPOSITION OF”.

El patrón tiene longitud 3 y se repitió 7.076 veces en el texto, por lo que ocupa la posición 47º en la lista de patrones más frecuentes en este escenario de estudio.

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p596	p31	1286	3	492	**3.690**
2	p75	p2	1286	3	4.801	**2.412**
3	p683	1300	p18	3	297	**2.227**
4	p148	1300	1123	2	3.196	**1.598**
5	p98	1411	1213	2	3.162	**1.581**
6	p126	p6	1309	3	2.582	**1.297**
7	p2206	p148	1248	3	151	**1.132**
8	p109	p46	1110	3	2.022	**1.016**
9	p706	1300	p99	3	129	**967**
10	p2174	1224	1293	2	1.909	**954**
11	p215	p214	p112	18	327	**912**
12	p184	1286	1110	2	1.786	**893**
13	p2238	p148	p262	4	82	**820**
14	p251	1110	p75	4	1.077	**807**
15	p2699	1411	p787	3	105	**787**
16	p1242	p148	1237	3	103	**772**
17	p2716	p111	1293	3	84	**630**
18	p112	p35	p100	6	729	**546**
19	p968	p48	1286	3	70	**525**
20	p2582	1237	p2174	3	69	**517**
21	p1046	p18	1286	3	66	**495**
22	p401	p4	1286	4	925	**462**
23	p478	1110	1286	2	873	**436**
24	p2191	p98	1300	3	58	**435**
25	p339	1224	1411	2	867	**433**
26	p324	1224	1309	2	804	**402**
27	p705	1286	1151	2	791	**395**
28	p468	p11	p109	9	386	**390**
29	p1217	1300	p307	4	39	**390**
30	p2183	p81	p184	6	516	**387**

Figura 143. Patrones del dominio con mayor ponderación del experimento 4

El patrón del dominio con la mayor ponderación es p596, cuya secuencia es la siguiente:

- **p596**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 492 veces en el texto, por lo que ocupa la posición 28º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.5. Experimento 5: frecuencia = 1, con semántica

8.5.1 Estudio de patrones generados

Se han generado un total de 142.064 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.738
7	p8	1144	1119	2	53.571
8	p15	1144	1213	2	49.414
9	p7	p2	1144	3	45.765
10	p9	1123	1144	2	43.189
11	p10	1224	p1	3	38.738
12	p11	p4	p4	6	32.092
13	p12	1144	1237	2	29.788
14	p13	1144	1248	2	28.730
15	p14	1230	1108	2	27.146
16	p53	1110	1144	2	26.738
17	p36	p1	1213	3	23.365
18	p17	1119	1237	2	23.184
19	p21	p6	1237	3	19.231
20	p20	1144	1108	2	18.575
21	p26	p1	1119	3	17.721
22	p18	1123	1110	2	17.379
23	p19	1144	1230	2	16.449
24	p27	1144	1151	2	16.017
25	p22	p1	1237	3	15.380
26	p25	p1	1248	3	14.989
27	p28	1144	1123	2	14.885
28	p24	1110	p7	4	14.727
29	p16	1123	1123	2	14.002
30	p29	1144	1229	2	12.287
31	p31	1151	1110	2	12.025
32	p63	1110	p1	3	11.867
33	p33	1224	p3	4	11.412
34	p23	p1	p3	5	11.411
35	p30	p1	p2	4	10.099

36	p81	1108	1213	2	9.801
37	p34	p1	1108	3	9.267
38	p37	p8	1110	3	9.081
39	p105	p14	1213	3	8.829
40	p46	1224	1119	2	8.610
41	p56	1230	1151	2	8.411
42	p38	p2	1248	3	8.365
43	p42	p1	1123	3	8.230
44	p73	1213	1144	2	8.186
45	p32	p1	1230	3	8.175
46	p40	1144	1228	2	8.110
47	p47	1230	1103	2	8.084
48	p35	p10	1237	4	8.072
49	p39	1151	1108	2	8.018
50	p91	p3	1213	4	8.009
51	p44	1224	1103	2	7.103
52	p75	1110	1248	2	7.100
53	p50	1224	p17	3	7.001
54	p59	p1	1151	3	6.760
55	p43	1144	1197	2	6.719
56	p66	1123	p1	3	6.566
57	p48	1158	1237	2	6.559
58	p65	p3	1119	4	6.550
59	p54	p11	p11	12	6.517
60	p51	p6	1119	3	6.350
61	p57	1224	p5	5	6.342
62	p111	1119	1213	2	6.290
63	p45	1144	1286	2	6.288
64	p82	p4	1144	4	6.233
65	p52	p2	p1	4	6.198
66	p55	1223	1144	2	6.135
67	p96	1119	1144	2	5.785
68	p78	1144	1158	2	5.701
69	p112	1103	1213	2	5.626
70	p161	1230	1144	2	5.469
71	p70	p3	1123	4	5.379
72	p61	p2	1119	3	5.336
73	p90	1151	1144	2	5.256
74	p72	p3	1248	4	5.157
75	p69	1223	p1	3	5.121
76	p60	p2	p2	4	5.107
77	p49	1144	1236	2	5.090
78	p68	1144	p9	3	5.063
79	p62	1213	1223	2	5.053
80	p71	1108	1229	2	5.023

81	p76	p5	p1	6	4.953
82	p67	1144	1103	2	4.870
83	p79	p8	1237	3	4.869
84	p77	p2	1286	3	4.801
85	p93281	1144	p53	3	4.741
86	p109	p4	p1	5	4.546
87	p115	1108	1221	2	4.476
88	p139	1230	p15	3	4.329
89	p85	1119	1110	2	4.241
90	p87	1223	1103	2	4.210
91	p103	p5	p5	8	4.175
92	p41	p16	p16	4	4.140
93	p84	p6	1248	3	4.137
94	p74	1221	1144	2	4.025
95	p97	1108	1144	2	3.910
96	p83	p3	1237	4	3.884
97	p128	1103	1144	2	3.865
98	p93282	p53	p53	4	3.732
99	p92	1411	1213	2	3.611
100	p138	1229	1108	2	3.603

Figura 144. Patrones más repetidos del experimento 5

Figura 145. Gráfico de los patrones más frecuentes del experimento 5

8.5.2 Estudio de los patrones del dominio

De los 142.064 patrones generados en este escenario, 19.933 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 10,53% de los patrones pertenecen al dominio de estudio.

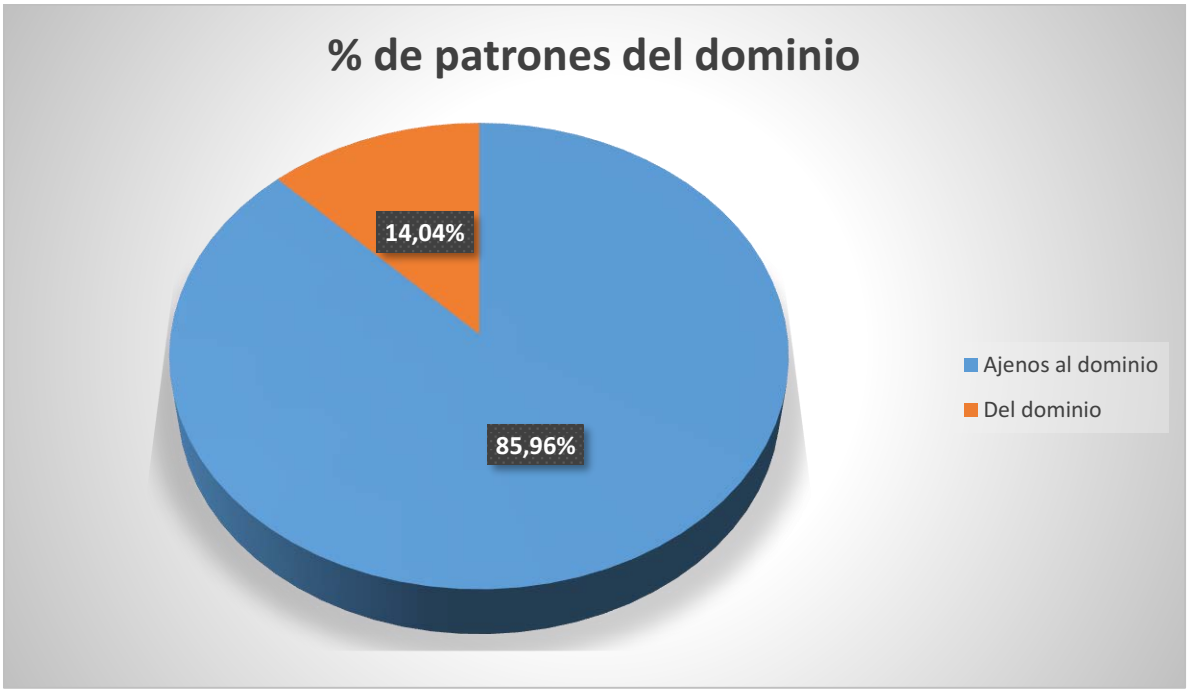

Figura 146. Proporción de patrones en el experimento 5

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p45	1144	1286	2	6.288
2	p77	p2	1286	3	4.801
3	p92	1411	1213	2	3.611
4	p150	1300	1123	2	3.205
5	p123	p6	1309	3	2.582
6	p189	1144	1411	2	2.279
7	p110	p45	1110	3	2.042
8	p281	1144	1309	2	2.019
9	p72381	1224	1293	2	1.907
10	p184	1286	1110	2	1.787
11	p101	1291	1144	2	1.145

12	p402	p1	1286	3	1.113
13	p251	1110	p77	4	1.097
14	p417	1103	1286	2	994
15	p456	p3	1286	4	986
16	p387	p4	1286	4	935
17	p334	1224	1411	2	924
18	p441	1110	1286	2	916
19	p539	1286	1151	2	847
20	p318	1224	1309	2	813
21	p3648	1144	1288	2	773
22	p113	p35	p101	6	732
23	p72406	1293	1144	2	700
24	p474	p6	1411	3	684
25	p315	1286	1144	2	684
26	p668	1300	p9	3	673
27	p835	1309	1144	2	585
28	p293	1144	1564	2	576
29	p72393	p82	p184	6	516
30	p454	1411	1237	2	506
31	p72517	1144	1293	2	458
32	p560	1221	1411	2	453
33	p653	p31	1286	3	452
34	p1105	p1	1309	3	445
35	p2077	1288	1144	2	397
36	p834	1300	1144	2	389
37	p72431	p15	1293	3	380
38	p360	p92	p123	5	366
39	p667	1300	p18	3	363
40	p800	1224	1300	2	338
41	p93298	p93286	1286	4	337
42	p1184	1309	1213	2	328
43	p72922	p53	1286	3	328
44	p773	1309	1110	2	324
45	p2081	1363	1123	2	321
46	p72413	1293	1411	2	320
47	p734	1345	1228	2	319
48	p1803	1123	1309	2	318
49	p874	1224	1291	2	307
50	p1271	1291	p1	3	302
51	p1823	1224	1294	2	301
52	p72414	p1	1411	3	291
53	p9846	1286	p1	3	289
54	p916	1224	1558	2	288
55	p102727	p3625	1286	4	281
56	p725	p6	1526	3	273

57	p769	p13	1286	3	269
58	p1591	p6	1308	3	268
59	p1350	p5	1286	5	264
60	p82875	p9	1293	3	256
61	p611	p30	1286	5	253
62	p1182	1248	1286	2	252
63	p999	p55	1309	3	250
64	p72449	1293	p1	3	247
65	p1995	p10	1309	4	247
66	p1095	1218	1309	2	235
67	p915	1224	1526	2	235
68	p72961	1224	1269	2	234
69	p641	1144	1553	2	230
70	p757	1144	1283	2	228
71	p2082	1411	1144	2	224
72	p479	1564	1213	2	222
73	p1009	p25	1286	4	220
74	p4410	p150	p3	5	218
75	p982	p45	1144	3	215
76	p917	1224	1564	2	213
77	p2880	1123	1286	2	212
78	p72408	p150	1248	3	211
79	p1474	1218	1411	2	201
80	p1376	1224	1553	2	200
81	p5390	1411	1248	2	199
82	p2932	1221	1309	2	198
83	p1725	p75	p77	5	197
84	p5295	1288	p1	3	194
85	p1804	1144	p92	3	192
86	p1206	p150	1237	3	191
87	p1480	1224	1459	2	190
88	p4592	p76	1286	7	189
89	p72617	p82	1286	5	189
90	p892	p55	1411	3	186
91	p72627	p1	1288	3	185
92	p1599	p3	1309	4	183
93	p685	1300	p102	3	180
94	p5316	1411	1110	2	177
95	p72758	1293	1248	2	170
96	p72652	p12	1293	3	166
97	p73370	p6	1293	3	165
98	p2884	1144	1308	2	163
99	p72428	p72369	1286	6	162
100	p2059	1224	1556	2	162

Figura 147. Patrones del dominio más repetidos del experimento 5

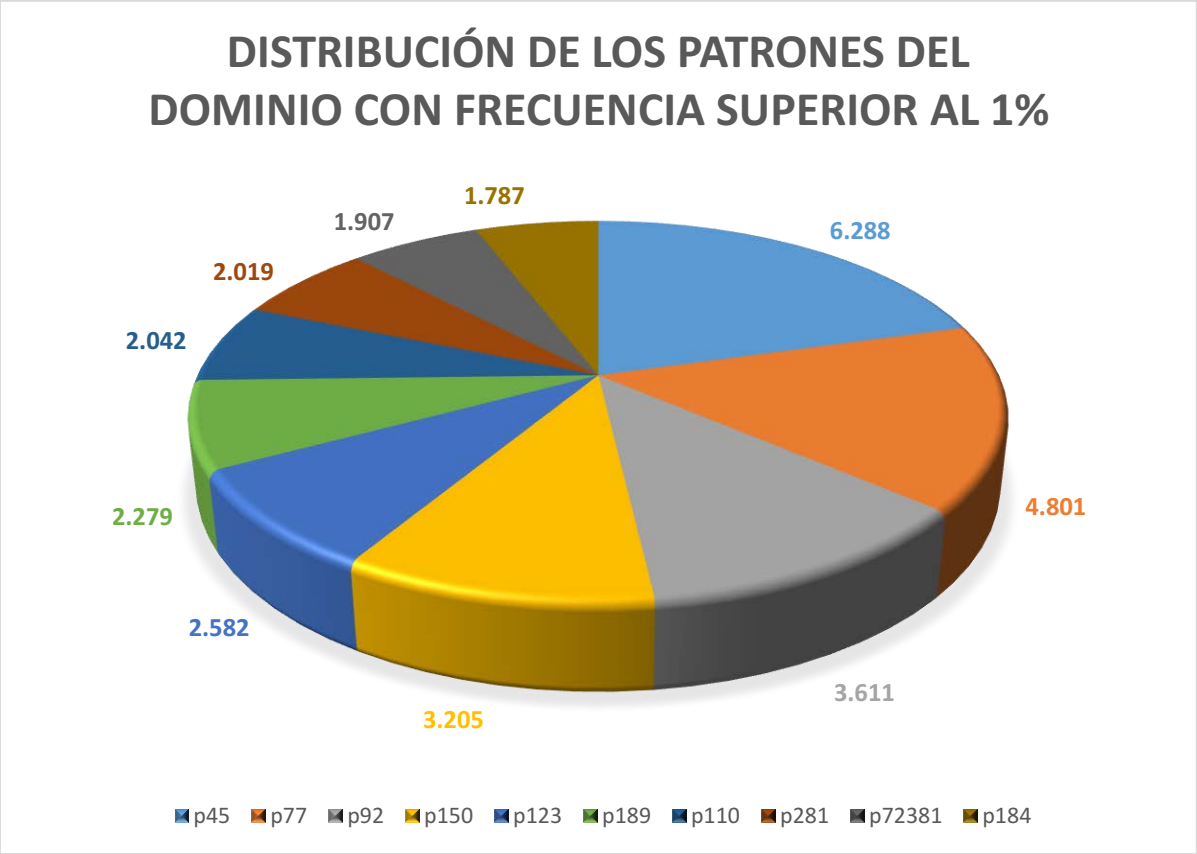

Figura 148. Gráfico de patrones del dominio en el experimento 5

8.5.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 222 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1213	PREPOSITION	1.801
2	1108	VERB	1.777
3	1218	QUANTIFIER DETERMINER	1.645
4	1151	ADVERB	1.389
5	1119	NOUN	1.364
6	1144	UNCLASSIFIED NOUN	1.144
7	1123	NUMBER	1.136
8	1103	ADJECTIVE	1.125
9	1110	SYMBOL	754
10	1248	AND LINKING	745

Figura 149. Categorías más frecuentes en el experimento 5

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	483
2	1411	Mutation	428
3	1300	Exons	420
4	1309	Genes	395
5	1293	Genes, Neurofibromatosis 1	350
6	1288	Genes, Tumor Suppressor	295
7	1291	Genes, Wilms Tumor	251
8	1294	Genes, Neurofibromatosis 2	188
9	1308	Alleles	181
10	1526	Phenotype	165

Figura 150. Categorías de genética más frecuentes en el experimento 5

Limitando los resultados a sólo los conceptos asociados a la sordera, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	245
2	1553	Deafness	186
3	1561	Hearing Loss, Sensorineural	180
4	1558	Hearing Loss, Functional	118
5	1559	Hearing Loss, High-Frequency	60
6	1563	Wolfram Syndrome	55
7	1555	Hearing Loss, Bilateral	42
8	1556	Hearing Loss, Central	32
9	1566	Hearing Loss, Mixed Conductive-Sensorineural	17
10	1565	Hearing Loss, Unilateral	17

Figura 151. Categorías de sordera más frecuentes en el experimento 5

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

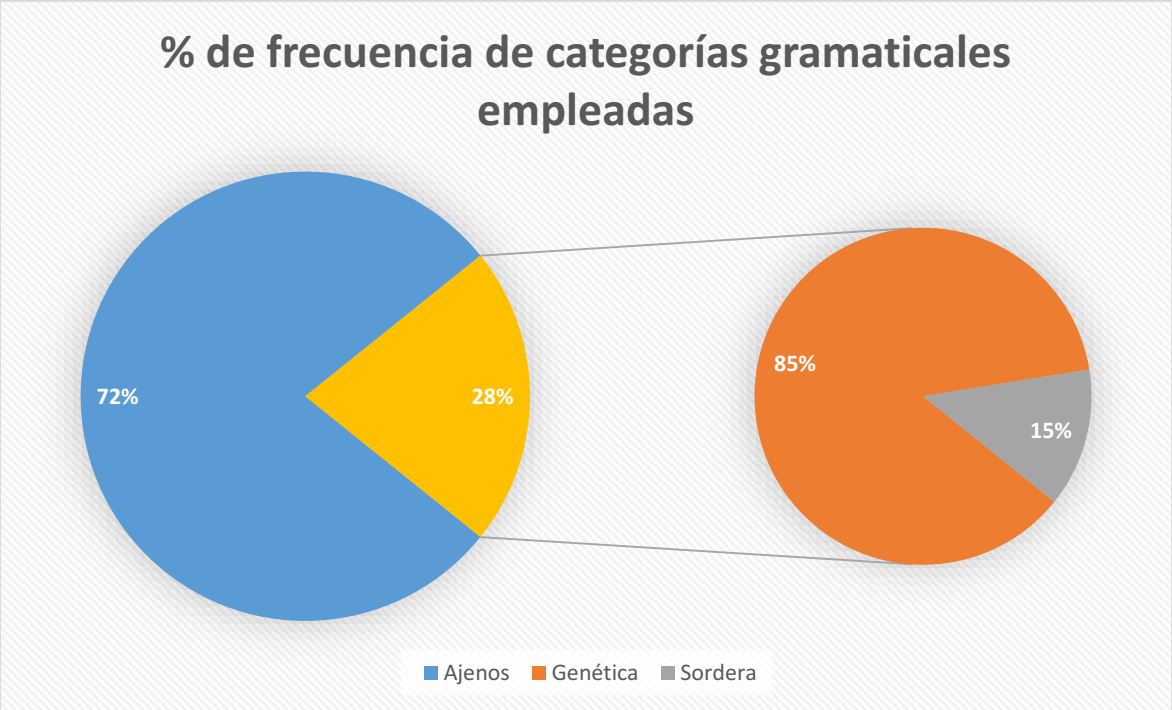

Figura 152. Proporción de categorías gramaticales en el experimento 5

Aproximadamente el 72% de elementos que forman los patrones son ajenos a nuestro dominio. El 28% forman parte del dominio, de los cuales el 85% están asociados a la "Genética" y el 15% restante a la "Sordera".

8.5.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p105	p14	1213	3	8.829	**66.217**
2	p50	1224	p17	3	7.001	**52.507**
3	p41	p16	p16	4	4.140	**41.400**
4	p104	p41	p41	8	1.619	**32.380**
5	p114	p14	1236	3	3.571	**26.782**
6	p116	p14	1229	3	3.462	**25.965**
7	p4	p1	1110	3	93.593	**23.866**
8	p288	p104	p104	16	552	**22.080**
9	p244	p47	1213	3	2.443	**18.322**
10	p160	p18	1123	3	2.211	**16.582**
11	p11	p4	p4	6	32.092	**16.366**
12	p14	1230	1108	2	27.146	**13.573**

13	p167	1224	p48	3	1.764	**13.230**
14	p187	p18	p18	4	1.245	**12.450**
15	p7	p2	1144	3	45.765	**11.670**
16	p17	1119	1237	2	23.184	**11.592**
17	p10	1224	p1	3	38.738	**9.878**
18	p21	p6	1237	3	19.231	**9.663**
19	p287	p47	1229	3	1.167	**8.752**
20	p18	1123	1110	2	17.379	**8.689**
21	p89128	p288	p288	32	105	**8.400**
22	p24	1110	p7	4	14.727	**7.363**
23	p342	p288	1123	17	173	**7.352**
24	p16	1123	1123	2	14.002	**7.001**
25	p54	p11	p11	12	6.517	**6.647**
26	p366	1223	p17	3	848	**6.360**
27	p259	p31	1203	3	840	**6.300**
28	p31	1151	1110	2	12.025	**6.012**
29	p36	p1	1213	3	23.365	**5.958**
30	p319	1233	p14	3	778	**5.835**
31	p126708	p126707	p10139	60	36	**5.400**
32	p481	p47	1228	3	707	**5.302**
33	p587	p122	1108	3	656	**4.920**
34	p109250	p3626	p3626	4	492	**4.920**
35	p81	1108	1213	2	9.801	**4.900**
36	p424	1203	p56	3	642	**4.815**
37	p234	1230	p39	3	642	**4.815**
38	p505	p14	1228	3	628	**4.710**
39	p126707	p89128	p288	48	39	**4.680**
40	p37	p8	1110	3	9.081	**4.563**
41	p26	p1	1119	3	17.721	**4.518**
42	p658	p14	1151	3	587	**4.402**
43	p46	1224	1119	2	8.610	**4.305**
44	p616	1233	p116	4	423	**4.230**
45	p56	1230	1151	2	8.411	**4.205**
46	p38	p2	1248	3	8.365	**4.203**
47	p581	1110	p31	3	558	**4.185**
48	p47	1230	1103	2	8.084	**4.042**
49	p35	p10	1237	4	8.072	**4.036**
50	p504	p14	1166	3	537	**4.027**
51	p39	1151	1108	2	8.018	**4.009**
52	p22	p1	1237	3	15.380	**3.921**
53	p853	p39	1213	3	513	**3.847**
54	p25	p1	1248	3	14.989	**3.822**
55	p1106	1103	p17	3	508	**3.810**
56	p1640	1233	p105	4	381	**3.810**
57	p109252	p109250	p109250	8	190	**3.800**

58	p44	1224	1103	2	7.103	**3.551**
59	p75	1110	1248	2	7.100	**3.550**
60	p306	p102	1123	3	458	**3.435**
61	p653	p31	1286	3	452	**3.390**
62	p2	1144	1110	2	135.563	**3.389**
63	p2074	1240	p105	4	335	**3.350**
64	p533	p227	1151	3	441	**3.307**
65	p48	1158	1237	2	6.559	**3.279**
66	p72371	p46	1237	3	430	**3.225**
67	p51	p6	1119	3	6.350	**3.190**
68	p912	1213	p44	3	425	**3.187**
69	p111	1119	1213	2	6.290	**3.145**
70	p2925	1218	p48	3	416	**3.120**
71	p486	p47	1236	3	411	**3.082**
72	p484	1203	p39	3	410	**3.075**
73	p802	1228	p248	3	406	**3.045**
74	p63	1110	p1	3	11.867	**3.026**
75	p925	p31	1213	3	402	**3.015**
76	p494	p14	1221	3	391	**2.932**
77	p1003	p44	p17	4	287	**2.870**
78	p651	p46	1248	3	381	**2.857**
79	p33	1224	p3	4	11.412	**2.853**
80	p980	p263	p114	5	227	**2.837**
81	p112	1103	1213	2	5.626	**2.813**
82	p1832	1240	p14	3	373	**2.797**
83	p190	1221	p95	3	368	**2.760**
84	p493	p14	p62	4	274	**2.740**
85	p667	1300	p18	3	363	**2.722**
86	p72439	p56	1108	3	360	**2.700**
87	p72459	p56	p81	4	270	**2.700**
88	p1137	p46	1213	3	358	**2.685**
89	p61	p2	1119	3	5.336	**2.681**
90	p544	p39	1229	3	350	**2.625**
91	p109263	p109252	p109252	16	65	**2.600**
92	p1849	p16	1123	3	341	**2.557**
93	p60	p2	p2	4	5.107	**2.553**
94	p62	1213	1223	2	5.053	**2.526**
95	p30	p1	p2	4	10.099	**2.524**
96	p1619	1203	p115	3	336	**2.520**
97	p548	p14	1248	3	335	**2.512**
98	p71	1108	1229	2	5.023	**2.511**
99	p694	p31	1151	3	328	**2.460**
100	p79	p8	1237	3	4.869	**2.446**

Figura 153. Patrones con mayor ponderación del experimento 5

El patrón con la mayor ponderación es p43, cuya secuencia es la siguiente:

- **p105**: “p14 + 1213” = “1230 + 1108 + 1213” = “VERB TO BE + VERB + PREPOSITION”

El patrón tiene longitud 3 y se repitió 8.829 veces en el texto, por lo que ocupa la posición 39º en la lista de patrones más frecuentes en este escenario de estudio.

Algunos otros patrones ponderados de interés encontrados en la lista son:

- **p366**: “1223 + p17” = “1223 + 1119 + 1237” = “INDEFINITE ARTICLE + NOUN + PREPOSITION OF”.
- **p616**: “1233 + p116” = “1233 + p14 + 1229” = “1233 + 1230 + 1108 + 1229” = “VERB TO HAVE + VERB TO BE + VERB + PREPOSITION TO”.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p653	p31	1286	3	452	**3.390**
2	p667	1300	p18	3	363	**2.722**
3	p77	p2	1286	3	4.801	**2.412**
4	p92	1411	1213	2	3.611	**1.805**
5	p150	1300	1123	2	3.205	**1.602**
6	p72408	p150	1248	3	211	**1.582**
7	p1206	p150	1237	3	191	**1.432**
8	p685	1300	p102	3	180	**1.350**
9	p123	p6	1309	3	2.582	**1.297**
10	p72430	p150	p266	4	112	**1.120**
11	p72792	p112	1293	3	147	**1.102**
12	p110	p45	1110	3	2.042	**1.026**
13	p1012	p18	1286	3	136	**1.020**
14	p2442	p47	1286	3	133	**997**
15	p72381	1224	1293	2	1.907	**953**
16	p82222	1411	p497	3	125	**937**
17	p213	p212	p113	18	327	**912**

18	p72521	1237	p72381	3	120	**900**
19	p184	1286	1110	2	1.787	**893**
20	p251	1110	p77	4	1.097	**822**
21	p2654	p85	1286	3	103	**772**
22	p914	1223	p92	3	102	**765**
23	p1183	1300	p306	4	73	**730**
24	p72796	p92	p72381	4	73	**730**
25	p1726	p75	1286	3	88	**660**
26	p73309	p17	p72381	4	64	**640**
27	p1270	1291	p429	3	84	**630**
28	p2432	1363	p102	3	80	**600**
29	p3681	1203	p417	3	79	**592**
30	p1521	p150	1213	3	77	**577**

Figura 154. Patrones del dominio con mayor ponderación del experimento 5

El patrón del dominio con la mayor ponderación es p653, cuya secuencia es la siguiente:

- **p653**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 452 veces en el texto, por lo que ocupa la posición 33º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.6. Experimento 6: frecuencia = 5, con semántica

8.6.1 Estudio de patrones generados

Se han generado un total de 14.382 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.738
7	p8	1144	1119	2	53.542
8	p15	1144	1213	2	49.409
9	p7	p2	1144	3	45.765
10	p9	1123	1144	2	43.189
11	p10	1224	p1	3	38.738
12	p11	p4	p4	6	32.092
13	p12	1144	1237	2	29.788
14	p13	1144	1248	2	28.730
15	p14	1230	1108	2	27.134
16	p53	1110	1144	2	26.738
17	p36	p1	1213	3	23.365
18	p17	1119	1237	2	23.142
19	p21	p6	1237	3	19.231
20	p20	1144	1108	2	18.517
21	p26	p1	1119	3	17.712
22	p18	1123	1110	2	17.379
23	p19	1144	1230	2	16.449
24	p27	1144	1151	2	16.007
25	p22	p1	1237	3	15.380
26	p25	p1	1248	3	14.989
27	p28	1144	1123	2	14.885
28	p24	1110	p7	4	14.727
29	p16	1123	1123	2	14.002
30	p29	1144	1229	2	12.287
31	p31	1151	1110	2	12.006
32	p63	1110	p1	3	11.867
33	p33	1224	p3	4	11.412
34	p23	p1	p3	5	11.411
35	p30	p1	p2	4	10.099

36	p81	1108	1213	2	9.668
37	p34	p1	1108	3	9.212
38	p37	p8	1110	3	9.079
39	p105	p14	1213	3	8.771
40	p46	1224	1119	2	8.562
41	p56	1230	1151	2	8.411
42	p38	p2	1248	3	8.365
43	p42	p1	1123	3	8.230
44	p32	p1	1230	3	8.175
45	p73	1213	1144	2	8.153
46	p40	1144	1228	2	8.110
47	p35	p10	1237	4	8.072
48	p47	1230	1103	2	8.063
49	p91	p3	1213	4	7.991
50	p39	1151	1108	2	7.909
51	p75	1110	1248	2	7.100
52	p44	1224	1103	2	7.016
53	p50	1224	p17	3	6.988
54	p59	p1	1151	3	6.760
55	p43	1144	1197	2	6.719
56	p66	1123	p1	3	6.566
57	p48	1158	1237	2	6.559
58	p65	p3	1119	4	6.541
59	p54	p11	p11	12	6.517
60	p57	1224	p5	5	6.342
61	p51	p6	1119	3	6.315
62	p45	1144	1286	2	6.288
63	p82	p4	1144	4	6.233
64	p52	p2	p1	4	6.198
65	p111	1119	1213	2	6.195
66	p55	1223	1144	2	6.135
67	p96	1119	1144	2	5.726
68	p78	1144	1158	2	5.701
69	p112	1103	1213	2	5.565
70	p161	1230	1144	2	5.469
71	p70	p3	1123	4	5.379
72	p61	p2	1119	3	5.336
73	p90	1151	1144	2	5.236
74	p72	p3	1248	4	5.157
75	p69	1223	p1	3	5.121
76	p60	p2	p2	4	5.107
77	p49	1144	1236	2	5.090
78	p68	1144	p9	3	5.063
79	p62	1213	1223	2	5.045
80	p71	1108	1229	2	4.956

81	p76	p5	p1	6	4.953
82	p79	p8	1237	3	4.869
83	p67	1144	1103	2	4.856
84	p77	p2	1286	3	4.801
85	p11455	1144	p53	3	4.741
86	p109	p4	p1	5	4.546
87	p115	1108	1221	2	4.439
88	p139	1230	p15	3	4.329
89	p85	1119	1110	2	4.210
90	p103	p5	p5	8	4.175
91	p87	1223	1103	2	4.144
92	p41	p16	p16	4	4.140
93	p84	p6	1248	3	4.137
94	p74	1221	1144	2	4.025
95	p83	p3	1237	4	3.884
96	p128	1103	1144	2	3.807
97	p97	1108	1144	2	3.792
98	p11456	p53	p53	4	3.732
99	p92	1411	1213	2	3.601
100	p114	p14	1236	3	3.571

Figura 155. Patrones más repetidos del experimento 6

Figura 156. Gráfico de los patrones más frecuentes del experimento 6

8.6.2 Estudio de los patrones del dominio

De los 14.382 patrones generados en este escenario, 2.115 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 14,71% de los patrones pertenecen al dominio de estudio.

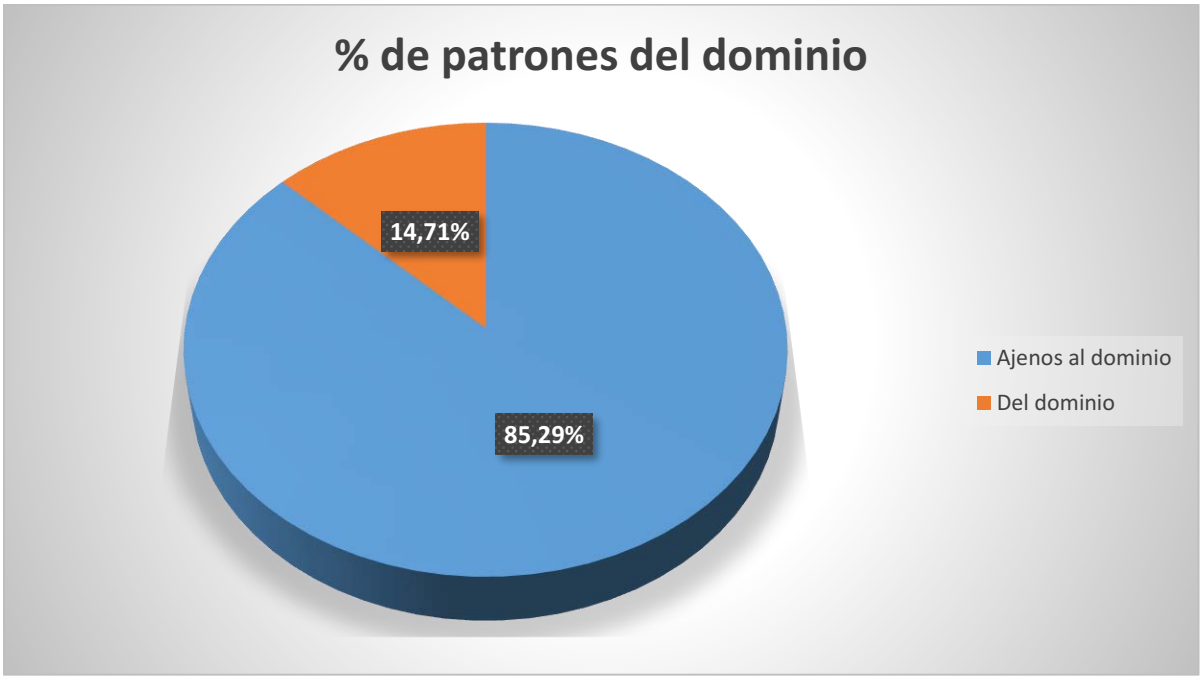

Figura 157. Proporción de patrones en el experimento 6

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p45	1144	1286	2	6.288
2	p77	p2	1286	3	4.801
3	p92	1411	1213	2	3.601
4	p150	1300	1123	2	3.205
5	p123	p6	1309	3	2.582
6	p189	1144	1411	2	2.279
7	p110	p45	1110	3	2.042
8	p281	1144	1309	2	2.019
9	p9957	1224	1293	2	1.907
10	p184	1286	1110	2	1.787
11	p101	1291	1144	2	1.143
12	p402	p1	1286	3	1.113

13	p251	1110	p77	4	1.097
14	p417	1103	1286	2	989
15	p456	p3	1286	4	986
16	p387	p4	1286	4	935
17	p334	1224	1411	2	924
18	p441	1110	1286	2	916
19	p539	1286	1151	2	845
20	p318	1224	1309	2	809
21	p3648	1144	1288	2	771
22	p113	p35	p101	6	730
23	p9982	1293	1144	2	698
24	p315	1286	1144	2	684
25	p474	p6	1411	3	684
26	p668	1300	p9	3	673
27	p835	1309	1144	2	585
28	p293	1144	1564	2	572
29	p9969	p82	p184	6	516
30	p454	1411	1237	2	498
31	p560	1221	1411	2	453
32	p653	p31	1286	3	452
33	p10096	1144	1293	2	451
34	p1105	p1	1309	3	442
35	p2077	1288	1144	2	397
36	p834	1300	1144	2	385
37	p10009	p15	1293	3	367
38	p360	p92	p123	5	366
39	p667	1300	p18	3	363
40	p11472	p11460	1286	4	337
41	p10520	p53	1286	3	326
42	p800	1224	1300	2	322
43	p734	1345	1228	2	319
44	p773	1309	1110	2	319
45	p1803	1123	1309	2	318
46	p9989	1293	1411	2	315
47	p1184	1309	1213	2	314
48	p2081	1363	1123	2	311
49	p874	1224	1291	2	304
50	p1271	1291	p1	3	299
51	p1823	1224	1294	2	294
52	p9991	p1	1411	3	289
53	p9846	1286	p1	3	287
54	p916	1224	1558	2	286
55	p12204	p3625	1286	4	281
56	p725	p6	1526	3	269
57	p769	p13	1286	3	264

58	p1591	p6	1308	3	258
59	p1350	p5	1286	5	256
60	p611	p30	1286	5	253
61	p11227	p9	1293	3	252
62	p999	p55	1309	3	250
63	p1182	1248	1286	2	244
64	p10027	1293	p1	3	243
65	p1995	p10	1309	4	240
66	p915	1224	1526	2	235
67	p10560	1224	1269	2	232
68	p757	1144	1283	2	228
69	p641	1144	1553	2	220
70	p982	p45	1144	3	215
71	p2082	1411	1144	2	214
72	p1009	p25	1286	4	213
73	p9984	p150	1248	3	211
74	p917	1224	1564	2	209
75	p4410	p150	p3	5	209
76	p479	1564	1213	2	206
77	p2880	1123	1286	2	202
78	p2932	1221	1309	2	198
79	p1095	1218	1309	2	198
80	p1376	1224	1553	2	196
81	p5390	1411	1248	2	196
82	p1804	1144	p92	3	190
83	p4592	p76	1286	7	187
84	p1725	p75	p77	5	187
85	p10201	p82	1286	5	186
86	p1206	p150	1237	3	186
87	p5295	1288	p1	3	181
88	p1480	1224	1459	2	181
89	p685	1300	p102	3	180
90	p892	p55	1411	3	178
91	p1599	p3	1309	4	176
92	p5316	1411	1110	2	174
93	p10349	1293	1248	2	170
94	p10211	p1	1288	3	170
95	p10006	p9945	1286	6	162
96	p2059	1224	1556	2	159
97	p10239	p12	1293	3	158
98	p1474	1218	1411	2	157
99	p2884	1144	1308	2	157
100	p1378	1248	p77	4	157

Figura 158. Patrones del dominio más repetidos en el experimento 6

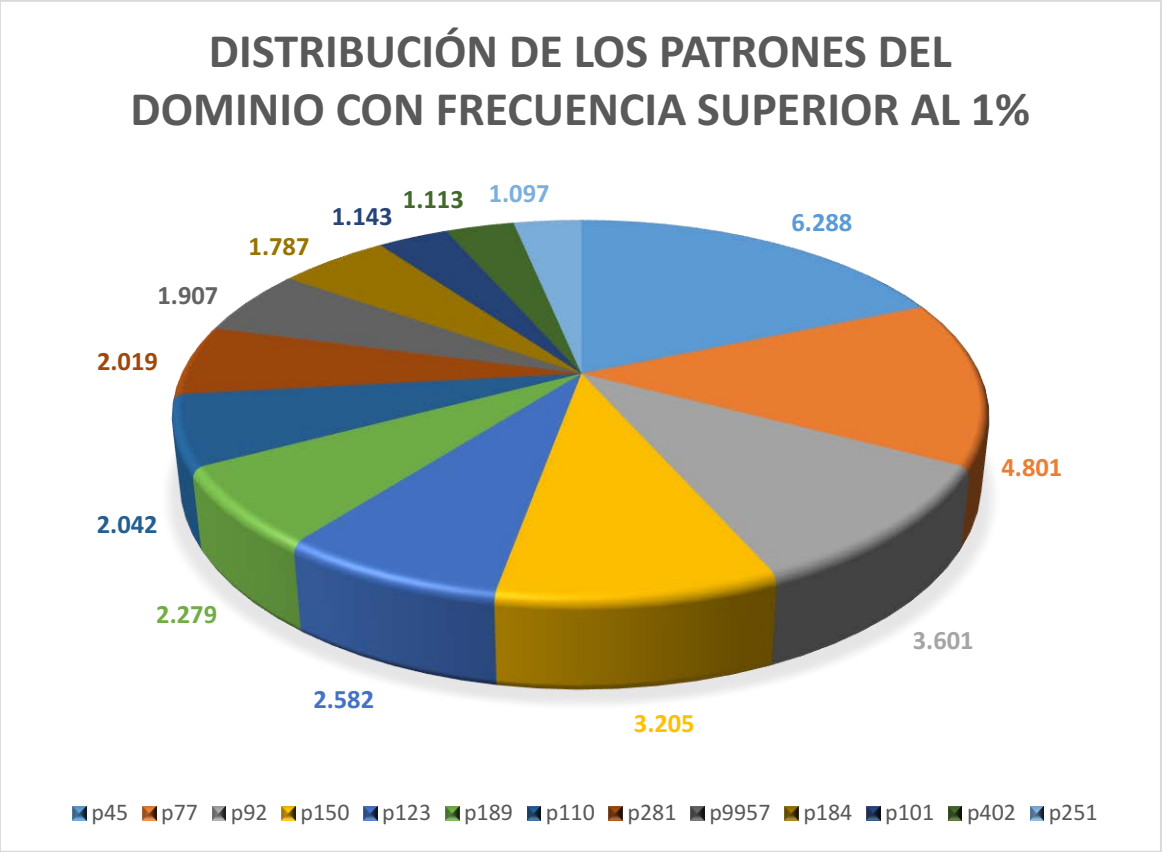

Figura 159. Gráfico de patrones del dominio en el experimento 6

8.6.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 222 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	429
2	1213	PREPOSITION	407
3	1123	NUMBER	369
4	1108	VERB	354
5	1119	NOUN	353
6	1151	ADVERB	347
7	1103	ADJECTIVE	256
8	1218	QUANTIFIER DETERMINER	252
9	1110	SYMBOL	240
10	1248	AND LINKING	220

Figura 160. Categorías más frecuentes en el experimento 6

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	155
2	1411	Mutation	99
3	1309	Genes	94
4	1300	Exons	88
5	1293	Genes, Neurofibromatosis 1	77
6	1294	Genes, Neurofibromatosis 2	42
7	1291	Genes, Wilms Tumor	42
8	1288	Genes, Tumor Suppressor	42
9	1363	Introns	39
10	1526	Phenotype	29

Figura 161. Categorías de genética más frecuentes en el experimento 6

Limitando los resultados a sólo los conceptos asociados a la sordera, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	56
2	1553	Deafness	33
3	1561	Hearing Loss, Sensorineural	31
4	1558	Hearing Loss, Functional	23
5	1563	Wolfram Syndrome	9
6	1555	Hearing Loss, Bilateral	7
7	1559	Hearing Loss, High-Frequency	7
8	1556	Hearing Loss, Central	6
9	1567	Usher Syndromes	4
10	1566	Hearing Loss, Mixed Conductive-Sensorineural	3

Figura 162. Categorías de sordera más frecuentes en el experimento 6

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

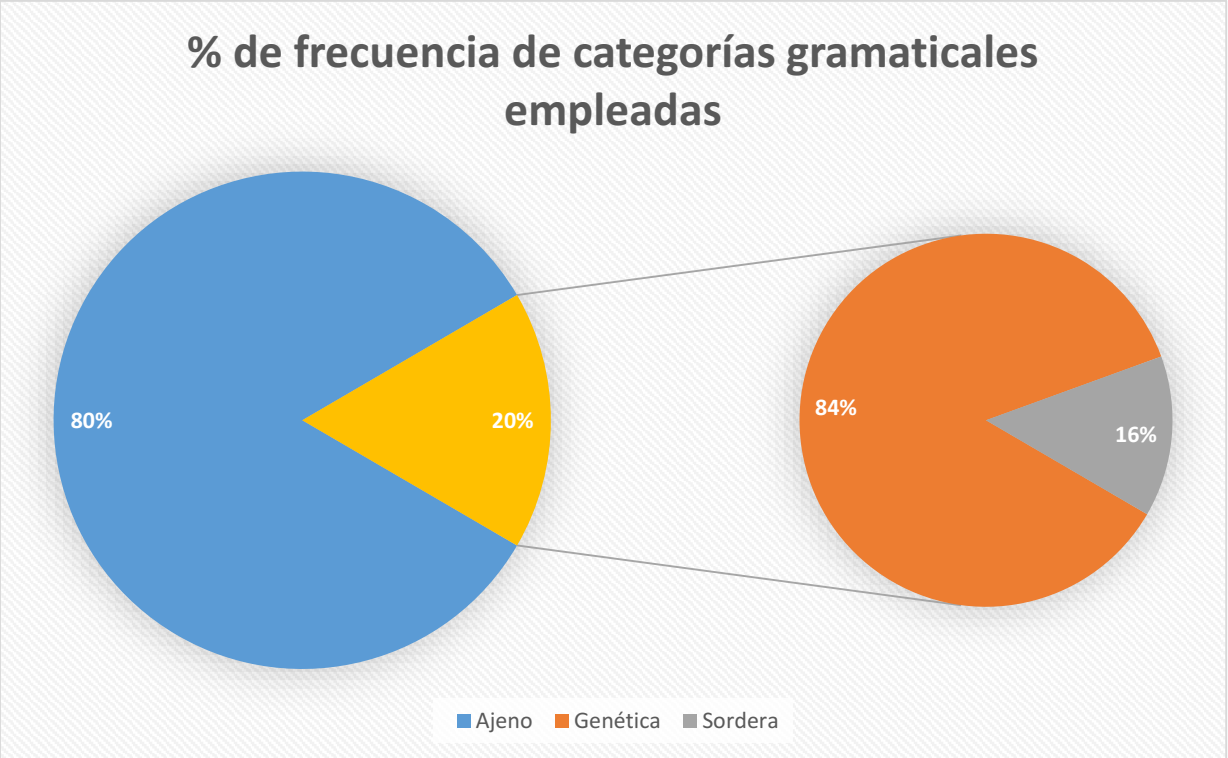

Figura 163. Proporción de categorías gramaticales en el experimento 6

Aproximadamente el 80% de elementos que forman los patrones son ajenos a nuestro dominio. El 20% forman parte del dominio, de los cuales el 84% están asociados a la "Genética" y el 16% restante a la "Sordera".

8.6.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p105	p14	1213	3	8.771	**65.782**
2	p50	1224	p17	3	6.988	**52.410**
3	p41	p16	p16	4	4.140	**41.400**
4	p104	p41	p41	8	1.608	**32.160**
5	p114	p14	1236	3	3.571	**26.782**
6	p116	p14	1229	3	3.462	**25.965**
7	p4	p1	1110	3	93.593	**23.866**
8	p288	p104	p104	16	552	**22.080**
9	p244	p47	1213	3	2.412	**18.090**
10	p160	p18	1123	3	2.211	**16.582**
11	p11	p4	p4	6	32.092	**16.366**

12	p14	1230	1108	2	27.134	**13.567**
13	p167	1224	p48	3	1.764	**13.230**
14	p187	p18	p18	4	1.245	**12.450**
15	p7	p2	1144	3	45.765	**11.670**
16	p17	1119	1237	2	23.142	**11.571**
17	p10	1224	p1	3	38.738	**9.878**
18	p21	p6	1237	3	19.231	**9.663**
19	p18	1123	1110	2	17.379	**8.689**
20	p287	p47	1229	3	1.153	**8.647**
21	p11221	p288	p288	32	99	**7.920**
22	p24	1110	p7	4	14.727	**7.363**
23	p342	p288	1123	17	169	**7.182**
24	p16	1123	1123	2	14.002	**7.001**
25	p54	p11	p11	12	6.517	**6.647**
26	p215	1151	p206	8	327	**6.540**
27	p366	1223	p17	3	844	**6.330**
28	p259	p31	1203	3	840	**6.300**
29	p31	1151	1110	2	12.006	**6.003**
30	p36	p1	1213	3	23.365	**5.958**
31	p319	1233	p14	3	778	**5.835**
32	p13605	p13604	p13602	60	36	**5.400**
33	p481	p47	1228	3	697	**5.227**
34	p12509	p3626	p3626	4	488	**4.880**
35	p81	1108	1213	2	9.668	**4.834**
36	p234	1230	p39	3	642	**4.815**
37	p424	1203	p56	3	640	**4.800**
38	p505	p14	1228	3	628	**4.710**
39	p13604	p11221	p288	48	39	**4.680**
40	p37	p8	1110	3	9.079	**4.562**
41	p26	p1	1119	3	17.712	**4.516**
42	p658	p14	1151	3	584	**4.380**
43	p46	1224	1119	2	8.562	**4.281**
44	p587	p122	1108	3	565	**4.237**
45	p616	1233	p116	4	423	**4.230**
46	p56	1230	1151	2	8.411	**4.205**
47	p38	p2	1248	3	8.365	**4.203**
48	p581	1110	p31	3	558	**4.185**
49	p35	p10	1237	4	8.072	**4.036**
50	p47	1230	1103	2	8.063	**4.031**
51	p504	p14	1166	3	534	**4.005**
52	p39	1151	1108	2	7.909	**3.954**
53	p22	p1	1237	3	15.380	**3.921**
54	p25	p1	1248	3	14.989	**3.822**
55	p12511	p12509	p12509	8	190	**3.800**
56	p1640	1233	p105	4	368	**3.680**

57	p853	p39	1213	3	479	**3.592**
58	p1106	1103	p17	3	475	**3.562**
59	p75	1110	1248	2	7.100	**3.550**
60	p44	1224	1103	2	7.016	**3.508**
61	p306	p102	1123	3	458	**3.435**
62	p653	p31	1286	3	452	**3.390**
63	p2	1144	1110	2	135.563	**3.389**
64	p48	1158	1237	2	6.559	**3.279**
65	p533	p227	1151	3	436	**3.270**
66	p9947	p46	1237	3	425	**3.187**
67	p51	p6	1119	3	6.315	**3.173**
68	p111	1119	1213	2	6.195	**3.097**
69	p484	1203	p39	3	410	**3.075**
70	p486	p47	1236	3	406	**3.045**
71	p63	1110	p1	3	11.867	**3.026**
72	p802	1228	p248	3	402	**3.015**
73	p925	p31	1213	3	393	**2.947**
74	p494	p14	1221	3	391	**2.932**
75	p912	1213	p44	3	388	**2.910**
76	p2074	1240	p105	4	290	**2.900**
77	p33	1224	p3	4	11.412	**2.853**
78	p651	p46	1248	3	379	**2.842**
79	p2925	1218	p48	3	377	**2.827**
80	p1003	p44	p17	4	279	**2.790**
81	p112	1103	1213	2	5.565	**2.782**
82	p493	p14	p62	4	274	**2.740**
83	p667	1300	p18	3	363	**2.722**
84	p10017	p56	1108	3	360	**2.700**
85	p61	p2	1119	3	5.336	**2.681**
86	p10037	p56	p81	4	268	**2.680**
87	p190	1221	p95	3	351	**2.632**
88	p12525	p12511	p12511	16	65	**2.600**
89	p544	p39	1229	3	343	**2.572**
90	p60	p2	p2	4	5.107	**2.553**
91	p1849	p16	1123	3	337	**2.527**
92	p30	p1	p2	4	10.099	**2.524**
93	p62	1213	1223	2	5.045	**2.522**
94	p548	p14	1248	3	335	**2.512**
95	p980	p263	p114	5	201	**2.512**
96	p71	1108	1229	2	4.956	**2.478**
97	p79	p8	1237	3	4.869	**2.446**
98	p77	p2	1286	3	4.801	**2.412**
99	p401	p18	1248	3	321	**2.407**
100	p1832	1240	p14	3	316	**2.370**

Figura 164. Patrones con mayor ponderación del experimento 6

El patrón con la mayor ponderación es p105, cuya secuencia es la siguiente:

- **p105**: “p14 + 1213” = “1230 + 1108 + 1213” = “VERB TO BE + VERB + PREPOSITION”

El patrón tiene longitud 3 y se repitió 8.771 veces en el texto, por lo que ocupa la posición 39º en la lista de patrones más frecuentes en este escenario de estudio.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p653	p31	1286	3	452	**3.390**
2	p667	1300	p18	3	363	**2.722**
3	p77	p2	1286	3	4.801	**2.412**
4	p92	1411	1213	2	3.601	**1.800**
5	p150	1300	1123	2	3.205	**1.602**
6	p9984	p150	1248	3	211	**1.582**
7	p1206	p150	1237	3	186	**1.395**
8	p685	1300	p102	3	180	**1.350**
9	p123	p6	1309	3	2.582	**1.297**
10	p10008	p150	p266	4	112	**1.120**
11	p10385	p112	1293	3	137	**1.027**
12	p110	p45	1110	3	2.042	**1.026**
13	p9957	1224	1293	2	1.907	**953**
14	p1012	p18	1286	3	127	**952**
15	p10812	1411	p497	3	123	**922**
16	p213	p212	p113	18	327	**912**
17	p184	1286	1110	2	1.787	**893**
18	p10100	1237	p9957	3	118	**885**
19	p2442	p47	1286	3	111	**832**
20	p251	1110	p77	4	1.097	**822**
21	p1183	1300	p306	4	73	**730**
22	p10389	p92	p9957	4	66	**660**
23	p2654	p85	1286	3	82	**615**
24	p914	1223	p92	3	81	**607**
25	p113	p35	p101	6	730	**547**

26	p1726	p75	1286	3	73	**547**
27	p11800	p17	p9957	4	51	**510**
28	p1270	1291	p429	3	67	**502**
29	p417	1103	1286	2	989	**494**
30	p387	p4	1286	4	935	**467**

Figura 165. Patrones del dominio con mayor ponderación del experimento 6

El patrón del dominio con la mayor ponderación es p653, cuya secuencia es la siguiente:

- **p653**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 452 veces en el texto, por lo que ocupa la posición 33º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.7. Experimento 7: frecuencia = 10, con semántica

8.7.1 Estudio de patrones generados

Se han generado un total de 5.505 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.738
7	p8	1144	1119	2	53.534
8	p15	1144	1213	2	49.394
9	p7	p2	1144	3	45.765
10	p9	1123	1144	2	43.189
11	p10	1224	p1	3	38.738
12	p11	p4	p4	6	32.092
13	p12	1144	1237	2	29.788
14	p13	1144	1248	2	28.730
15	p14	1230	1108	2	27.125
16	p53	1110	1144	2	26.738
17	p36	p1	1213	3	23.333
18	p17	1119	1237	2	23.115
19	p21	p6	1237	3	19.231
20	p20	1144	1108	2	18.471
21	p26	p1	1119	3	17.695
22	p18	1123	1110	2	17.379
23	p19	1144	1230	2	16.449
24	p27	1144	1151	2	16.007
25	p22	p1	1237	3	15.380
26	p25	p1	1248	3	14.989
27	p28	1144	1123	2	14.885
28	p24	1110	p7	4	14.727
29	p16	1123	1123	2	14.002
30	p29	1144	1229	2	12.287
31	p31	1151	1110	2	11.963
32	p63	1110	p1	3	11.867
33	p33	1224	p3	4	11.412
34	p23	p1	p3	5	11.411
35	p30	p1	p2	4	10.099

36	p81	1108	1213	2	9.604
37	p34	p1	1108	3	9.183
38	p37	p8	1110	3	9.079
39	p105	p14	1213	3	8.748
40	p46	1224	1119	2	8.502
41	p56	1230	1151	2	8.405
42	p38	p2	1248	3	8.365
43	p42	p1	1123	3	8.230
44	p32	p1	1230	3	8.175
45	p40	1144	1228	2	8.110
46	p35	p10	1237	4	8.072
47	p73	1213	1144	2	8.060
48	p47	1230	1103	2	8.041
49	p91	p3	1213	4	7.976
50	p39	1151	1108	2	7.860
51	p75	1110	1248	2	7.100
52	p50	1224	p17	3	6.979
53	p44	1224	1103	2	6.886
54	p59	p1	1151	3	6.760
55	p43	1144	1197	2	6.719
56	p66	1123	p1	3	6.566
57	p48	1158	1237	2	6.559
58	p65	p3	1119	4	6.532
59	p54	p11	p11	12	6.517
60	p57	1224	p5	5	6.342
61	p51	p6	1119	3	6.292
62	p45	1144	1286	2	6.288
63	p82	p4	1144	4	6.233
64	p52	p2	p1	4	6.198
65	p55	1223	1144	2	6.135
66	p111	1119	1213	2	6.121
67	p96	1119	1144	2	5.707
68	p78	1144	1158	2	5.701
69	p112	1103	1213	2	5.494
70	p161	1230	1144	2	5.469
71	p70	p3	1123	4	5.379
72	p61	p2	1119	3	5.336
73	p90	1151	1144	2	5.213
74	p72	p3	1248	4	5.157
75	p69	1223	p1	3	5.114
76	p60	p2	p2	4	5.107
77	p49	1144	1236	2	5.090
78	p68	1144	p9	3	5.063
79	p62	1213	1223	2	5.027
80	p76	p5	p1	6	4.953

81	p71	1108	1229	2	4.923
82	p79	p8	1237	3	4.869
83	p67	1144	1103	2	4.856
84	p77	p2	1286	3	4.801
85	p4414	1144	p53	3	4.741
86	p109	p4	p1	5	4.546
87	p115	1108	1221	2	4.392
88	p139	1230	p15	3	4.316
89	p85	1119	1110	2	4.210
90	p103	p5	p5	8	4.175
91	p41	p16	p16	4	4.140
92	p84	p6	1248	3	4.137
93	p87	1223	1103	2	4.105
94	p74	1221	1144	2	4.025
95	p83	p3	1237	4	3.884
96	p4415	p53	p53	4	3.732
97	p128	1103	1144	2	3.727
98	p97	1108	1144	2	3.696
99	p92	1411	1213	2	3.574
100	p114	p14	1236	3	3.571

Figura 166. Patrones más repetidos del experimento 7

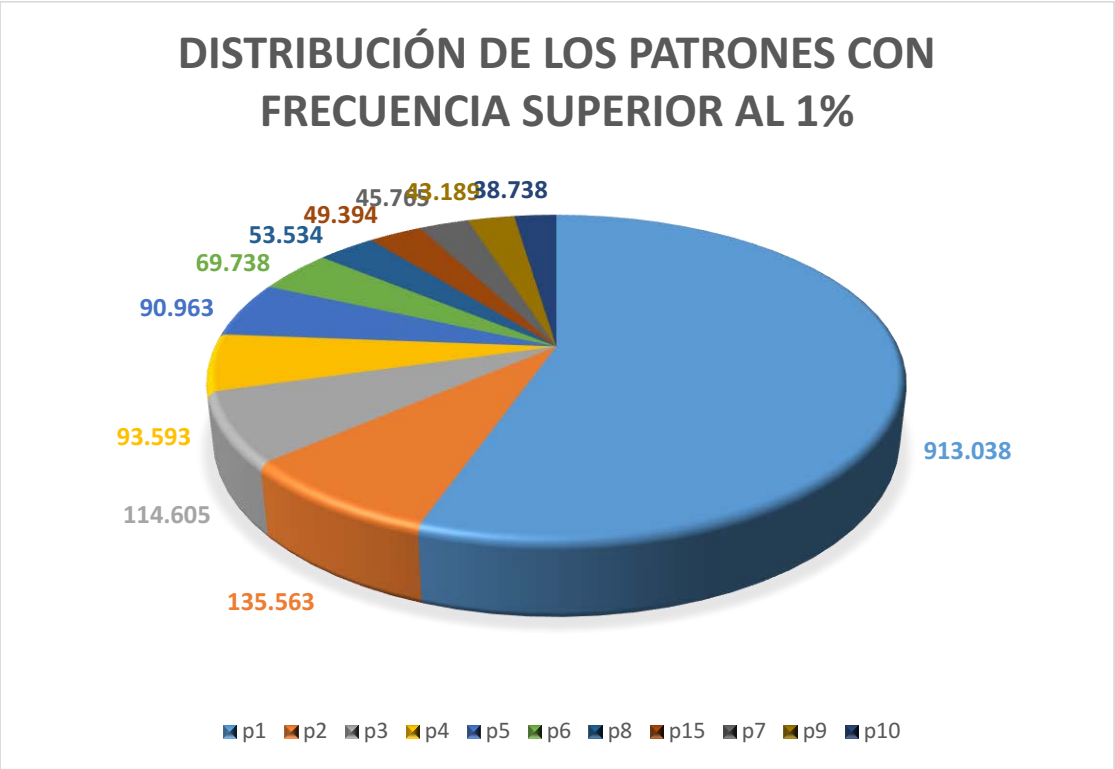

Figura 167. Gráfico de los patrones más frecuentes del experimento 7

8.7.2 Estudio de los patrones del dominio

De los 5.505 patrones generados en este escenario, 729 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 15,26% de los patrones pertenecen al dominio de estudio.

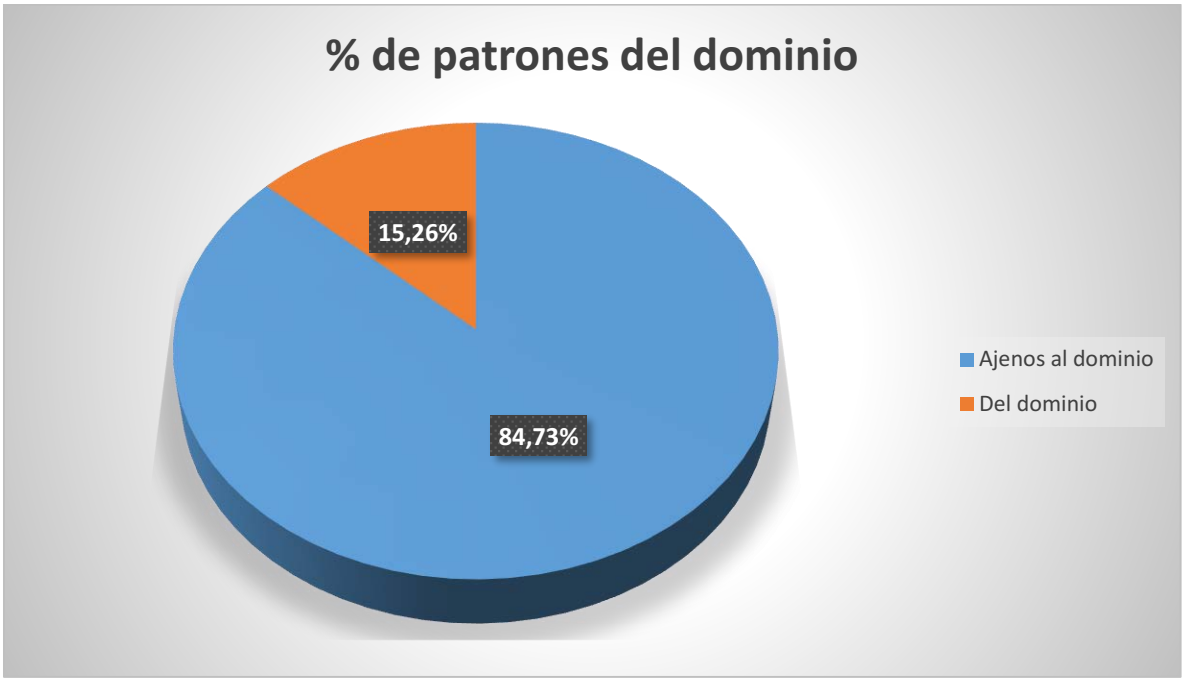

Figura 168. Proporción de patrones en el experimento 7

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p45	1144	1286	2	6.288
2	p77	p2	1286	3	4.801
3	p92	1411	1213	2	3.574
4	p150	1300	1123	2	3.195
5	p123	p6	1309	3	2.582
6	p189	1144	1411	2	2.279
7	p110	p45	1110	3	2.042
8	p281	1144	1309	2	2.019
9	p3817	1224	1293	2	1.897
10	p184	1286	1110	2	1.779
11	p101	1291	1144	2	1.143
12	p402	p1	1286	3	1.113

13	p251	1110	p77	4	1.097
14	p456	p3	1286	4	986
15	p417	1103	1286	2	964
16	p387	p4	1286	4	925
17	p334	1224	1411	2	915
18	p441	1110	1286	2	907
19	p539	1286	1151	2	845
20	p318	1224	1309	2	809
21	p3648	1144	1288	2	755
22	p113	p35	p101	6	730
23	p3842	1293	1144	2	698
24	p474	p6	1411	3	684
25	p668	1300	p9	3	666
26	p315	1286	1144	2	659
27	p835	1309	1144	2	585
28	p293	1144	1564	2	553
29	p3829	p82	p184	6	516
30	p454	1411	1237	2	498
31	p653	p31	1286	3	452
32	p3979	1144	1293	2	451
33	p1105	p1	1309	3	442
34	p560	1221	1411	2	437
35	p2077	1288	1144	2	388
36	p834	1300	1144	2	385
37	p3873	p15	1293	3	352
38	p667	1300	p18	3	343
39	p360	p92	p123	5	341
40	p4432	p4420	1286	4	337
41	p800	1224	1300	2	322
42	p773	1309	1110	2	311
43	p734	1345	1228	2	310
44	p1803	1123	1309	2	310
45	p874	1224	1291	2	304
46	p4343	p53	1286	3	301
47	p3849	1293	1411	2	300
48	p2081	1363	1123	2	294
49	p1271	1291	p1	3	289
50	p1823	1224	1294	2	285
51	p4759	p3625	1286	4	281
52	p3861	1286	p1	3	271
53	p1184	1309	1213	2	271
54	p3851	p1	1411	3	263
55	p1591	p6	1308	3	249
56	p916	1224	1558	2	247
57	p4386	p9	1293	3	246

58	p611	p30	1286	5	246
59	p769	p13	1286	3	243
60	p725	p6	1526	3	242
61	p1350	p5	1286	5	242
62	p999	p55	1309	3	234
63	p1182	1248	1286	2	226
64	p3893	1293	p1	3	225
65	p1995	p10	1309	4	225
66	p4324	1224	1269	2	210
67	p915	1224	1526	2	209
68	p757	1144	1283	2	203
69	p3844	p150	1248	3	203
70	p641	1144	1553	2	199
71	p4071	p150	p3	5	194
72	p2082	1411	1144	2	190
73	p982	p45	1144	3	190
74	p4105	p82	1286	5	186
75	p1376	1224	1553	2	181
76	p1009	p25	1286	4	178
77	p1804	1144	p92	3	177
78	p2880	1123	1286	2	176
79	p2932	1221	1309	2	175
80	p4009	p76	1286	7	170
81	p3967	1411	1248	2	167
82	p917	1224	1564	2	167
83	p685	1300	p102	3	165
84	p1480	1224	1459	2	165
85	p3869	p3805	1286	6	162
86	p4118	p1	1288	3	160
87	p1725	p75	p77	5	160
88	p892	p55	1411	3	159
89	p479	1564	1213	2	158
90	p1206	p150	1237	3	155
91	p4062	1288	p1	3	149
92	p1095	1218	1309	2	143
93	p1599	p3	1309	4	141
94	p4154	p12	1293	3	140
95	p2059	1224	1556	2	138
96	p4588	p6	1293	3	131
97	p4175	1411	p73	3	130
98	p2884	1144	1308	2	130
99	p2943	1224	1561	2	127
100	p1579	p11	p45	8	125

Figura 169. Patrones del domino más repetidos en el experimento 7

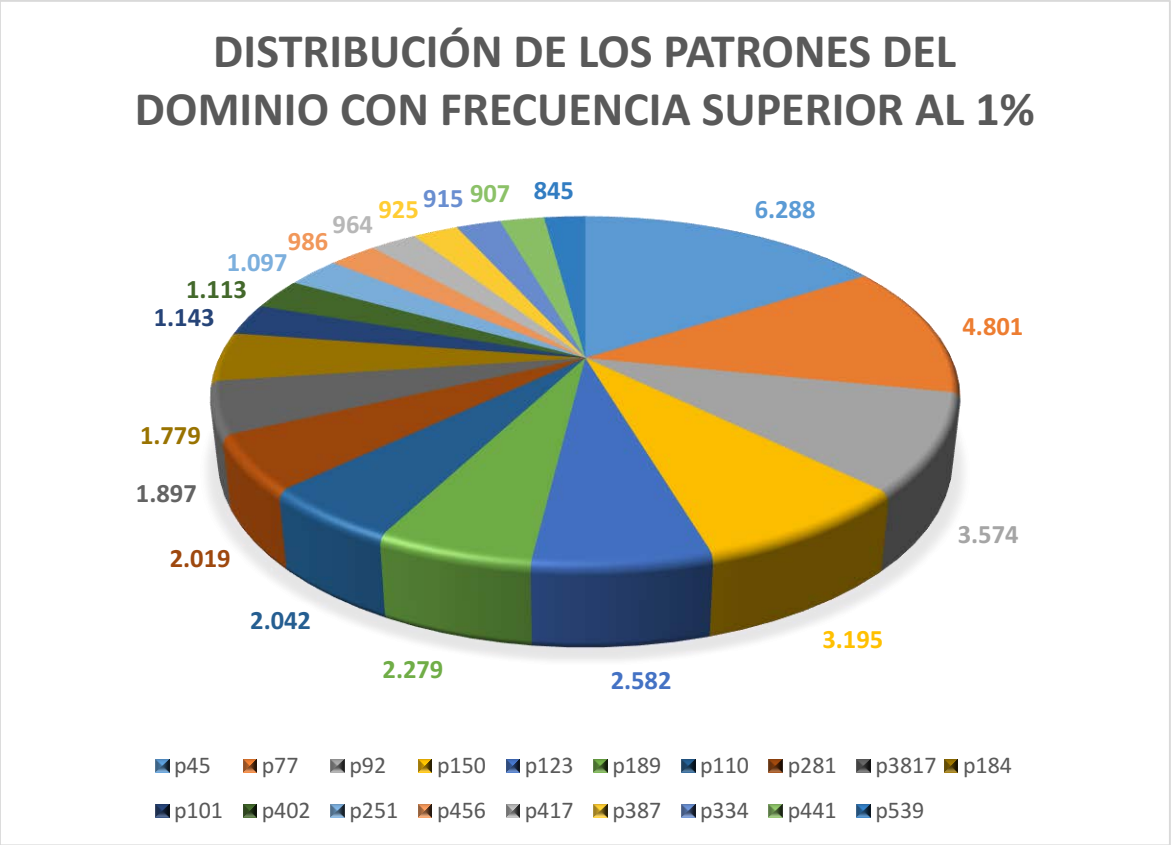

Figura 170. Gráfico de patrones del dominio en el experimento 7

8.7.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 108 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	293
2	1213	PREPOSITION	225
3	1123	NUMBER	209
4	1119	NOUN	190
5	1151	ADVERB	184
6	1108	VERB	179
7	1110	SYMBOL	140
8	1248	AND LINKING	131
9	1103	ADJECTIVE	117
10	1224	DEFINITE ARTICLE	110

Figura 171. Categorías más frecuentes en el experimento 7

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	81
2	1411	Mutation	47
3	1309	Genes	44
4	1300	Exons	41
5	1293	Genes, Neurofibromatosis 1	38
6	1288	Genes, Tumor Suppressor	21
7	1294	Genes, Neurofibromatosis 2	16
8	1291	Genes, Wilms Tumor	15
9	1363	Introns	13
10	1269	Phenotype	11

Figura 172. Categorías de sordera más frecuentes en el experimento 7

Limitando los resultados a sólo los conceptos asociados a la sordera, las categorías empleadas han sido las siguientes, ordenadas por número de repeticiones:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	22
2	1553	Deafness	12
3	1561	Hearing Loss, Sensorineural	11
4	1558	Hearing Loss, Functional	7
5	1559	Hearing Loss, High-Frequency	4
6	1556	Hearing Loss, Central	2
7	1563	Wolfram Syndrome	2
8	1555	Hearing Loss, Bilateral	1

Figura 173. Categorías de sordera más frecuentes en el experimento 7

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

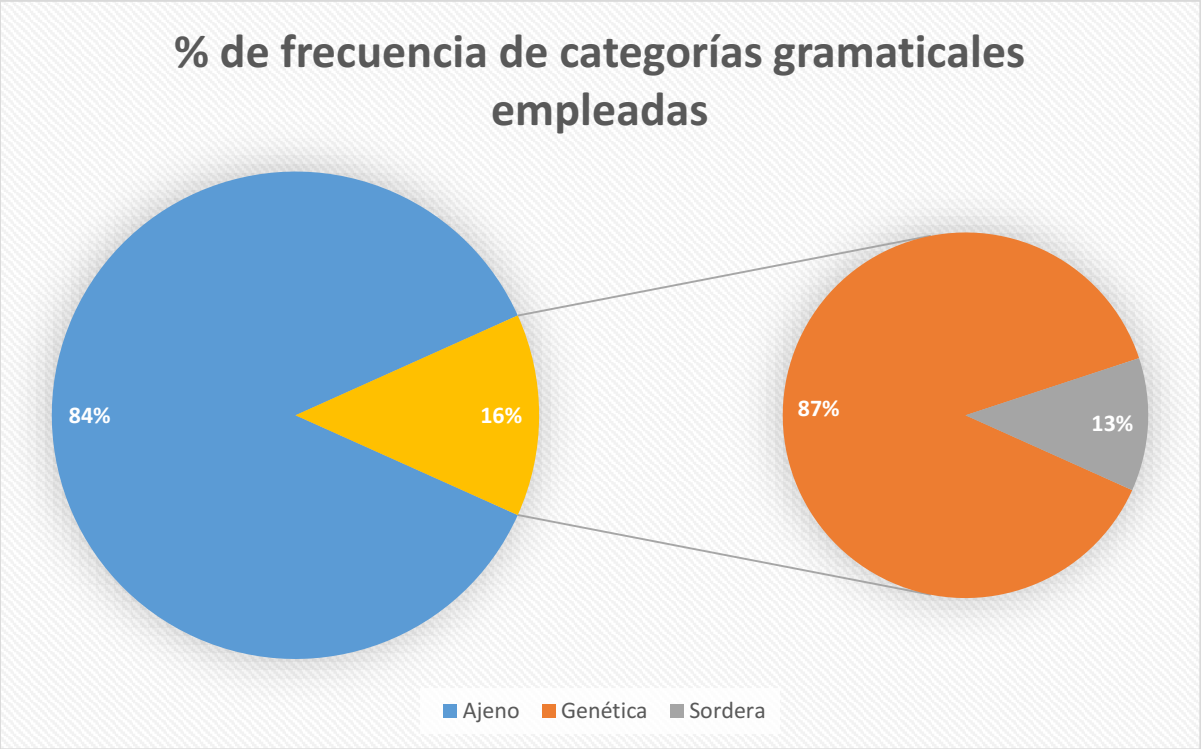

Figura 174. Proporción de categorías gramaticales en el experimento 7

Aproximadamente el 84% de elementos que forman los patrones son ajenos a nuestro dominio. El 16% forman parte del dominio, de los cuales el 87% están asociados a la "Genética" y el 13% restante a la "Sordera".

8.7.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p105	p14	1213	3	8.748	**65.610**
2	p50	1224	p17	3	6.979	**52.342**
3	p41	p16	p16	4	4.140	**41.400**
4	p104	p41	p41	8	1.602	**32.040**
5	p114	p14	1236	3	3.571	**26.782**
6	p116	p14	1229	3	3.462	**25.965**
7	p4	p1	1110	3	93.593	**23.866**
8	p288	p104	p104	16	545	**21.800**
9	p244	p47	1213	3	2.370	**17.775**
10	p160	p18	1123	3	2.211	**16.582**
11	p11	p4	p4	6	32.092	**16.366**

12	p14	1230	1108	2	27.125	**13.562**
13	p167	1224	p48	3	1.764	**13.230**
14	p187	p18	p18	4	1.245	**12.450**
15	p7	p2	1144	3	45.765	**11.670**
16	p17	1119	1237	2	23.115	**11.557**
17	p10	1224	p1	3	38.738	**9.878**
18	p21	p6	1237	3	19.231	**9.663**
19	p18	1123	1110	2	17.379	**8.689**
20	p287	p47	1229	3	1.147	**8.602**
21	p4371	p288	p288	32	99	**7.920**
22	p24	1110	p7	4	14.727	**7.363**
23	p342	p288	1123	17	169	**7.182**
24	p16	1123	1123	2	14.002	**7.001**
25	p54	p11	p11	12	6.517	**6.647**
26	p215	1151	p206	8	327	**6.540**
27	p259	p31	1203	3	840	**6.300**
28	p366	1223	p17	3	838	**6.285**
29	p31	1151	1110	2	11.963	**5.981**
30	p36	p1	1213	3	23.333	**5.949**
31	p319	1233	p14	3	770	**5.775**
32	p5231	p5229	p5227	60	36	**5.400**
33	p481	p47	1228	3	688	**5.160**
34	p4854	p3626	p3626	4	488	**4.880**
35	p234	1230	p39	3	642	**4.815**
36	p81	1108	1213	2	9.604	**4.802**
37	p424	1203	p56	3	640	**4.800**
38	p505	p14	1228	3	628	**4.710**
39	p5229	p4371	p288	48	39	**4.680**
40	p37	p8	1110	3	9.079	**4.562**
41	p26	p1	1119	3	17.695	**4.512**
42	p658	p14	1151	3	574	**4.305**
43	p46	1224	1119	2	8.502	**4.251**
44	p38	p2	1248	3	8.365	**4.203**
45	p56	1230	1151	2	8.405	**4.202**
46	p616	1233	p116	4	416	**4.160**
47	p581	1110	p31	3	548	**4.110**
48	p35	p10	1237	4	8.072	**4.036**
49	p47	1230	1103	2	8.041	**4.020**
50	p504	p14	1166	3	534	**4.005**
51	p39	1151	1108	2	7.860	**3.930**
52	p22	p1	1237	3	15.380	**3.921**
53	p25	p1	1248	3	14.989	**3.822**
54	p4857	p4854	p4854	8	190	**3.800**
55	p587	p122	1108	3	503	**3.772**
56	p75	1110	1248	2	7.100	**3.550**

57	p44	1224	1103	2	6.886	**3.443**
58	p1640	1233	p105	4	344	**3.440**
59	p306	p102	1123	3	458	**3.435**
60	p653	p31	1286	3	452	**3.390**
61	p2	1144	1110	2	135.563	**3.389**
62	p1106	1103	p17	3	447	**3.352**
63	p48	1158	1237	2	6.559	**3.279**
64	p533	p227	1151	3	436	**3.270**
65	p853	p39	1213	3	431	**3.232**
66	p3807	p46	1237	3	425	**3.187**
67	p51	p6	1119	3	6.292	**3.161**
68	p111	1119	1213	2	6.121	**3.060**
69	p486	p47	1236	3	406	**3.045**
70	p63	1110	p1	3	11.867	**3.026**
71	p802	1228	p248	3	402	**3.015**
72	p484	1203	p39	3	400	**3.000**
73	p494	p14	1221	3	391	**2.932**
74	p33	1224	p3	4	11.412	**2.853**
75	p651	p46	1248	3	379	**2.842**
76	p112	1103	1213	2	5.494	**2.747**
77	p3883	p56	1108	3	360	**2.700**
78	p61	p2	1119	3	5.336	**2.681**
79	p493	p14	p62	4	266	**2.660**
80	p190	1221	p95	3	351	**2.632**
81	p1003	p44	p17	4	260	**2.600**
82	p4880	p4857	p4857	16	65	**2.600**
83	p912	1213	p44	3	344	**2.580**
84	p667	1300	p18	3	343	**2.572**
85	p544	p39	1229	3	343	**2.572**
86	p60	p2	p2	4	5.107	**2.553**
87	p30	p1	p2	4	10.099	**2.524**
88	p62	1213	1223	2	5.027	**2.513**
89	p925	p31	1213	3	334	**2.505**
90	p1849	p16	1123	3	330	**2.475**
91	p3903	p56	p81	4	247	**2.470**
92	p71	1108	1229	2	4.923	**2.461**
93	p79	p8	1237	3	4.869	**2.446**
94	p77	p2	1286	3	4.801	**2.412**
95	p548	p14	1248	3	318	**2.385**
96	p34	p1	1108	3	9.183	**2.341**
97	p401	p18	1248	3	306	**2.295**
98	p115	1108	1221	2	4.392	**2.196**
99	p139	1230	p15	3	4.316	**2.168**
100	p2925	1218	p48	3	289	**2.167**

Figura 175. Patrones con mayor ponderación del experimento 7

El patrón con la mayor ponderación es p105, cuya secuencia es la siguiente:

- **p105**: “p14 + 1213” = “1230 + 1108 + 1213” = “VERB TO BE + VERB + PREPOSITION”

El patrón tiene longitud 3 y se repitió 8.748 veces en el texto, por lo que ocupa la posición 39º en la lista de patrones más frecuentes en este escenario de estudio.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p653	p31	1286	3	452	**3.390**
2	p667	1300	p18	3	343	**2.572**
3	p77	p2	1286	3	4.801	**2.412**
4	p92	1411	1213	2	3.574	**1.787**
5	p150	1300	1123	2	3.195	**1.597**
6	p3844	p150	1248	3	203	**1.522**
7	p123	p6	1309	3	2.582	**1.297**
8	p685	1300	p102	3	165	**1.237**
9	p1206	p150	1237	3	155	**1.162**
10	p110	p45	1110	3	2.042	**1.026**
11	p3871	p150	p266	4	95	**950**
12	p3817	1224	1293	2	1.897	**948**
13	p213	p212	p113	18	327	**912**
14	p184	1286	1110	2	1.779	**889**
15	p4206	1411	p497	3	116	**870**
16	p4192	p112	1293	3	115	**862**
17	p3983	1237	p3817	3	110	**825**
18	p251	1110	p77	4	1.097	**822**
19	p1012	p18	1286	3	100	**750**
20	p1183	1300	p306	4	73	**730**
21	p2654	p85	1286	3	74	**555**
22	p113	p35	p101	6	730	**547**
23	p914	1223	p92	3	73	**547**
24	p417	1103	1286	2	964	**482**
25	p387	p4	1286	4	925	**462**
26	p334	1224	1411	2	915	**457**
27	p441	1110	1286	2	907	**453**

28	p1213	p92	1224	3	60	**450**
29	p3836	p92	1300	3	58	**435**
30	p539	1286	1151	2	845	**422**

Figura 176. Patrones del dominio con mayor ponderación del experimento 7

El patrón del dominio con la mayor ponderación es p653, cuya secuencia es la siguiente:

- **p653**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 452 veces en el texto, por lo que ocupa la posición 31º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

8.8. Experimento 8: frecuencia = 20, con semántica

8.8.1 Estudio de patrones generados

Se han generado un total de 2.775 patrones en este escenario. A continuación se adjuntan los 100 patrones que más se han repetido en el texto.

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p1	1144	1144	2	913.038
2	p2	1144	1110	2	135.563
3	p3	p1	1144	3	114.605
4	p4	p1	1110	3	93.593
5	p5	p1	p1	4	90.963
6	p6	1224	1144	2	69.738
7	p8	1144	1119	2	53.485
8	p15	1144	1213	2	49.374
9	p7	p2	1144	3	45.765
10	p9	1123	1144	2	43.189
11	p10	1224	p1	3	38.738
12	p11	p4	p4	6	32.092
13	p12	1144	1237	2	29.788
14	p13	1144	1248	2	28.730
15	p14	1230	1108	2	27.125
16	p53	1110	1144	2	26.738
17	p36	p1	1213	3	23.275
18	p17	1119	1237	2	23.083
19	p21	p6	1237	3	19.231
20	p20	1144	1108	2	18.459
21	p26	p1	1119	3	17.695
22	p18	1123	1110	2	17.379
23	p19	1144	1230	2	16.449
24	p27	1144	1151	2	16.007
25	p22	p1	1237	3	15.380
26	p25	p1	1248	3	14.989
27	p28	1144	1123	2	14.885
28	p24	1110	p7	4	14.727
29	p16	1123	1123	2	14.002
30	p29	1144	1229	2	12.287
31	p31	1151	1110	2	11.920
32	p63	1110	p1	3	11.867
33	p33	1224	p3	4	11.412
34	p23	p1	p3	5	11.411
35	p30	p1	p2	4	10.099

36	p81	1108	1213	2	9.542
37	p34	p1	1108	3	9.156
38	p37	p8	1110	3	9.079
39	p105	p14	1213	3	8.724
40	p46	1224	1119	2	8.491
41	p56	1230	1151	2	8.405
42	p38	p2	1248	3	8.365
43	p42	p1	1123	3	8.230
44	p32	p1	1230	3	8.175
45	p40	1144	1228	2	8.110
46	p35	p10	1237	4	8.072
47	p73	1213	1144	2	8.015
48	p47	1230	1103	2	7.985
49	p91	p3	1213	4	7.937
50	p39	1151	1108	2	7.740
51	p75	1110	1248	2	7.100
52	p50	1224	p17	3	6.886
53	p44	1224	1103	2	6.782
54	p59	p1	1151	3	6.760
55	p43	1144	1197	2	6.719
56	p66	1123	p1	3	6.566
57	p48	1158	1237	2	6.559
58	p65	p3	1119	4	6.532
59	p54	p11	p11	12	6.517
60	p57	1224	p5	5	6.342
61	p45	1144	1286	2	6.288
62	p82	p4	1144	4	6.233
63	p52	p2	p1	4	6.198
64	p51	p6	1119	3	6.181
65	p55	1223	1144	2	6.135
66	p111	1119	1213	2	5.934
67	p96	1119	1144	2	5.707
68	p78	1144	1158	2	5.701
69	p161	1230	1144	2	5.469
70	p112	1103	1213	2	5.465
71	p70	p3	1123	4	5.379
72	p61	p2	1119	3	5.336
73	p90	1151	1144	2	5.202
74	p72	p3	1248	4	5.157
75	p69	1223	p1	3	5.114
76	p60	p2	p2	4	5.107
77	p49	1144	1236	2	5.090
78	p68	1144	p9	3	5.063
79	p62	1213	1223	2	4.963
80	p76	p5	p1	6	4.953

81	p71	1108	1229	2	4.923
82	p79	p8	1237	3	4.869
83	p67	1144	1103	2	4.856
84	p77	p2	1286	3	4.801
85	p2332	1144	p53	3	4.741
86	p109	p4	p1	5	4.546
87	p115	1108	1221	2	4.392
88	p139	1230	p15	3	4.277
89	p85	1119	1110	2	4.198
90	p103	p5	p5	8	4.175
91	p84	p6	1248	3	4.137
92	p41	p16	p16	4	4.092
93	p87	1223	1103	2	4.059
94	p74	1221	1144	2	4.006
95	p83	p3	1237	4	3.884
96	p2333	p53	p53	4	3.732
97	p128	1103	1144	2	3.727
98	p97	1108	1144	2	3.696
99	p114	p14	1236	3	3.571
100	p89	p4	1248	4	3.568

Figura 177. Patrones más repetidos del experimento 8

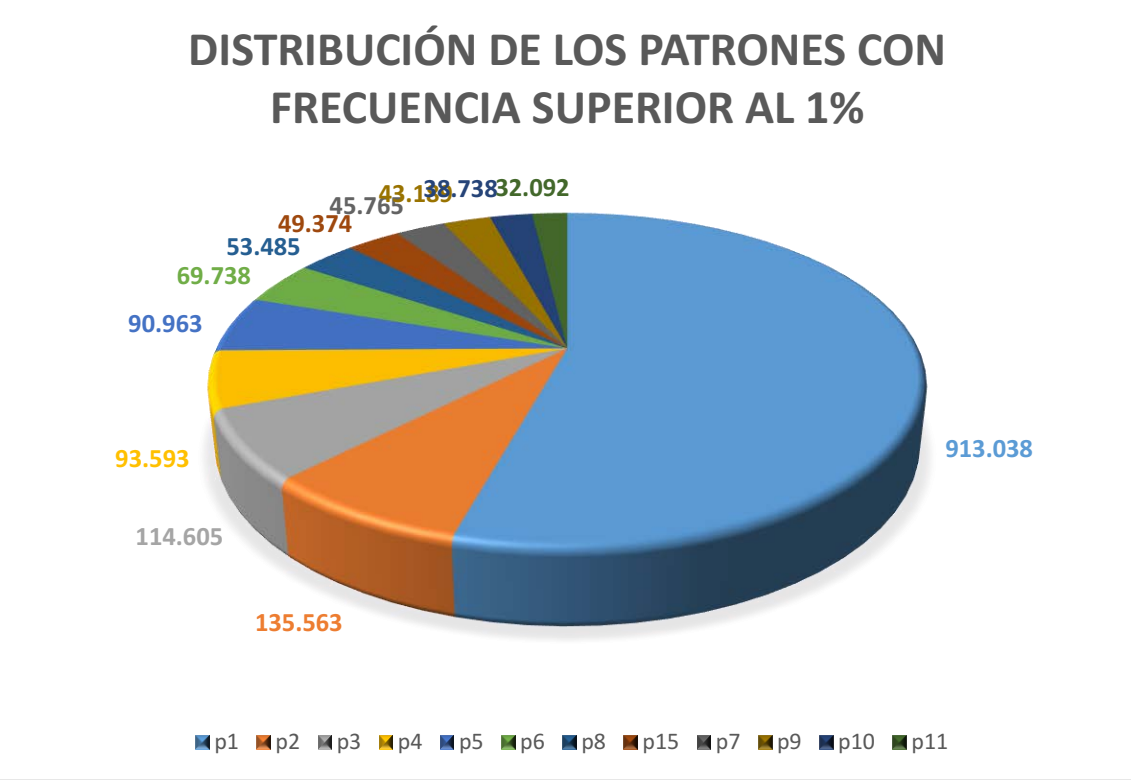

Figura 178. Gráfico de los patrones más frecuentes del experimento 8

8.8.2 Estudio de los patrones del dominio

De los 2.775 patrones generados en este escenario, 310 son patrones simples o compuestos que contienen al menos un concepto insertado en la base de datos. Por tanto, el 12,57% de los patrones pertenecen al dominio de estudio.

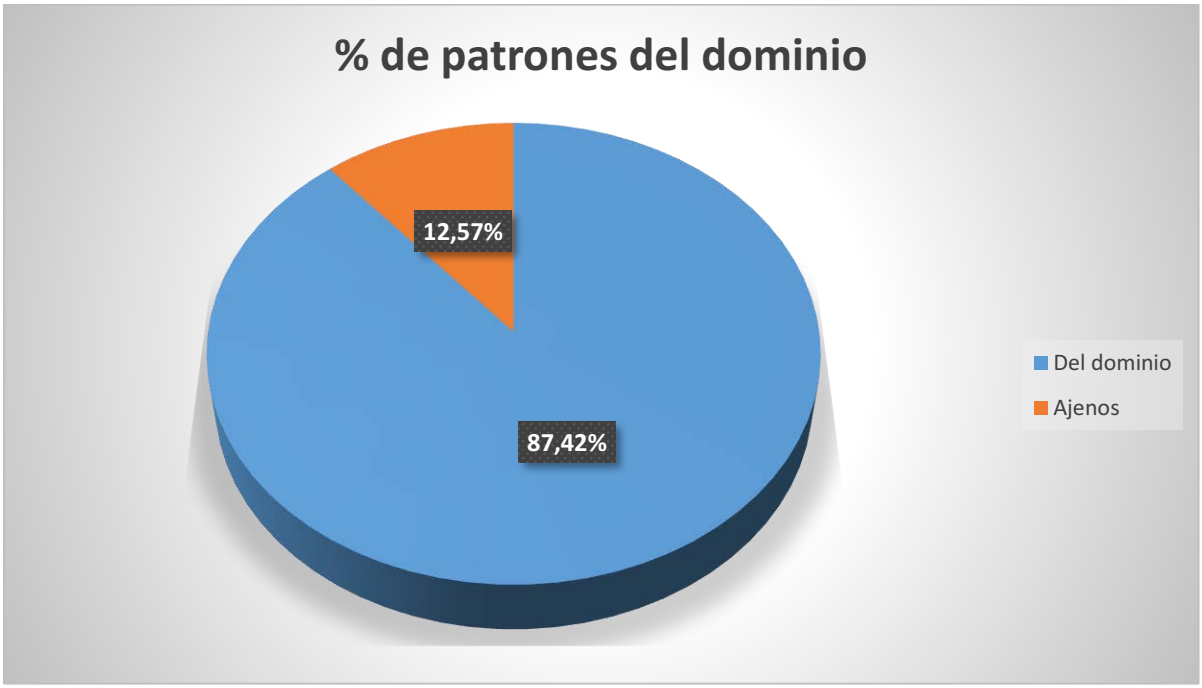

Figura 179. Proporción de patrones en el experimento 8

A continuación se adjunta una tabla con los 100 patrones del dominio más frecuentes:

Posición	Patrón	Izquierda	Derecha	Longitud	Repeticiones
1	p45	1144	1286	2	6.288
2	p77	p2	1286	3	4.801
3	p92	1411	1213	2	3.469
4	p150	1300	1123	2	3.195
5	p123	p6	1309	3	2.582
6	p189	1144	1411	2	2.279
7	p110	p45	1110	3	2.020
8	p281	1144	1309	2	2.000
9	p2099	1224	1293	2	1.897
10	p184	1286	1110	2	1.779
11	p101	1291	1144	2	1.125
12	p402	p1	1286	3	1.113

13	p251	1110	p77	4	1.077
14	p456	p3	1286	4	974
15	p387	p4	1286	4	925
16	p334	1224	1411	2	915
17	p441	1110	1286	2	907
18	p417	1103	1286	2	854
19	p539	1286	1151	2	845
20	p318	1224	1309	2	809
21	p113	p35	p101	6	730
22	p2232	1144	1288	2	707
23	p474	p6	1411	3	684
24	p2124	1293	1144	2	652
25	p315	1286	1144	2	643
26	p668	1300	p9	3	630
27	p835	1309	1144	2	555
28	p293	1144	1564	2	524
29	p2111	p82	p184	6	516
30	p454	1411	1237	2	442
31	p653	p31	1286	3	437
32	p2264	1144	1293	2	421
33	p560	1221	1411	2	395
34	p1105	p1	1309	3	382
35	p2077	1288	1144	2	341
36	p2354	p2338	1286	4	337
37	p834	1300	1144	2	317
38	p667	1300	p18	3	308
39	p2081	1363	1123	2	294
40	p2515	p2220	1286	4	281
41	p1803	1123	1309	2	279
42	p360	p92	p123	5	278
43	p800	1224	1300	2	277
44	p2133	1293	1411	2	272
45	p2161	p15	1293	3	269
46	p2344	p53	1286	3	262
47	p1823	1224	1294	2	261
48	p773	1309	1110	2	257
49	p2146	1286	p1	3	256
50	p2135	p1	1411	3	235
51	p2424	p9	1293	3	234
52	p734	1345	1228	2	233
53	p1350	p5	1286	5	211
54	p1271	1291	p1	3	205
55	p2299	1224	1269	2	199
56	p1591	p6	1308	3	198
57	p916	1224	1558	2	196

58	p2184	1293	p1	3	184
59	p2346	p150	p3	5	181
60	p2305	p82	1286	5	174
61	p725	p6	1526	3	172
62	p874	1224	1291	2	169
63	p641	1144	1553	2	165
64	p2156	p2087	1286	6	162
65	p1995	p10	1309	4	161
66	p611	p30	1286	5	153
67	p2126	p150	1248	3	150
68	p1804	1144	p92	3	147
69	p479	1564	1213	2	138
70	p917	1224	1564	2	135
71	p999	p55	1309	3	133
72	p2242	1411	p73	3	130
73	p685	1300	p102	3	129
74	p769	p13	1286	3	129
75	p1184	1309	1213	2	126
76	p1182	1248	1286	2	125
77	p2664	1144	1328	2	124
78	p2082	1411	1144	2	121
79	p2665	p2664	p66	5	120
80	p2699	1123	1286	2	119
81	p1725	p75	p77	5	117
82	p2270	p112	1293	3	115
83	p2238	p1	1288	3	115
84	p1009	p25	1286	4	114
85	p2245	p76	1286	7	112
86	p2561	1411	1248	2	108
87	p2292	1411	p497	3	103
88	p1206	p150	1237	3	102
89	p2120	1294	1411	2	102
90	p2625	1144	1294	2	100
91	p915	1224	1526	2	99
92	p2130	1294	1144	2	99
93	p2460	1293	1248	2	99
94	p1041	1300	p1	3	99
95	p1379	1309	1119	2	98
96	p726	p4	p45	5	94
97	p1599	p3	1309	4	93
98	p2529	1144	1410	2	92
99	p2275	p6	1288	3	92
100	p2550	p154	p1803	5	92

Figura 180. Patrones del domino más repetidos del experimento 8

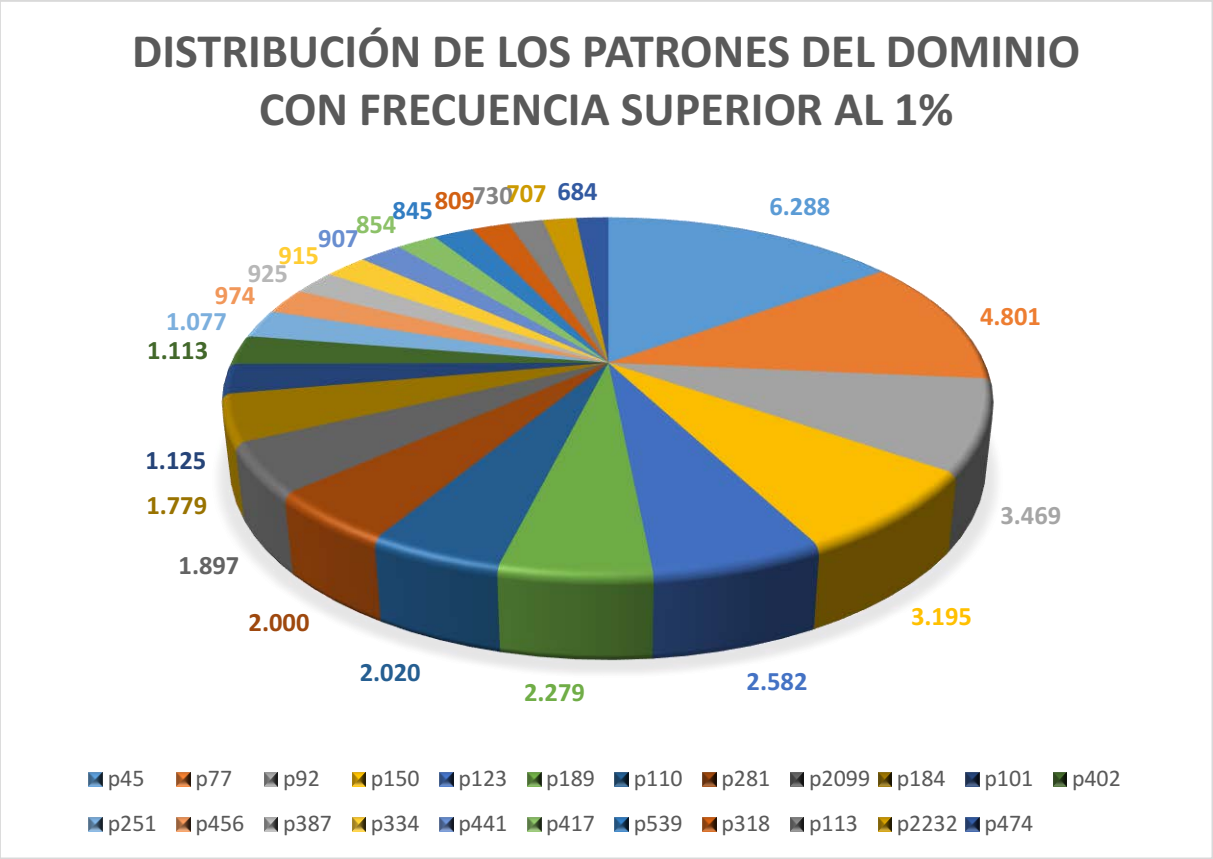

Figura 181. Gráfico de patrones del dominio en el experimento 8

8.8.3 Estudio de categorías gramaticales

De las 371 categorías gramaticales que ha empleado la herramienta, 89 han aparecido al menos una vez en el conjunto de todos los patrones. A continuación se adjuntarán las 10 categorías más frecuentes:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1144	UNCLASSIFIED NOUN	216
2	1123	NUMBER	138
3	1213	PREPOSITION	126
4	1151	ADVERB	117
5	1119	NOUN	115
6	1108	VERB	110
7	1110	SYMBOL	96
8	1248	AND LINKING	82
9	1224	DEFINITE ARTICLE	80
10	1103	ADJECTIVE	72

Figura 182. Categorías más frecuentes en el experimento 8

Limitando los resultados a sólo los conceptos asociados a la genética, las 10 categorías más frecuentes son:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1286	Genes, vif	48
2	1411	Mutation	27
3	1309	Genes	22
4	1300	Exons	21
5	1293	Genes, Neurofibromatosis 1	19
6	1288	Genes, Tumor Suppressor	9
7	1291	Genes, Wilms Tumor	8
8	1294	Genes, Neurofibromatosis 2	6
9	1526	Phenotype	6
10	1312	Genes, Dominant	6

Figura 183. Categorías de sordera más frecuentes en el experimento 8

Limitando los resultados a sólo los conceptos asociados a la sordera, las categorías empleadas han sido las siguientes, ordenadas por número de repeticiones:

Posición	Elemento	Nombre Elemento	Repeticiones
1	1564	Hearing Loss	12
2	1553	Deafness	8
3	1561	Hearing Loss, Sensorineural	7
4	1563	Wolfram Syndrome	2
5	1559	Hearing Loss, High-Frequency	2
6	1556	Hearing Loss, Central	2
7	1558	Hearing Loss, Functional	1

Figura 184. Categorías de sordera más frecuentes en el experimento 8

Al sumar todas las repeticiones de las categorías gramaticales del mismo grupo (ajenos, perteneciente a “genética”, perteneciente a “sordera”), la distribución ha quedado de la siguiente forma:

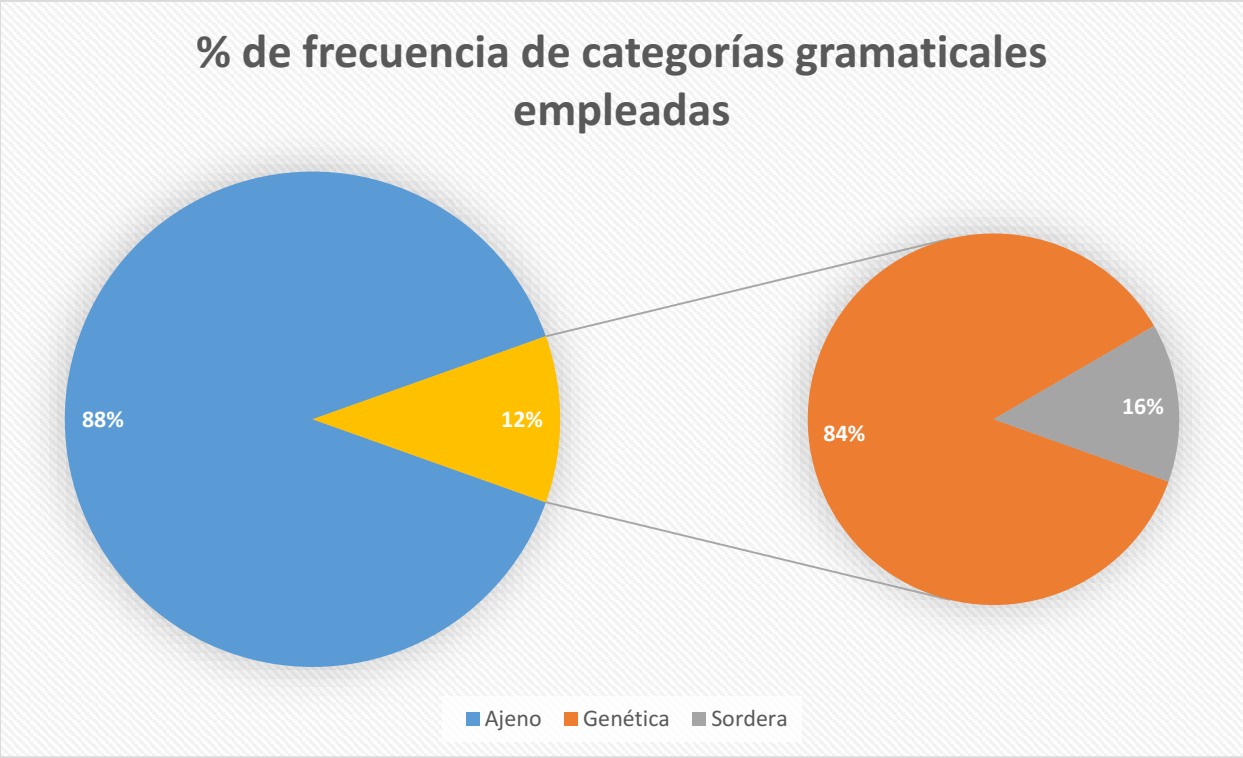

Figura 185. Proporción de categorías gramaticales en el experimento 8

Aproximadamente el 88% de elementos que forman los patrones son ajenos a nuestro dominio. El 12% forman parte del dominio, de los cuales el 84% están asociados a la "Genética" y el 16% restante a la "Sordera".

8.8.4 Estudio de los patrones con ponderación

Tras aplicar la fórmula de ponderación en los resultados del análisis en este escenario de estudio, los 100 patrones más valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p105	p14	1213	3	8.724	**65.430**
2	p50	1224	p17	3	6.886	**51.645**
3	p41	p16	p16	4	4.092	**40.920**
4	p104	p41	p41	8	1.569	**31.380**
5	p114	p14	1236	3	3.571	**26.782**
6	p116	p14	1229	3	3.462	**25.965**
7	p4	p1	1110	3	93.593	**23.866**
8	p288	p104	p104	16	531	**21.240**
9	p244	p47	1213	3	2.294	**17.205**
10	p160	p18	1123	3	2.211	**16.582**
11	p11	p4	p4	6	32.092	**16.366**

12	p14	1230	1108	2	27.125	**13.562**
13	p167	1224	p48	3	1.764	**13.230**
14	p187	p18	p18	4	1.245	**12.450**
15	p7	p2	1144	3	45.765	**11.670**
16	p17	1119	1237	2	23.083	**11.541**
17	p10	1224	p1	3	38.738	**9.878**
18	p21	p6	1237	3	19.231	**9.663**
19	p18	1123	1110	2	17.379	**8.689**
20	p287	p47	1229	3	1.136	**8.520**
21	p24	1110	p7	4	14.727	**7.363**
22	p342	p288	1123	17	169	**7.182**
23	p16	1123	1123	2	14.002	**7.001**
24	p2695	p288	p288	32	86	**6.880**
25	p54	p11	p11	12	6.517	**6.647**
26	p366	1223	p17	3	838	**6.285**
27	p259	p31	1203	3	825	**6.187**
28	p31	1151	1110	2	11.920	**5.960**
29	p36	p1	1213	3	23.275	**5.935**
30	p319	1233	p14	3	770	**5.775**
31	p2704	p2702	p2700	60	36	**5.400**
32	p481	p47	1228	3	688	**5.160**
33	p2545	p2544	p2544	4	488	**4.880**
34	p234	1230	p39	3	642	**4.815**
35	p424	1203	p56	3	640	**4.800**
36	p81	1108	1213	2	9.542	**4.771**
37	p505	p14	1228	3	628	**4.710**
38	p2702	p2695	p288	48	39	**4.680**
39	p37	p8	1110	3	9.079	**4.562**
40	p26	p1	1119	3	17.695	**4.512**
41	p658	p14	1151	3	574	**4.305**
42	p46	1224	1119	2	8.491	**4.245**
43	p38	p2	1248	3	8.365	**4.203**
44	p56	1230	1151	2	8.405	**4.202**
45	p35	p10	1237	4	8.072	**4.036**
46	p47	1230	1103	2	7.985	**3.992**
47	p581	1110	p31	3	528	**3.960**
48	p22	p1	1237	3	15.380	**3.921**
49	p39	1151	1108	2	7.740	**3.870**
50	p504	p14	1166	3	515	**3.862**
51	p25	p1	1248	3	14.989	**3.822**
52	p2548	p2545	p2545	8	190	**3.800**
53	p616	1233	p116	4	365	**3.650**
54	p75	1110	1248	2	7.100	**3.550**
55	p44	1224	1103	2	6.782	**3.391**
56	p2	1144	1110	2	135.563	**3.389**

57	p48	1158	1237	2	6.559	**3.279**
58	p653	p31	1286	3	437	**3.277**
59	p1106	1103	p17	3	436	**3.270**
60	p2089	p46	1237	3	425	**3.187**
61	p51	p6	1119	3	6.181	**3.105**
62	p306	p102	1123	3	405	**3.037**
63	p63	1110	p1	3	11.867	**3.026**
64	p533	p227	1151	3	398	**2.985**
65	p111	1119	1213	2	5.934	**2.967**
66	p33	1224	p3	4	11.412	**2.853**
67	p802	1228	p248	3	374	**2.805**
68	p112	1103	1213	2	5.465	**2.732**
69	p2173	p56	1108	3	360	**2.700**
70	p486	p47	1236	3	360	**2.700**
71	p61	p2	1119	3	5.336	**2.681**
72	p493	p14	p62	4	266	**2.660**
73	p190	1221	p95	3	351	**2.632**
74	p494	p14	1221	3	343	**2.572**
75	p60	p2	p2	4	5.107	**2.553**
76	p30	p1	p2	4	10.099	**2.524**
77	p62	1213	1223	2	4.963	**2.481**
78	p71	1108	1229	2	4.923	**2.461**
79	p79	p8	1237	3	4.869	**2.446**
80	p77	p2	1286	3	4.801	**2.412**
81	p587	p122	1108	3	320	**2.400**
82	p484	1203	p39	3	317	**2.377**
83	p34	p1	1108	3	9.156	**2.334**
84	p667	1300	p18	3	308	**2.310**
85	p1849	p16	1123	3	303	**2.272**
86	p651	p46	1248	3	301	**2.257**
87	p544	p39	1229	3	293	**2.197**
88	p115	1108	1221	2	4.392	**2.196**
89	p853	p39	1213	3	289	**2.167**
90	p139	1230	p15	3	4.277	**2.149**
91	p85	1119	1110	2	4.198	**2.099**
92	p42	p1	1123	3	8.230	**2.098**
93	p32	p1	1230	3	8.175	**2.084**
94	p84	p6	1248	3	4.137	**2.078**
95	p87	1223	1103	2	4.059	**2.029**
96	p1108	p288	p41	20	40	**2.000**
97	p91	p3	1213	4	7.937	**1.984**
98	p2758	p2548	p2548	16	48	**1.920**
99	p2333	p53	p53	4	3.732	**1.866**
100	p912	1213	p44	3	238	**1.785**

Figura 186. Patrones con mayor ponderación del experimento 8

El patrón con la mayor ponderación es p105, cuya secuencia es la siguiente:

- **p105**: “p14 + 1213” = “1230 + 1108 + 1213” = “VERB TO BE + VERB + PREPOSITION”

El patrón tiene longitud 3 y se repitió 8.724 veces en el texto, por lo que ocupa la posición 39º en la lista de patrones más frecuentes en este escenario de estudio.

Al filtrar los resultados de forma que sólo se obtengan patrones que contengan al menos un concepto del dominio, los 30 patrones mejor valuados son los siguientes:

Pos.	Patrón	Izquierda	Derecha	Longitud	Repeticiones	Ponderación
1	p653	p31	1286	3	452	**3.390**
2	p667	1300	p18	3	343	**2.572**
3	p77	p2	1286	3	4.801	**2.412**
4	p92	1411	1213	2	3.574	**1.787**
5	p150	1300	1123	2	3.195	**1.597**
6	p3844	p150	1248	3	203	**1.522**
7	p123	p6	1309	3	2.582	**1.297**
8	p685	1300	p102	3	165	**1.237**
9	p1206	p150	1237	3	155	**1.162**
10	p110	p45	1110	3	2.042	**1.026**
11	p3871	p150	p266	4	95	**950**
12	p3817	1224	1293	2	1.897	**948**
13	p213	p212	p113	18	327	**912**
14	p184	1286	1110	2	1.779	**889**
15	p4206	1411	p497	3	116	**870**
16	p4192	p112	1293	3	115	**862**
17	p3983	1237	p3817	3	110	**825**
18	p251	1110	p77	4	1.097	**822**
19	p1012	p18	1286	3	100	**750**
20	p1183	1300	p306	4	73	**730**
21	p2654	p85	1286	3	74	**555**
22	p113	p35	p101	6	730	**547**
23	p914	1223	p92	3	73	**547**
24	p417	1103	1286	2	964	**482**
25	p387	p4	1286	4	925	**462**
26	p334	1224	1411	2	915	**457**
27	p441	1110	1286	2	907	**453**

28	p1213	p92	1224	3	60	**450**
29	p3836	p92	1300	3	58	**435**
30	p539	1286	1151	2	845	**422**

Figura 187. Patrones del dominio con mayor ponderación del experimento 8

El patrón del dominio con la mayor ponderación es p653, cuya secuencia es la siguiente:

- **p653**: “p31 + 1286” = “1151 + 1110 + 1286” = “ADVERB + SYMBOL + Genes, vif”.

El patrón tiene longitud 3 y se repitió 437 veces en el texto, por lo que ocupa la posición 31º en la lista de patrones del dominio más frecuentes en este escenario de estudio.

9. Conclusiones

En esta sección se comentarán los resultados obtenidos tras completar la experimentación con los ocho escenarios planteados.

9.1. Suma total de todos los patrones

Durante la extracción de los resultados por cada escenario, hemos podido comprobar que existe una tendencia en el número total de patrones generados tras completar cada estudio.

Experimento	Configuración	Número de patrones
Experimento 1	Frec. = 1, NO sem	140.909
Experimento 2	Frec. = 5, NO sem	14.891
Experimento 3	Frec. = 10, NO sem	5.707
Experimento 4	Frec. = 20, NO sem	2.847
Experimento 5	Frec. = 1, SÍ sem	142.064
Experimento 6	Frec. = 5, SÍ sem	14.382
Experimento 7	Frec. = 10, SÍ sem	5.505
Experimento 8	Frec. = 20, SÍ sem	2.775

Figura 188. Suma total de patrones en cada escenario

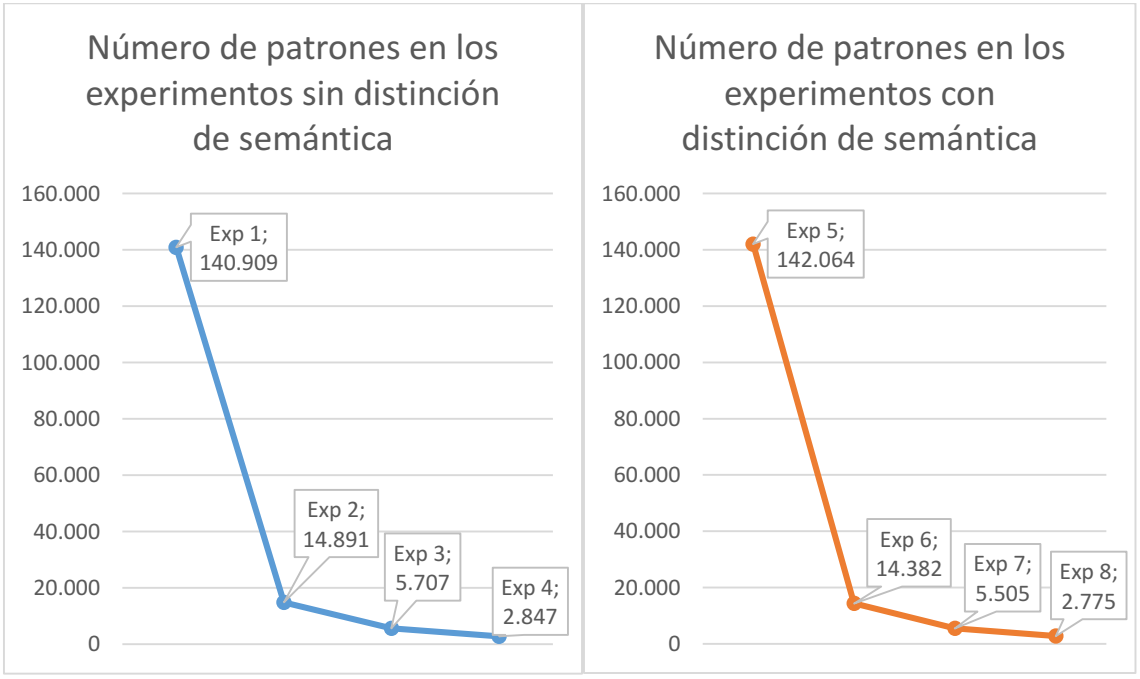

Figura 189. Gráficos de suma total de patrones

Como era de esperar, la frecuencia mínima para la generación de un patrón ha influido en el número total de patrones generados. A partir de las gráficas extraídas, se puede concluir que la frecuencia mínima es inversamente proporcional al número de patrones.

Este hecho tiene una explicación sencilla: al aumentar la frecuencia mínima, se obliga a todas las posibles secuencias lingüísticas a tener un mayor número de repeticiones para que se consideren un patrón. Por tanto, en el caso más relajado (frecuencia mínima a 1) hay más posibilidades de formar patrones. En el caso más estricto (frecuencia mínima a 20) es más difícil llegar a dicho número de repeticiones, por lo que el número total se reduce considerablemente.

Utilizando los valores 1, 5, 10 y 20, se puede apreciar una tendencia exponencial en el número de patrones generados. Los experimentos 1 y 2 superan en 10 veces a los experimentos 5 y 6, respectivamente. Sin embargo, en el resto de experimentos, la tendencia empieza a estabilizarse.

Por último, cabe destacar que la diferenciación por semántica ha influido poco en el número de patrones. Comparando los experimentos 1 y 5 (frecuencia mínima a 1), se puede apreciar que el uso de la semántica ha permitido generar aproximadamente un 0,8% más de patrones. Sin embargo, en el resto de parejas de experimentos, se observa una disminución 3% de media al aplicar semántica.

Es posible que, en los casos más relajados, la semántica permita a la herramienta detectar patrones que de otro modo no sería capaz de reconocer, aumentando el espacio de resultados. Sin embargo, en los experimentos más estrictos, es más difícil encontrar secuencias con una frecuencia mínima de aparición ya que además de las categorías gramaticales la semántica debe coincidir.

9.2. Patrones más frecuentes

En los ocho escenarios, se extrajeron los 100 patrones más frecuentes junto con su número de repeticiones. Una observación rápida permite verificar que los patrones más frecuentes son siempre los mismos, con sólo una ligera diferencia de repeticiones a partir del 70º patrón.

Un número considerable de los patrones del top 100 emplean la categoría gramatical 1144 o referencian a un patrón con dicha categoría. Como recordatorio, el identificador 1144 corresponde a “Unclassified Noun” (sustantivo sin clasificar), una categoría que representa a todos los términos que no aparecen en la ontología (es decir, en la tabla Vocabulary). Ya que sólo se incluyeron los términos relacionados con la sordera genética, era de esperar que la mayoría de patrones generados tengan que ver con sustantivos que la herramienta no reconoce.

El patrón más frecuente en todos los experimentos es el primero en ser creado (p1), con la secuencia “1144 + 1144” (UN. NOUN + UN. NOUN). Asimismo, algunos de los patrones más frecuentes referencian a p1, como, por ejemplo:

- p4: “p1 + 1144” = “1144 + 1144 + 1144” = “UN. NOUN + UN. NOUN + UN. NOUN”.
- p4: “p1 + 1110” = “1144 + 1144 + 1110” = “UN. NOUN + UN. NOUN + SYMBOL”.
- p5: “p1 + p1” = “1144 + 1144 + 1144 + 1144” = “UN. NOUN + UN. NOUN + UN. NOUN + UN. NOUN”.

Algunos ejemplos de patrones frecuentes que no emplean la categoría 1144 son:

- “1230 + 1108” = “VERB TO BE + VERB”
- “1123 + 1110” = “NUMBER + SYMBOL”
- “1119 + 1237” = “NOUN + PREPOSITION OF”
- “1230 + 1151” = “VERB TO BE + ADVERB”
- “1108 + 1213” = “VERB + PREPOSITION”

También cabe mencionar que en los 100 patrones más frecuentes no se emplea ninguna categoría gramatical que corresponda al dominio de estudio.

9.3. Número de patrones correspondientes al dominio

Una vez obtenidos los patrones tras cada estudio, se filtró la búsqueda para generar un subconjunto de patrones cuyos elementos contengan categorías asociadas al dominio de estudio o referencien a patrones que sí lo hagan.

Este subconjunto es una minoría comparado con los resultados totales:

Experimento	Configuración	Número de patrones del dominio	% de patrones del dominio
Experimento 1	Frec. = 1, NO sem	20.666	14,66%
Experimento 2	Frec. = 5, NO sem	2.121	14,24%
Experimento 3	Frec. = 10, NO sem	720	12,61%
Experimento 4	Frec. = 20, NO sem	300	10,53%
Experimento 5	Frec. = 1, SÍ sem	19.933	14,04%
Experimento 6	Frec. = 5, SÍ sem	2.115	14,71%
Experimento 7	Frec. = 10, SÍ sem	729	15,26%
Experimento 8	Frec. = 20, SÍ sem	310	12,57%

Figura 190. Suma de patrones del domino en cada escenario

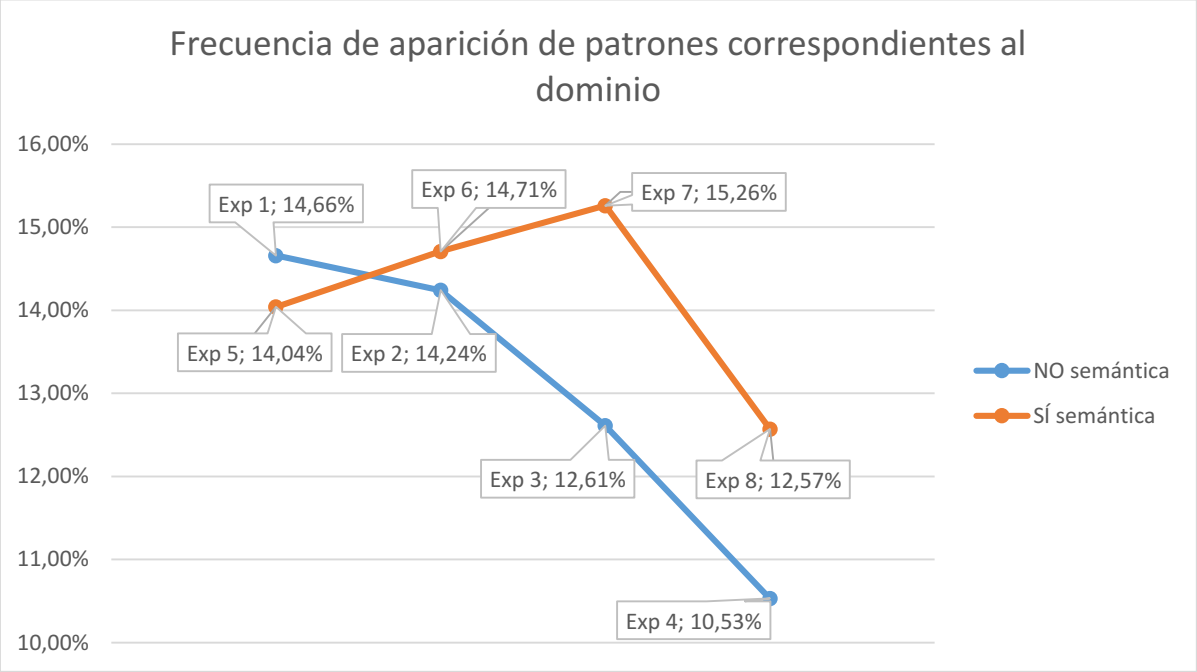

Figura 191. Gráfico de suma de patrones del dominio

De media, entre 10% y el 16% de los patrones generados en cada escenario tienen al menos una categoría gramatical del dominio en su secuencia. Si bien parece que una frecuencia mínima mayor reduce el número de patrones del dominio, no se dispone de suficiente información para reconocer una tendencia.

Otra particularidad de los resultados es que la diferencia por considerar la semántica es muy pequeña, o prácticamente despreciable en los casos más estrictos. Por ejemplo, entre los experimentos 3 y 7 (con frecuencia mínima 20) sólo tiene una diferencia de 9 patrones generados. Esto también sucede si consideramos todos los patrones del experimento.

9.4. Patrones del dominio más frecuentes

Utilizando un procedimiento similar al estudio con todos los patrones, por cada escenario se extrajo los 100 patrones más frecuentes que guardan una relación con el dominio.

De nuevo, se puede observar que, con sólo algunas diferencias en el número de repeticiones, los patrones más frecuentes se repiten en todos los escenarios. Se pueden empezar a percibir diferencias a partir del 50º patrón.

Además, la categoría gramatical 1144 (Unclassified Noun, “Sustantivo sin clasificar”) vuelve a ser muy recurrente en los patrones del dominio más comunes. En este caso la categoría aparece directamente en uno de los elementos del patrón o como referencia a otros patrones ajenos al dominio. Por ejemplo:

- p46: “1144 + 1286” = “UN. NOUN + Genes, vif”
- p75: “p2 + 1286” = “1144 + 1110 + 1286” = “UN. NOUN + SYMBOL + Genes, vif”.
- p23: “1300 + p9” = “1300 + 1123 + 1144” = “Exons + NUMBER + UN. NOUN

Otra particularidad detectada es que la categoría gramatical 1286 (Genes, vif) se repite con mucha frecuencia como elemento situado a la derecha de los patrones. Esta categoría suele aparecer al final de varios patrones de longitudes 2 a 6. Por ejemplo:

- p460: “p3 + 1286” = “p1 + 1144 + 1286” = “1144 + 1144 +1144 + 1286” = “UN. NOUN + UN. NOUN + UN. NOUN + Genes, vif”.

Además, otro tipo de patrones frecuentes son secuencias de longitud 2 en donde se combina una categoría del dominio y una categoría ajena. Por ejemplo:

- “1224 + 1411” = “DEFINITE ARTICLE + Mutation”
- “1123 + 1309” = “NUMBER + Genes”
- “1363 + 1123” = “Introns + NUMBER”

- “1286 + 1151” = “Genes, vif + ADVERB”
- “1345 + 1228” = “Genes, Essential + PREPOSITION FOR”

9.5. Categorías gramaticales más frecuentes

Por cada escenario, se contaron las veces que se repite cada categoría gramatical en el conjunto completo de patrones de cada escenario y se compilaron en ficheros aparte para su estudio.

A continuación, se adjunta la tabla de categorías ajenas al dominio que han aparecido al menos una vez en el subconjunto de 10 categorías más frecuentes en cada escenario:

Identificador	Nombre Categoría
1218	QUANTIFIER DETERMINER
1213	PREPOSITION
1151	ADVERB
1119	NOUN
1123	NUMBER
1108	VERB
1144	UNCLASSIFIED NOUN
1103	ADJECTIVE
1110	SYMBOL
1248	AND LINKING
1224	DEFINITE ARTICLE

Figura 192. Categorías gramaticales más frecuentes en el análisis

Las categorías que han aparecido en el subconjunto de 10 categorías más frecuentes relacionados con la “Genética” son las siguientes:

Identificador	Nombre Categoría
1286	Genes, vif
1411	Mutation
1300	Exons
1309	Genes
1293	Genes, Neurofibromatosis 1
1288	Genes, Tumor Suppressor
1291	Genes, Wilms Tumor
1526	Phenotype

1308	Alleles
1294	Genes, Neurofibromatosis 2
1363	Introns
1312	Genes, Dominant

Figura 193. Categorías de genética más frecuentes en el análisis

Por último, las categorías que han aparecido en el subconjunto de 10 categorías más frecuentes relacionados con la "Sordera" son:

Identificador	Nombre Categoría
1564	Hearing Loss
1553	Deafness
1561	Hearing Loss, Sensorineural
1558	Hearing Loss, Functional
1563	Wolfram Syndrome
1559	Hearing Loss, High-Frequency
1555	Hearing Loss, Bilateral
1556	Hearing Loss, Central
1562	Presbycusis
1565	Hearing Loss, Unilateral
1566	Hearing Loss, Mixed Conductive-Sensorineural
1567	Usher Syndromes

Figura 194. Categorías de sordera más frecuentes en el análisis

9.6. Porcentaje de categorías gramaticales del dominio empleadas

A continuación, se sumó el número de repeticiones totales de cada categoría gramatical y se calculó el número de veces que aparecen los conceptos del dominio de estudio en tanto por ciento. Asimismo, se obtuvo la proporción entre repeticiones de conceptos de "Genética" frente a "Sordera":

Experimento	Configuración	% dominio	% "Genética"	% "Sordera"
Experimento 1	Frec. = 1, NO sem	31%	85%	15%
Experimento 2	Frec. = 5, NO sem	17%	82%	18%
Experimento 3	Frec. = 10, NO sem	16%	85%	15%
Experimento 4	Frec. = 20, NO sem	12%	83%	17%
Experimento 5	Frec. = 1, SÍ sem	28%	85%	15%
Experimento 6	Frec. = 5, SÍ sem	20%	84%	16%
Experimento 7	Frec. = 10, SÍ sem	16%	87%	13%
Experimento 8	Frec. = 20, SÍ sem	12%	84%	16%

Figura 195. Proporción de tipos de patrones

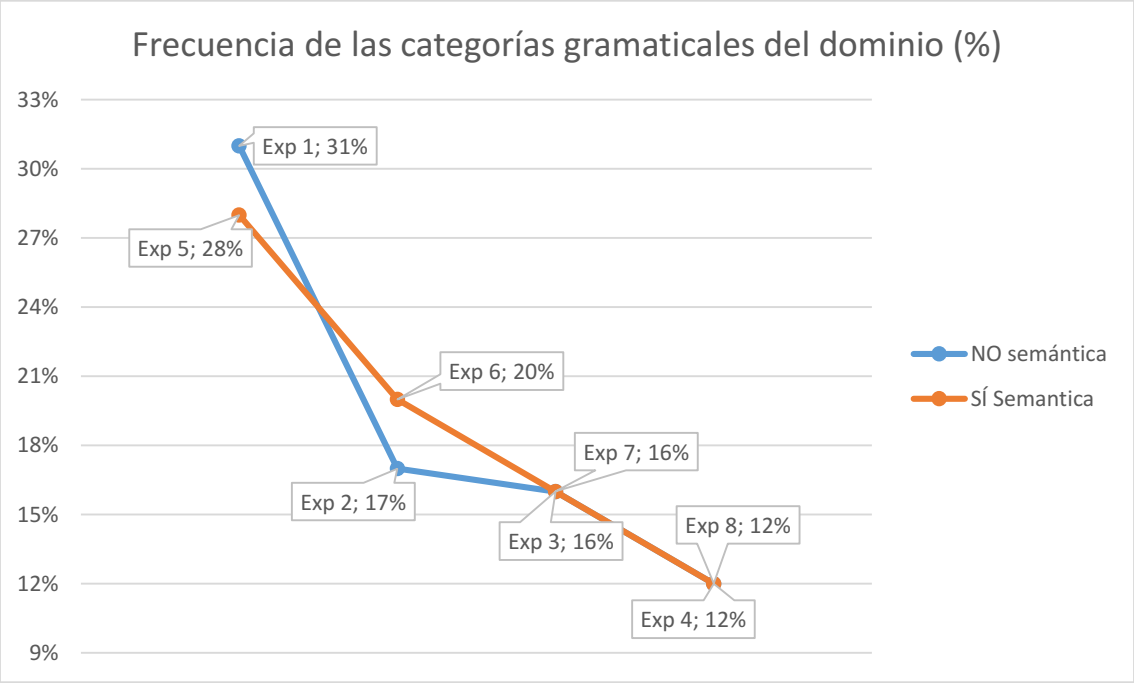

Figura 196. Gráfico de proporción de tipo de patrones

El uso de las categorías gramaticales del dominio de estudio tiende a decrecer cuando mayor sea la frecuencia mínima del experimento.

9.7. Estudio con ponderación de patrones

A partir de los patrones extraídos, se aplicó la fórmula de ponderación (especificada en el apartado 8) para determinar las secuencias de categorías gramaticales con mayor peso en el estudio.

En los ocho casos de estudio, los resultados son muy similares entre sí. Sin embargo, conviene destacar que el patrón con el valor de ponderación más alto es diferente según si se considera o se ignora la semántica.

En los escenarios en donde se ignoraba la semántica, la secuencia gramatical más valiosa es:

***pN**: "1224 + 1119 + 1237" = "DEFINITE ARTICLE + NOUN + PREPOSITION OF"

Los casos sin semántica ofrecen la mayor ponderación a una estructura muy frecuente en el inglés en donde predomina el sustantivo de la frase. Un ejemplo de sentencia con este patrón sería *"The mayority of…"*.

Por otro lado, la secuencia gramatical más valiosa en los escenarios que consideran la semántica es:

***pS**: "1230 + 1108 + 1213" = "VERB TO BE + VERB + PREPOSITION"

La secuencia mejor ponderada consiste en la estructura básica de un verbo que emplea "*to be*" como elemento auxiliar, seguido de una preposición. Un ejemplo de esta sentencia sería "*…was walking at…*".

Ambas secuencias estaban contenidas en patrones que superaban en aproximadamente 12.000 puntos con respecto al según patrón mejor valorado:

Experimento	Configuración	Ponderación del patrón mejor valorado	Diferencia con respecto al segundo patrón	Media varianza entre patrones sucesivos (top 100)
Experimento 1	Frec.=1, NO sem	53.070	+11.640	517
Experimento 2	Frec.=5, NO sem	53.070	+11.640	513
Experimento 3	Frec.=10, NO sem	53.070	+11.640	513
Experimento 4	Frec.=20, NO sem	53.070	+12.120	517
Experimento 5	Frec.=1, SÍ sem	66.217	+13.710	644
Experimento 6	Frec.=5, SÍ sem	65.782	+13.372	641
Experimento 7	Frec.=10, SÍ sem	65.610	+13.268	641
Experimento 8	Frec.=20, SÍ sem	65.430	+13.785	643

Figura 197. Varianzas de la ponderación de los patrones por experimento

Sin embargo, también hay que destacar que, en los escenarios en donde se ignoraba la gramática, la secuencia **pS* aparece con un alto grado de ponderación. Como se puede apreciar en los resultados correspondientes, el patrón que contenía dicha secuencia ocupaba la 4ª posición en la lista de ponderación.

El fenómeno inverso también se aprecia en los escenarios en donde se considera la semántica. La secuencia **pN* siempre ocupa la 2ª posición en las listas correspondientes a estos experimentos.

Además de **pN* y **pS*, se han encontrado otras secuencias comunes en el inglés, como:

- "1224 + 1158 + 1237" = "DEFINITE ARTICLE + NOT GROUPING NOUN + PREPOSITION OF".
- "1230 + 1108 + p55" = "1230 +1108 + 1213 + 1223" = "VERB TO BE + VERB + PREPOSITION + INDEFINITE ARTICLE".
- "1233 + 1230 + 1108 + 1229" = "VERB TO HAVE + VERB TO BE + VERB + PREPOSITION TO".

En las listas de 100 patrones mejor ponderados también se pudo encontrar algunas secuencias gramaticales que emplean un concepto del dominio:

- ***pD1:** "1300 + 1123 + 1110" = "Exons + INDEFINITE ARTICLE + SYMBOL".
- ***pD2:** "1151 + 1110 + 1286" = "ADVERB + SYMBOL + Genes, vif".

Cabe destacar que el patrón del dominio **pD2* siempre ha tenido el mejor valor de ponderación en los ocho escenarios de estudio.

Otro aspecto interesante es la alta frecuencia de preposiciones en los patrones mejor ponderados, en particular "of" (y posibles sinónimos como "about", "like", "regarding"…), cuyo identificador es 1237. Esto era de esperar: el número de preposiciones en el inglés es muy limitado en comparación a otras categorías sintácticas, por lo que se pueden asociar en la base de datos de la herramienta con facilidad.

A lo largo del tiempo, es de esperar que un idioma crezca. Esto influye sobre todo en las palabras categorizadas como nombres y verbos. Por ejemplo, en los años 50 no se empleaban palabras y expresiones como "*smartphone"* o "*[to] google (something)*". Sin embargo, otras categorías como las preposiciones se han mantenido constante en tamaño.

También es conveniente saber que el uso de preposiciones es bastante frecuente en la prosa académica inglesa. De acuerdo a los resultados obtenidos por el estudio de Douglas Biber et al (1999), por cada millón de

palabras en los textos de este tipo, 150.000 son preposiciones, sólo superado por sustantivos (300.000). [47]

Por último, cabe considerar que los resultados de los patrones más significantes ofrecerían secuencias más largas e interesantes si se expandiera la base de datos de la herramienta. Ya que la fórmula empleada para la ponderación penaliza considerablemente el uso de nombres sin clasificar, muchos de los patrones encontrados por la herramienta son secuencias parciales de una oración completa, ya que su conocimiento de sustantivos es limitado. Si se adquiriesen más nombres, ya sean del dominio de estudio, de otros dominios o de carácter generalista, aumentarían las posibilidades de formar patrones que representasen frases completas.

10. Conclusiones finales

Una vez completados todos los experimentos y extraído los resultados que ofrece cada uno, se han llegado a una serie de conclusiones acerca del estudio de patrones en el campo de la sordera genética:

1. En primer lugar, se puede considerar que los resultados ofrecidos por herramienta BoilerPlates son coherentes. Si bien es cierto que la mayoría de patrones generados utilizan la categoría 1144 (“Unclassified Noun”) en la mayoría de los patrones más frecuentes, se han podido detectar secuencias que son comunes en el inglés, como “VERB TO BE + VERB” o “NOUN + PREPOSITION OF”.

2. Una dificultad encontrada durante el proceso de análisis fue la gran cantidad de documentos a procesar por parte de la herramienta. Debido a limitaciones de tamaño, se tuvo que dividir todos los documentos en catorce lotes, complicando en gran medida la extracción de resultados. Podría haber sido posible limitar el efecto de la separación por lotes si se hubiese considerado emplear menos lotes de mayor tamaño, como por ejemplo 7 lotes de 100 documentos.

3. La variable de frecuencia mínima ha sido especialmente influyente en el proceso de análisis. Un valor alto reducía considerablemente el número de patrones generados, por lo que sólo aparecerían en los resultados aquellas secuencias lingüísticas que se repiten con mucha frecuencia en todos los documentos.

4. En cuanto a la diferenciación por semántica:

    - La variación en el número de patrones encontrados según si se ha aplicado semántica o no ha resultado ser baja. La mayor diferencia en tanto por ciento se pudo encontrar en la pareja de experimentos 3 a 8 (frecuencia mínima = 10), en donde el caso sin semántica tenía un 3,6% (202) más patrones que en el caso con semántica.

    - También cabe destacar que, exceptuando la pareja de experimentos con frecuencia mínima en 1, el número de patrones descubiertos fue siempre menor en los escenarios con semántica si se compara con su homólogo que la ignora.

    - Al clasificar los patrones según el número de repeticiones, los resultados obtenidos fueron muy similares entre sí, si sólo se considera el conjunto de los 100 patrones más frecuentes.

- Sin embargo, la mayor variación en los resultados se encontró cuando se realizó la ponderación de los patrones. Al aplicar o ignorar semántica, la secuencia gramatical más relevante alternaba entre “artículo + nombre + preposición” y “verbo (to be) + verbo + preposición”.

- Además, considerando las listas de 100 patrones mejor valorados, se pudo encontrar diferencias muy considerables entre parejas de experimentos que emplean semántica o no.

Se puede concluir, por tanto, que el uso de la semántica ha influido en los resultados.

5. De todos los patrones detectados en cada escenario, entre el 10% y el 16% tenían secuencias que emplean los conceptos del dominio de estudio. Estos resultados tienen sentido, ya que los términos que son específicos del dominio es sólo un subconjunto del resto de palabras que se pueden emplear en un documento que trata de la sordera genética.

6. Al evaluar los conceptos del MeSH (las categorías gramaticales en la base de datos) relacionados con el concepto general de “Genética”, los términos más frecuentes son los relacionados con:

    - **El gen Vif del VIH** (“Genes, vif”): “*DNA sequences that form the coding region for the vif (virion infectivity factor) protein that is important for the generation of infectious virions in human immunodeficiency virus (HIV)*.” [48]

    - **Neurofibromatosis, tipos 1 y 2** (“Genes, Neurofibromatosis”). Es una enfermedad relacionada con las mutaciones de los genes NF1 [49] y NF2 [50].

    - **Las mutaciones** (“Mutation”): “The changing of the structure of a gene, resulting in a variant form that may be transmitted to subsequent generations […]”. [51]

    - **Los alelos** (“Alleles”): “One of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.” [52]

    - **Los exones** (“Exons”): “A segment of a DNA or RNA molecule containing information coding for a protein or peptide sequence.” [53]

- **Los intrones** (“Introns”): “A segment of a DNA or RNA molecule which does not code for proteins and interrupts the sequence of genes”. [54]

- **El fenotipo** (“Phenotype”): “The set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.” [55]

Los términos que se encuentran con mucha frecuencia son los relacionados con los cambios o trastornos de los genes. También cabe destacar que el concepto “Genes, vif” siempre ha sido el más repetido, por lo que es posible que el gen Vif tenga mucha relevancia en el dominio.

7. En cuanto a los conceptos del MeSH relacionados con la sordera, se puede apreciar que se han empleado casi todas las posibles variantes que se insertaron en la base de datos (sordera neurosensorial, bilateral, unilateral, de alta frecuencia…).

   Cabe destacar también la aparición en los resultados del síndrome de Wolfram (“Wolfram Syndrome”) y, en menor medida, el síndrome de Usher (“Usher Syndrome”), ambos con origen genético. Sin embargo, estos dos conceptos tenían una frecuencia de aparición mucho menor que los conceptos más generalistas de la sordera. Una posible causa de esto es que ambos síndromes se consideran raros (4 ó 5 entre 100.000 personas para Usher [56]; 1 entre 500.000 personas para Wolfram tipo 1 [57]), por lo que encontrar documentación para estas afecciones es más complicado.

10.1. Nuevas líneas de trabajo

Una vez llegado a este punto del estudio, se han considerado líneas de trabajo que pueden partir de este proyecto.

1. En primer lugar, debido a falta de tiempo, no ha sido posible probar más configuraciones utilizando nuevos valores para la frecuencia mínima. Se podrían repetir los experimentos empleando valores como 50, 100 ó 200. De esta manera, se filtrarán aún más los resultados, de forma que sólo se registran los patrones con una frecuencia muy superior al resto.

2. La semántica ha influido durante el proceso de generación de patrones, si bien esta diferencia sólo se pudo apreciar de forma considerable al realizar el estudio con ponderado. Una posibilidad de trabajo futuro sería

volver a realizar la agrupación por semánticas utilizando una configuración diferente, para verificar si la diferenciación por semántica influye en el dominio de estudio o no.

3. La categoría gramatical más empleada es “Unclassified Noun”, debido a que existen muchas palabras en los documentos que la base de datos no es capaz de reconocer. Con el objetivo de obtener resultados más precisos y limitar el uso de esta categoría, se podría ampliar la terminología que acepta BoilerPlates, tanto en el resto de campos de la medicina como campos más generalistas.

4. Hasta ahora, los programas empleados para la extracción, procesado e inserción de los conceptos del MeSH a la base de datos son programas Java separados, lo que puede hacer difícil su uso para un usuario que no conoce su implementación. Se podría realizar mejoras en este aspecto, como generar un programa unificado con una interfaz gráfica que permita extraer los conceptos del tesauro y procesarlos para su inserción de una forma cómoda y rápida. Asimismo, se podría intentar automatizar el proceso de extracción de resultados una vez concluido el proceso de análisis, realizando las consultas SQL necesarias sin necesidad de ser ejecutado de forma manual.

11. Referencias

[1] Hipertextual: “2020, cuando el zettabyte dejó obsoleto al gigabyte”. https://hipertextual.com/2014/05/emc-world-2014-joe-tucci Última consulta: 27 de enero de 2017.

[2] Palme, J., 1984. You Have 134 Unread mail! Do You Want to Read Them Now? IFIP WG 6.5 Working conference on computer-based document services, Nottingham

[3] Schultze, U. and Vandenbosch, B. (1998). Information Overload in a Groupware Environment: Now you see it, now you don’t, Journal of Organisational Computing and Electronic Commerce 8(2): 127–148.

[4] Lewis, D., 1996. Dying for Information, Reuters Business Information.

[5] Sanderson, Mark and Croft, W. Bruce. The History of Information Retrieval Research. http://ciir-publications.cs.umass.edu/getpdf.php?id=1066

[6] Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, United Kingdom: Cambridge UniversityPress.

[7] Greengrass, Ed, 2000. Information Retrieval: A Survey. https://www.csee.umbc.edu/csee/research/cadip/readings/IR.report.120600.book.pdf

[8] Korfhage, Robert R., 1997. Information Storage and Retrieval.

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

[10] F. J. Martín Mateos, J. L. Ruiz Reina, 2012. “Procesamiento del Lenguaje Natural”. Dpto.Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla. URL: http://www.cs.us.es/cursos/ia2/temas/tema-06.pdf. Última consulta: 28 de enero de 2017.

[11] Rapaport, William J., 2005. The Turing Test. Department of Computer Science and Engineering, Department of Philosophy, and Center for Cognitive Science State University of New York at Buffalo, Buffalo, NY 14260-2000. https://pdfs.semanticscholar.org/bd48/66e584ac513befe5e3f4a9f24c5cc9e9405b.pdf

[12] Hutchins, John W. The Georgetown-IBM experiment demonstrated in January 1954. http://www.hutchinsweb.me.uk/AMTA-2004.pdf

[13] Jones, Karen, 2001. Natural language processing: a historical review. Computer Laboratory, University of Cambridge

[14] Página web: Molto. http://www.molto-project.eu/ Última visita: 28 de enero de 2017.

[15] Anjali, M. K. and Babu, Anto P., 2014. Ambiguities in Natural Language Processing. Department of Information Technology, Kannur University, Kerala, India. https://www.ijircce.com/upload/2014/sacaim/59_Paper%2027.pdf

[16] Neubig, Graham. NLP Programming Tutorial 4 - Word Segmentation. Nara Institute of Science and Technology (NAIST). http://www.phontron.com/slides/nlp-programming-en-03-ws.pdf Última visita: 28 de enero de 2017.

[17] Tesauro de la UNESCO. Definición de “Tesauro”. http://vocabularies.unesco.org/browser/thesaurus/es/page/concept12263 Última visita: 29 de enero de 2017.

[18] ANSI/NISO, 1993. Guidelines for the Construction, Format, and Maintenance of Monolingual Thesauri.

[19] Slype, Georges van., 1991. Los lenguajes de indización: Concepción, construcción y utilización en los sistemas documentales. Madrid, Pirámide.

[20] Diccionario de la Real Academia Española. Definición de “ontología”. http://dle.rae.es/?id=R5B0YYh. Última visita: 30 de enero de 2017.

[21] Gruber T., 1993. A translation approach to portable ontologies. Knowledge Acquisition

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

[23] Uschold M., Gruninger M., 1996. Ontologies: Principles, Methods and Applications. Knowledge Engineering Re-view, Vol. 11, No. 2, pp 93-115.

[24] Gruninger, M. y Lee, J., 2002. Ontology Applications and Design. Communications of the ACM. 45(2), pp. 39-41.

[25] Guarino, N., 1998. Formal Ontology in Information Systems. In Proceedings of FOIS'98, Trento, Italy, IOS Press, Amsterdam.

[26] Página web de la NLM. Fact Sheet - The National Library of Medicine. https://www.nlm.nih.gov/pubs/factsheets/nlm.html Última visita: 30 de enero de 2017.

[27] Página web de la NLM. Fact Sheet - Medical Subject Headings (MeSH®). https://www.nlm.nih.gov/pubs/factsheets/mesh.html Última visita: 30 de enero de 2017.

[28] Página de entrada a MeSH, versión 2016. https://meshb-prev.nlm.nih.gov/#/fieldSearch Última visita: 16 de agosto de 2016.

[29] Página web: MeSH – Vista general de los árboles de terminología. https://www.nlm.nih.gov/mesh/2016/mesh_browser/MeSHtree.A.html#link_id

[30] Página web: MeSH – Definición del concepto “Hearing Loss, Sudden”. https://meshb-prev.nlm.nih.gov/#/record/ui?ui=D003639 Última visita: 16 de agosto de 2016.

[31] Página web: Historia de Eclipse. http://www.ibm.com/developerworks/rational/library/nov05/cernosek/ Última visita: 2 de febrero de 2017.

[32] Página web: documentación de Eclipse. http://help.eclipse.org/neon/index.jsp?topic=%2Forg.eclipse.platform.doc.isv%2Fguide%2Fint_eclipse.htm Última visita: 2 de febrero de 2017.

[33] Eclipse Foundation. Eclipse Public License - v 1.0. http://www.eclipse.org/legal/epl-v10.html Última visita: 2 de febrero de 2017.

[34] Página web. Listado de Licencias de Software Libre incompatibles con GNU. https://www.gnu.org/philosophy/license-list.html#GPLIncompatibleLicenses Última visita: 2 de febrero de 2017.

[35] Página web. Preguntas más frecuentes de la EPL, pregunta 3: "¿cómo difiere la EPL con la CPL?" http://www.eclipse.org/legal/eplfaq.php Última visita: 2 de febrero de 2017.

[36] Oracle (1997). Java Code Conventions. http://www.oracle.com/technetwork/java/codeconventions-150003.pdf

[37] Página web. The Windows Club: History & Evolution Of Microsoft Office Software. http://www.thewindowsclub.com/history-evolution-microsoft-office-software Última visita: 2 de febrero de 2017.

[38] Términos y condiciones de uso de Microsoft Office 2013. https://img.labnol.org/di/microsoft-office-license.pdf Última visita: 2 de febrero de 2017.

[39] Moreno, Valentín, Pablo Miguel Suárez, Anabel Fraga, Juan Llorens y Eugenio Parra. 2013. Método de generación de patrones semánticos. PCT/ES2013/070638, issued 2013.

[40] Parra, Eugenio. 2016. Metodología orientada a la optimización automática de la calidad de los requisitos. PhD

[41] Arroyo Minguela, Leticia. 2015 Extracción de patrones sintáctico-semánticos de documentos de patentes. Proyecto fin de carrera. Ingeniería Técnica en Informática de Gestión. Escuela Politécnica Superior.

[42] de la O Maestro, Nuria. 2015. Evaluación de un sistema de procesamiento del lenguaje natural de la banca. Proyecto Final de Carrera. Ingeniería Técnica en Informática de Gestión. Escuela Politécnica Superior.

[43] Rodríguez Barberena, Valeria. Evaluation of a natural language processing system in public health. Trabajo Final de Máster. Máster en Ciencia y Tecnología Informática. Escuela Politécnica Superior.

[44] MeSH: consulta del término "Hearing Loss". https://meshb.nlm.nih.gov/#/record/ui?ui=D034381 Última visita: 16 de febrero de 2017.

[45] MeSH: consulta del término “Genetic Phenomena”. https://meshb.nlm.nih.gov/#/record/ui?ui=D055614 Última visita: 16 de febrero de 2017.

[46] MeSH: página de descarga de ficheros. https://www.nlm.nih.gov/mesh/download_mesh.html Última consulta: 16 de agosto de 2016.

[47] Biber, Douglas et al, 1999. Longman Grammar of Spoken and Written English. Longman.

[48] MeSH: Definición de “Genes, Vif”. https://meshb-prev.nlm.nih.gov/#/record/ui?ui=D016341 Última consulta: 5 de febrero de 2017.

[49] Biblioteca Nacional de EE.UU – Genetics Home Reference. Entrada de “neurofibromatosis type 1”. https://ghr.nlm.nih.gov/condition/neurofibromatosis-type-1#genes Última visita: 5 de febrero de 2017.

[50] Biblioteca Nacional de EE.UU – Genetics Home Reference. Entrada de “neurofibromatosis type 2”. https://ghr.nlm.nih.gov/condition/neurofibromatosis-type-2#genes Última visita: 5 de febrero de 2017.

[51] Diccionario de Oxford. Definición de “Mutation” (segunda entrada). http://www.oxforddictionaries.com/es/definicion/ingles_americano/mutation Última visita: 5 de febrero de 2017.

[52] Diccionario de Oxford. Definición de “Allele”. http://www.oxforddictionaries.com/es/definicion/ingles_americano/allele?q=Allele Última visita: 5 de febrero de 2017.

[53] Diccionario de Oxford. Definición de “Exon”. http://www.oxforddictionaries.com/es/definicion/ingles/exon Última visita: 5 de febrero de 2017.

[54] Diccionario de Oxford. Definición de “Intron”. http://www.oxforddictionaries.com/es/definicion/ingles/intron Última visita: 5 de febrero de 2017.

[55] Diccionario de Oxford. Definición de “Phenotype”. http://www.oxforddictionaries.com/es/definicion/ingles/phenotype Última visita: 5 de febrero de 2017.

[56] Biblioteca Nacional de EE.UU – Genetics Home Reference. Entrada de “Usher syndrome”. https://ghr.nlm.nih.gov/condition/usher-syndrome Última visita: 5 de febrero de 2017.

[57] Biblioteca Nacional de EE.UU – Genetics Home Reference. Entrada de “Wolfram syndrome”. https://ghr.nlm.nih.gov/condition/wolfram-syndrome Última visita: 5 de febrero de 2017.

Anexo A. Lista de semánticas incorporadas a Rqa Quality Analyzer

En este anexo se adjuntan todas las semánticas que se han tenido en consideración a lo largo del proyecto.

En primer lugar, las semánticas genéricas ya insertadas previamente en la base de datos son las siguientes:

Identificador	Nombre semántica	Dominio
1021	- ASSOCIATION (OK)	Ajeno
1022	- ACTION (OK)	Ajeno
1023	- Action-Objective/Finish (OK pero borrar?)	Ajeno
1024	- Action-Destination (OK pero borrar?)	Ajeno
1025	Cause-Effect	Ajeno
1026	Concept-Attribute	Ajeno
1027	Agent-Instrument	Ajeno
1028	- Action-Agent (OK pero borrar?)	Ajeno
1029	Date	Ajeno
1030	Origin	Ajeno
1031	Exemplification	Ajeno
1032	Meronymy	Ajeno
1033	Holonymy	Ajeno
1034	Antonymy	Ajeno
1035	Hyponymy	Ajeno
1036	Enumerate	Ajeno
1037	- CONSTRAINT (OK)	Ajeno
1038	- VALUE (OK)	Ajeno
1039	- RANGE (value limitations) (OK)	Ajeno
1040	- RANGE >= MINIMUM (OK)	Ajeno
1041	- RANGE > MINIMUM (OK)	Ajeno
1042	- RANGE <= (MAXIMUM) (OK)	Ajeno
1043	- RANGE < MAXIMUM (OK)	Ajeno
1044	Stereotype	Ajeno
1045	- RANGE = EQUAL (OK)	Ajeno
1046	Metadata	Ajeno
1047	Generalisation-Specialisation	Ajeno
1048	Hierarchy	Ajeno
1049	Occurrence	Ajeno
1050	Idiomatic	Ajeno
1051	Realization	Ajeno
1052	Inverse Generalisation-Specialisation	Ajeno
1053	Asymmetric Generic Relationship	Ajeno
1054	DataType	Ajeno
1055	Generic Relationship	Ajeno
1056	LinkObject	Ajeno

1057	Flow	Ajeno
1058	Annotate	Ajeno
1059	Ortographic Equivalence	Ajeno
1060	Equivalence	Ajeno
1061	Link	Ajeno
1062	Sequence	Ajeno
1063	Synonymy	Ajeno
1064	Linguistic	Ajeno
1065	Dependency	Ajeno
1066	Composition Association (Hyperonymy)	Ajeno
1067	Inverse Composition Association	Ajeno
1068	Aggregation Association	Ajeno
1069	Inverse Aggregation Association	Ajeno
1070	Semantic Structural Association	Ajeno
1071	- Transition (OK)	Ajeno
1072	- Sustainability (OK)	Ajeno
1073	- Condition (OK)	Ajeno
1074	- Functionality for Requirements (OK)	Ajeno
1075	- CAPABILITY (OK)	Ajeno
1076	- Communication (OK)	Ajeno
1077	- Design (OK)	Ajeno
1078	- Run (OK)	Ajeno
1079	- Destroy (OK but remove?)	Ajeno
1080	- Deploy (OK but remove?)	Ajeno
1081	- Evaluate (OK)	Ajeno
1082	- Code (OK)	Ajeno
1083	- Teach (OK)	Ajeno
1084	- Close (OK but remove?)	Ajeno
1085	- Automate (OK)	Ajeno
1086	- Provide (OK)	Ajeno
1087	- Reset (OK)	Ajeno
1088	- Begin (OK but remove?)	Ajeno
1089	- To Pressure (OK)	Ajeno
1090	- Config (OK)	Ajeno
1091	- Disconnect (OK)	Ajeno
1092	- Connect (OK)	Ajeno
1093	- Specify (OK)	Ajeno
1094	- Deny (OK)	Ajeno
1095	- Support (OK)	Ajeno
1096	- Launch (OK)	Ajeno
1097	- Allow (OK)	Ajeno
1098	- Ask (OK but remove?)	Ajeno
1099	- Document (OK)	Ajeno
1100	- Search (OK)	Ajeno
1101	- Operation (OK)	Ajeno
1102	- Access (OK)	Ajeno
1103	- Receive (OK but remove?)	Ajeno
1104	- Send (info) (OK)	Ajeno

1105	- Notify (OK)	Ajeno
1106	- Load (Fill) (OK)	Ajeno
1107	- Synchronize (OK)	Ajeno
1108	- Verify (OK)	Ajeno
1109	- Visualization (OK)	Ajeno
1110	- Generate (OK)	Ajeno
1111	- Select (OK but Remove?)	Ajeno
1112	- Modify (OK)	Ajeno
1113	- Save (OK)	Ajeno
1114	- Remove (OK)	Ajeno
1115	- Add (OK)	Ajeno
1116	Anchestors	Ajeno
1117	Theme/Content	Ajeno
1118	Entity-Requirement/Condition	Ajeno
1119	Value-Price	Ajeno
1120	Surface	Ajeno
1121	Distance	Ajeno
1122	Temperature	Ajeno
1123	Author-Object	Ajeno
1124	Product-Material	Ajeno
1125	- AGENT (OK)	Ajeno
1126	- STAKEHOLDER (OK)	Ajeno
1127	- SYSTEM FUNCTION (OK)	Ajeno
1128	- SYSTEM NOT FUNCTION (OK)	Ajeno
1129	- MODAL (OK)	Ajeno
1130	- MODAL COMPULSORY (OK)	Ajeno
1131	- MODAL OPTIONAL (OK)	Ajeno
1132	- MODAL FUTURE (OK)	Ajeno
1133	- MODAL ASSIGNEMENT (OK)	Ajeno
1134	- DUTY ACTION (OK)	Ajeno
1135	- STATE (MODE) ASSIGNEMENT (OK)	Ajeno
1136	- PORTABILITY (OK)	Ajeno
1137	- RANGE ALL (OK)	Ajeno
1138	- RANGE ANY (OK)	Ajeno
1139	- RANGE LITTLE-FEW-SOME (OK)	Ajeno
1140	- RANGE MUCH-MANY (OK)	Ajeno
1141	- RANGE ENOUGH (OK)	Ajeno
1142	- RANGE NO VALUE (OK)	Ajeno
1143	- RATE (OK)	Ajeno
1144	- UNIT (OK)	Ajeno
1145	- RANGE BETWEEN (OK)	Ajeno
1146	- RANGE DURING (Sustainability) (OK)	Ajeno
1147	- REQUIREMENT (OK)	Ajeno
1148	- COMPULSORY REQUIREMENT (OK)	Ajeno
1149	- OPTIONAL REQUIREMENT (OK)	Ajeno
1150	- FUTURE REQUIREMENT (OK)	Ajeno
1151	- STAKEHOLDER COMPULSORY REQUIREMENT (OK)	Ajeno

1152	- SYSTEM FUNCTIONAL COMPULSORY REQUIREMENT (OK)	Ajeno
1153	- SYSTEM NOT FUNCTIONAL COMPULSORY REQUIREMENT (OK)	Ajeno
1154	- RANGE SIMULTANEUSLY (OK)	Ajeno
1155	- ACTION ACTIVATION (OK)	Ajeno
1156	- WHEN ACTIVATION (OK)	Ajeno
1157	- WHILE ACTIVATION (OK)	Ajeno
1158	- PROPERTY (OK)	Ajeno
1159	- STAKEHOLDER OPTIONAL REQUIREMENT (OK)	Ajeno
1160	- SYSTEM FUNCTIONAL OPTIONAL REQUIREMENT (OK)	Ajeno
1161	- SYSTEM NOT FUNCTIONAL OPTIONAL REQUIREMENT (OK)	Ajeno
1162	- SYSTEM NOT FUNCTIONAL FUTURE REQUIREMENT (OK)	Ajeno
1163	- SYSTEM FUNCTIONAL FUTURE REQUIREMENT (OK)	Ajeno
1164	- STAKEHOLDER FUTURE REQUIREMENT (OK)	Ajeno
1165	- IF ACTIVATION (OK)	Ajeno

Figura 198. Semánticas ajenas al dominio

A continuación, se adjuntan todas las semánticas que fueron generadas durante el proceso de agrupación de conceptos del MeSH. Con el propósito de facilitar la identificación de cada semántica, se introdujo un nombre de forma manual que define a grandes rasgos el significado y contexto de los conceptos asociados a él.

Identificador	Nombre semántica	Dominio
1166	Types of Genes 1	Genética
1167	Types of Genes 2	Genética
1168	Types of Genes 3	Genética
1169	Exon	Genética
1170	Microbe	Genética
1171	Operon	Genética
1172	Gene Rearrangement	Genética
1173	Genetic processes	Genética
1174	DNA breaks	Genética
1175	Mutation	Genética
1176	Recombination	Genética
1177	Gene Expression Regulation	Genética
1178	Genome	Genética
1179	Mutation 2	Genética

1180	Polymorphism	Genética
1181	Biological Evolution	Genética
1182	General	Genética
1183	Deafness	Sordera

Figura 199. Semánticas del dominio generadas

En total se generaron 17 semánticas con los conceptos asociados a la genética y 1 semántica para todos los conceptos de sordera.

Anexo B. Categorías gramaticales de Rqa Quality Analyzer

A continuación, se adjuntan todas las categorías gramaticales que aparecen en la tabla “Rules_Families” de la base de datos Rqa Quality Analizer.

Antes de empezar el proyecto, la herramienta contaba con las siguientes categorías gramaticales de carácter general:

Identificador	Nombre
1099	INVARIANT
1100	LINKING
1102	PURPOSE/GOAL
1103	ADJECTIVE
1106	ADVERBIAL PHRASE
1108	VERB
1109	LLCHART VARIABLE
1110	SYMBOL
1113	PRONOUN
1114	RECOVERABLE PRONOUN
1119	NOUN
1123	NUMBER
1144	UNCLASSIFIED NOUN
1151	ADVERB
1152	TIME ADVERB
1153	PLACE ADVERB
1155	DETERMINER
1156	ARTICLE (OK)
1158	NOT GROUPING NOUN
1159	REQUs Domain NOT GROUPING NOUN (OK)
1160	MEASUREMENT UNIT
1166	PHRASAL VERB BASE
1168	ABSOLUTE VERB
1169	PHRASAL VERB PARTICLE
1170	PHRASAL VERB
1191	STOPWORD
1193	OR LINKING
1197	RELATIVE PRONOUN
1203	PERSONAL PRONOUN
1213	PREPOSITION
1215	SOFTWARE
1216	UTTERANCE DETERMINER
1218	QUANTIFIER DETERMINER
1219	NUMBER DETERMINER
1220	PARTITIVE DETERMINER
1221	DEMONSTRATIVE DETERMINER
1222	POSSESSIVE DETERMINER

1223	INDEFINITE ARTICLE
1224	DEFINITE ARTICLE
1225	NEGATION
1226	VERB TO DO
1228	PREPOSITION FOR
1229	PREPOSITION TO
1230	VERB TO BE
1232	UNCLASSIFIED VERB
1233	VERB TO HAVE
1236	PREPOSITION BY
1237	PREPOSITION OF
1240	MODAL VERB
1241	UNCLASSIFIED ADJECTIVE
1242	TIME ADVERBIAL PHRASE
1243	PREPOSITIONAL LINKING PHRASE
1244	PREPOSITIONAL LOCATION
1247	CAUSAL CONECTOR
1248	AND LINKING
1251	REQUIREMENT/CONDITION
1254	ARTICLE
1255	PUNTUACTION MARK
1256	SENTENCE BREAKER
1257	ACRONYM
1258	UNCLASSIFIED ADVERB

Figura 200. Categorías gramaticales ajenas al dominio

Las categorías gramaticales asociadas a la genética son las siguientes:

Identificador	Nombre
1259	Proto-Oncogenes
1260	Genes, ras
1261	Genes, fms
1262	Genes, mos
1263	Genes, myc
1264	Genes, abl
1265	Genes, src
1266	Genes, jun
1267	Genes, fos
1268	Genes, erbB-2
1269	Genes, erbB-1
1270	Genes, erbA
1271	Genes, erbB
1272	Genes, bcl-2
1273	Genes, bcl-1
1274	Genes, myb
1275	Genes, sis

1276	Genes, rel
1277	Genes, Viral
1278	Genes, gag
1279	Genes, pol
1280	Genes, env
1281	Genes, nef
1282	Genes, tat
1283	Genes, rev
1284	Genes, vpr
1285	Genes, vpu
1286	Genes, vif
1287	Genes, pX
1288	Genes, Tumor Suppressor
1289	Genes, p53
1290	Genes, Retinoblastoma
1291	Genes, Wilms Tumor
1292	Genes, DCC
1293	Genes, Neurofibromatosis 1
1294	Genes, Neurofibromatosis 2
1295	Genes, MCC
1296	Genes, APC
1297	Genes, BRCA1
1298	Genes, p16
1299	Genes, BRCA2
1300	Exons
1301	Immunoglobulin Switch Region
1302	Expressed Sequence Tags
1303	Transcription Initiation Site
1304	Gene Components
1305	Regulatory Elements, Transcriptional
1306	VDJ Exons
1307	Hinge Exons
1308	Alleles
1309	Genes
1310	Genes, araC
1311	Genes, Bacterial
1312	Genes, Dominant
1313	Genes, Fungal
1314	Genes, Homeobox
1315	Genes, MHC Class II
1316	Genes, Immunoglobulin
1317	Genes, Lethal
1318	Genes, MHC Class I
1319	Genes, Overlapping
1320	Genes, Recessive
1321	Genes, Regulator
1322	Multigene Family

1323	Genes, Switch
1324	Genes, Synthetic
1325	Major Histocompatibility Complex
1326	Oncogenes
1327	Pseudogenes
1328	Genes, Suppressor
1329	Genes, RAG-1
1330	Genes, Protozoan
1331	Genes, Helminth
1332	Genes, Plant
1333	Genes, Insect
1334	Genes, Immediate-Early
1335	Genes, Reporter
1336	Genes, cdc
1337	Transgenes
1338	Genes, MDR
1339	Genes, T-Cell Receptor
1340	Genes, T-Cell Receptor alpha
1341	Genes, T-Cell Receptor beta
1342	Genes, T-Cell Receptor gamma
1343	Genes, T-Cell Receptor delta
1344	Genes, Archaeal
1345	Genes, Essential
1346	Genes, Duplicate
1347	Genes, rRNA
1348	Genes, sry
1349	Nested Genes
1350	Genes, Transgenic, Suicide
1351	Genes, Mating Type, Fungal
1352	Genes, X-Linked
1353	Genes, Y-Linked
1354	Genes, Mitochondrial
1355	Genes, Developmental
1356	Genes, Immunoglobulin Heavy Chain
1357	Genes, Immunoglobulin Light Chain
1358	Genes, Neoplasm
1359	Genes, Modifier
1360	Genes, Chloroplast
1361	Genes, Microbial
1362	Attachment Sites, Microbiological
1363	Introns
1364	Lac Operon
1365	Minor Histocompatibility Loci
1366	Operon
1367	Replicon

1368	rRNA Operon
1369	Sequence Tagged Sites
1370	Minor Lymphocyte Stimulatory Loci
1371	Regulon
1372	Replication Origin
1373	5' Flanking Region
1374	3' Flanking Region
1375	Genome Components
1376	Quantitative Trait Loci
1377	DNA Sequence, Unstable
1378	Chromosome Fragile Sites
1379	Genetic Loci
1380	t-Complex Genome Region
1381	Achaete-Scute Complex Genome Region
1382	Gene Rearrangement
1383	Gene Rearrangement, B-Lymphocyte
1384	Gene Rearrangement, B-Lymphocyte, Heavy Chain
1385	Gene Rearrangement, B-Lymphocyte, Light Chain
1386	Gene Rearrangement, T-Lymphocyte
1387	Gene Rearrangement, alpha-Chain T-Cell Antigen Receptor
1388	Gene Rearrangement, beta-Chain T-Cell Antigen Receptor
1389	Gene Rearrangement, gamma-Chain T-Cell Antigen Receptor
1390	Gene Rearrangement, delta-Chain T-Cell Antigen Receptor
1391	Immunoglobulin Class Switching
1392	V(D)J Recombination
1393	Dosage Compensation, Genetic
1394	Genomic Imprinting
1395	Gene Silencing
1396	RNA Interference
1397	Epigenesis, Genetic
1398	X Chromosome Inactivation
1399	Chromosomal Position Effects
1400	Epigenetic Repression
1401	CRISPR-Cas Systems
1402	DNA Damage
1403	DNA Breaks, Double-Stranded
1404	DNA Breaks, Single-Stranded
1405	DNA Fragmentation
1406	DNA Breaks

1407	Chromosome Breakpoints
1408	Mutation Accumulation
1409	Silent Mutation
1410	Mosaicism
1411	Mutation
1412	Frameshift Mutation
1413	Point Mutation
1414	Germ-Line Mutation
1415	Loss of Heterozygosity
1416	Mutation, Missense
1417	Base Pair Mismatch
1418	Allelic Imbalance
1419	Chimerism
1420	Haploinsufficiency
1421	Conjugation, Genetic
1422	Crossing Over, Genetic
1423	Gene Conversion
1424	Recombination, Genetic
1425	Sister Chromatid Exchange
1426	Transformation, Genetic
1427	Gene Transfer, Horizontal
1428	Oncogene Fusion
1429	Gene Fusion
1430	Homologous Recombination
1431	Ectopic Gene Expression
1432	Enzyme Induction
1433	Enzyme Repression
1434	Epistasis, Genetic
1435	Gene Amplification
1436	Gene Expression Regulation
1437	Transcriptional Activation
1438	Gene Expression Regulation, Bacterial
1439	Gene Expression Regulation, Fungal
1440	Gene Expression Regulation, Viral
1441	Gene Expression Regulation, Enzymologic
1442	Gene Expression Regulation, Neoplastic
1443	Gene Expression Regulation, Leukemic
1444	Gene Expression Regulation, Plant
1445	Gene Expression Regulation, Developmental
1446	Gene Expression Regulation, Archaeal

1447	Bacteriocin Plasmids
1448	Cosmids
1449	F Factor
1450	Genetic Code
1451	Genetic Vectors
1452	Hemolysin Factors
1453	Lactose Factors
1454	Plasmids
1455	R Factors
1456	Templates, Genetic
1457	Genomic Library
1458	Gene Library
1459	Genome, Human
1460	Reading Frames
1461	Open Reading Frames
1462	Genome
1463	Genome, Viral
1464	Genome, Bacterial
1465	Genome, Fungal
1466	Genome, Protozoan
1467	Genome, Plant
1468	Genome, Archaeal
1469	Plant Tumor-Inducing Plasmids
1470	Genetic Structures
1471	Genome, Insect
1472	Genome, Helminth
1473	Gene Regulatory Networks
1474	Genome, Plastid
1475	Genome, Chloroplast
1476	Genome, Mitochondrial
1477	Metagenome
1478	Exome
1479	Genome Size
1480	Karyotype
1481	Genome, Microbial
1482	Suppression, Genetic
1483	Mutagenesis
1484	Gene Deletion
1485	Sequence Deletion
1486	Gene Duplication
1487	Somatic Hypermutation, Immunoglobulin
1488	INDEL Mutation
1489	Sequence Inversion
1490	Polymorphism, Genetic
1491	Polymorphism, Restriction Fragment Length
1492	Polymorphism, Single-Stranded Conformational

1493	Polymorphism, Single Nucleotide
1494	Genomic Structural Variation
1495	DNA Copy Number Variations
1496	Biological Evolution
1497	Lysogeny
1498	Selection, Genetic
1499	Gene Expression
1500	Virus Integration
1501	Evolution, Molecular
1502	Sex Determination Processes
1503	Genetic Processes
1504	Heredity
1505	Genetic Speciation
1506	Self-Fertilization
1507	Mutation Rate
1508	Reproductive Isolation
1509	Endoreduplication
1510	Genetic Background
1511	Antibody Diversity
1512	Antigenic Variation
1513	Consanguinity
1514	Diploidy
1515	Extrachromosomal Inheritance
1516	Gene Frequency
1517	Gene Pool
1518	Genotype
1519	Haploidy
1520	Haplotypes
1521	Heterozygote
1522	Homozygote
1523	Hybrid Vigor
1524	Genetic Linkage
1525	Lod Score
1526	Phenotype
1527	Phylogeny
1528	Ploidies
1529	Sex Ratio
1530	Genetic Variation
1531	Linkage Disequilibrium
1532	Gene Dosage
1533	Founder Effect
1534	Genetic Heterogeneity
1535	Quantitative Trait, Heritable
1536	Penetrance
1537	Multifactorial Inheritance
1538	Gene Order
1539	Inheritance Patterns

1540	Genetic Load
1541	Genetic Drift
1542	Gene Flow
1543	Genetic Phenomena
1544	Genetic Fitness
1545	Hemizygote
1546	Endophenotypes
1547	Genetic Pleiotropy
1548	Gene-Environment Interaction
1549	Ecotype
1550	DNA Transformation Competence
1551	Sympatry
1552	Serogroup

Figura 201. Categorías gramaticales asociadas a la genética

Las categorías gramaticales asociadas a la sordera son las siguientes:

Identificador	Nombre
1553	Deafness
1554	Hearing Loss, Sudden
1555	Hearing Loss, Bilateral
1556	Hearing Loss, Central
1557	Hearing Loss, Conductive
1558	Hearing Loss, Functional
1559	Hearing Loss, High-Frequency
1560	Hearing Loss, Noise-Induced
1561	Hearing Loss, Sensorineural
1562	Presbycusis
1563	Wolfram Syndrome
1564	Hearing Loss
1565	Hearing Loss, Unilateral
1566	Hearing Loss, Mixed Conductive-Sensorineural
1567	Usher Syndromes
1568	Deaf-Blind Disorders

Figura 202. Categorías gramaticales asociadas a la sordera

Anexo C. Requisitos del proyecto

C.1. Formato de un requisito

R<TS-XXX>			
Fecha	MM/DD/AAAA	**Fuente**	<Fuente>
Prioridad	<Prioridad>	**Necesidad**	<Necesidad>
Objetivo	<Objetivo>		
Descripción	<Descripción>		
Precondiciones	<Precondición 1> <Precondición 2> …		
Trazabilidad	<Requisito A> <Requisito B> …		

Figura 203. Ejemplo de estructura de un requisito

- **R<TS-XXX>** representa el identificador del requisito. Tiene el siguiente significado:
  - T es el tipo de requisito. Puede tener uno de los siguientes valores:
    - F: Funcional
    - N: No funcional
  - S representa la sección del proyecto al que pertenece el requisito. Puede tomar los siguientes valores:
    - E corresponde al proceso de extracción de terminología.
    - A corresponde al proceso de agrupación de semánticas.
    - D corresponde al proceso de adquisición y adaptación de documentos.
    - B corresponde al uso de BoilerPlates.
    - R corresponde a la extracción de resultados.
  - XXX es un número que identifica al requisito de forma inequívoca.

- En el campo **Fecha** se introduce el día en el que el requisito fue creado.
- En el campo **Fuente** se introduce la fuente del requisito.
- El campo **Prioridad** puede tomar los valores Alta, Media o Baja.
- El campo **Necesidad** puede tomar los valores Alta, Media o Baja.
- En el campo **Objetivo** se introduce el objetivo del requisito.
- En el campo **Descripción** se introduce una descripción breve del requisito.
- En el campo **Precondiciones** se introduce las condiciones previas que deben producirse para la consecución del requisito.
- En el campo **Trazabilidad** se introducen los requisitos que están asociados a éste.

C.2. Especificación de requisitos

RFE-001			
Fecha	02/03/2016	**Fuente**	Consulta MeSH
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Extracción de terminología		
Descripción	El programa realizará la extracción de la terminología a partir del tesauro MeSH.		
Precondiciones	--		
Trazabilidad	RFE-002, RFE-003, RNE-001		

Figura 204. Requisito RFE-001

RFE-002			
Fecha	02/03/2016	**Fuente**	Consulta MeSH
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Adquisición de conceptos de “Genética”		
Descripción	El programa extraerá todos los conceptos asociados al subárbol de MeSH cuya raíz es el concepto “Genética”.		
Precondiciones	--		
Trazabilidad	RFE-001, RFE-003		

Figura 205. Requisito RFE-002

RFE-003			
Fecha	02/03/2016	**Fuente**	Consulta MeSH
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Adquisición de conceptos de “Sordera”		
Descripción	El programa extraerá todos los conceptos asociados al subárbol de MeSH cuya raíz es el concepto “Sordera”.		
Precondiciones	--		
Trazabilidad	RFE-001, RFE-002		

Figura 206. Requisito RFE-003

RFE-004			
Fecha	02/03/2016	**Fuente**	Consulta MeSH
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Campos a extraer por concepto		
Descripción	El programa obtendrá, por cada concepto del dominio, su nombre, su definición, sus términos […]		
Precondiciones	Conceptos del dominio encontrados y extraídos		
Trazabilidad	--		

Figura 207. Requisito RFE-004

RFE-005			
Fecha	02/06/2016	**Fuente**	Programación
Prioridad	Media	**Necesidad**	Media
Objetivo	Separación de conceptos		
Descripción	Los campos de información de cada concepto estarán separados por un salto de línea.		
Precondiciones	Conceptos del dominio encontrados y extraídos		
Trazabilidad	RFE-006, RFE-007		

Figura 208. Requisito RFE-005

RFE-006			
Fecha	02/06/2016	**Fuente**	Programación
Prioridad	Baja	**Necesidad**	Media
Objetivo	Separación de campos		
Descripción	Cada concepto estará separado mediante el uso de un identificador de separación.		
Precondiciones	Conceptos del dominio encontrados y extraídos		
Trazabilidad	RFE-005, RFE-007		

Figura 209. Requisito RFE-006

RFE-007			
Fecha	02/06/2016	**Fuente**	Programación
Prioridad	Baja	**Necesidad**	Baja
Objetivo	Separación de contenido en el mismo campo		
Descripción	Los campos de un concepto que contengan varios valores estarán separados por un separador.		
Precondiciones	Conceptos del dominio encontrados y extraídos		
Trazabilidad	RFE-005, RFE-006		

Figura 210. Requisito RFE-007

RFE-008			
Fecha	02/06/2016	**Fuente**	Programación
Prioridad	Media	**Necesidad**	Alta
Objetivo	Generación del fichero final		
Descripción	El programa generará un fichero de texto plano (.txt) con la información de todos los conceptos encontrados.		
Precondiciones	Conceptos del dominio formateados correctamente		
Trazabilidad	RNE-012		

Figura 211. Requisito RFE-008

RNE-009			
Fecha	02/04/2016	**Fuente**	Consulta MeSH
Prioridad	Media	**Necesidad**	Media
Objetivo	Uso de fichero MeSH		
Descripción	El fichero de entrada del algoritmo será el obtenido tras la consulta al MeSH el 3 de febrero de 2016 (desc16.xml).		
Precondiciones	--		
Trazabilidad	RFE-001		

Figura 212. Requisito RFE-009

RNE-010			
Fecha	02/04/2016	**Fuente**	Consulta MeSH
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Adquisición total de conceptos del dominio		
Descripción	La tasa de conceptos extraídos que correspondan a los subárboles del dominio de estudio será 100%.		
Precondiciones	--		
Trazabilidad	RFE-002, RFE-003		

Figura 213. Requisito RNE-010

RNE-011			
Fecha	02/04/2016	**Fuente**	Consulta MeSH
Prioridad	Baja	**Necesidad**	Alta
Objetivo	No adquisición de conceptos ajenos al dominio		
Descripción	La tasa de conceptos extraídos que no correspondan a los subárboles del dominio de estudio será 0%.		
Precondiciones	--		
Trazabilidad	RFE-002, RFE-003		

Figura 214. Requisito RNE-011

RNE-012			
Fecha	02/04/2016	**Fuente**	Consulta MeSH
Prioridad	Media	**Necesidad**	Baja
Objetivo	Duración del proceso de extracción		
Descripción	La duración total del proceso de extracción y de conceptos y generación del fichero de salida será inferior a 30 segundos.		
Precondiciones	--		
Trazabilidad	RFE-001, RFE-004, RFE-008		

Figura 215. Requisito RNE-012

RFA-013			
Fecha	03/08/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Proceso de agrupación en semánticas		
Descripción	El programa realizará la agrupación de los conceptos en semánticas.		
Precondiciones	Realizado proceso de extracción de conceptos		
Trazabilidad	RFA-014, RFA-015, RFA-016, RNA-019, RNA-020		

Figura 216. Requisito RFA-013

RFA-014			
Fecha	03/08/2016	**Fuente**	Análisis previo
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Fichero de salida del proceso de agrupación		
Descripción	El programa generará un fichero de texto plano (.txt) con los resultados de la agrupación por semánticas.		
Precondiciones	Completado el proceso de agrupación por semánticas		
Trazabilidad	RFA-013		

Figura 217. Requisito RFA-014

RFA-015			
Fecha	03/08/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Conceptos asignados a única semántica		
Descripción	Un concepto podrá pertenecer a una sola semántica.		
Precondiciones	--		
Trazabilidad	RFE-016		

Figura 218. Requisito RFA-015

RFA-016			
Fecha	03/20/2016	**Fuente**	Aná. posterior
Prioridad	Baja	**Necesidad**	Media
Objetivo	No repeticiones de términos		
Descripción	Los términos de un concepto no tendrán repeticiones.		
Precondiciones	--		
Trazabilidad	RFE-015		

Figura 219. Requisito RFA-016

RNA-017			
Fecha	03/08/2016	**Fuente**	Análisis previo
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Fichero de entrada del proceso de agrupación		
Descripción	El programa recibirá como entrada el fichero generado tras la extracción de los conceptos del dominio.		
Precondiciones	Realizado proceso de extracción de conceptos		
Trazabilidad	RFE-008, RNA-013		

Figura 220. Requisito RFA-017

RNA-018			
Fecha	03/20/2016	**Fuente**	Análisis previo
Prioridad	Baja	**Necesidad**	Baja
Objetivo	Consideración de sinónimos como términos		
Descripción	El programa considerará los sinónimos de cada concepto como términos.		
Precondiciones	--		
Trazabilidad	RFE-004, RNA-013		

Figura 221. Requisito RNA-018

RNA-019			
Fecha	03/04/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Mínimo número de conceptos por semántica		
Descripción	El mínimo número de conceptos para generar una semántica será 5.		
Precondiciones	--		
Trazabilidad	RNA-013		

Figura 222. Requisito RNA-019

RNA-020			
Fecha	03/04/2016	**Fuente**	Análisis previo
Prioridad	Baja	**Necesidad**	Alta
Objetivo	Tasa de conceptos sin clasificar		
Descripción	El número de conceptos que no queden asociados a alguna semántica será inferior al 2,5% del total.		
Precondiciones	--		
Trazabilidad	RNA-013		

Figura 223. Requisito RNA-020

RFD-021			
Fecha	12/01/2016	**Fuente**	Análisis inicial
Prioridad	Alta	**Necesidad**	Media
Objetivo	Conversión de formato de documentación		
Descripción	Los documentos del dominio serán transformados a ficheros de texto plano (.txt).		
Precondiciones	Documentación adquirida		
Trazabilidad	RND-022, RND-024		

Figura 224. Requisito RFD-021

RND-022			
Fecha	12/01/2016	**Fuente**	Análisis inicial
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Fuente de adquisición de documentación		
Descripción	Los documentos de estudio deberán ser obtenidos a partir de fuentes oficiales asociadas a las ciencias de la vida.		
Precondiciones	--		
Trazabilidad	RND-023		

Figura 225. Requisito RND-022

RND-023			
Fecha	12/01/2016	**Fuente**	Análisis inicial
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Formato inicial de documentación		
Descripción	Los documentos serán adquiridos con el formato PDF.		
Precondiciones	--		
Trazabilidad	RND-022		

Figura 226. Requisito RND-023

RND-024			
Fecha	12/01/2016	**Fuente**	Análisis inicial
Prioridad	Baja	**Necesidad**	Baja
Objetivo	Tasa de conversiones exitosas		
Descripción	La tasa de documentos transformados al formato .txt con éxito será superior al 70%.		
Precondiciones	Documentación adquirida		
Trazabilidad	RFD-021		

Figura 227. Requisito RND-024

RFB-025			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Inserción de terminología en BoilerPlates		
Descripción	La herramienta BoilerPlates recibirá como entrada el fichero que contiene los conceptos del dominio agrupados en semánticas.		
Precondiciones	Fichero de agrupación en semánticas generado		
Trazabilidad	RFA-014, RNB-030		

Figura 228. Requisito RFB-025

RFB-026			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Baja	**Necesidad**	Baja
Objetivo	Inserción de texto de análisis en BoilerPlates		
Descripción	La herramienta BoilerPlates recibirá como entrada los documentos del dominio convertidos al formato .txt.		
Precondiciones	Documentación adquirida y procesada		
Trazabilidad	RFD-021, RND-024, RNB-030		

Figura 229. Requisito RFB-026

RFB-027			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Realización del análisis de patrones		
Descripción	La herramienta BoilerPlates realizará la obtención de los patrones lingüísticos mediante el uso de los ficheros de entrada.		
Precondiciones	Documentación y terminología insertada en la herramienta		
Trazabilidad	RFB-025, RFB-026, RNB-031		

Figura 230. Requisito RFB-027

RFB-028			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Escenarios de extracción de patrones		
Descripción	La herramienta BoilerPlates realizará ocho procesos de análisis, que representarán los escenarios de estudio.		
Precondiciones	--		
Trazabilidad	RFB-027		

Figura 231. Requisito RFB-028

RFB-029			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Alta	**Necesidad**	Alta
Objetivo	Ficheros de salida de BoilerPlates		
Descripción	La herramienta BoilerPlates generará como salida una copia de la base de datos que contenga los patrones extraídos en el proceso de análisis.		
Precondiciones	Análisis de patrones completado		
Trazabilidad	RFB-027, RFB-028		

Figura 232. Requisito RFB-029

RNB-030			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Baja	**Necesidad**	Baja
Objetivo	Herramienta de gestión de BBDD		
Descripción	La herramienta de gestión y manipulación de bases de datos será Microsoft Access.		
Precondiciones	--		
Trazabilidad	RFB-025, RFB-026, RFB-029		

Figura 233. Requisito RNB-030

RNB-031			
Fecha	05/05/2016	**Fuente**	BoilerPlates
Prioridad	Media	**Necesidad**	Alta
Objetivo	Tiempo de completitud del análisis		
Descripción	El tiempo necesario para completar el estudio de los ocho escenarios será inferior a 30 días.		
Precondiciones	--		
Trazabilidad	RFB-027, RFB-028		

Figura 234. Requisito RNB-031

RFR-032			
Fecha	07/16/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Extracción del número total de patrones		
Descripción	Por cada escenario de estudio se obtendrá el número total de patrones.		
Precondiciones	Ficheros de salida de patrones generados		
Trazabilidad	RFB-028, RFB-029		

Figura 235. Requisito RFR-032

RFR-033			
Fecha	07/16/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Extracción del número total de patrones		
Descripción	Por cada escenario de estudio se obtendrán los 100 patrones más frecuentes en el proceso de análisis.		
Precondiciones	Ficheros de salida de patrones generados		
Trazabilidad	RFB-028, RFB-029		

Figura 236. Requisito RFR-033

RFR-034			
Fecha	07/16/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Extracción del número total de patrones		
Descripción	Por cada escenario de estudio se obtendrá el número de patrones que contengan al menos un término del dominio de estudio.		
Precondiciones	Ficheros de salida de patrones generados		
Trazabilidad	RFB-028, RFB-029		

Figura 237. Requisito RFR-034

RFR-035			
Fecha	07/16/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Extracción del número total de patrones		
Descripción	Por cada escenario de estudio se obtendrán los 100 patrones que contengan al menos un término más frecuentes en el proceso de análisis.		
Precondiciones	Ficheros de salida de patrones generados		
Trazabilidad	RFB-028, RFB-029		

Figura 238. Requisito RFR-035

RFR-036			
Fecha	07/16/2016	**Fuente**	Análisis previo
Prioridad	Media	**Necesidad**	Media
Objetivo	Extracción del número total de patrones		
Descripción	Por cada escenario de estudio se obtendrán los conceptos más frecuentes en el proceso de análisis.		
Precondiciones	Ficheros de salida de patrones generados		
Trazabilidad	RFB-028, RFB-029		

Figura 239. Requisito RFR-036

Anexo D. Conversor de PDF a TXT

Para realizar la conversión de los documentos de su formato original (pdf) al formato admitido por BoilerPlates (txt), se ha hecho uso de la herramienta **pdf2txt**, desarrollada en el lenguaje Java por David Catalán.

A continuación, se definirá su composición y funcionamiento:

- **Código fuente**. Se puede encontrar en la carpeta src. Está compuesto por dos clases:
    - **PdfConverter**. Con la clase propia. Tiene dos métodos para realizar el proceso de *parseado*: el método "Estándar" y el adaptado para el analizador semántico.
    - **PDFTextStripper_Own**: Es una adaptación de la clase de pdfbox, para mejorar el proceso de separación por párrafos.

- **Paquetes Java incluidos:**
    - Librerías libres de APACHE
    - commons-logging-1.2.jar
    - fontbox-1.8.8.jar
    - pdfbox-1.8.8.jar
    - pdf2txt.jar (es el paquete resultante de la compilación del código fuente indicado previamente.

- **Para compilar la herramienta:**
    - **En el entorno de programación Eclipse**. El paquete está preparado para ser importado desde Eclipse. Al ejecutar la herramienta, se realizará su compilación automáticamente.
    - **En la terminal de Windows (cmd)**. `compile.cmd`
    - **En la terminal de Unix**. `compile.sh`

- **Parámetros de entrada.** Son necesarios tres parámetros:
    - **Directorio de entrada**. Ruta en donde se encuentran los ficheros en formato pdf que serán convertidos.

- **Directorio de salida**. Ruta en donde se encontrarán los ficheros en formato txt tras realizar la conversión.

- **Número de palabras**. Es el número mínimo de palabras para considerar la existencia de un párrafo en el texto. Si el número de palabras de un párrafo es inferior a este valor, no se tiene en consideración. Este parámetro no se utiliza si se emplea la versión "Standard" del programa.

- **Para ejecutar la herramienta**:

  - **En el entorno de programación Eclipse**. Incorporar el launcher "PdfConverter.launch".

  - **En la terminal de Windows (cmd)**. `pdf2txt.cmd`

  - **En la terminal de Unix**. `pdf2txt.sh`

Anexo E. Resumen del proyecto en inglés

E.1. Introduction

In this project, we will attempt to solve the following problem: is it possible to make a computer recognize linguistic patterns in a document? For example, when reading the sentence:

"*The patient had a temperature of 102F at Saturday*"

"Determiner + Noun + Verb + Preposition + Noun + Preposition + Noun + Preposition + Noun"

Is it possible for the computer to understand it both syntactically and semantically? Moreover, once this information is acquired, is it possible to generate patterns that allow recognizing similar sentences throughout a document?

With this problem as a starting point, this Final Degree Project was developed at the Carlos III de Madrid College. This project is related to the studies of Natural Language Processing (NLP).

For this project, by using vocabulary from a specific domain, the computer will read a number of documents and automatically generate all linguistic patterns contained in them. Therefore, there are **two main objectives**:

- Acquire the terminology from a specific domain.
- Perform the linguistic pattern learning process by reading and analysing a number of documents from the same domain.

The chosen domain for this project was "**Genetic Deafness**" (medicine), and the language used was **English**.

This annex is a summary of the project, which covers the followed work procedure and the results of the experimentation, as well as the conclusions reached after the process was completed.

E.2. Motivation

There are two potential motivations for this project:

- **Effective information extraction from a text document**. When a human reads a sentence, he or she extracts a piece of the information that he or she considers useful. See the previous example: "*The patient had a temperature of 102F at Saturday*". From this sentence, we can retain the information related to the temperature (102F) and/or the time when the measure was taken (at Saturday).

  For a computer, this "understanding" process is difficult. There are no trivial algorithms that, after reading the sentence, can storage the knowledge received about the patient's temperature and the day the measure was taken. A recurrent solution for this problem is to use templates and file formats, which allows the computer to receive all the information without actually understanding it. This, however, forces the information to be pre-processed so it can meet software specifications, which consumes more time and effort.

  A previous learning of linguistic patterns may solve this problem. If the computer is able to recognize the example sentence above as a learned pattern, then it would realize that the seventh word indicates the temperature (102F) and the last one indicates the day (Saturday).

  Therefore, the pattern analysis could be useful to acquire all relevant information from a document of variable length.

- **Writing style guide development**. In several fields, like law or medicine, there are specific writing styles to follow. Therefore, the results from this project in a specific domain can be useful in a didactic way. For example: which word families are more frequent in the domain? Which linguistic sequences are more common? And so on.

  By studying the linguistic patterns from a domain, we can contribute to the learning of the writing style expected in a document, in order to show all the information in an organized state.

  This is also helpful for computers: when the writing style is standardized, the pattern recognition is easier and, therefore, the information retrieval is easier and more efficient.

E.3. State of the art

E.3.1 Information Retrieval

Since the dawn of the digital era, the amount of information available for a human has increased exponentially. And, as of today, this amount is growing faster than ever.

All this data is presented to the user in several ways, such as digital newspapers, websites, books, and so on. However, this leads us to a problem: with all the information we can find, is it possible to extract the information we require in an effective way?

This is known as the "information overload" problem. The massive amount of data and the availability of the tools used to extract it make it difficult to acquire what we need. And yet, this is a vital process in our society. Several authors, such as Lewis D. (1996), stated that being able to process a huge amount of information is crucial for modern life.

From this problem, the discipline of Information Retrieval (IR) arose, which studies the algorithms and techniques used to efficiently extract information from databases or any digital source. The origins of IR date back to the 1940's, with the first computers capable of performing data queries. Nowadays, it has become more relevant due to the rise of the Internet.

E.3.2 Natural Language Processing

Natural Language Processing (NLP) is a discipline of Artificial Intelligence related to the research and development of mechanisms that allow effective communication between a computer and human language. NLP is used nowadays in automatic translation, speech recognition and text summaries, among other systems.

The first steps of NLP date back in the 1950's, with the publication of the famous "Turing Test" (known as "*Computing machinery and intelligence*"), in which Alan Turing argued that computer could possibly "have intelligence". In 1956, the "Georgetown Experiment", a collaboration between IBM and the Georgetown College, was presented. In the experiment, 60 sentences written in Russian were accurately translated into English.

These successes made the governments interested in this discipline, and began founding the research. It showed great premise at first. However, as automatic translation was growing more difficult, certain problems such as ambiguity arose and the limitations of the computer technology of that time, the research slowed down to the point that the founding was severely decreased.

In the last years, NLP is becoming relevant again, thanks to projects like “Google Translate” or the EU based “Molto”. Nowadays, a large amount of text can be processed with considerable accuracy.

One of the main challenges NLP has tried to overcome since its beginnings is the problem of ambiguity. In human language, it is common for a word to have multiple meanings, and even one sentence could be interpreted in several ways. A typical example of ambiguity is, for example: “The passer-by helps dog bite victim”. Is the passer-by helping a dog bite someone, or is the passer-by helping someone who was bitten by a dog? A computer could try to interpret the actual meaning of the sentence, but it is unlikely that it will succeed.

E.4. Work methodology and procedures

The methodology of this project is based on the usage of concepts related to genetic deafness in order to recognize the most frequent patterns from a number of documents while facing the problems and limitations from the Natural Language Processing.

In this project, the **BoilerPlates** tool has been used. This tool, developed by the Carlos III de Madrid College research team, can perform the analysis of linguistic patterns given a previously learned set of concepts and grammar categories. Its database already contains vocabulary that is not related to any domain (such as the verb “to be”, prepositions, adverbs…). However, to effectively use this tool in a certain domain, the database must be updated with domain-specific concepts prior to the pattern extraction.

The methodology used for this project is divided in the following steps:

1. **Acquisition and processing of domain documents**. In order to perform an accurate research about linguistic patterns in the domain of genetic deafness, acquisition of documents that are related to this topic is required.

   To solve this, a batch of scientific publications were requested to the Hospital Ramón y Cajal Genetic Services. Once the documents were received, an external tool was used to adapt the content to a format BoilerPlates accepts. Further explanation can be found in section 7.1.

2. **Extraction of related terminology**. To perform the analysis, the specific terminology of the domain is also required. This information must be retrieved from an official medicine organization.

The chosen organization was the US National Library of Medicine (NLM). Founded in 1836, it is the largest library of medical terminology of the world. Its contents can be access online and free of charge with the MeSH thesaurus, introduced in paper in 1960. Once the vocabulary related to the “Genetics” and “Deafness” were located, a parser was developed in order to extract all the information with near perfect accuracy in a legible format. Further explanation can be found in section 7.2.

3. **Adaptation of terminology to BoilerPlates**. Once the terminology was acquired, it had to be processed in order to be accepted by BoilerPlates.

   First, all domain concepts must be clustered in semantic groups in order to perform the semantic analysis. For this reason, an algorithm was developed to automatically generate a variable number of groups in which the concepts were categorized according to its identification string at the MeSH thesaurus. An in-depth explanation of the algorithm can be found in section 7.3.

   Once the clustering process was complete, the concepts were introduced in the BoilerPlates’ databases. More information about the databases used for the linguistic analysis can be found in section 7.4.

4. **Using the BoilerPlates tool**. Once the clustered terminology in BoilerPlates were inserted and the documents related to genetic deafness were processed, the analysis was ready to begin.

   Eight different scenarios were identified for this study, each one using a unique setting. Two variable were modified for each analysis: the minimal frequency to accept a sequence as a pattern, and considering (or ignoring) concept semantics. For each scenario, a batch of results were obtained. The specific values for each scenario will be mentioned later on this summary.

   See section 7.5 for a detailed explanation on how to use the BoilerPlates user interface, and refer to sections 7.6 and 7.7 to learn more about the pattern identification process.

5. **Extraction of results**. Once the pattern identification process was complete for all scenarios, the results were adapted in order to show the information in a legible manner. This is further explained at section 7.8.

6. **Exposition and analysis of results.** The results for each scenario can be found at section 8. The conclusions reached after the analysis of all scenarios are explained at section 9. A brief summary of the results can be found later on this summary.

E.5. Results of the experimentation

Eight scenarios were identified for this study:

1. Minimal frequency **1**, **WITHOUT** semantic distinction.
2. Minimal frequency **5**, **WITHOUT** semantic distinction.
3. Minimal frequency **10**, **WITHOUT** semantic distinction.
4. Minimal frequency **20**, **WITHOUT** semantic distinction.
5. Minimal frequency **1**, **WITH** semantic distinction.
6. Minimal frequency **5**, **WITH** semantic distinction.
7. Minimal frequency **10**, **WITH** semantic distinction.
8. Minimal frequency **20**, **WITH** semantic distinction.

Once the scenarios were individually analysed, the combined results were as it follows:

E.5.1 Number of patterns identified

As expected, the lower the minimal frequency (MF) value chosen, the more patterns are considered in the identification process.

The growth of the number of patterns found is inversely exponential: in the experiments with MF established to 1, about 140.000 patterns were found. When raising the MF to 5, the number drops at 14.500 on average, which is about a 90% decrease on patterns found. On the experiments with MF set to 10 and 20, the number of patterns were approximately 5.600 and 2.800 respectively, which is a considerable drop but less severe than before.

When increasing the MF value, we force all possible linguistic sequences to have a higher number of repetitions in order to be considered a pattern. Therefore, in the most relaxed case (MF set to 1), all sequences can become a pattern as long as it has at least one repetition throughout the text. In the strictest case (MF set to 20) each sequence needs at least 20 repetitions to become a pattern. For that reason, the total number of patterns decreases considerably.

Applying semantics to the pattern extraction did a minor difference on the number of patterns. When MF was set to 1 and semantic distinction was activated, the total number increased by about 1.100 patterns (0.8% increase). However, when MF was set to 5, 10 and 20 the number decreased slightly.

It could be possible that, on the most relaxed cases, the semantic analysis allows identification of patterns that would not be detected otherwise. In stricter experiments, it is harder to find sequences with the required minimal

frequency, since the semantic group for each component in the sequence must match.

E.5.2 Most frequent patterns

For each scenario, the 100 most frequent patterns were extracted, as shown in section 7 of this document. A quick observation of the results shows that the most frequent patterns tend to be the same.

Unfortunately, the majority of the identified patterns have one or several concepts categorized as “1144” (Unclassified Noun). This represents all terms that are not recognized by BoilerPlates. Therefore, the most frequent patterns were sequences like “Un. Noun + Un. Noun + Symbol” and so on.

However, after a filtered search, frequent patterns in which all terms are known have been found. These are simple sequences in the English language, such as “Verb to be + Verb” or “Noun + Preposition (of)”. This implies that the pattern extraction was actually successful, although the results would improve greater if we were to add more terminology.

E.5.3 Patterns related to genetic deafness

After analysing the results as a whole set of patterns, a subset was created for each scenario, containing only those patterns which has at least one concept related to genetic deafness (“domain patterns”).

This subset was a minority in every studied case. For all scenarios, about 10 to 16% of the patterns contained at least one concept from the domain. This was predictable, as most of the words contained in one document about a certain topic are not terminology from the domain, such as prepositions, numbers or unrelated nouns.

In addition to this, most of the domain patterns have at least one concept unclassified, like “Un. Noun + Genes, vif” or “Exons + Number + Un. Noun”.

However, several valuable domain patterns were found, such as “Definite Article + Mutation”, “Number + Genes” or “Genes, Essential + Preposition (for)”.

E.5.4 Most frequent concepts

The next step on the study was to analyse what kind of concepts were the most frequent in the domain. To do this, the repetitions of all concepts in the patterns were counted, and then compiled in a top 10 list for each scenario.

On the one hand, when considering only terms related to “genetics”, most of the domain patterns found had a term which was related to either a specific gen or a transformation (mutation) of a gen.

On the other hand, when considering only terms related to “deafness”, several types of deafness appeared, such as bilateral, unilateral, high-frequency, etc.

A later experiment determined the percentage of all repetitions of domain concepts in the patterns for one scenario. In the most relaxed cases (MF set to 1), about 30% of all concepts in the pattern were related to genetic deafness, while in the strictest cases (MT set to 20) this value decreases to about 12%.

In addition to this, the proportion between “genetics” concepts and “deafness” concepts used in the patterns is rather stable: in all scenarios, about 85% of all domain concepts appearing in the results were related to “genetics”, while the remaining 15% were the ones corresponding to “deafness”.

E.5.5 Weighted study

The final experimentation performed to the combined results was to apply a “weighting formula” to all identified patterns for each scenario. The aim of this experiment was to highlight the longest patterns and those patterns in which all concepts were known (that is, they do not have any “unclassified nouns”). The formula is explained at the beginning of section 8.

Once the formula was applied, several common English word sequences appeared as the most valuable patterns. For example, on the scenarios in which the semantic distinction was disabled, the pattern with the highest weight value was “Definite Article + Noun + Preposition (of)”.

On the other hand, on the scenarios in which the semantic distinction was activated, the most valuable pattern was “Verb to be + Verb + Preposition”.

In addition to these patterns, other valuable sequences were:

- “Definite Article + Not Grouping Noun + Preposition (of)”.
- “Verb to be + verb + preposition + Indefinite article”.
- “Verb to have + Verb to be + verb + preposition (to)”.

There were even some domain patterns with a high weight value in at least one scenario, such as “Exons + Indefinite article + Symbol”.

One last aspect worth mentioning is that there was a high frequency of prepositions in the most valuable patterns, such as “of”, “like” or “about”. This was expected; in the English language, the number of prepositions is very low and stable when compared to other syntactic categories, such as verbs or nouns, and, therefore, it is easier to use them in patterns.

Over time, a language is expected to grow. This is particularly relevant in nouns and verbs. For example, in the 1950's words and expressions like "smartphone" or "to google [something]" were not used. However, other categories, such as prepositions, are most of the time constant in size.

E.6. Final remarks

Once the analysis of the results was completed, we reached the following conclusions regarding the pattern identification and studio in the domain of genetic deafness:

1. The results offered by the BoilerPlates are coherent. Although most generated patterns had at least one undefined concept, several sequences that can be found in the English language were discovered, such as "Verb to be + Verb". Albeit less common, there were also patterns that used concepts from the domain.

2. The minimal frequency value made a considerable impact in the analysis. The higher the value, the lower the number of identified patterns, since a sequence must reach a bigger number of repetitions until it is considered a pattern.

3. On the matter of semantic distinction:

    - With respect to percentage, the number of patterns discovered on the scenarios that applied semantic distinction was similar to those scenarios that did not. It also worth noting that, when using very low MF values, using semantics can contribute to increasing pattern detections.

    - The first experiments showed that there was little difference whether applying semantics or not. However, when the weighted study was completed, it was discovered that semantics played a considerable role during the analysis, such as which patterns were considered the most valuable.

4. Of all patterns discovered, around 10% to 16% had at least one domain concept. This was expected, since the specific terminology of the domain is only a small subset of all the possible words that can be used in a document about genetic deafness.

5. The most frequent concepts that are related to "genetics" were those that are about gene manipulation and mutation.

6. Most concepts related to “deafness” were used at least once in all the scenarios, such as neurosensory, unilateral, bilateral, etc. However, the concepts “Wolfram Syndrome” and “Usher Syndrome” were the least used concepts from this subset.

E.6.1 Future lines of work

There have been considerations about future lines of work that may use this project as a starting point:

1. Due to lack of time, only eight case scenarios were tested in this project. It could be possible to repeat all the experimentation using a considerably higher MF value, such as 50, 100 or 200.

2. Semantic distinction was proven influential on the analysis, but only after the weighted study was completed. Another possible project would be repeating the concept clustering process in order to generate a new set of semantics to use in the process.

3. The most common grammatical category used in the patterns was “Unclassified noun”, due to lack of vocabulary in BoilerPlates ‘databases. Aside from the domain-specific concepts, more terms from all fields could be included in the databases, in order to generate results that are more significant.

4. As of today, the algorithms and parsers used in this project to extract, manipulate and insert the MeSH concepts into BoilerPlates are separated Java classes and packages, which makes difficult for an external user to perform all steps defined in this project. There could be improvements on this matter, such as implementing a single program, which contains all algorithms linked together, and a graphic interface, that could extract all concepts from the thesaurus and pre-process them in a simple, comfortable manner.