# Agrupación de entidades semánticas en oraciones

Montón Munuera, Lidia

Curso 2013-2014

Director: LEO WANNER

GRADO EN INGENIERÍA INFORMÁTICA

Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Trabajo de Fin de Grado

# Agradecimientos

Primero de todo quiero agradecer a Nadjet Bouayad el haberme aceptado para trabajar en su propuesta de Trabajo de Fin de Grado, sin ella no hubiese encontrado un reto tan interesante como el planteado dentro de la Generación de Lenguaje Natural. Asimismo, he de agradecer a Leo Wanner que haya accedido finalmente a ser mi tutor en este trabajo cuando se le solicitó de forma improvista y que haya podido guiarme y ayudarme en el desarrollo del mismo.

Quiero dedicar este trabajo a mi familia, que han mostrado una enorme paciencia y apoyo incondicional incluso en mis momentos más difíciles. Especialmente a mi pareja, Sergi, por las críticas y opiniones que han sido claves para mejorar este trabajo.

## Resumen

En este trabajo se propone un módulo del sistema de planificación de textos. El objetivo es formular una solución para la agrupación en oraciones de unidades de contenido pertenecientes a una base de conocimiento. La solución se basa en la representación de las unidades de contenido en términos de etiquetas semánticas, agrupadas según los datos extraídos de un corpus con textos de referencia. Este punto de vista se ha utilizado en pocos trabajos, ya que generalmente las agrupaciones para formar oraciones se realizan en términos de $Sujeto + Predicado + Objeto$, y no dinámicamente como se plantea en este trabajo. Mediante los experimentos realizados se demuestra que pueden obtenerse resultados satisfactorios en la planificación de textos utilizando este método.

## Abstract

In this work a text planning system module is proposed. The objective is to provide a solution for grouping of concepts in sentences, given a knowledge base. The solution is based on the annotation of sentences of a training corpus with semantic role labels and subsequent learning of which semantic roles usually go together and how large sequences of semantic roles usually are. This approach is novel, since normally the composition of concepts into sentences follows a rather simple pattern of $Subject + Predicate + Object$, which is not dynamic, as is proposed in the present work. Experiments demonstrate that satisfactory results can be obtained in text planning using the proposed method.

# Índice

Agradecimientos ..... iii
Resumen ..... v
Lista de figuras ..... ix
Lista de tablas ..... x
1 Introducción ..... 1
2 Hoja de ruta del trabajo ..... 3
3 Estado de la cuestión ..... 5
3.1 Generación de Lenguaje Natural (GLN) ..... 5
3.1.1 Visión global de sistemas de generación de textos ..... 5
3.1.2 Planificación de textos ..... 6
3.2 Ontologías ..... 10
3.3 Uso de Plantillas en PLN ..... 12
3.3.1 Extracción de información ..... 12
3.3.2 Semántica de casos ..... 13
3.4 Conclusiones del Estado de la Cuestión ..... 15
4 Descripción del conjunto de datos utilizados ..... 17
4.1 El corpus de entrenamiento ..... 17
4.2 Etiquetas ..... 17
4.3 Herramienta de anotación ..... 19
4.3.1 RDFaCE ..... 19
4.3.2 Brat Rapid Annotation Tool ..... 20
5 Enfoque tomado para la agrupación de contenido en oraciones ..... 25
5.1 Reflexiones preliminares ..... 25
5.2 Método ..... 27
5.2.1 Extracción de plantillas ..... 27
5.2.2 Medidas ..... 28
5.2.3 Algoritmo de extracción de datos ..... 31
5.2.4 Algoritmo de agrupación de unidades de contenido ..... 32
6 Experimentos y Resultados ..... 39
6.1 Algoritmo de baseline ..... 39
6.2 Métricas de evaluación ..... 40
6.3 Análisis de los resultados ..... 41

6.4 Detalle de cotas superiores e inferiores ..... 43
6.4.1 Análisis de la precisión y cobertura..... 44
6.4.2 Análisis del tipo de anotaciones ..... 46
7 Conclusiones y Trabajo futuro ..... 49
8 Referencias ..... 51
9 Anexos..... 53
9.1 Plantillas extraídas ..... 53
9.2 Clasificación de etiquetas en detalle ..... 62

## Lista de figuras

2.1 Diagrama de Gantt ..... 4
3.1 Esquema de GLN ..... 6
3.2 Ejemplo de relaciones RTS ..... 7
3.3 Sistema de E. Hovy ..... 8
3.4 Grafo de la plantilla Identificación ..... 9
3.5 Predicados de la plantilla Identificación ..... 10
3.6 Relaciones entre recursos en RDF ..... 11
3.7 Representación de FRED ..... 12
3.8 Plantilla de EXIT ..... 13
3.9 Demo de anotación de FrameNet ..... 14
3.10 Ejemplo etiquetado de FrameNet ..... 15
4.1 Frase de la biografía de A. Schmidt ..... 17
4.2 Oraciones de B. Stanwyck y A. Boshle ..... 19
4.3 Demo de RDFaCE ..... 20
4.4 Cargar un documento en Brat ..... 21
4.5 Etiquetas en Brat ..... 21
4.6 Comparación del procesado de la biografía ..... 22
4.7 Fichero Ryan_Adams.ann ..... 23
5.1 Ejemplo de tripletas ..... 26
5.2 Frase de S. Soni ..... 27
5.3 Ejemplo de registro de plantillas ..... 28
5.4 Ejemplo de identificación de bigramas ..... 30
5.5 Etiquetas de niveles diferentes ..... 31
5.6 Ejemplo de plantillas ..... 31
5.7 Clasificación en plantillas ..... 32
5.8 Ejemplo de núcleos ..... 34
5.9 Clasificación de etiquetas nivel 1 ..... 36
6.1 Ejemplo de Precisión y Cobertura ..... 40
6.2 Promedio de Valor-F ..... 43
6.3 Frases anotadas de K. McPhee ..... 47
6.4 Frases anotadas de F. Gehry ..... 48

## Lista de tablas

Tabla 6-1 Resultados de A. A. Allen ..... 41
Tabla 6-2 Agrupaciones realizadas y Valor-F ..... 42
Tabla 6-3 Resultados con valor-F promedio bajo ..... 44
Tabla 6-4 Resultados con valor-F promedio alto ..... 45
Tabla 6-5 Etiquetas de K. McPhee y F. Gehry ..... 46
Tabla 9-1 Etiquetas de A. E. Sutherland ..... 62
Tabla 9-2 Pendientes tras las plantillas ..... 63
Tabla 9-3 Pendientes tras agrupaciones ..... 64

# 1 Introducción

Independientemente de quién seas, a la hora de escribir tu biografía compartes ciertos aspectos con el resto del mundo: El nacimiento y, algún día lo más lejano posible, la muerte. Más allá de estos datos comunes para todos, cada vida es una historia diferente. Sin embargo, hay ciertos elementos que podemos agrupar en conjuntos: Personas con las que te relacionas, lugares que visitas u organizaciones a las que perteneces, entre otros. Estas características invitan a pensar que es posible organizar los componentes de un texto, una biografía en este caso, en oraciones basándonos en estas categorías, teniendo en cuenta qué datos pueden relacionarse entre ellos y cómo pueden agruparse. La idea de trabajar con conceptos organizados de una base de conocimiento ha sido analizada por estudios como el publicado por F. Draicchio et al., [1] en el cual las ontologías, entendidas como estructuras jerárquicas de conceptos, presentes en el texto se enriquecen con información adicional semántica utilizando "etiquetas" y mediante las anotaciones se puede mapear cómo deben estructurarse las unidades de conocimiento.

A partir de una base de conocimiento con información relativa a una persona, hemos planteado cómo ir seleccionando los diferentes conceptos que se encuentran dentro de ella y agruparlos de manera que el resultado final sea un conjunto de oraciones estructuradas con sentido. Es precisamente lo que hacemos los humanos al hablar: Sabemos lo que queremos decir, todo lo que queremos transmitir, pero decidimos aplicar un cierto criterio que nos obliga a agrupar todos esos datos en frases. En lugar de expresarlo todo en una única oración, las personas decidimos utilizar otras más cortas con fragmentos del discurso.

En general, hasta ahora las técnicas de agrupación de unidades de contenido en oraciones han sido simplificadas realizando agrupaciones con la estructura $Sujeto + Predicado + Objeto$, por lo que se han realizado pocos trabajos orientados a estrategias no triviales para la agrupación de conceptos en oraciones. Por ejemplo: *María come una manzana*, es una oración porque contiene palabras con estructura gramatical y transmite un mensaje. La diferencia principal en este trabajo es que se realiza un enfoque empírico del problema de la agrupación de unidades de contenido en una frase: Se pretende demostrar que un sistema puede aprender qué conceptos poner en una oración a partir del estudio de cómo las personas forman frases mediante textos de ejemplo. Por lo tanto, los conceptos presentes en el texto se agrupan dinámicamente, sin establecer estructuras predefinidas.

Para lograr la agrupación de todo el contenido en oraciones diferentes, es necesario primero identificar y clasificar las unidades de contenido del texto según su función, proceso explicado en el apartado 4. Posteriormente se aplica el algoritmo de agrupación de unidades de contenido, cuyos detalles se exponen en el apartado 5. Los experimentos de comprobación de los resultados del algoritmo realizados para verificar su corrección se encuentran, junto con sus resultados, en el apartado 6. Finalmente, se pueden encontrar las conclusiones del trabajo y algunas propuestas de mejoras en el futuro en la sección 7.

## 2 Hoja de ruta del trabajo

El desarrollo del trabajo se puede dividir en las siguientes fases:

1. Documentación y planteamiento: Lectura y análisis de estudios previos en los que se trataran temas relacionados con la Planificación de Textos. Las primeras lecturas han sido relativas a la Generación de Lenguaje Natural, para tener una idea general de los sistemas, y posteriormente se han ido escogiendo estudios más específicos relacionados con métodos de planificación de textos que permitieran tener claro el problema a resolver en este trabajo y cómo enfocarlo.

2. Diseño propuesta: Una vez establecido el ámbito y objetivo del trabajo, es necesario establecer el método que se aplicará para resolver el problema planteado, así como los datos necesarios (en el apartado 4 se detallan qué datos se han utilizado y en el 5 se especifica la metodología seguida).

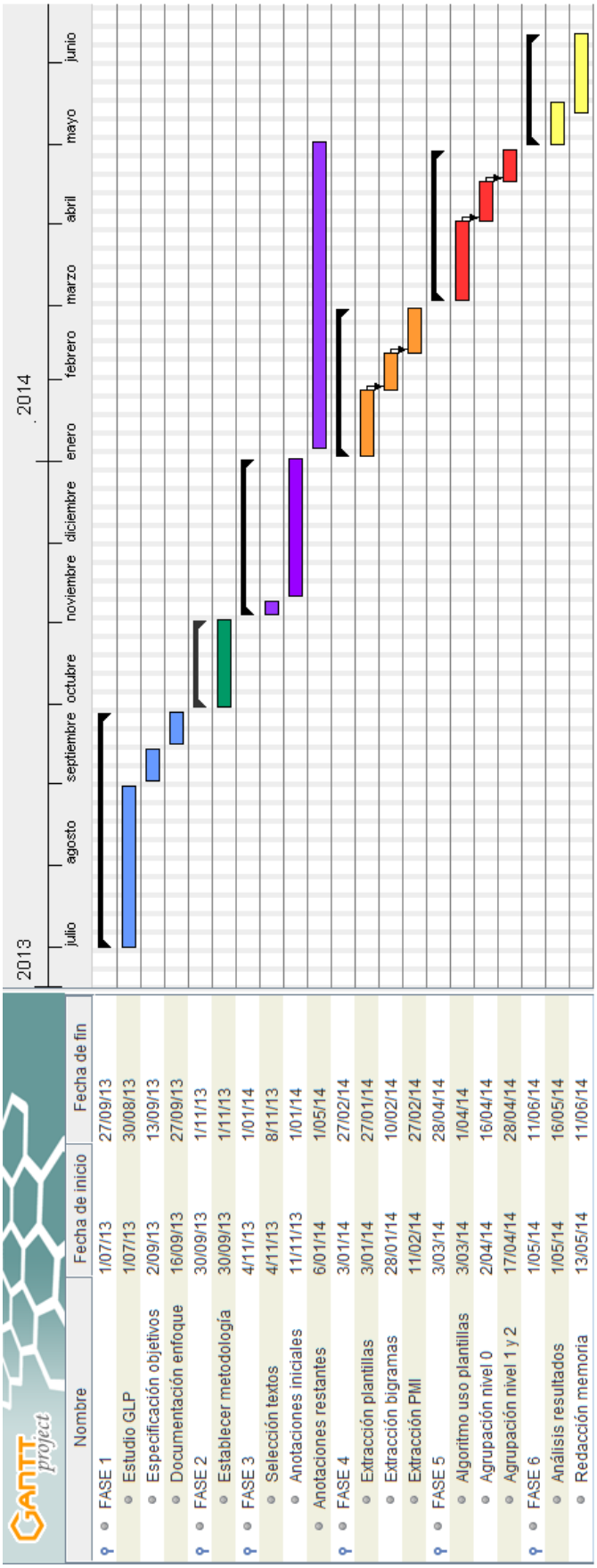
3. Anotación del corpus: La tarea de etiquetado de los conceptos presentes en los textos del corpus de entrenamiento se ha ido desarrollando también en los puntos siguientes, puesto que es una labor tediosa y extensa debido a que queríamos anotar unos 100 textos. Sin embargo, se requieren unas anotaciones mínimas para poder realizar comprobaciones en los algoritmos diseñados según se van elaborando.

4. Implementar algoritmo de obtención de datos: Se requiere un código que permita extraer las unidades de contenido del corpus de entrenamiento necesarias como referencia para el algoritmo encargado de realizar las agrupaciones de conceptos (plantillas, bigramas y PMI, tal como se explicará en el apartado 5).

5. Implementar algoritmo de resolución: Una vez se dispone de todos los datos necesarios, debe diseñarse el algoritmo que realizará las agrupaciones de conceptos en oraciones e implementarlo. Este algoritmo puede desglosarse en 3 partes según las agrupaciones que realiza: Primero aplica todas las plantillas que sea posibles y después realiza agrupaciones de etiquetas sin plantilla (pertenecientes a diferentes niveles jerárquicos). Los detalles concretos de este algoritmo están en el apartado 5.2.3.

6. Análisis de resultados y redacción de memoria: Ejecución del algoritmo y comprobación de sus resultados, los cuales se analizan en la memoria del trabajo. Asimismo, en esta fase se escribe también el resto de información necesaria para documentar el estudio.

Estas fases se pueden observar representadas gráficamente en el diagrama de Gantt de la imagen 2.1.




**2.1 Diagrama de Gantt**

# 3 Estado de la cuestión

Como se ha indicado previamente, este trabajo se encuentra dentro del campo del Procesamiento del Lenguaje Natural, el cual se define como un *área de investigación y aplicación que explora cómo los ordenadores pueden usarse para entender y manipular el lenguaje natural* [2]. El presente trabajo se encuentra en uno de los ámbitos de PLN: La generación de textos a partir de representaciones semánticas como se observan en el caso de ontologías. Nos centramos en la tarea de Planificación de Textos, la cual tiene como objetivo determinar qué contenido es necesario en un texto y cómo se debe estructurar para que tenga sentido, para lo que hay diferentes formas de plantearlo. En concreto, el problema abordado consiste en desarrollar una técnica para la agrupación de unidades semánticas en oraciones utilizando técnicas de aprendizaje.

A continuación se detalla la base teórica sobre la que se sustenta este trabajo, situada dentro del ámbito de la Generación del Lenguaje Natural, exponiendo la definición de las "ontologías" (desde el punto de vista de "estructuras jerárquicas de conocimientos") y la justificación del uso de Plantillas como solución en problemas como el planteado en este trabajo.

## 3.1 Generación de Lenguaje Natural (GLN)

En este apartado se definen las tareas realizadas en la Generación del Lenguaje Natural, dando una visión general de estos sistemas y destacando aquellos aspectos relevantes para el presente trabajo. Asimismo, se contrasta la visión "top-down" y "bottom-up" de Planificación de Textos, tarea en la cual se centra nuestro trabajo.

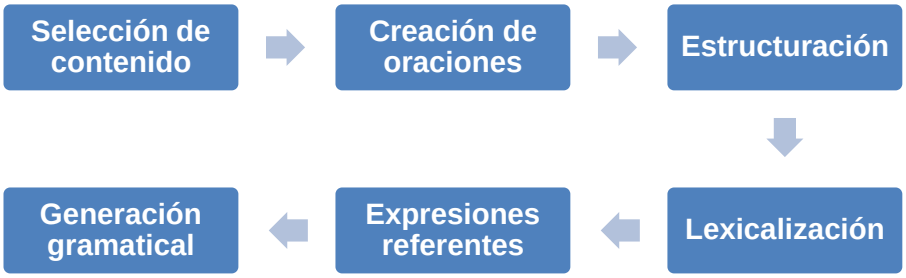
### 3.1.1 Visión global de sistemas de generación de textos

Un sistema de generación de textos debe ser construido teniendo en cuenta las diferentes tareas que se deben llevar a cabo en este proceso hasta obtener el texto final. Las actividades básicas en GLN son [3] [4] :

1. Selección de contenido: Decidir qué información debe incluirse en el texto. Se seleccionan aquellas unidades de contenido del fichero de entrada que se comunicarán en el texto, agrupándolas en mensajes que después se utilizaran en las fases siguientes de generación del lenguaje para darles forma.
2. Agregación de oraciones: Agrupar los mensajes en frases, o bien exponer mensajes diferentes en oraciones separadas.
3. Estructuración del contenido: Se impone un orden y una estructura en el conjunto de oraciones. En el texto debe haber un orden coherente que permita leerlo con facilidad y no transmitir información al azar.
4. Lexicalización: Decidir con qué palabras deben expresarse los conceptos y relaciones entre los mensajes.
5. Expresiones referentes: Generación de anáforas, es decir, sustituir cuando sea posible nombres de entidades por pronombres que hagan mención a ellas.

6. Generación gramatical: Asegurar la adecuación del texto en términos sintácticos, morfológicos y ortográficos.

En la imagen 3.1 se puede observar una representación esquemática del proceso de Generación del Lenguaje Natural explicado.




**3.1 Esquema de GLN**

Según el criterio del diseñador del sistema de generación de textos, algunas de estas tareas pueden ser realizadas simultáneamente.

Las tareas de un sistema de GLN tradicionalmente se agrupan en tres pasos clave:

1. Planificación de textos: Incluye la selección de contenido, la agrupación en oraciones y la estructuración del discurso.
2. Planificación de oraciones: Engloba el trabajo realizado a nivel de frase, el cual realiza las tareas 4-5 especificadas previamente (lexicalización y uso de anáforas).
3. Finalmente, se realiza la generación gramatical en un paso final que, tal como se ha explicado, procesa la sintaxis, morfología y ortografía del texto obtenido.

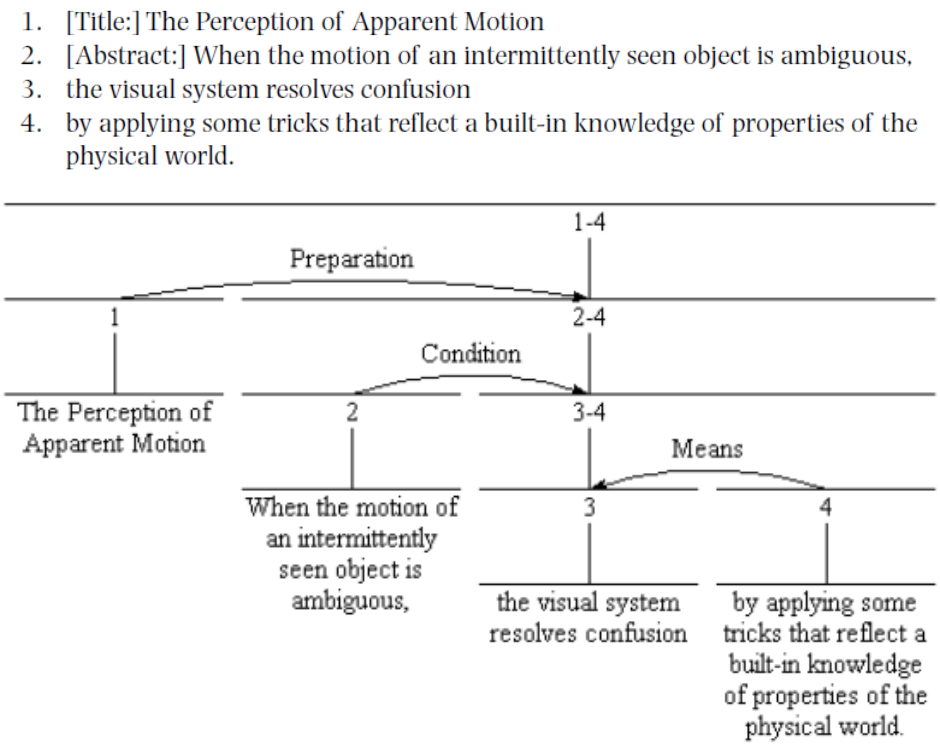
En este estudio, el trabajo se basa en la primera área, a nivel de planificación de textos. El objetivo es determinar qué decir y organizarlo coherentemente, mientras que se deja fuera del alcance del proyecto la especificación de qué términos exactos utilizar. A nivel de oración, solamente se analiza qué conceptos poner en ella, pero no se llega a realizar la tarea de dar estructura interna a las oraciones.

### 3.1.2 Planificación de textos

La planificación de textos puede plantearse desde dos puntos de vista opuestos: “top-down” y “bottom-up”.

El planteamiento bottom-up a menudo se basa en la Teoría de la Estructura Retórica (en inglés “Rhetorical Structure Theory”, con siglas RST), la cual propone representar jerárquicamente los fragmentos de un texto en una estructura de árbol empezando la organización de los elementos “desde abajo”. Así pues, en RST es básico establecer cómo los elementos anotados del texto se estructuran y combinan entre ellos mediante relaciones discursivas, lo cual permite asegurar que el texto creado es coherente. Muchas de estas relaciones contienen dos partes primarias: El Núcleo, que forma el centro, y el Satélite, que desempeña el papel de calificativo [5].

A continuación se muestra un ejemplo de representación de RST (imagen 3.2), donde se aprecia esta relación entre las unidades presentes en un texto. En ella se observan las relaciones RTS de un texto (compuesto por el título y el resumen), correspondiente a un ejemplo extraído del estudio de M. Taboada y W. Mann [6]. Cada flecha representa una relación que va desde el satélite hacia el núcleo, indicando sobre ella el tipo de relación. Por ejemplo, en la relación "Condition" el núcleo lo forman las unidades 3-4 y el satélite es la unidad 2.




**3.2 Ejemplo de relaciones RTS**

RST se formuló inicialmente como una teoría que permitía explicar cómo estaban planificados los textos, y fue E. Hovy quien formalizó la teoría para la generación de textos. En la imagen 3.3 se observa un ejemplo de un sistema propuesto por él basado en RST para la generación.

```
(GOAL (BMB SPEAKER HEARER (SEQUENCE-OF E1 ?NEXT))

((ENROUTE E1)                ((POSITION P1)              ((SHIP K1)
 (ACTOR E1 K1)                (HEADING P1 H1)             (NAME K1 KNOX)
 (DESTINATION E1 S1)          (LATITUDE P1 79)            (READINESS K1 C1)
 (NEXT-ACTION E1 A1)          (LONGITUDE P1 18))         ((PORT S1)
 (LOCATION E1 P1))           ((HEADING H1)                (NAME S1 SASEBO))
((ARRIVE A1)                  (COURSE H1 195))           ((DATE T1)
 (ACTOR A1 K1)               ((LOAD L1)                   (DAY T1 24)
 (TIME A1 T1))                (ACTOR L1 K1)               (MONTH T1 4))
 (NEXT-ACTION A1 L1))         (STARTTIME L1 T2)          ((DATE T2)
((READINESS-STATUS C1)        (ENDTIME L1 T3))            (DAY T2 25)
 (NAME C1 C4))                                            (MONTH T2 4))
                                                         ((DATE T3)
                                                          (DAY T3 28)
                                                          (MONTH T3 4))
                                  |
                                  |
                              SEQUENCE
                               /     \
                              /       \
                           CIRC      SEQUENCE
                           /  \        /  \
                        ELAB   ELAB   A1   L1
                        /  \   /  \
                      E1   C1  P1  H1

Knox, which is C4, is en route to Sasebo. It is at 79N 18E heading
SSW. It will arrive on 4/24, and will load for four days.
```

**3.3 Sistema de E. Hovy**

En el ejemplo de Hovy mostrado en la imagen 3.3 se observa cómo el diagrama de la estructura RST de los elementos de un párrafo tiene aspecto de árbol, el cual se ha formado poniendo a la izquierda el núcleo y a la derecha el satélite de cada relación. En la parte superior de la imagen están las afirmaciones que enlazan los diferentes elementos, y el texto de la parte inferior es el resultado obtenido.

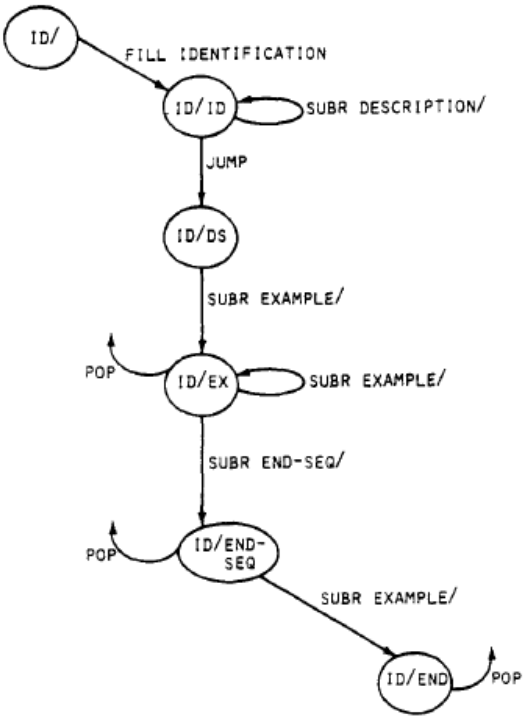
El método bottom-up permite empezar la estructura del discurso seleccionando conceptos individuales y buscando relaciones binarias con otros conceptos. Es decir, teniendo en mente la organización de un árbol de jerarquías se puede considerar que RST empieza la estructura trabajando con pequeñas partes del árbol y termina juntándolas para completarlo.

En contraste con este enfoque de planificación de textos, se encuentra el “top-down”, el cual realiza la estructuración de las unidades de contenido basándose en el objetivo del discurso. La aproximación top-down más relevante para el presente trabajo es la propuesta por McKeown, en la cual se utilizan “Plantillas” dentro de las cuales se introduce la información que se desea transmitir. Así pues, mientras que en los métodos bottom-up se estructuran los conceptos de modo dinámico y el discurso se construye recursivamente mediante composiciones sucesivas de relaciones, al utilizar plantillas para estructurar el contenido desde el principio se determina qué se quiere transmitir y cómo.

K. McKeown [7], basa su trabajo en el conocimiento semántico de la información relevante para determinar cómo estructurar lo que se quiere decir. Identifica como unidad básica del discurso lo que llama "afirmaciones retóricas", las cuales describen la función de cada oración. Algunos ejemplos ilustrativos de las afirmaciones retóricas de su trabajo son:

1. *Afirmación atributiva: "Mary tiene un abrigo rosa".*
2. *Afirmación equivalente: "Los vinos descritos como 'fabulosos' son buenos vinos provenientes de un buen lugar".*
3. *Afirmación de especificación: "Mary es bastante grande. Ella pesa 180 kilos."*

El sistema utilizado contiene plantillas, las cuales son representaciones predefinidas de estructuras de párrafos que se usan como patrones donde colocar las afirmaciones retóricas. El primer generador de textos implementado en sus estudios, TEXT, utiliza un total de 4 plantillas (Identificar, Describir, Comparar-Contrastar y Atributivo) para generar pequeñas descripciones de barcos. En la siguiente imagen (3.4) se puede observar el grafo de estados correspondiente a la creación de una de estas plantillas, en concreto la de Identificación. Este grafo empieza seleccionando una primera afirmación de identificación. En pasos posteriores se pueden añadir afirmaciones adicionales (indicado con "subr") o bien salir del grafo (instrucción "pop").




**3.4 Grafo de la plantilla Identificación**

Un ejemplo concreto de McKeown en el que se aprecia la identificación del tipo de afirmación dentro de la plantilla de Identificación, es el siguiente (imagen 3.5):

"**Eltville** (Germany) (1) An important wine village of the Rheingau region. (2) The vineyards make wines that are emphatically of the Rheingau style, (3) with a considerable weight for a white wine. (4) Taubenberg, Sonnenberg and Langenstuck are among vineyards of note." (Paterson [28]).

Classification of example

1. Identification (class & attribute)
2. Attributive
3. Amplification
4. Particular-illustration

**3.5 Afirmaciones de la plantilla Identificación**

Teniendo en cuenta el tipo de afirmación y observando cómo los párrafos están estructurados, se realiza una identificación de las plantillas más comunes utilizados en diferentes textos (la variedad de textos evita que reflejen el estilo de una única persona) para describir cómo comunicar información con un propósito concreto [7].

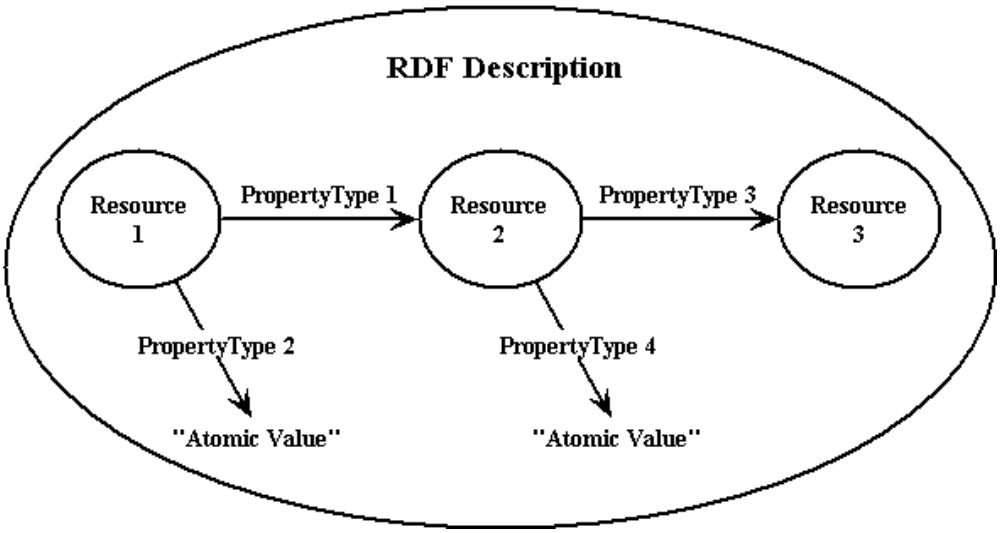
## 3.2 Ontologías

La información extraída del texto se puede representar en lo que se denomina "ontologías", las cuales permiten realizar una descripción formal del contenido de un texto en términos de conceptos y relaciones entre ellos.

La noción de "ontología" existe desde que los filósofos griegos del siglo IV A.C. tuvieron la necesidad de crear un término que hiciera referencia a conceptos o ítems del ámbito del conocimiento. En concreto, Aristóteles inició la discusión sobre qué hay más allá de la física en su libro de metafísica. Para ellos una ontología era una conceptualización de una cierta visión del mundo, independientemente del lenguaje usado (la idea es la misma pese a explicarla en inglés o italiano, o incluso según el vocabulario utilizado) [8]. En un artículo de 1999 en la revista "Intelligent Systems and their Applications" se expuso la importancia de las ontologías, entendidas como *un conjunto de conceptos o términos que pueden ser usados para describir alguna área del conocimiento o construir una representación del mismo*. La virtud de utilizar ontologías es que facilitan la compartición de conocimientos, ya que si se usan como base para diferentes sistemas se puede encontrar una terminología común que facilite su compartición y re-uso [9].

Actualmente la definición de "ontología" tiene en Inteligencia Artificial diversas interpretaciones, además de la filosófica [10]. En el presente trabajo se coincide en la definición de ontología como una *especificación de una conceptualización*. Una conceptualización es una visión simplificada del mundo, en la cual tenemos un conjunto de objetos y su relación entre ellos. Así pues, las ontologías permiten organizar los datos en un esquema conceptual que podemos procesar mediante un programa que extraiga la información que deseamos. Mediante las ontologías es posible trabajar con conceptos, en lugar de palabras clave, en los sistemas de recuperación de información.

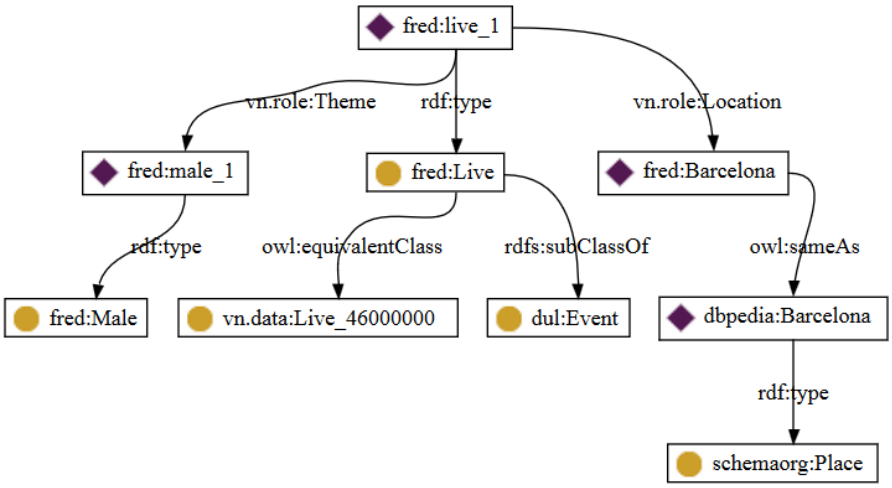
La representación estándar de las ontologías hoy en día es el RDF (siglas en inglés de "Resource Description Framework", Marco de Descripción de Recursos), el cual es un modelo basado en la sintaxis XML (siglas en inglés de "Extensible Markup Language") que define el formato de los datos para poderlos procesar por máquinas y clasificar semánticamente la información. El lenguaje RDF permite representar los conocimientos mediante metadatos, que son datos interrelacionados entre sí de modo que aportan información unos de otros. Esto implica que las consultas en la base de datos son más completas que las basadas en palabras clave, puesto que se incluye la noción de recursos: Objetos que pueden estar relacionados con otros recursos mediante direcciones URI ("Uniform Resource Identifier", identificador único de objetos). Esta relación puede representarse gráficamente en un esquema en el que se muestren las relaciones semánticas entre los recursos, lo cual puede interpretarse como un grafo dirigido [11]. En la imagen 3.6 se puede observar un esquema ilustrativo de las relaciones entre recursos en RDF[1]




3.6 Relaciones entre recursos en RDF

En el trabajo de F. Draicchio et al. [1], se tiene en cuenta cómo las ontologías se relacionan a nivel de oración, superponiendo a la representación RDF etiquetas semánticas del contenido. Es un planteamiento diferente al habitual, puesto que crea representaciones semánticas de los conceptos, algo de lo que no está dentro del alcance de RDF. En sus estudios han realizado la herramienta llamada FRED, la cual procesa texto y genera un gráfico con la representación RDF e información semántica de las ontologías presentes en él. Por ejemplo, utilizando la demo online de este programa con la oración *He lives in Barcelona*, se obtiene la representación gráfica de las ontologías RDF que se observa en la imagen 3.7.

[1] Imagen obtenida de la dirección Web http://www.dlib.org/dlib/may98/miller/05miller.html.




3.7 Representación de FRED

Como se puede observar en la imagen 3.7, para cada elemento se identifica de qué tipo es ("Male", "Event" y "Place"), además de establecer las relaciones entre ellos.

## 3.3 Uso de Plantillas en PLN

En el Procesamiento del Lenguaje Natural (PLN) habitualmente se han utilizado plantillas para realizar algunas de sus tareas. A continuación se detallan aquellas tareas que también se han desarrollado en este trabajo y que han influido en las plantillas creadas. Mediante la Extracción de Información se han obtenido datos que han permitido crear las plantillas que definen qué conceptos pueden agruparse en forma de oraciones, mientras que en la Semántica de Casos se justifica su uso en la generación de textos.

### 3.3.1 Extracción de información

La tarea de Extracción de Información pertenece al ámbito de Procesamiento del Lenguaje Natural, y está basada en el reconocimiento, etiquetado y extracción de la información en una representación estructurada de ciertos elementos de información como por ejemplo personas, lugares u organizaciones de grandes colecciones de texto [12]. Habitualmente se relaciona con documentos basados en texto, a pesar que de que en principio la extracción de información se puede aplicar también a imágenes, vídeos o ficheros de audio. Su objetivo es identificar aspectos relevantes del mensaje que se quiere transmitir, lo cual requiere seleccionar piezas clave de los documentos (frases o palabras). Esta tarea es compleja para las máquinas debido a que el lenguaje natural es ambiguo, redundante y está estrechamente ligado al contexto [13].

La información extraída de un texto habitualmente se almacena de forma estructurada, predefiniendo una serie de plantillas que contendrán información del dominio. A modo de ejemplo, se muestra la plantilla utilizada en EXIT (imagen 3.8), un sistema de extracción de información de textos relativos a la compra-venta de inmuebles, para la frase *Don Fernando García Cueto vende la casa situada en la calle Foguerer nº4 a Don Andrés Iritigoya López* [14] (los atributos no rellenados se deben a que en la oración original no se aporta esa información):

| | |
|---|---|
| ***<PLANTILLA-0001>*** : = | **<PERSONA-0002>** : = |
| *DOC_NR*: 0001 | *NOMBRE*: "ANDRÉS IRITIGOYA LÓPEZ" |
| *CONTENIDO*: <COMPRA_VENTA-0001> | *DNI*: |
| ***<COMPRA_VENTA-0001>*** : = | *DIRECCION*: |
| *VENDEDOR*: <PERSONA-0001> | **<INMUEBLE-0001>** : = |
| *COMPRADOR*: <PERSONA-0002> | *IDENTIF.*: "0001" |
| *OBJETO_COMP*: <INMUEBLE-0001> | *TIPO*: "CASA" |
| ***<PERSONA-0001>*** : = | *VIA*: "CALLE" |
| *NOMBRE*: "FERNANDO GARCÍA CUETO" | *DIREC.*: "FOGUERER Nº 4 |
| *DNI*: | *LINDA*: |
| *DIRECCION*: | *REGISTRO*: |

**3.8 Plantilla de EXIT**

Gracias al uso de estas plantillas en Extracción de la Información se pueden recuperar directamente los datos relevantes presentes en un texto.

### 3.3.2 Semántica de casos

Esta teoría lingüística inicialmente desarrollada por Charles J. Fillmore se basa en la idea de que el mensaje de una oración varía según la estructura de la misma, la cual viene definida por un patrón. Los patrones son conjuntos de unidades de contenido que se relacionan semánticamente entre sí, de modo que las palabras tienen significados diferentes según el patrón al que pertenezcan, como por ejemplo "instrumento", "agente de la oración", "lugar", etc. [15] La semántica de casos pretende entender y describir el significado de las palabras y construcciones gramaticales, enlazándolos asumiendo que el modo en la que una palabra concreta está situada dentro de una oración remarca aspectos específicos del patrón en el que esta palabra pertenece. A continuación se presenta un ejemplo ilustrativo de Fillmore [16]:

*Imaginemos dos hermanos gemelos idénticos, Mark y Mike, ambos en descansando en camillas de un hospital exactamente en la misma posición. Cuando una enfermera entra en la habitación de Mark, dice: "Veo que Mark es capaz de levantarse ya". Posteriormente, entra en la habitación de Mike y comenta: "Veo que Mike es capaz de levantarse ya".*

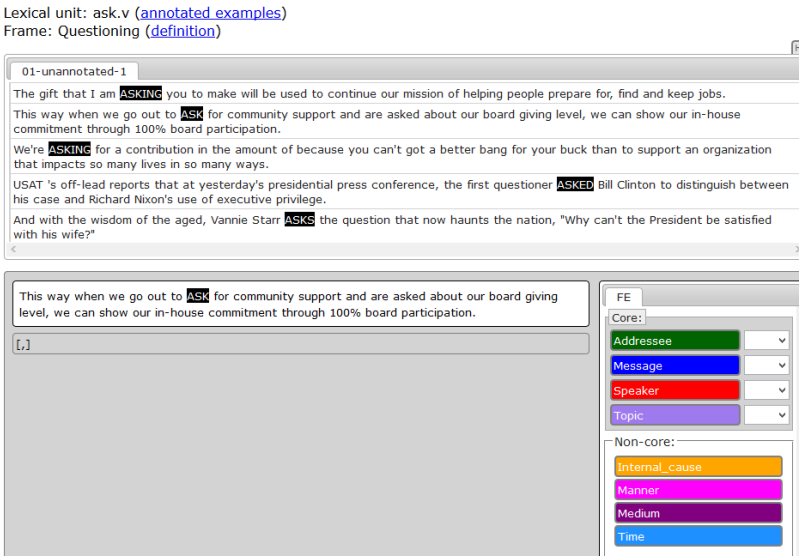
En este ejemplo, claramente se puede observar que ambas oraciones transmiten un mensaje diferente. Sin embargo, este significado simplemente está relativizado a la escena, y se puede ver que las frases pertenecen a un mismo patrón, por ejemplo, se puede decir que "Mark" y "Mike" son "pacientes". Por lo tanto, la información transmitida en estas dos oraciones puede considerarse la misma que "Veo que <paciente> es capaz de levantarse ya". Así pues, lo que se interpreta como

“significado” de una frase, está estrechamente ligado qué conceptos se escogen para formarla y no tanto a las palabras concretas [16].

Este enfoque de detectar patrones en los conceptos dispuestos en las oraciones es similar al aplicado en el presente trabajo, en el cual se han clasificado ontologías según su significado para distinguir diferentes funciones de las mismas en las frases. Para Fillmore los patrones son un instrumento lingüístico y están vinculados a la realización lingüística, en cambio en este estudio se trabaja a nivel de conocimientos y se utilizan las etiquetas como un vínculo entre ellos. La propuesta de Fillmore ha sido utilizada también en otros estudios y proyectos, como es el caso de FrameNet.

El proyecto FrameNet está creando un recurso léxico online basándose en la semántica de casos y utilizando datos del “British National Corpus”. Su objetivo es obtener descripciones semánticas y sintácticas de unidades léxicas desde el punto de vista de los patrones explicados anteriormente. Estas descripciones se registran mediante anotaciones en los textos de corpus por parte de humanos utilizando herramientas específicas para esta función [17].

A continuación se muestra una captura de pantalla de la demo de anotación del proyecto Framenet (imagen 3.9), la cual se encuentra en su página web. En esta demostración se trabaja con la unidad léxica “Ask” (en inglés, “preguntar”), que pertenece al marco de Cuestionamiento. En el recuadro superior se observan las frases de ejemplo, y mediante los botones de colores inferiores se pueden etiquetar los componentes de la oración seleccionada.




**3.9 Demo de anotación de FrameNet**

También se puede observar en la imagen 3.10 una captura de pantalla de un ejemplo ya etiquetado de FrameNet. El documento procesado pertenece al "American National Corpus". Las palabras mostradas en azul son las originales del texto, mientras que las que se muestran en negro y más pequeñas son las anotaciones realizadas.

**FrameNet Data**

**American National Corpus Texts - Berlitz History of Greece** Lexical Unit Index Frame Index

1. *a* BRIEF$_{\text{Duration_description}}$ HISTORY$_{\text{History}}$

2. PREHISTORIC$_{\text{Temporal_collocation}}$ *man* IN$_{\text{Locative_relation}}$ *Asia Minor ( now* MODERN$_{\text{Temporal_collocation}}$ *Turkey ) or Greece* COULD$_{\text{Capability}}$ LOOK$_{\text{Perception_active}}$ *out across the Aegean toward the horizon and* SEE$_{\text{Perception_experience}}$ *the* faint silhouette OF$_{\text{Partitive}}$ *land . Their* curiosity PUSHED$_{\text{Subjective_influence}}$ *them to* BUILD$_{\text{Building}}$ VESSELS$_{\text{Vehicle}}$ *that were* STRONG$_{\text{Exertive_force}}$ ENOUGH$_{\text{Sufficiency}}$ *to* FORD$_{\text{Intentional_traversing}}$ *the* open SEAS$_{\text{Natural_features}}$ *and* REACH$_{\text{Arriving}}$ *these* ISLANDS$_{\text{Natural_features}}$ *'* MARKING$_{\text{Sign}}$ *the* START$_{\text{Temporal_subregion}}$ *of the* LONG$_{\text{Duration_description}}$ legacy *of Mediterranean seafaring .*

**3.10 Ejemplo etiquetado de FrameNet**

## 3.4 Conclusiones del Estado de la Cuestión

Una vez expuesta la base teórica de este trabajo, a continuación se procede a concretar cómo se han aplicado estos enfoques.

Para el experimento realizado en el presente trabajo se ha trabajado con etiquetas semánticas, como se ilustran en FRED, asumiendo que a través de FRED se puede establecer una proyección bidireccional entre ontologías y conjuntos de etiquetas semánticas. Por lo tanto, en este punto de vista cada etiqueta es una instancia de una ontología.

Este etiquetado simplifica la tarea de extracción de la información, puesto que el ordenador no necesita identificar la información relevante y se eliminan problemas de ambigüedad del texto o de entendimiento del contexto. Como las anotaciones en los textos han sido realizadas mediante un humano, la tarea de extracción de información ha requerido implementar un método de discriminación y selección de los datos una vez etiquetados, y además se asegura la calidad de los datos de entrenamiento.

Se han utilizado dos métodos diferentes para la agrupación de unidades de contenido en oraciones, cada uno basado en un enfoque de planificación de textos opuesto:

- Inicialmente, se crean plantillas a partir del corpus de entrenamiento y después se "rellenan" con los datos de la base de conocimiento. Este sería un punto de vista top-down en sintonía con la propuesta de McKeown.
- Posteriormente, se trabaja con aquellas unidades de conocimiento que no ha sido posible hacer encajar en las plantillas. En este caso, se utiliza un enfoque bottom-up del problema, trabajando individualmente con los conceptos como en RTS.

El objetivo de las plantillas es guiar la agrupación de la información de un repositorio de conocimiento en unidades (frases) de un modo similar al propuesto por McKeown (y apoyándose en la idea de los casos semánticos de Fillmore; véase también los comentarios sobre FrameNet), el cual es un punto de vista reciente y diferente respecto a la creación tradicional de plantillas donde se establecían de un modo predefinido. Tal como se ha interpretado en este trabajo, el contenido de una plantilla no tiene una estructura específica, sino que es posible combinar los conceptos de diferentes modos para crear una oración. Por ejemplo, es lo mismo *Pedro vive en Madrid*, que *En Madrid vive Pedro*, lo importante son los componentes de la oración y su relación, y no tanto el orden en el que se expresan. Así pues, las plantillas indican qué agrupaciones de conceptos se pueden realizar a nivel de oración.

Las plantillas se han creado a partir de la anotación de conceptos de las oraciones en el corpus (etiquetados según su significado), puesto que el objetivo es agrupar la información en oraciones y no en párrafos. Así pues, una plantilla podría ser, para el ejemplo anterior *Pedro vive en Madrid*, $Persona + Evento + Lugar$ (el orden de los elementos es irrelevante). De esta manera, se obtienen plantillas flexibles para las cuales no se requiere un conocimiento previo por parte del usuario del ámbito de la base de conocimiento, sino que dependen únicamente del corpus.

Asimismo, una vez agrupados todos los conceptos posibles dinámicamente mediante plantillas, se aplica un punto de vista bottom-up para terminar de organizar aquellos elementos "sobrantes" que no han podido encajar con el resto utilizando plantillas. Por lo tanto, también se ha adaptado el punto de vista RST para poder trabajar a nivel de conceptos y tener en cuenta las relaciones semánticas encontradas entre ellos en el corpus de aprendizaje. Mediante este enfoque se realizan dos tipos de agrupaciones:

1. Entre todos los conceptos que no se han podido agrupar con el resto mediante plantillas, se buscan agrupaciones nuevas. Es un punto de vista bottom-up porque se analizan las relaciones entre conceptos individualmente y se van realizando grupos nuevos.
2. Una vez terminadas las agrupaciones, se aplica de nuevo este tipo de enfoque para poner los conceptos que quedan pendientes de agrupar en los conjuntos ya existentes. En esta ocasión, se comprueba si una unidad de conocimiento puede encajar con cada una de las agrupaciones de conceptos y se busca la mejor opción.

Los detalles de este método, en el cual se utiliza inicialmente el enfoque bottom-up y posteriormente top-down, se explican en el apartado 5.2. Asimismo, en se ha incluido como Anexo un ejemplo concreto de clasificación de etiquetas en el que se aprecia el uso de ambos enfoques (apartado 9.2).

De un modo similar al planteado en el proyecto FrameNet, en el presente trabajo se ha almacenado información respecto a los patrones de conceptos presentes en los textos, pretendiendo demostrar que hay una relación entre las diferentes etiquetas agrupadas dentro de una misma oración. Gracias a esta relación se pueden crear agrupaciones coherentes automáticamente por un sistema de planificación de textos.

# 4 Descripción del conjunto de datos utilizados

El algoritmo encargado de realizar la agrupación de unidades de contenido en oraciones requiere de datos obtenidos previamente. A continuación se explica qué datos se han extraído de los textos de ejemplo y cómo se ha organizado la información a partir de ellos.

## 4.1 El corpus de entrenamiento

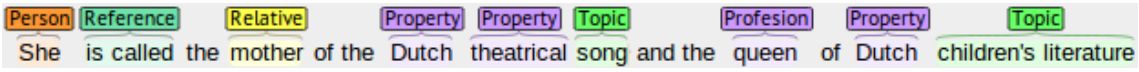
El corpus es una colección de textos usada como referencia en la lingüística, por ejemplo para realizar estudios sobre diferentes aspectos de la redacción de los documentos. Cuanto mayor es el corpus, más fiables son los resultados obtenidos en los estudios que se basan en él.

El tipo de texto utilizado en este trabajo han sido biografías de personas conocidas extraídas de la base de datos DBPedia [18], de las cuales se han escogido un total de 100 personajes con profesiones diferentes. El criterio de selección aplicado ha sido buscar esencialmente artistas, científicos, escritores o políticos, debido a que se ha identificado un estilo de escritura diferente entre ellos; siendo las vidas de personajes históricos más complejas que el resto en cuanto a relaciones sociales, los textos de científicos contienen explicaciones técnicas, y los escritores y artistas tienen biografías en las que se exponen sus obras. Se han descartado biografías de deportistas por su simplicidad en las explicaciones, ya que la mayoría de ellas se centran en enumerar los equipos y ligas en las cuales ha jugado el protagonista, lo cual no nos aporta información interesante en relación con la redacción del texto.

## 4.2 Etiquetas

La obtención de unidades de contenido del texto se realiza mediante el previo etiquetado de las entidades semánticas que forman las oraciones de los textos del corpus, por lo que una etiqueta representa un fragmento de contenido de una frase. No hay establecida una relación lineal entre las palabras y las etiquetas, de tal forma que en una sola etiqueta se pueden encontrar una o más palabras, y puede haber palabras del texto que no han sido etiquetadas debido a que no aportaban información relevante (como por ejemplo artículos o conjunciones). Además, según el contexto una misma palabra puede tener un significado diferente, lo cual dificulta el etiquetado automático de la misma.

En la imagen 4.1 se muestra un ejemplo en el que se reflejan las palabras no etiquetadas. Como se puede observar, palabras como “of”, “the” y “and” no aportan información necesaria a ser almacenada y procesada, por lo que no hace falta que sean etiquetadas.

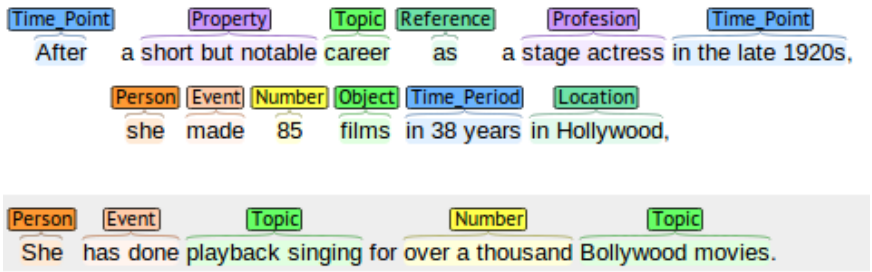



4.1 Frase de la biografía de A. Schmidt

Se ha escogido un conjunto de etiquetas que permita recoger toda la información clave de las oraciones y clasificarla según su contenido. Las etiquetas son:

- Person: Identifica las personas a las que se hace mención.
- Birth: Contiene la información relativa al nacimiento de la persona protagonista en la biografía.
- Circumstance: Engloba aquella información que ha causado alguna consecuencia, puede ser que esta etiqueta incluya otras dentro que detallen la situación expuesta.
- Death: Similar a "Birth", pero con la muerte del personaje.
- Description: Incluye información que detalla algo mencionado en la misma frase.
- Event: Identifica los eventos o verbos de la oración.
- Event_Establish: En este caso, especifica aquellos verbos que suponen el inicio de algún hecho.
- Location: Determina los lugares expuestos en la oración. En el caso de detallar una localización, por ejemplo del modo "Madrid, España", estos dos lugares se consideran como uno solo.
- Number: Etiqueta para los números (tanto cardinales como ordinales).
- Object: Todos aquellos objetos (p.ej. "mesa", "piano", "ventana", etc.) de la oración son identificados con esta etiqueta.
- Organization: Señala las instituciones citadas.
- Ownership: Identifica palabras que denotan propiedad.
- Parenthesis: Sirve para marcar aquellos datos que están entre paréntesis dentro de una frase.
- Profesion: Determina las profesiones.
- Property: Contiene todas aquellas palabras que indican propiedades o atributos de otros datos.
- Time_period: Información relativa a un periodo de tiempo.
- Time_point: En este caso, se marcan los datos referentes a un momento específico (incluye también las fechas).
- Topic: Identifica aquella información "abstracta", como ideas o títulos de obras.
- Relative: Permite diferenciar las palabras que denotan parentesco.
- Set: Engloba aquella información relativa a grupos de personas.
- Reference: Especifica los datos que permiten hacer referencia a otros, como por ejemplo "también conocido como…", "del mismo modo que…",…
- Reference_Relation: Tiene los verbos que establecen vínculos fuertes entre conjuntos de datos, como es el caso de "ser" o "tener".
- Related: Informa de datos que están relacionados entre ellos mediante mención específica de dicha relación.

A continuación se presenta a modo de evidencia una imagen que refleja cómo algunas de estas etiquetas se aplican en dos oraciones distintas (imagen 4.2).




**4.2 Oraciones de B. Stanwyck y A. Boshle**

En las oraciones de la imagen 4.2 se pueden observar diferentes formas de empleo de algunas etiquetas, como por ejemplo el uso de "Time_Point" en una sola palabra ("After") o para un momento determinado ("in the late 1920s").

## 4.3 Herramienta de anotación

En este apartado se explica qué instrumento se ha utilizado para etiquetar la información contenida en el texto, basándose el listado de etiquetas que hemos establecido previamente, el cual se ha explicado en el apartado anterior.

Mediante la herramienta de anotación se agiliza el trabajo a realizar en la anotación de entidades, ya que permite almacenar rápidamente los datos que de otro modo habría que ir registrando manualmente. Esto se debe a que se puede trabajar sobre el texto directamente seleccionando los datos y clasificándolos, lo cual evita tener que escribirlos uno a uno en un documento aparte.

Se han probado dos herramientas de anotación distintas: Inicialmente RDFaCE [19] y posteriormente Brat Annotation Tool [20].

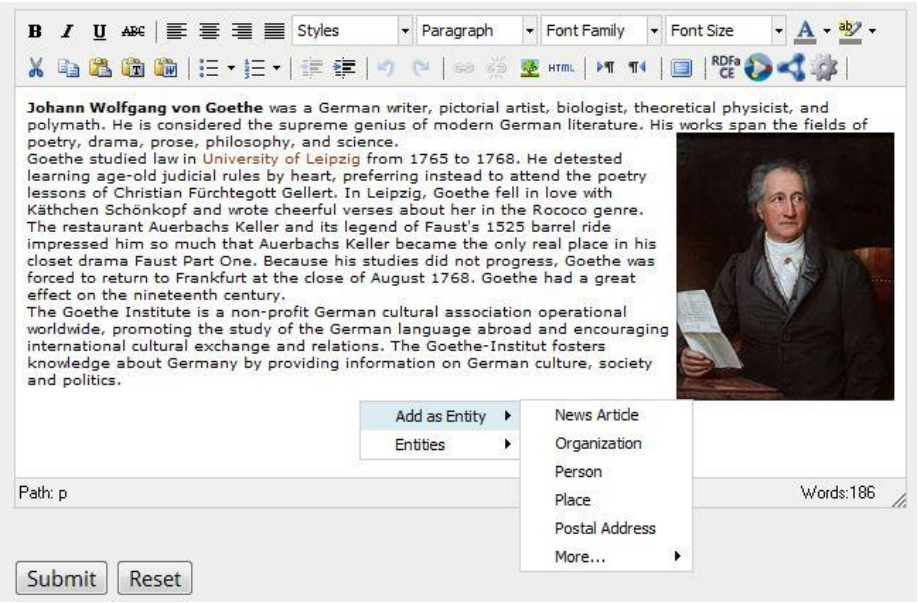
### 4.3.1 RDFaCE

Su objetivo principal es combinar el paradigma WYSIWYG (What You See Is What You Get, "lo que ves es lo que tienes") con la anotación semántica. Su interfaz contiene un listado de etiquetas disponibles, las cuales se pueden utilizar para ir clasificando la selección de texto que deseemos. Asimismo, incluye la opción de analizar el texto mediante consultas en algunas bases de datos que, en caso de encontrar coincidencias, permite realizar anotaciones automáticamente (especialmente útil para personas, lugares y organizaciones, que son las más fáciles de encontrar).

La página web de RDFaCE incluye una exhaustiva descripción de la misma, una demo on-line y la opción de descargar la versión completa gratuitamente [19].

Después de tratar de etiquetar algunos textos mediante esta herramienta, se decidió descartar su uso debido a dificultades en su manipulación y fallos en algunos botones de la interfaz del programa instalado en nuestro equipo, los cuales no se pudieron solventar para llegar a trabajar correctamente.

En la imagen 4.3 correspondiente a demo de la página web de RDFaCE, se observa el texto de la biografía de Johan Wolfgang von Goethe, así como una imagen suya que también puede etiquetarse. En pulsar el botón derecho del ratón nos aparece la opción de añadir como entidad la información seleccionada, escogiendo alguna de las etiquetas de la lista.




**4.3 Demo de RDFaCE**

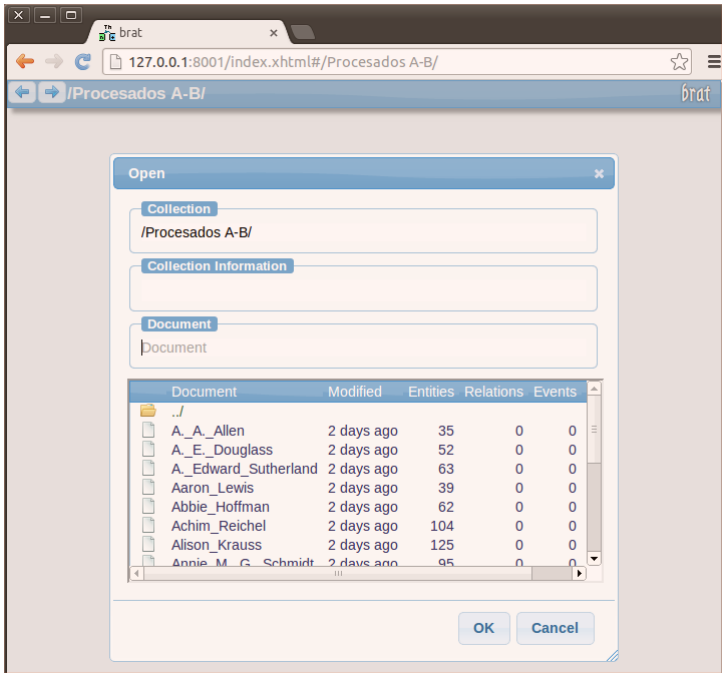
## 4.3.2 Brat Rapid Annotation Tool

Brat es una herramienta *open source* basada en la web, la cual debe ejecutarse localmente (en *localhost*) para poder cargar los documentos y trabajar sobre ellos. Permite, además de realizar anotaciones, establecer relaciones entre las mismas, entre otras funciones explicadas en su página web [20].

Se encontró esta herramienta más fácil de utilizar que la anterior y con explicaciones detalladas de su funcionamiento en la web, y enseguida se vio cómo adaptarla a los requerimientos de este trabajo (el listado de etiquetas se puede modificar fácilmente mediante un fichero de configuración), por lo que ha sido la que finalmente se ha utilizado para el etiquetado de todos los textos del corpus.

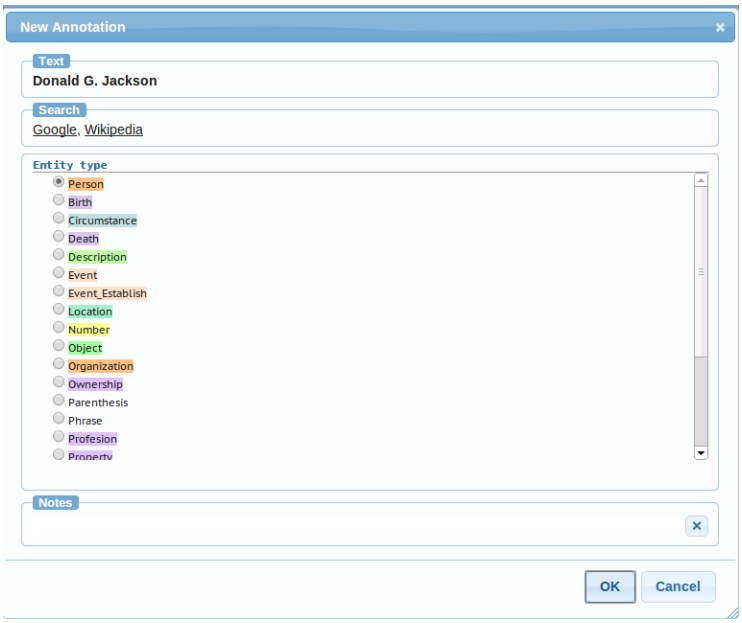
En la siguiente imagen (Imagen 4.4) se puede observar la opción de carga de un documento en Brat. Están listados los ficheros de texto, de los cuales se selecciona el deseado y se pulsa el botón "OK" para empezar a etiquetar. Los resultados se guardan en un documento que tiene el mismo nombre, pero extensión .ann (debe ser creado vacío previamente). El valor numérico que se aprecia en la columna

“Entities” es el número de etiquetas de la biografía almacenadas en su correspondiente fichero .ann.




4.4 Cargar un documento en Brat

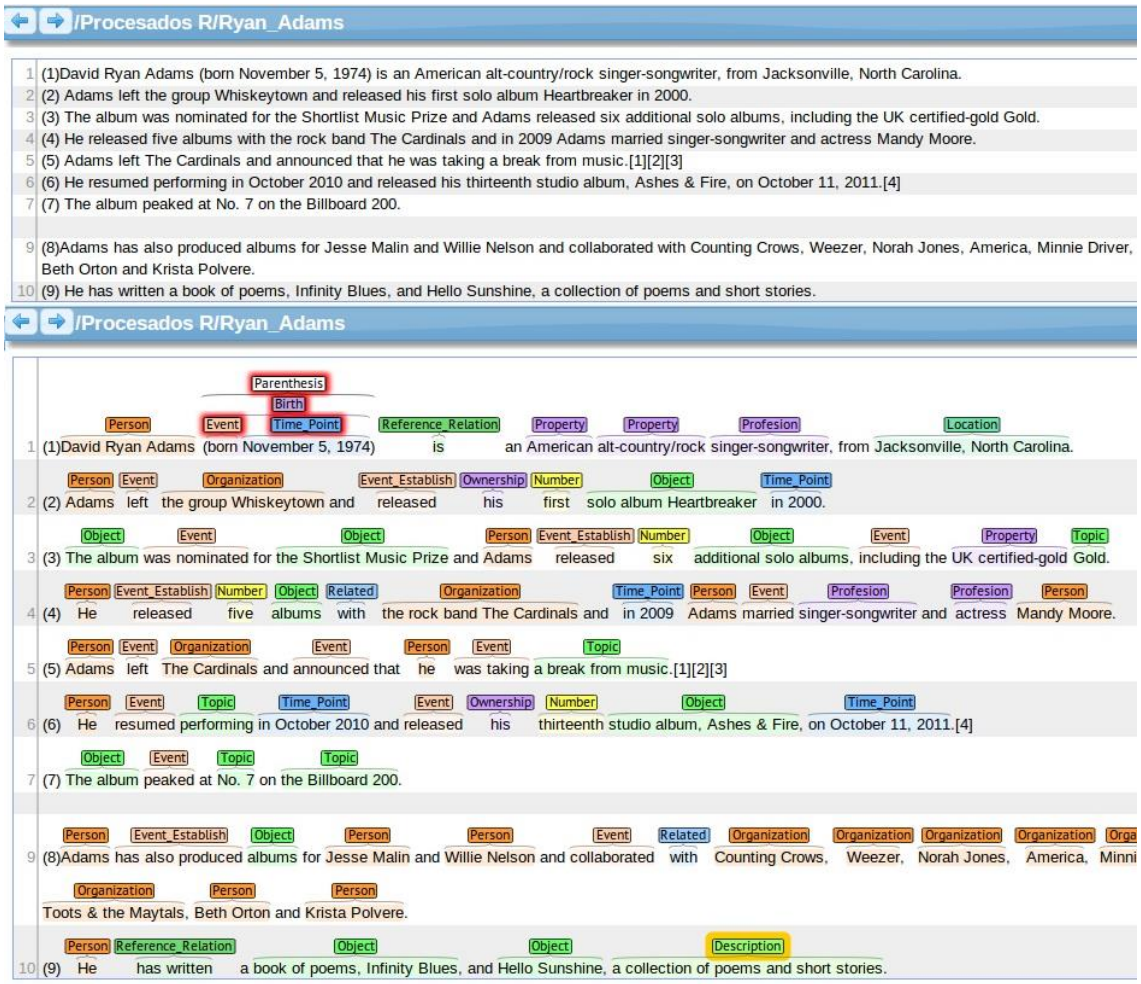
En la imagen 4.5 se ha capturado el listado de etiquetas en la configuración de Brat de nuestro trabajo. En el apartado superior “Text” se puede observar el texto seleccionado (en este caso el nombre de Donald G. Jackson), el cual se puede anotar seleccionando la etiqueta deseada del listado. Dicho listado lo establece el usuario en un fichero aparte llamado “annotation.conf”, y es posible personalizar los colores de las etiquetas en el fichero “visual.conf” según las propias preferencias.




4.5 Etiquetas en Brat

El único inconveniente que se encontró con esta herramienta fue que no se pudo identificar la codificación de los caracteres utilizada para determinar la posición de las etiquetas en el fichero de anotaciones, por lo que para el valor de inicio y de fin de las etiquetas establecido por esta herramienta no parecía tener sentido (y en la página web no se hacía mención a este problema). La solución ha sido utilizar una etiqueta “Phrase” que registra el inicio y el final de las frases, de modo que para detectar qué etiquetas hay dentro de una frase se compara el inicio y final de las mismas con el de la frase.

A continuación, en la imagen 4.6, se puede observar la biografía de R. Adams cargada en Brat antes y después de realizar las anotaciones (sin las etiquetas de frases):




**4.6 Comparación del procesado de la biografía**

Esta herramienta considera el solapamiento de etiquetas como erróneo, por este motivo en la imagen 4.6 podemos ver que en la etiqueta “Parenthesis” y en las incluidas en ella aparece un sombreado rojo alrededor del nombre. Precisamente por esta razón, se han omitido las etiquetas “Phrase” de la imagen anterior y poder observar mejor el etiquetado.

Los datos etiquetados se guardan en un fichero con extensión .ann, que tiene el aspecto de la imagen 4.7. En ella se puede observar el identificador de la anotación, el tipo de etiqueta, la posición inicial que ocupa en el texto, la posición final y el texto que contiene la etiqueta.

```
Ryan_Adams.ann
T1      Person 3 19     David Ryan Adams
T2      Event 21 25     born
T3      Time_Point 26 42        November 5, 1974
T4      Birth 21 42     born November 5, 1974
T5      Parenthesis 20 43       (born November 5, 1974)
T6      Reference_Relation 44 46        is
T7      Property 50 58  American
T8      Property 59 75  alt-country/rock
T9      Profesion 76 93 singer-songwriter
T10     Location 100 128        Jacksonville, North Carolina
```

**4.7 Fichero Ryan_Adams.ann**

Pese a que Brat muestre las etiquetas solapadas remarcadas en rojo, los datos se guardan en el fichero de anotaciones correctamente. Es gracias a la posición inicial y final de cada etiqueta como se detecta cuándo una se encuentra dentro de otra. En la imagen anterior, se aprecia cómo “Birth” (T4) empieza y acaba en posiciones que implican la inclusión de “Event” (T2) y “Time_Point” (T3) en ella.

# 5 Enfoque tomado para la agrupación de contenido en oraciones

Tal como se ha explicado en el apartado 4, para la obtención de datos se han seleccionado 100 textos en los cuales se ha realizado la anotación de los conceptos mediante la herramienta Brat. Del listado de etiquetas obtenido, se pueden detectar ciertas agrupaciones dentro de oraciones que son comunes en diferentes textos (mientras otras suceden en casos concretos), las cuales se han considerado "agrupaciones comunes" que pueden ser utilizadas como plantillas a la hora de realizar agrupaciones por parte de un algoritmo de planificación de textos. Asimismo, además de estas agrupaciones comunes dentro de diferentes oraciones, se ha extraído información respecto al PMI ("Pointwise Mutual Information") entre etiquetas y un listado de bigramas.

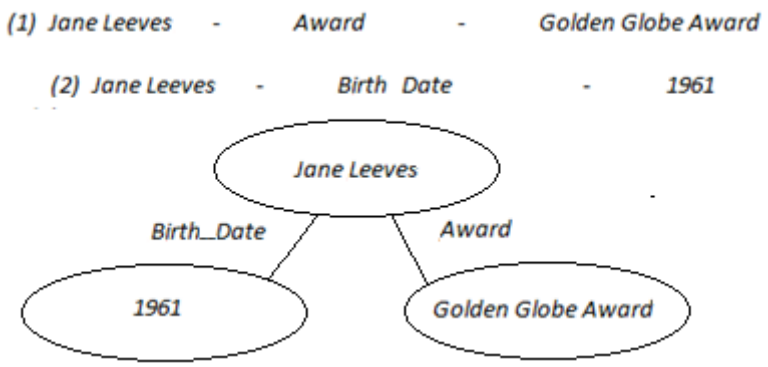
En la preparación de los datos para la planificación de textos, es necesario también tener un listado con los conceptos anotados del texto que se desea procesar. Se comprueba qué "plantillas comunes" pueden aplicarse con este listado, y una vez seleccionadas se finaliza la planificación con la información adicional extraída del corpus (PMI y bigramas).

En los siguientes subapartados se puede encontrar una descripción detalla de este proceso.

## 5.1 Reflexiones preliminares

Dentro del "Marco de Descripción de Recursos" (RDF), comentado en el apartado 3.2, es común representar las ontologías mediante tripletas, por lo que fue el método de representación de información planteado inicialmente. Las tripletas establecen enlaces entre un Sujeto (como podría ser la persona de quien trata la biografía), una Propiedad (como por ejemplo "autor", "fecha", "familiar",…) y un Objeto (o "propiedad del sujeto") [21].

En la imagen 5.1 se observa un ejemplo de tripletas en el cual la tripleta (1) indica que la persona Jane Leeves ha ganado el premio Golden Globe Award, y la tripleta (2) indica que la persona Jane Leeves ha nacido en la fecha 1961. Estos datos pueden representarse gráficamente, como se demuestra en la misma imagen.




**5.1 Ejemplo de tripletas**

La información enlazada mediante tripletas se almacena en lo que se denominan "bases de datos relacionales", y para operar con este tipo de datos se utiliza el lenguaje SPARQL de un modo similar a las consultas SQL en bases de datos. El objetivo era el mismo, crear agrupaciones de contenido dentro de la base de conocimiento relativa a la biografía de una persona, pero recuperando la información mediante estas consultas.

El motivo principal para descartar este modo de trabajo fue que las tripletas de las que se disponía no tenían un formato correcto para trabajar con ellas. En concreto, después de estudiar la documentación de referencia se observó que el formato de una tripleta relaciona tres conceptos entre ellos, tal como se ha explicado previamente; sin embargo, en el conjunto de datos con el que se trabajaba se vio que una sola tripleta incluía más información de la que debería. Es decir, siguiendo con el ejemplo anterior, en la relación "Jane Leeves + Award" además de indicar el nombre del premio también se incluía el año, el motivo, el lugar, etc. Por lo tanto, no cumplía estrictamente con la definición de lo que es una tripleta, y se tuvo que descartar seguir trabajando con este tipo de información.

Mientras se buscaban nuevos datos en formato de tripletas, se consideraron también otros puntos de vista que permitieran alcanzar el mismo objetivo, y se decidió que era interesante utilizar como datos "etiquetas" del texto. Pese a que hay diversos estudios en el ámbito de generación de textos basados en el uso de tripletas, actualmente se está investigando también el uso de "plantillas" con información respecto a las unidades de contenido presentes en oraciones, que ha sido el enfoque que se ha tomado finalmente. Mediante este método no es necesario utilizar información adicional a los textos, puesto que se pueden identificar directamente en las oraciones los conceptos y anotarlos en un documento aparte.

## 5.2 Método

Los pasos seguidos para la agrupación de conceptos con el enfoque tomado se pueden resumir en:

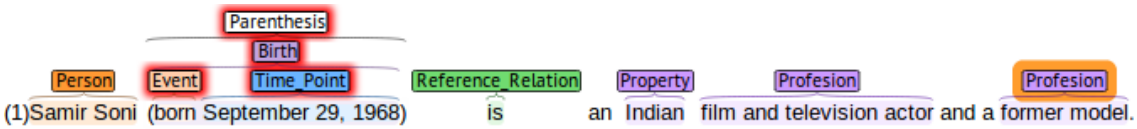
1. Anotación del corpus.
2. Extracción de plantillas.
3. Extracción de medidas (PMI y bigramas).
4. Generar la estructura del discurso

En los siguientes apartados se detalla la implementación de estos pasos y se definen los algoritmos utilizados tanto para extraer unidades de contenido del corpus como para realizar la agrupación de unidades de contenido presentes en la base de conocimiento.

### 5.2.1 Extracción de plantillas

Se ha extraído del corpus de entrenamiento un listado de las etiquetas contenidas en cada oración, el cual permite representar cómo se han estructurado los conceptos en los textos. En la organización de anotaciones en frases es importante considerar las jerarquías entre etiquetas. Las jerarquías se establecen teniendo en cuenta si una etiqueta está dentro de otra, lo cual se registra mediante un valor numérico que indica el nivel de "profundidad" en el que encontramos una etiqueta. Por ejemplo, el fragmento de texto que se encuentra entre paréntesis se anota por lo menos dos veces con etiquetas de diferentes niveles jerárquicos: La etiqueta que informa de un paréntesis en la oración y las etiquetas que detallan la información de dentro de este paréntesis, las cuales consideramos que están "dentro" de la primera.

En la frase anotada extraída de la biografía de Samir Soni (imagen 5.2) podemos observar cómo el fragmento "born September 29, 1968" está anotado con diversas etiquetas, cada una de las cuales pertenece a un nivel jerárquico diferente, siendo "Parenthesis" la que engloba a todas las demás.




5.2 Frase de S. Soni

Mediante estos datos se pretende identificar a nivel de texto cómo se distribuyen la información entre las oraciones. Debido a que el objetivo es observar la organización del contenido de las frases, la colocación original de las etiquetas dentro de la oración no se ha registrado, de modo que dos locuciones con las mismas etiquetas (y los mismos niveles jerárquicos) se consideran iguales pese a que los conceptos estén dispuestos en un orden diferente. Asimismo, también se ha tenido en cuenta que hay

oraciones que pueden ser un subconjunto de otras, en el sentido en el que todas las etiquetas que contienen se encuentran a su vez en otra frase junto con más etiquetas.

El resultado final es una lista de “contenedores” o “plantillas” de oraciones, cada uno de los cuales aparece una o más veces en el corpus. A continuación se comenta un ejemplo simplificado de la obtención de este listado, representado en la imagen 5.3:

| 1. Obtención de datos: | 2. Orden alfabético: |
|---|---|
| (1) [e2, 0][e3, 0][e3, 0][e1, 0][e6, 0] | (1) [e1, 0] [e2, 0][e3, 0][e3, 0] [e6, 0] |
| (2) [e4, 1][e2, 1][e4, 1][e5, 0][e3, 0] | (2) [e2, 1][e3, 0] [e4, 1] [e4, 1][e5, 0] |
| (3) [e3, 0][e2, 0][e3, 0] | (3) [e2, 0] [e3, 0] [e3, 0] |

**5.3 Ejemplo de registro de plantillas**

En la imagen 5.3 cada línea de la lista izquierda representa una oración procesada del corpus y en ella encontramos las etiquetas que la forman junto con el llamado “nivel jerárquico” de cada una, que indica si se encuentran contenidas dentro de otra (tiene valor 1 en caso de estar dentro de una etiqueta de nivel 0). Las etiquetas se ordenan alfabéticamente para detectar si hay estructuras repetidas (en la lista derecha se observan los elementos ordenados). Vemos que en el ejemplo todas las etiquetas de la plantilla 3 se encuentran en la 1, así que es un subconjunto de la primera aunque en la frase original no aparecieran en el mismo orden. Por lo tanto, podemos considerar que la plantilla 3 se encuentra dos veces en el corpus.

A la hora de agrupar las etiquetas de un texto, el listado de plantillas permite realizar una distribución inicial de etiquetas en algunos grupos. Cabe destacar que se han tenido en cuenta las plantillas de etiquetas que se han encontrado repetidas veces en los textos de entrenamiento, ya que el resto de plantillas se han considerado menos relevantes al producirse sólo en textos concretos del corpus de entrenamiento.

### 5.2.2 Medidas

Para la distribución del contenido de la base de conocimiento que se está procesando, el primer paso a realizar es buscar qué plantillas se pueden aplicar que encajen exactamente con los conceptos anotados y permitan tener pocas “etiquetas sobrantes”. Es decir, teniendo el listado de plantillas explicado en el punto anterior, se buscan cuáles es posible utilizar en función del total de etiquetas que hay disponibles para esta base de conocimiento, con lo cual se establece una primera agrupación de etiquetas. Las etiquetas para las que no se ha encontrado ninguna combinación de plantillas en las que se puedan poner deben ser agrupadas posteriormente mediante información adicional obtenida del corpus. A continuación se explica qué otros datos se han extraído de los textos de entrenamiento para ubicar correctamente el contenido anotado que queda pendiente.

### a) Pointwise Mutual Information

El *Pointwise Mutual Information* (PMI) es una medida que representa mediante un valor numérico la correlación entre una palabra y otra, que en este trabajo se ha aplicado para representar numéricamente la relación entre tipos de etiquetas en el corpus. Su cálculo se realiza de la siguiente manera [22]:

Siendo $x$ e $y$ dos variables independientes discretas aleatorias, su correlación dada la probabilidad de cada una de ellas $p(x)$ y $p(y)$, y la conjunta de ambas $p(x,y)$, se calcula mediante la fórmula

$$pmi(x,y) = \log\left(\frac{p(x,y)}{p(x) * p(y)}\right)$$

Tal como explican Church y Hanks en su estudio [23], si hay una asociación entre $x$ e $y$, su valor de $p(x,y)$ será mayor que $p(x) * p(y)$, con lo cual su PMI será mayor que 0. En el caso contrario, si tienen una distribución en la cual cuando aparece $x$ no está $y$ y viceversa, entonces su probabilidad conjunta será mucho menor y se obtendrá un PMI menor que 0. Si no hay una relación interesante entre $x$ e $y$, entonces $p(x,y)$ será muy similar a $p(x) * p(y)$, haciendo que el cálculo del logaritmo dé 0.

Se ha normalizado el valor de PMI para obtener resultados entre -1 y +1 mediante la fórmula:

$$npmi(x,y) = \frac{pmi(x,y)}{-\log[p(x,y)]}$$

Aplicando estos valores al problema planteado en este trabajo con las etiquetas, se ha establecido que para una etiqueta $a$ y otra etiqueta $b$, las probabilidades se extraen con los siguientes valores:

$$p(a) = \frac{nº\ frases\ en\ las\ que\ se\ encuentra\ la\ etiqueta\ a}{nº\ de\ frases\ totales\ del\ corpus}$$

$$p(a,b) = \frac{nº\ frases\ en\ las\ que\ se\ encuentran\ la\ etiqueta\ a\ y\ la\ b}{nº\ de\ frases\ totales\ del\ corpus}$$

En resumen, mediante estos cálculos se puede saber si dos etiquetas sienten “atracción”, “repulsión” o si son independientes una de la otra. Utilizando este valor se puede asegurar que las frases creadas contienen componentes asociados entre ellos comprobando que los valores de PMI son positivos.

El valor de PMI permite entonces trabajar con las etiquetas que en una primera instancia no se han podido agrupar mediante las plantillas. Se puede aplicar tanto para crear agrupaciones nuevas de etiquetas, como para comprobar en qué agrupación ya creada podemos añadir cada una de las etiquetas pendientes de unir. Los pasos que sigue el algoritmo se explican más adelante en detalle, en el apartado 5.2.3, donde se puede apreciar el uso del PMI.

## b) Bigramas

Los bigramas son secuencias de dos elementos que se suceden en una oración. Registrar esta información permite detectar restricciones en cuanto al orden en el que tienen que aparecer las anotaciones de una oración, especialmente teniendo en cuenta que hay ciertas etiquetas que en ningún caso pueden ir una tras la otra porque la oración carecería de sentido, como por ejemplo "birth + birth" (en todo caso sería una sola etiqueta que englobaría toda la información) o "reference_relation + death".

Es importante tener en cuenta estas restricciones para crear agrupaciones de etiquetas en las cuales todas ellas "encajen" y poder descartar aquellas etiquetas que no puedan combinarse de ninguna manera con el resto. A la hora de comprobar si una etiqueta encaja con los elementos de una plantilla, es posible que el valor de PMI obtenido sea positivo (indicando que sí hay atracción entre la etiqueta y la plantilla), pero es mediante los bigramas como se sabe si realmente es posible combinar correctamente los elementos agrupados por esta plantilla y la nueva etiqueta que queremos añadir.

En el cálculo de bigramas del corpus de entrenamiento se ha tenido en cuenta también el nivel jerárquico de las etiquetas, debido a que hay ocasiones en las que una etiqueta engloba a otras. Por ejemplo, *Parenthesis* contiene información que consideramos que "está dentro suyo". Esto significa que las etiquetas de dentro del paréntesis no forman bigramas con las de fuera, sino únicamente entre ellas. Del mismo modo, hemos considerado que la etiqueta final de una oración no tiene relación con la primera de la frase siguiente.

(1) $e_1$ Parenthesis $e_6$
$e_3$ $e_4$ $e_5$

(2) $e_7$ $e_8$ $e_9$

Bigramas: ($e_1$, Parenthesis), (Parenthesis, $e_6$), ($e_3$, $e_4$), ($e_4$, $e_5$), ($e_7$, $e_8$), ($e_8$, $e_9$)

**5.4 Ejemplo de identificación de bigramas**

En el ejemplo de la imagen 5.4 vemos dos oraciones, (1) y (2); en la primera de ellas se observan 6 etiquetas, de las cuales una es "Parenthesis" y engloba a otras tres (e3, e4 y e5). Las etiquetas e1 y e6 no forman bigramas con las tres etiquetas de dentro del paréntesis debido a la diferencia de nivel de jerarquía, pero sí se forman los bigramas (e1, Parenthesis) y (Parenthesis, e6). Asimismo, entre e6 y e7 no hay un bigrama porque pertenecen a oraciones diferentes.

### 5.2.3 Algoritmo de extracción de datos

La extracción de unidades de contenido del corpus de entrenamiento se ha realizado mediante un programa en Python que procesa todos los ficheros con extensión .ann uno por uno y guarda el inicio y final de las frases y de las anotaciones (consultar el apartado 4.2.2 sobre la herramienta Brat para más detalles del fichero de anotaciones). Una vez almacenada esta información, el proceso sigue los siguientes pasos:

1. Se determina qué etiquetas hay en cada frase a partir de sus posiciones de inicio y final.
2. Para comprobar la "jerarquía" entre etiquetas, es decir, cuáles están dentro de otras, se asigna a cada una de ellas un nivel de profundidad. Las etiquetas que tienen nivel 0 son aquellas que no se encuentran dentro de otras, después hay algunas nivel 1 que sí están englobadas por otras, y finalmente existen etiquetas nivel 2 que significa que están a un nivel más profundo. Por ejemplo, una etiqueta nivel 2 puede ser un "Time_Point" que aporte información sobre el nacimiento, por lo que se encontraría dentro de "Birth", y a su vez puede que esté dentro de un paréntesis. En la imagen 5.5 se muestra un ejemplo de esta situación:

```
Time_Point 39 53        March 27, 1911
Parenthesis 39 114      (March 27, 1911, in Sulphur Rock, Arkansas - June 11, 1970, in San Francisco)
Birth 39 80     March 27, 1911, in Sulphur Rock, Arkansas
```

**5.5 Etiquetas de niveles diferentes**

En esta imagen podemos observar una etiqueta que se encuentra a nivel 2 de profundidad (Time_Point), junto con las etiquetas nivel 1 y nivel 0 que la incluyen ("Birth" y "Parenthesis"). Estos datos corresponden al fichero de extensión .ann que se procesa para extraer la información.

3. Mediante la información relativa a las etiquetas contenidas dentro de una frase y el nivel de profundidad de las mismas, extraemos los bigramas presentes en la oración, tal como se ha explicado en el apartado anterior.
4. Se crea el listado de plantillas encontradas en el corpus de entrenamiento a partir de la información de cada frase (etiquetas que la forman y niveles jerárquicos asociados). Una vez procesadas todas las frases, se comprueba si existen duplicados o si una plantilla es un subconjunto de otra.

   En la imagen 5.6 hay un ejemplo de dos posibles plantillas, siendo la superior (A) un subconjunto de la inferior (B). El número final representa cuántas veces se ha encontrado este contenedor en el corpus. En este ejemplo, vemos que solamente se han procesado dos oraciones, y debido a que (A) se encuentra también contenida en (B) (los elementos de A son un subconjunto de los elementos de B), se ha incrementado su contador.

```
(A)    [[['Event', 0], ['Person', 0], ['Topic', 0]], 2]

(B)    [[['Event', 0], ['Person', 0], ['Topic', 0], ['Time_Point', 0]], 1]
```

**5.6 Ejemplo de plantillas**

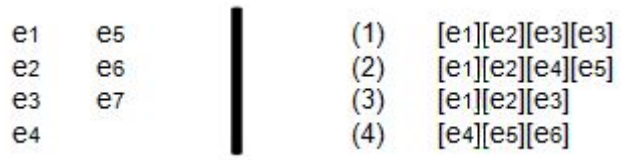
5. Se realizan cálculos estadísticos respecto al número de frases en las que aparece una etiqueta y también en cuántas frases aparecen dos etiquetas juntas (es decir, se calcula en cuántas frases aparece cada "par de etiquetas"). Estos cálculos permitirán saber la probabilidad de cada etiqueta y las probabilidades conjuntas.
6. Finalmente, con los cálculos estadísticos se obtiene para cada par de etiquetas el valor de su PMI.

### 5.2.4 Algoritmo de agrupación de unidades de contenido

Los datos obtenidos a partir del corpus de entrenamiento se han aplicado para agrupar las anotaciones de la biografía deseada, que desempeña el papel de base de conocimiento. Para este algoritmo es necesario un fichero con el listado de anotaciones (rellenado mediante la herramienta Brat) y el fichero con el resultado de la extracción de datos del corpus. Los pasos que sigue son los siguientes:

1) Almacenamiento de etiquetas de la biografía, resultados de bigramas, PMI y plantillas obtenidas (ignorando aquellas que tienen un número de usos bajo) a partir de los ficheros.
2) Se busca el nivel de profundidad de las etiquetas de la base de conocimiento para identificar cuáles se encuentran dentro de otras.
3) Se realiza una primera clasificación de etiquetas en plantillas. Estas plantillas están ordenadas de mayor tamaño a menor, y según se dispone de las etiquetas necesarias para "llenarlas", se van seleccionando. El resultado obtenido se optimiza para intentar "colocar" el mayor número de etiquetas en plantillas.

   Para ilustrar esta situación, se muestra en la imagen 5.7 un ejemplo de clasificación de etiquetas en plantillas:




**5.7 Clasificación en plantillas**

En la izquierda de la imagen 5.7 se ven las etiquetas e1…e7 pendientes de clasificar, y a la derecha están las 4 plantillas disponibles, ordenadas por tamaño. En seleccionar la primera plantilla, se puede observar que se dispone de e1, e2 y e3, pero se requeriría una segunda etiqueta e3 que no está en el listado de pendientes, por lo que esta plantilla se descarta al no poderse rellenar entera. Se seleccionaría por lo tanto la plantilla 2 con los elementos e1, e2, e4 y e5, entonces estos 4 elementos dejan de estar pendientes y se sigue trabajando con e3, e6 y e7. Se descartarían las dos plantillas siguientes (3 y 4) porque faltan elementos que se han puesto en la plantilla seleccionada previamente y con los tres elementos disponibles ahora no se pueden rellenar.

Una vez obtenido este resultado, al realizar la optimización se observa que seleccionando la plantilla 2 (como hemos hecho) sobran 3 etiquetas, pero seleccionando las plantillas 3 y 4 solamente se tendría una etiqueta (e7) como sobrante. Por lo tanto, el resultado final será escoger estas dos plantillas y descartar coger solamente la 2.

4) Ahora se dispone de algunas etiquetas agrupadas, pero se debe seguir trabajando con las etiquetas que no se han podido poner en plantillas. Ignorando por el momento las etiquetas de nivel jerárquico 1 y 2, se intentan agrupar las de nivel 0 según el valor del PMI: Si es positivo, significa que tienen atracción entre ellas y se pueden agrupar, mientras que si es negativo implica que deben ir separadas. Es posible ajustar el valor buscado de PMI de manera que se escojan valores más altos que 0, puesto que si dos etiquetas tienen un PMI cercano a 0 implica que son independientes una de otra y no que tengan atracción. Por lo tanto, para asegurar que las agrupaciones de etiquetas están formadas por conceptos con un mínimo de atracción entre ellos, el valor de PMI debe ser superior a 0.1 (o más, según el nivel de rigor deseado).

5) Una vez establecidos los "núcleos de etiquetas" de nivel 0 (agrupaciones de etiquetas en las que todas tienen atracción hacia una de ellas, que forma el "centro" del núcleo), se comprueba si se pueden crear conjuntos con ellos mediante los datos de bigramas y teniendo en cuenta que para crear una agrupación como mínimo debe haber tres etiquetas. El mínimo de etiquetas viene establecido por el menor tamaño de las plantillas, las cuales nunca tienen menos de 3 elementos.

   El cálculo empieza con el núcleo que tiene un PMI mayor entre sus componentes, para el cual analiza si entre las etiquetas que lo forman existe alguna combinación de bigramas que permita su creación. Puede suceder:

   a. Se encuentra una forma de "encajar" las etiquetas. Entonces formamos una agrupación nueva con las mismas y las eliminamos del listado de "etiquetas pendientes". Recalculamos los núcleos para tener en cuenta esta actualización.

   b. No hay ninguna manera de ordenar las etiquetas. Eliminamos el elemento que tiene menor atracción respecto al resto y recalculamos el valor de PMI de este núcleo. Actualizamos la información del total de núcleos, y seguimos con el siguiente de la lista.

   A continuación se muestra un ejemplo de creación de núcleos a partir de 4 etiquetas (imagen 5.8). A partir de cada una de ellas se establece un núcleo y se calcula el PMI, representando los valores del modo:

   *[etiqueta_centro, [(etiqueta_añadida, PMI respecto el núcleo)], PMI total]*

   Si una etiqueta no tiene atracción con ninguna otra (como e4 en el ejemplo), por defecto se le establece un valor de -1 y será procesada posteriormente.

Etiquetas: $e_1$, $e_2$, $e_3$, $e_4$

Núcleos:

[$e_1$, [($e_2$, '0.54'), ($e_3$, '0.26')], 0.4]  [$e_3$, [($e_1$, '0.26')], 0.26]

[$e_2$, [($e_1$, '0.54')], 0.54]  [$e_4$, [], -1]]

**5.8 Ejemplo de núcleos**

Explicación del ejemplo en la imagen 5.8:

Pese a que el segundo núcleo, e2+e1, es el de PMI mayor (0.54), al ser solamente dos etiquetas no se podrá formar una agrupación a partir de él. Por lo tanto, se elimina el elemento e1 de este núcleo y se actualiza el valor de PMI (que valdrá -1 debido a que no tiene más etiquetas). El siguiente núcleo más grande es el que tiene como centro e1 y su PMI vale 0.4, que sí podría formar una agrupación. Las etiquetas de este núcleo formarán una agrupación nueva en el caso que:

- Se encuentre un modo de ordenar los elementos por bigramas, y además
- e1 no tenga un valor de PMI mayor con alguna plantilla ya creada (lo cual implicaría que tiene mayor atracción hacia esa que hacia esta agrupación que se pretende crear),

Si se forma una agrupación nueva con estos tres elementos (e1, e2 y e3), se eliminan del listado de etiquetas pendientes y se actualizan los núcleos que fueran a utilizarlas (en este caso, solamente quedaba e3 relacionado con e1). Se seguirá iterando hasta que se hayan comprobado todos los núcleos. En este ejemplo solamente quedaría e4, que no tiene etiquetas con atracción hacia él, por lo que no se formaría ninguna agrupación nueva con esta etiqueta y se dejará para procesarla más adelante (en el paso siguiente).

6) Llegados a este punto, ya no se realizarán nuevas agrupaciones de etiquetas, sino que se añaden las etiquetas pendientes a los grupos ya existentes. Siguiendo con las posibles etiquetas nivel 0 que hayan quedado sueltas, se calcula el valor de PMI de cada una de ellas individualmente con todas las agrupaciones, y se ordena el resultado por atracción (PMI). Escogiendo el primer valor de este listado, se comprueba mediante bigramas si hay alguna manera de ordenar los elementos de la agrupación para incluir la etiqueta en ella (en la comprobación de bigramas solamente tenemos en cuenta etiquetas que tienen el mismo nivel jerárquico, y en este caso ignoramos las de nivel 1 ó 2):
    a. Si puede ponerse entre los elementos de la agrupación, se recalculan los valores de PMI de las etiquetas que siguen pendientes con todas las agrupaciones (incluida ésta última que acaba de modificarse), y se reordena el listado.

    b. Si no hay ninguna combinación de las etiquetas que permita dejarle un sitio en esa agrupación, se continúa por el siguiente elemento del listado. Este elemento puede ser el PMI de la misma etiqueta con otra agrupación, o bien de otra etiqueta con una agrupación cualquiera.

Para más aportar más detalles sobre los pasos 4-6 de este algoritmo, se ha adjuntado en Anexos la ejecución paso a paso de la agrupación de conceptos de una biografía y los resultados obtenidos (ver apartado 9.2).

7) Para la clasificación de etiquetas de nivel jerárquico 1 (las incluidas dentro de otras de nivel 0) se realiza un cálculo similar al anterior. Se ha extraído del corpus el listado de etiquetas que requieren de otras dentro (*Parenthesis*, *Birth* y *Death*) y también el de etiquetas que pueden o no contener a otras (*Time_Period*, *Time_Point*, *Description*, *Circumstance*). Inicialmente, se comprueba que las etiquetas que necesitan de otras dentro no se encuentren vacías. Una vez éstas contienen como mínimo una etiqueta, deja de realizarse distinción entre los dos grupos y se realiza una única lista. Se calcula el valor de PMI de todas las etiquetas pendientes de clasificar de nivel 1 con todas las agrupaciones que contengan como mínimo una de las etiquetas que se encuentran en el listado anterior, y se ordenan los resultados de mayor atracción a menor. A continuación se empieza tratando el valor mayor de PMI entre una etiqueta y una agrupación, y se procede a comprobar si se pueden juntar, teniendo en cuenta que es posible que haya más de una etiqueta de nivel 0 donde se puede poner:

    a. En el caso que solamente haya una etiqueta que pertenece a este listado, es decir, sólo hay una etiqueta que pueda hacer de "recipiente" de la de nivel 1 que se pretende incluir:

        i. Si no hay ninguna etiqueta de nivel 1 en la agrupación, no se tienen problemas en poner ésta.

        ii. Si hay otra etiqueta de nivel 1, se comprueba si tienen algún bigrama en común que las permita juntar.

        iii. Si ya hay más de una etiqueta de nivel 1, debe revisarse que todas ellas se puedan poner ordenadas mediante bigramas.

    b. Si se tiene más de una etiqueta que pueda actuar de recipiente para las de nivel 1:

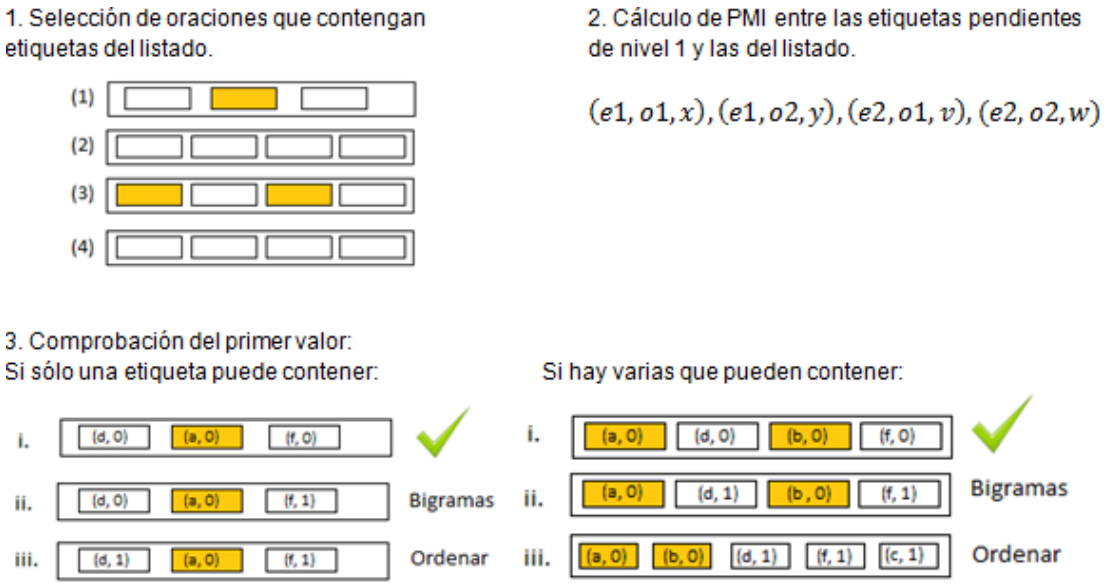
        i. Si hay menos elementos nivel 1 que elementos nivel 0 que pueden contenerlos, entonces la actual se añade sin problemas.

        ii. Si hay exactamente el mismo número de etiquetas de nivel 1 que etiquetas nivel 0 que pertenezcan al listado de posibles recipientes, es suficiente encontrar un solo bigrama entre las etiquetas de nivel 1 para juntarlas. Esto implicaría que esas dos etiquetas para las cuales hay un bigrama pueden ir juntas dentro de la misma etiqueta de nivel 0.

        iii. En el caso que haya más etiquetas nivel 1 que de las nivel 0 que las puedan contener, hay que buscar si hay combinaciones aceptables entre las etiquetas de nivel 1 de manera que puedan hacerse agrupaciones entre ellas dentro de las etiquetas de nivel jerárquico 0 que hacen de "recipientes".

Similarmente al tratamiento de las etiquetas pendientes de nivel 0, en este caso también cuando logramos poner la etiqueta en una agrupación recalculamos los valores de PMI entre las etiquetas pendientes y las frases. Del mismo modo, si no logramos poner la etiqueta, pasamos a procesar el elemento siguiente del listado.

En la imagen 5.9 se ha realizado una representación gráfica de este problema.




5.9 Clasificación de etiquetas nivel 1

En el ejemplo de la imagen 5.9 se han representado mediante cajas las agrupaciones de etiquetas, dentro de las cuales se encuentran marcadas de amarillo las etiquetas de nivel 0 que pueden contener otras. En el punto 1 se observa que se seleccionan las agrupaciones 1 y 3 porque tienen etiquetas de este tipo, y en el punto 2 se muestran los valores calculados de PMI de las etiquetas de nivel 1 pendientes de clasificar con estas agrupaciones (en este ejemplo se han considerado las etiquetas e1 y e2, mientras que o1 y o2 representan las agrupaciones). Seleccionando el mayor valor de PMI, se comprueba qué tipo de estructura tiene la agrupación, el cual será alguno de los reflejados en el punto 3. Si no hay etiquetas nivel 1 (estructura i), se puede poner la etiqueta actual de nivel 1 sin problemas. En este ejemplo, el mayor valor de PMI corresponde a la etiqueta e1 con la agrupación 1, la cual si consideramos que no contiene etiquetas de nivel 1 vemos que tiene la estructura i y solamente una etiqueta que pueda contener otra. Sin embargo, podemos encontrarnos con otros casos:

- En la estructura ii, se debe comprobar si existe un bigrama entre las etiquetas de nivel 1 (las que haya en la agrupación y la que se quiere poner).
- En la estructura iii, se tienen que ordenar las etiquetas nivel 1 dentro de las de nivel 0 para confirmar que se pueden tener todas, y si se tienen varias nivel 0 donde ponerlas hay que revisar todas las combinaciones posibles: En el ejemplo de la imagen deberían buscarse bigramas entre d, f, c y la etiqueta

que se quiere añadir y asegurar que se pueden formar dos grupos entre ellas, uno de los cuales estaría dentro de la etiqueta a y el otro en b.

8) Finalmente, queda pendiente clasificar las etiquetas de nivel jerárquico 2. Para estas etiquetas, solamente hace falta tener en cuenta aquellas agrupaciones que incluyen etiquetas de nivel 1 que pueden contener otras. Una vez seleccionadas las agrupaciones correspondientes, se aplica el mismo cálculo que en las etiquetas de nivel 1 (paso 7 del algoritmo) para clasificarlas.

# 6 Experimentos y Resultados

Con el objetivo de analizar la corrección del algoritmo de agrupación de conceptos, se ha ejecutado este algoritmo sobre 10 biografías aleatorias y se ha estudiado el resultado obtenido. Se ha contrastado la clasificación de conceptos propuesta del algoritmo con la organización en la biografía original calculando los valores de precisión y cobertura. Asimismo, se ha establecido un segundo algoritmo utilizado de baseline, el cual realiza la agrupación de elementos de un modo mucho más simple, y su relación respecto a la organización de la biografía original sirve como referencia para contrastar los valores obtenidos en el algoritmo desarrollado en este trabajo.

## 6.1 Algoritmo de baseline

Se ha realizado la clasificación de elementos de la biografía mediante un segundo algoritmo, el cual también se ha utilizado en el experimento para tener valores de referencia y poder observar el funcionamiento del algoritmo diseñado originalmente. Este algoritmo aportará información adicional a la obtenida en los cálculos del Valor-F (detallado en el apartado siguiente).

El algoritmo de baseline sigue los pasos siguientes:

1. Calcula el promedio de etiquetas por frase en el corpus.
2. Agrupa los elementos del texto en conjuntos del tamaño identificado en el paso 1. Para ello, no tiene en cuenta ningún tipo de restricción o dato que aporte información respecto a las relaciones entre unidades de conocimiento a clasificar, sino que simplemente va realizando agrupaciones siguiendo el orden en el que vienen listadas las anotaciones.
3. Termina cuando se queda sin más elementos que clasificar.

Así pues, teniendo en cuenta que en el corpus se ha identificado un promedio de 10 unidades semánticas por frase, este algoritmo crea conjuntos de este número de elementos. En algunos casos, la última agrupación creada no llega a los 10, ya que se queda sin más elementos que poner y debe dejar el conjunto con un tamaño más pequeño que el resto.

## 6.2 Métricas de evaluación

El análisis de los resultados de algoritmos de clasificación como los utilizados se realiza utilizando las medidas de precisión y cobertura aplicadas a cada una de las agrupaciones realizadas.

- Precisión: Mide la capacidad del algoritmo para seleccionar en una agrupación los mismos componentes (reflejados por las anotaciones) que forman una oración en la biografía original. Para calcular la precisión de una agrupación se aplica la fórmula:

$$precisión = \frac{nº\ etiquetas\ comunes\ en\ la\ oración\ y\ la\ agrupación}{longitud\ de\ la\ agrupación\ generada}$$

  Por lo tanto, en tener una precisión de 1 significa que todo lo que se ha seleccionado como contenido de la agrupación pertenece a una única oración en el texto. Es decir, el agrupamiento realizado está formado por un subconjunto de elementos de la oración. Asimismo, si se obtiene una precisión de 0 será porque no coincide ni una sola etiqueta en la agrupación respecto a la frase con la que se está comparando.

- Cobertura: Mide la capacidad del algoritmo para seleccionar en una agrupación sólo las anotaciones que forman una oración en el texto original. Su cálculo se realiza con la fórmula:

$$cobertura = \frac{nº\ etiquetas\ comunes\ en\ la\ oración\ y\ la\ agrupación}{longitud\ de\ la\ oración}$$

  En este caso, una cobertura de 1 implica que en el agrupamiento de contenido realizado se han seleccionado todas las etiquetas de una oración. Por lo tanto, dicha oración está formada por un subconjunto de elementos de la agrupación.

En el caso de obtener tanto la precisión como la cobertura con valor 1, será porque se ha realizado una agrupación de elementos que corresponde exactamente con el contenido de una oración.

Se muestra en la imagen 6.1 un ejemplo numérico de cálculo de precisión y cobertura, cuya explicación se añade debajo de ella.

| | | | |
|---|---|---|---|
| Agrupación: | e1 e2 e3 | | |
| O1: | e1 e2 e3 e4 | p1 = 3/3 | c1 = 3/4 |
| O2: | e1 e2 e4 e5 e6 | p2 = 2/3 | c2 = 2/5 |

6.1 Ejemplo de Precisión y Cobertura

En la imagen 6.1 se observa a la izquierda la agrupación de 3 elementos (e1, e2 y e3) y las dos oraciones O1 y O2, y a la derecha se muestran los resultados de los cálculos. La precisión de este agrupamiento respecto a O1 será 1 porque todos los

elementos seleccionados están dentro de la oración, pero el valor de su cobertura es 0.75 debido a que en la frase original hay un elemento más que no se ha incluido en la agrupación. Entre las etiquetas seleccionadas y O2 se puede observar que solamente hay dos etiquetas en común (e1 y 2e), por lo que los resultados de precisión y cobertura son peores respecto a los valores obtenidos con O1.

Como valor único ponderado entre la precisión y la cobertura, se ha aplicado el Valor-F, que se calcula con la fórmula:

$$F = 2 \cdot \frac{Precisión \cdot Cobertura}{Precisión + Cobertura}$$

Es posible añadir pesos a esta fórmula para favorecer la precisión o la cobertura, pero en este caso se han considerado ambas igual de relevantes.

Así pues, una vez ejecutado el algoritmo de clasificación de entidades semánticas en agrupaciones se ha procedido a calcular el Valor-F de cada uno de estos conjuntos respecto a las oraciones originales de la biografía. Para cada grupo de elementos se ha guardado el valor más alto de Valor-F respecto a las frases originales, y después se ha calculado el promedio de todos los Valor-F registrados. Este valor final es el que da una idea de lo parecido que es el resultado del algoritmo al texto de la biografía escrito por un humano. En el apartado siguiente se comentan en detalle los resultados obtenidos en la evaluación de las ejecuciones de este algoritmo en biografías aleatorias.

## 6.3 Análisis de los resultados

En este apartado se analizan los resultados obtenidos en las 10 biografías seleccionadas. Antes de presentar los resultados de estas biografías, se desglosan los cálculos de una de ellas para observar en detalle el procedimiento seguido.

El texto que se ha escogido como ejemplo para mostrar los valores detallados del solapamiento entre la biografía y las agrupaciones es el correspondiente a A. A. Allen, debido a que es el más breve de todos y contiene menos oraciones y etiquetas. El texto original está separado en 4 frases, y en los dos algoritmos se han obtenido 4 agrupaciones con los resultados que se muestran en la tabla 6-1 (P = Precisión, C=Cobertura, F=Valor-F) (en las tablas y figuras siguientes nos referimos al algoritmo de agrupación propuesto en este trabajo como “Algoritmo complejo”):

| | Agrupación1 | | | Agrupación2 | | | Agrupación3 | | | Agrupación4 | | | Promedio Valor-F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | C | F | P | C | F | P | C | F | P | C | F | |
| A. Complejo | 2/4 | 2/5 | 0.44 | 5/7 | 5/7 | 0.71 | 3/6 | 3/6 | 0.5 | 12/15 | 12/14 | 0.83 | 0.62 |
| A. Baseline | 5/10 | 5/7 | 0.59 | 4/10 | 4/6 | 0.5 | 7/10 | 7/14 | 0.58 | 1/2 | 1/5 | 0.29 | 0.49 |

**Tabla 6-1 Resultados de A. A. Allen**

Cabe destacar que las agrupaciones realizadas por los algoritmos se comparan con aquella frase de la biografía con la que se obtiene el mayor Valor-F y se descartan el resto de comparaciones.

Como se puede observar en la tabla 6-1, en ambos casos se ha obtenido el mismo número de agrupaciones que el de frases de la biografía original, debido esencialmente a que hay pocos elementos que clasificar (en ejecuciones con más etiquetas habitualmente no coincide el número de agrupaciones). Sin embargo, el promedio de Valor-F obtenido en los dos algoritmos es diferente, puesto que los conjuntos creados no son los mismos. En el caso del algoritmo de baseline tres agrupaciones tienen 10 elementos (tal como se aprecia en el denominador del cálculo de la precisión), mientras que para el último conjunto solamente quedaban 2 elementos. En cambio, en el algoritmo diseñado en este trabajo el número de elementos en una agrupación varía en función de los mismos.

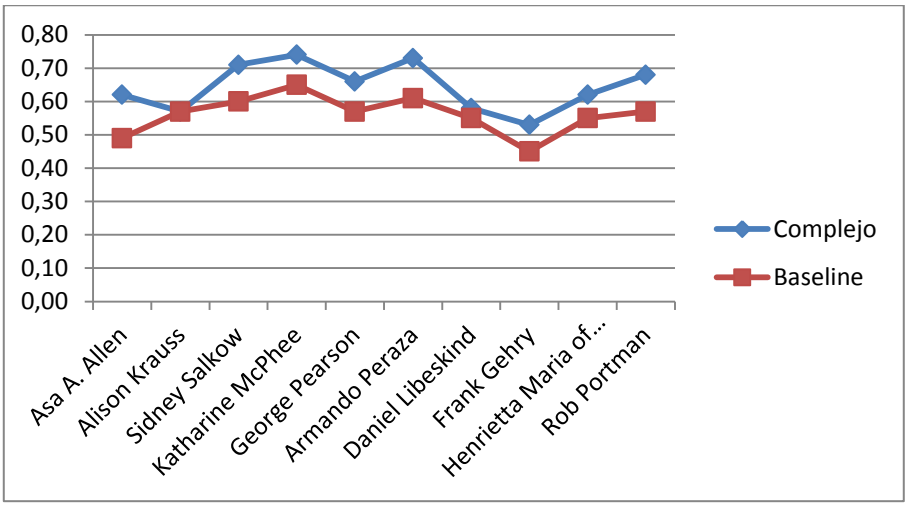
A continuación se muestra en la tabla 6-2 el promedio del Valor-F obtenido en ambos algoritmos para todas las biografías procesadas en este experimento, así como la comparación de agrupaciones creadas respecto al número de oraciones en el texto. En la última fila se ha añadido el valor promedio de los Valor-F obtenidos en cada caso.

| Biografía | Frases | Algoritmo Complejo | | Algoritmo Baseline | |
|---|---|---|---|---|---|
| | | Agrupaciones | Valor-F | Agrupaciones | Valor-F |
| Asa A. Allen | 4 | 4 | 0,62 | 4 | 0,49 |
| Alison Krauss | 9 | 15 | 0,57 | 12 | 0,57 |
| Sidney Salkow | 11 | 14 | 0,71 | 10 | 0,60 |
| Katharine McPhee | 13 | 13 | 0,74 | 10 | 0,65 |
| George Pearson | 15 | 15 | 0,66 | 12 | 0,57 |
| Armando Peraza | 8 | 8 | 0,73 | 8 | 0,61 |
| Daniel Libeskind | 6 | 8 | 0,58 | 7 | 0,55 |
| Frank Gehry | 6 | 12 | 0,53 | 10 | 0,45 |
| Henrietta Maria of France | 9 | 12 | 0,62 | 10 | 0,55 |
| Rob Portman | 14 | 17 | 0,68 | 13 | 0,57 |
| Promedio | | | 0,64 | | 0,56 |

**Tabla 6-2 Agrupaciones realizadas y Valor-F**

En el número de agrupaciones creadas se puede apreciar que el algoritmo diseñado tiende a crear más agrupaciones o, visto desde otro punto de vista, crea agrupaciones menores que el número de elementos de las oraciones originales.

A continuación se muestran los promedios del Valor-F de estas biografías gráficamente, en azul para el algoritmo implementado en este estudio y en rojo para el algoritmo de baseline:




**6.2 Promedio de Valor-F**

Como se aprecia en los valores de la imagen 6.2, los resultados obtenidos en el algoritmo diseñado en este trabajo son en todos los casos mejores respecto a los resultados del algoritmo de baseline (solamente en una ocasión se obtiene el mismo valor). En el primer algoritmo se han obtenido resultados acotados entre 0,53 y 0,74, mientras que en el baseline el Valor-F de las biografías se sitúa entre 0,45 y 0,65, un umbral inferior.

## 6.4 Detalle de cotas superiores e inferiores

A continuación se procederá a analizar con más detalle los valores menores y mayores obtenidos en el experimento anterior para el algoritmo diseñado en este trabajo.

Se han detectado 3 biografías para las cuales la agrupación de unidades sintácticas tiene un solapamiento bajo respecto al documento de texto original, puesto que el Valor-F obtenido en el experimento se encuentra por debajo de 0.6. Los nombres de los personajes de estas biografías son: F. Gehry, A. Krauss y D. Libeskind, con valores 0.53, 0.57 y 0.58 respectivamente. Por otro lado, se observan 3 personas cuyas agrupaciones de elementos de sus biografías tienen un promedio de Valor-F por encima de 0.7: K. McPhee, A. Peraza y S. Salkow, que en concreto tienen respectivamente 0.74, 0.73 y 0.71.

### 6.4.1 Análisis de la precisión y cobertura

El motivo de estas diferencias en los resultados puede deberse a diversos factores, el primero a destacar es el número de agrupaciones creadas respecto al número de oraciones del texto. Se puede observar que para F. Gehry (personaje para el que se ha obtenido el valor menor) se han creado 12 conjuntos de elementos, mientras que en su biografía hay 6 frases. Asimismo, para A. Krauss también se han creado prácticamente el doble de agrupaciones respecto al número de oraciones del texto (17 conjuntos generados de 9 oraciones). En el caso de D. Libeskind se observan el mismo número de agrupaciones que de frases presentes en el texto.

El siguiente factor a resaltar es la cobertura y precisión en las agrupaciones de estos textos, las cuales se muestran en la tabla 6-3 junto con el valor-F obtenido (Precisión (P), Cobertura (C) y Valor-F (F)):

| | F. Gehry | | | A. Krauss | | | D. Libeskind | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | C | F | P | C | F | P | C | F |
| A1 | 5/9 | 5/5 | 0,71 | 5/9 | 5/8 | 0,59 | 4/8 | 4/10 | 0,44 |
| A2 | 4/6 | 4/8 | 0,57 | 5/9 | 5/8 | 0,59 | 5/5 | 5/10 | 0,67 |
| A3 | 4/6 | 4/8 | 0,57 | 5/7 | 5/12 | 0,53 | 9/11 | 9/19 | 0,60 |
| A4 | 3/5 | 3/8 | 0,46 | 7/9 | 7/17 | 0,54 | 5/7 | 5/10 | 0,59 |
| A5 | 4/8 | 4/8 | 0,50 | 3/8 | 3/8 | 0,38 | 4/4 | 4/9 | 0,62 |
| A6 | 5/5 | 5/16 | 0,48 | 7/10 | 7/11 | 0,67 | 3/4 | 3/9 | 0,46 |
| A7 | 1/3 | 1/5 | 0,25 | 7/9 | 7/11 | 0,70 | 5/10 | 5/10 | 0,50 |
| A8 | 2/4 | 2/5 | 0,44 | 6/8 | 6/12 | 0,60 | 10/16 | 10/11 | 0,74 |
| A9 | 5/5 | 5/16 | 0,48 | 5/7 | 5/12 | 0,53 | | | |
| A10 | 3/9 | 3/5 | 0,43 | 3/5 | 3/11 | 0,37 | | | |
| A11 | 17/18 | 17/27 | 0,76 | 3/4 | 3/11 | 0,40 | | | |
| A12 | 10/17 | 10/12 | 0,69 | 8/10 | 8/14 | 0,67 | | | |
| A13 | | | | 7/8 | 7/14 | 0,64 | | | |
| A14 | | | | 4/8 | 4/8 | 0,50 | | | |
| A15 | | | | 8/9 | 8/11 | 0,80 | | | |
| Promedio | 0,64 | 0,51 | 0,53 | 0,71 | 0,50 | 0,57 | 0,74 | 0,51 | 0,58 |

**Tabla 6-3 Resultados con valor-F promedio bajo**

Cada fila de la tabla 6-3 representa una agrupación: En F. Gehry se han realizado 12, para A. Krauss 15 y con D. Libeskind 8. Tal como se puede observar, el valor de la cobertura para los tres casos es inferior al valor de la precisión. Esto significa que en general el número de etiquetas comunes entre las agrupaciones realizadas y las oraciones del texto es menor respecto a los elementos de las frases que respecto a los conjuntos creados.

Un motivo por el que puede provocarse esta situación es porque las agrupaciones tienen menos elementos que las oraciones, por lo que es más fácil que coincidan más de ellos respecto al total. Por ejemplo, en la segunda agrupación de G. Gehry presente en la tabla anterior se observan los valores 4/6 para precisión y 4/8 para la cobertura, lo cual implica:

$$precisión = \frac{4\ elementos\ comunes\ entre\ agrupación\ y\ frase}{6\ elementos\ en\ la\ agrupación} = 0.67$$

$$cobertura = \frac{4\ elementos\ comunes\ entre\ agrupación\ y\ frase}{8\ elementos\ en\ la\ frase} = 0.50$$

Como se observa en este ejemplo, el valor de la precisión es mayor debido a que 4 respecto 6 da un valor más alto que 4 respecto a 8. Esta situación se corresponde con la observación realizada previamente relativa a que se han generado más agrupaciones de elementos que el número de oraciones del texto, por lo que las primeras tienen en general menos elementos. En este punto se podría plantear la hipótesis que generar agrupaciones mayores mejoraría los resultados obtenidos en el experimento, puesto que podrían encontrarse más elementos comunes entre agrupaciones y frases.

Para contrastar esta información, a continuación se presentan los resultados detallados de los 3 personajes para los cuales se han obtenido los mejores valores (Precisión (P), Cobertura (C) y Valor-F (F)):

| | K. McPhee | | | A. Peraza | | | S. Salkow | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | C | F | P | C | F | P | C | F |
| A1 | 6/6 | 6/6 | 1 | 6/8 | 6/11 | 0,63 | 3/3 | 3/7 | 0,60 |
| A2 | 4/5 | 4/6 | 0,73 | 5/7 | 5/9 | 0,63 | 3/3 | 3/7 | 0,60 |
| A3 | 5/5 | 5/9 | 0,71 | 5/6 | 5/9 | 0,67 | 3/3 | 3/7 | 0,60 |
| A4 | 6/7 | 6/9 | 0,75 | 11/11 | 11/11 | 1 | 6/7 | 6/7 | 0,86 |
| A5 | 4/5 | 4/8 | 0,62 | 4/9 | 4/8 | 0,44 | 4/7 | 4/7 | 0,57 |
| A6 | 5/9 | 5/5 | 0,71 | 8/10 | 8/8 | 0,89 | 5/7 | 5/6 | 0,77 |
| A7 | 5/9 | 5/5 | 0,71 | 11/15 | 11/11 | 0,85 | 5/6 | 5/6 | 0,83 |
| A8 | 5/8 | 5/8 | 0,62 | 8/14 | 8/9 | 0,70 | 6/7 | 6/6 | 0,92 |
| A9 | 5/8 | 5/8 | 0,62 | | | | 6/7 | 6/6 | 0,92 |
| A10 | 5/7 | 5/5 | 0,83 | | | | 6/7 | 6/6 | 0,92 |
| A11 | 5/7 | 5/5 | 0,83 | | | | 7/7 | 7/22 | 0,48 |
| A12 | 6/8 | 6/9 | 0,71 | | | | 3/6 | 3/6 | 0,50 |
| A13 | 8/12 | 8/10 | 0,73 | | | | 4/11 | 4/6 | 0,47 |
| A14 | | | | | | | 12/13 | 12/15 | 0,86 |
| Promedio | 0,74 | 0,78 | 0,74 | 0,73 | 0,76 | 0,73 | 0,81 | 0,69 | 0,71 |

**Tabla 6-4 Resultados con valor-F promedio alto**

Cada fila de la tabla 6-4 representa una agrupación: En K. McPhee se han realizado 13, para A. Peraza 8 y con S. Salkow 14. En este caso se observa que para K. McPhee y A. Peraza el promedio de precisión de todas las agrupaciones es

ligeramente inferior que el promedio de cobertura, pero para S. Salkow vuelve a darse el caso que la precisión es bastante mayor que la cobertura, como se ha observado para los 3 casos con valores bajos de promedio de Valor-F. Para McPhee y Peraza se ha obtenido el mismo número de agrupaciones que de oraciones formadas en la biografía, mientras que en S. Salkow se han hecho 14 conjuntos de elementos frente a los 11 del texto original, lo cual puede tener relación con que en cada personaje el valor de la cobertura sea mejor que el de la precisión.

### 6.4.2 Análisis del tipo de anotaciones

Las agrupaciones de elementos del texto vienen determinadas por el tipo de etiquetas presentes en él, puesto que a partir de su tipo se aplican los cálculos relativos a plantillas, PMI y bigramas utilizados en el algoritmo (detallados en el apartado 5). Por lo tanto, las anotaciones presentes en el texto causan la creación de agrupaciones tan diferentes como las que se encuentran por ejemplo en los resultados de K. McPhee, donde se puede apreciar (en la tabla 6-4, gracias al cálculo de precisión) una agrupación de 12 elementos y otra de 5. Así pues, a continuación se procede a comparar el tipo de etiquetas utilizado en esta biografía (es en la que se ha obtenido el mejor resultado en la generación de agrupaciones), con las etiquetas presentes en el texto de F. Gehry (sus agrupaciones tienen un Valor-F total más bajo).

En la tabla 6-5 se listan las anotaciones pertenecientes a las biografías de K. McPhee (derecha) y F. Gehry (izquierda). Para McPhee hay un total de 88 etiquetas de nivel jerárquico 0, 3 de nivel 1 (que se encontrarían dentro de otras de nivel 0) y 2 de nivel 2 (las cuales están dentro de etiquetas de nivel 1). Entre las anotaciones de Gehry se identifican 72 de nivel jerárquico 0, 17 de nivel 1 y 3 de nivel 2.

| **Nivel 0** | | **Nivel 1** | |
|---|---|---|---|
| Topic | 17 | Property | 1 |
| Object | 13 | Birth | 1 |
| Event | 10 | Topic | 1 |
| Time_Point | 10 | | **3** |
| Organization | 7 | | |
| Event_Establish | 6 | | |
| Number | 6 | | |
| Profesion | 5 | | |
| Person | 4 | | |
| Ownership | 4 | | |
| Property | 3 | **Nivel 2** | |
| Reference_Relation | 3 | Event | 1 |
| Reference | 1 | Time_Point | 1 |
| Description | 1 | | **2** |
| Parenthesis | 1 | | |
| **Total** | **88** | | |

| **Nivel 0** | | **Nivel 1** | |
|---|---|---|---|
| Topic | 20 | Topic | 4 |
| Location | 12 | Profesion | 1 |
| Property | 9 | Set | 1 |
| Event | 8 | Time_Point | 2 |
| Ownership | 6 | Event | 2 |
| Person | 4 | Property | 3 |
| Reference_Relation | 3 | Object | 1 |
| Object | 3 | Birth | 1 |
| Reference | 2 | Number | 2 |
| Profesion | 2 | | **17** |
| Organization | 2 | **Nivel 2** | |
| Parenthesis | 1 | Event | 1 |
| Time_Point | 1 | Time_Point | 1 |
| Event_Establish | 1 | Person | 1 |
| Description | 1 | | **3** |
| **Total** | **72** | | |

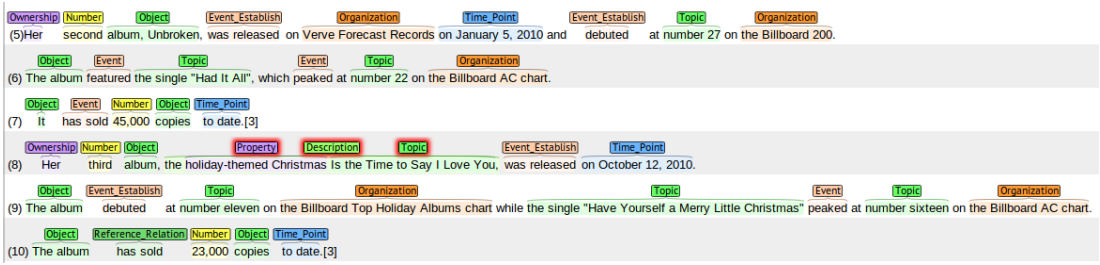
**Tabla 6-5 Etiquetas de K. McPhee y F. Gehry**

En estos datos se aprecian dos características que pueden ser claves para explicar la diferencia de los resultados obtenidos:

1. El número de etiquetas de nivel jerárquico 1 y de nivel 2 es inferior para aquél texto en el cual se han obtenido mejores resultados (en la tabla 6-5, a la izquierda). Por lo tanto, es más probable que se realice una agrupación con ellos similar a la oración de la biografía que los contiene, dando una precisión y cobertura mayor.
2. En ambos casos hay un número similar de etiquetas de nivel jerárquico 0, sin embargo, en el caso de F. Gehry se observa que un 45% corresponden a únicamente dos tipos de etiquetas ("Topic" supone un 28% del total y "Location" un 17%). En cambio, para K. McPhee el 40% del total de anotaciones de nivel jerárquico 0 está formado por tres tipos de etiquetas ("Topic" aparece un 19% de las ocasiones, "Object" un 15% y "Event" supone un 11% del total).

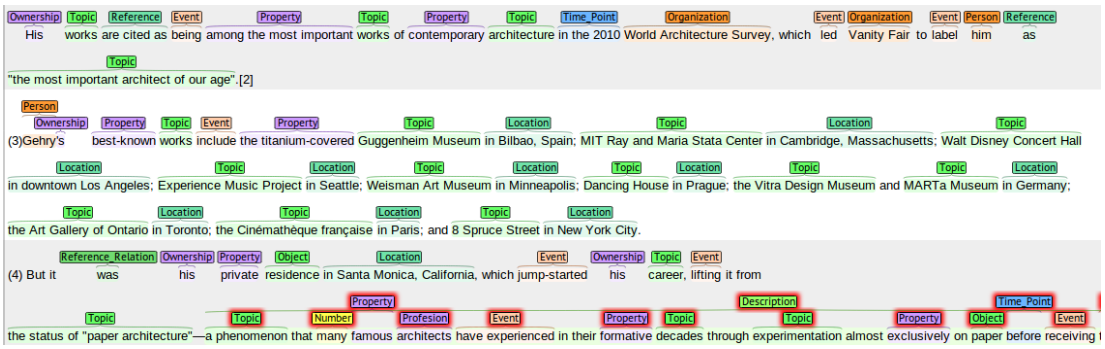
En cuanto a la estructura de las oraciones de la biografía, cabe destacar que en el caso del arquitecto F. Gehry se han escrito oraciones largas en las que se enumeran sus trabajos: Los nombres de edificios ("Topic") y los lugares de construcción ("Location"). En cambio, en la biografía de K. McPhee cada oración transmite un mensaje de un modo simple, sin enumeraciones ni frases compuestas. Por lo tanto, es más fácil reproducir la estructura de la biografía de McPhee que la de Gehry. Estas estructuras se pueden comparar en las imágenes 6.3 y 6.4 a continuación.

En la imagen 6.3 vemos un fragmento de la biografía de K. McPhee con anotaciones. Cada línea corresponde con una frase, las cuales se observa que son breves y por lo general solamente contienen una cláusula.




**6.3 Frases anotadas de K. McPhee**

En la imagen 6.4 tenemos un fragmento de la biografía de F. Gehry con anotaciones. En ella se observan tres oraciones (la última aparece cortada), las cuales son largas y contienen enumeraciones de elementos (véase que el color de las etiquetas de la segunda frase repite dos tonos verdes varias veces).




**6.4 Frases anotadas de F. Gehry**

En las otras dos biografías para las cuales los resultados obtenidos han sido bajos (correspondientes a D. Libeskind y A. Krauss), se ha comprobado que en el caso de Libeskind el personaje también es arquitecto y se repite la situación en la cual se realizan oraciones largas con enumeraciones de sus trabajos. Para A. Krauss la descripción de su trabajo es ligeramente diferente, puesto que es artista musical, sin embargo las frases contienen enumeraciones en cuanto a trabajos realizados y premios ganados, sin explicaciones detalladas de los mismos, y son oraciones complejas en cuanto a que contienen más de una proposición.

Las oraciones de las biografías en las cuales el algoritmo de agrupación de conceptos ha obtenido mejores resultados (A. Peraza y S. Salkow) tienen estructuras diferentes a las anteriores. En el caso de Peraza, la biografía se centra en la gente con la cual el personaje se ha relacionado durante su vida y no especifica obras concretas realizadas por él o premios ganados. A diferencia de las enumeraciones realizadas para los arquitectos, este texto se ha redactado de manera que las menciones a personajes se reparten entre diferentes oraciones, y no hay una única frase o dos que los contengan todos. Para Salkow el estilo de redactado de la biografía es similar a una historia y no se aprecian las enumeraciones presentes en los otros textos, dando explicaciones más detalladas de diferentes aspectos de su vida.

Así pues, se aprecia que el contenido de las biografías es clave para la organización posterior del mismo. Es decir, si el texto original contiene variedad de tipo de anotaciones (un número similar de “Person”, “Topic”, “Event”, etc.) y oraciones simples sin enumeraciones de logros, es más probable generar agrupaciones similares a las frases.

# 7 Conclusiones y Trabajo futuro

En este proyecto se ha propuesto un componente de Planificación de Textos de un sistema de Generación de Lenguaje Natural. Hemos trabajado con etiquetas semánticas, las cuales se han agrupado mediante plantillas (creadas dinámicamente a partir de un corpus de entrenamiento) en un enfoque top-down y posteriormente se ha aplicado un punto de vista bottom-up para terminar de clasificar los conocimientos individualmente en agrupaciones nuevas o dentro de las ya existentes. Como se ha mostrado en la sección 1, las etiquetas semánticas pueden relacionarse directamente con una representación RDF de una base de conocimiento. Por lo tanto, este sistema de generación es más flexible que aquellos que utilizan plantillas estáticas predefinidas, y además utilizar RST aporta la seguridad de que toda la información de la base de conocimiento se encuentra dentro de las agrupaciones.

Los resultados obtenidos en los experimentos muestran que la propuesta funciona bien en el dominio de las biografías. Es especialmente útil en bases de conocimiento con variedad en los tipos de conceptos presentes, donde los textos creados tienen una estructura narrativa sin enumeraciones predominantes (se comunican hechos descriptivos, sin simplemente expresar secuencias de elementos). La agrupación de conceptos utilizando datos extraídos de un corpus de entrenamiento (plantillas, PMI y bigramas) tiene un resultado similar a las agrupaciones realizadas en oraciones por parte de un humano, por lo que se puede considerar que sí es posible encontrar una solución automatizada a este problema en planificación de textos.

Algunas pautas sugeridas de trabajo futuro son:

- Ampliar la información relativa a las relaciones de etiquetas introduciendo relaciones discursivas, y no centrarse únicamente en los bigramas existentes entre ellas.

- Hacer experimentos con bases de conocimiento FRED (ontologías RDF enriquecidas con etiquetas semánticas), lo cual sería la aplicación en un escenario real de la técnica propuesta en este trabajo.

- Aplicar el punto de vista utilizado en este estudio para otros dominios, más allá de los textos de biografías de famosos. Sería interesante comprobar los resultados en bases de conocimiento relativas a otros ámbitos, por ejemplo en biología o deportes, donde también se puedan agrupar unidades de conocimiento de un tema concreto.

- Identificar dentro de las oraciones del corpus un nivel jerárquico más, relativo a cláusulas sencillas. En este trabajo no se han realizado distinciones entre oraciones simples y compuestas, y se podría comprobar si mediante el análisis de frases simples en el corpus los resultados de la planificación de texto son más precisos (en cuanto a la similitud con los textos originales). Por ejemplo, en la frase *Hoy llueve y mañana hará sol*, pueden identificarse dos oraciones diferentes separadas por la conjunción “y”, en lugar de tratarlas como una sola.

- Se puede profundizar el trabajo relativo a las jerarquías y en cómo mejorar la precisión con la que se estructuran unos elementos dentro de otros. Se podría sacar información adicional sobre las relaciones entre etiquetas en términos de valores estadísticos sobre cómo se estructuran estas jerarquías en el corpus. El objetivo de este planteamiento sería mejorar la distribución de conceptos etiquetados de niveles superiores (nivel 1 ó 2, aquellos que deben estar dentro de otros) según su "relación jerárquica" o la probabilidad de que estén dentro de una etiqueta de nivel inferior concreta. Así, no solamente se dispondría de un listado de "etiquetas que pueden contener otras", sino que además habría información sobre qué etiquetas es más probable que estén dentro de otras concretas y ayudar en su estructuración.

En resumen, los resultados obtenidos en este enfoque han sido positivos, pero en el futuro aún pueden mejorarse trabajando en ciertos aspectos pendientes de revisar.

# 8 Referencias


[1] F. Draicchio, A. Gangemi, V. Presutti, and A. G. Nuzzolese, “FRED : From Natural Language Text to RDF and OWL in One Click,” *Semant. Web ESWC*, vol. 7955, pp. 263–267, 2013.

[2] G. G. Chowdhury, “Natural language Processing,” *Annu. Rev. Inf. Sci. Technol.*, vol. 37, pp. 51–89, 2003.

[3] E. Reiter and R. Dale, “Building Applied Natural Language Generation Systems,” *Natural Language Engineering*, 1997.

[4] N. Bouayad-Agha, G. Casamayor, and L. Wanner, “Natural Language Generation in the Context of the Semantic Web,” *Semant. Web – Interoperability, Usability, Appl.*, vol. 00, pp. 1–22, 2012.

[5] E. Hovy, “Automated discourse generation using discourse structure relations,” *Artif. Intell.*, 1993.

[6] M. Taboada and W. Mann, “Rhetorical structure theory: Looking back and moving ahead,” *Discourse Stud.*, vol. 8, no. 3, pp. 423–459, Jun. 2006.

[7] K. R. McKeown, “Discourse Strategies for Generating Natural-Language Text,” *Artif. Intell.*, vol. 27, pp. 1–41, 1985.

[8] N. Guarino, “Formal Ontology and Information Systems,” *Proceedings of FOIS’98*, Trento, Italy, pp. 3–15, 1998.

[9] W. Swartout, T. Russ, and R. Macgregor, “Ontologies,” *Intel. Syst. their Appl. IEEE*, vol. 1094, pp. 18–19, 1999.

[10] P. Giaretta and N. Guarino, “Ontologies and knowledge bases towards a terminological clarification,” 1995.

[11] S. Candan, H. Liu, and R. Suvarna, “Resource description framework: metadata and its applications,” *ACM SIGKDD Explor. Newsl.*, vol. 3, no. 1, pp. 6–19, 2001.

[12] S. U. Elizabeth D. Liddy, “Natural Language Processing,” in *Encyclopedia of Library and Information Science*, 2n ed., NY: Marcel Decker, Inc, 2001.

[13] H. Assal, J. Seng, and F. Kurfess, “Semantically-enhanced information extraction,” *Proc. from 2011 IEEE Aerosp. Conf. Big Sky*, pp. 1–14, 2011.

[14] F. Llopis, R. Muñoz, A. Suárez, and A. Montoyo, “EXIT: Propuesta de un sistema de extracción de información de textos notariales.,” Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

[15] G. Benito, “Revisión de las gramáticas de casos,” *Philol. hispalensis*, pp. 257–273, 1992.

[16] Fillmore, “Topics in Lexical Semantics,” *Curr. Issues Linguist. Theory*, pp. 76–138, 1977.

[17] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The Berkeley FrameNet Project,” *Int. Comput. Sci. Insitute*, pp. 86–90, 1997.

[18] DBpedia, “wiki.dbpedia.org.” [Online]. Available: http://wiki.dbpedia.org/About. [Accessed: 20-Jun-2013].

[19] “RDFaCE.” [Online]. Available: http://rdface.aksw.org/. [Accessed: 20-Oct-2013].

[20] S. Pyysalo, P. Stenetorp, G. Topić, and T. Ohta, “Brat rapid annotation tool,” 2010. [Online]. Available: http://brat.nlplab.org/index.html. [Accessed: 15-Nov-2013].

[21] W. Zhang and Z. Deng, “Enquiring Semantic Relations among RDF Triples,” *11th Int. Symp. Distrib. Comput. Appl. to Business, Eng. Sci.*, pp. 477–481, Oct. 2012.

[22] R. Fano, “Transmission of Information: A Statistical Theory of Communication,” *MIT Press, Cambridge, MA*, 1961.

[23] K. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, 1990.

# 9 Anexos

Como información complementaria a la aportada en este trabajo, se incluye a modo de Anexos el listado de plantillas extraídas del corpus y un ejemplo de la ejecución paso a paso del algoritmo de agrupación.

## 9.1 Plantillas extraídas

A continuación se pueden observar las estructuras de etiquetas extraídas del corpus de entrenamiento que han sido utilizadas en el algoritmo. Estas plantillas, explicadas en detalle en el apartado 5.2, tienen dos datos clave:

- Etiquetas, ordenadas alfabéticamente, y su nivel de profundidad correspondiente.
- Número de veces que se ha encontrado esa estructura en los textos de entrenamiento, incluso siendo un "subconjunto" de otra plantilla.

De los 100 textos que forman el corpus de entrenamiento, se han obtenido un total de 866 plantillas, pero para el algoritmo ejecutado en este trabajo se han utilizado las 200 más comunes, que son las que se listan en este Anexo. La decisión de escoger este número de etiquetas se tomó en base a que no se quería usar un número demasiado elevado y, además, se quería asegurar que las plantillas utilizadas eran comunes en diversos textos y no correspondían a textos concretos.

[[['Event', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 471]
[[['Event', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 451]
[[['Event', 0], ['Person', 0], ['Topic', 0]], 418]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0]], 378]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 344]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 340]
[[['Event', 0], ['Time_Point', 0], ['Topic', 0]], 277]
[[['Event', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 243]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0]], 241]
[[['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 228]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 223]
[[['Event', 0], ['Ownership', 0], ['Topic', 0], ['Topic', 0]], 218]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 205]
[[['Person', 0], ['Reference_Relation', 0], ['Topic', 0]], 205]
[[['Event', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 202]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 183]

[[['Event', 0], ['Person', 0], ['Property', 0], ['Topic', 0]], 179]
[[['Event', 0], ['Object', 0], ['Topic', 0], ['Topic', 0]], 174]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Topic', 0], ['Topic', 0]], 173]
[[['Event', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 172]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 171]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Time_Point', 0]], 165]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 162]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 160]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Time_Point', 0], ['Time_Point', 0]], 158]
[[['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 153]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 151]
[[['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0]], 139]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 136]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Topic', 0]], 127]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Topic', 0]], 127]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0]], 124]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 123]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0]], 122]
[[['Person', 0], ['Property', 0], ['Reference_Relation', 0], ['Topic', 0]], 120]
[[['Event', 0], ['Ownership', 0], ['Time_Point', 0], ['Topic', 0]], 117]
[[['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 115]
[[['Event', 0], ['Property', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 114]
[[['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 113]
[[['Event', 0], ['Ownership', 0], ['Property', 0], ['Topic', 0], ['Topic', 0]], 111]
[[['Event', 0], ['Location', 0], ['Person', 0], ['Time_Point', 0]], 110]
[[['Event', 0], ['Number', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 109]
[[['Event', 0], ['Location', 0], ['Person', 0], ['Topic', 0]], 109]
[[['Event', 0], ['Ownership', 0], ['Property', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 108]
[[['Person', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 107]
[[['Event', 0], ['Organization', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 106]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Person', 0], ['Person', 0], ['Topic', 0]], 106]

[[['Ownership', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 105]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 104]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Location', 0], ['Person', 0], ['Topic', 0]], 104]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 103]
[[['Event', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0]], 101]
[[['Event', 0], ['Person', 0], ['Reference', 0], ['Topic', 0]], 97]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Person', 0], ['Topic', 0]], 96]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 96]
[[['Event', 0], ['Person', 0], ['Reference', 0], ['Topic', 0], ['Topic', 0]], 95]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Property', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 94]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 91]
[[['Event_Establish', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 89]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 88]
[[['Event', 0], ['Person', 0], ['Property', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 87]
[[['Event_Establish', 0], ['Event_Establish', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 87]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Property', 0]], 85]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Topic', 0], ['Topic', 0]], 81]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 81]
[[['Event', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 81]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 80]
[[['Event', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 80]
[[['Organization', 0], ['Person', 0], ['Profesion', 0], ['Reference_Relation', 0]], 80]
[[['Ownership', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 78]
[[['Event', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0], ['Topic', 0]], 78]
[[['Event', 0], ['Location', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0]], 78]

[[['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 76]
[[['Event', 0], ['Number', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Person', 0]], 75]
[[['Event', 0], ['Object', 0], ['Ownership', 0], ['Time_Point', 0]], 73]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 72]
[[['Event', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 72]
[[['Event', 0], ['Object', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0]], 72]
[[['Organization', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Reference_Relation', 0]], 72]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 71]
[[['Organization', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Reference_Relation', 0]], 70]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 70]
[[['Ownership', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 69]
[[['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Topic', 0]], 68]
[[['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 67]
[[['Person', 0], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Topic', 0]], 67]
[[['Event', 0], ['Location', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 66]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 65]
[[['Location', 0], ['Location', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Reference_Relation', 0]], 65]
[[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Topic', 0]], 65]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 64]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0]], 63]
[[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Reference_Relation', 0], ['Topic', 0]], 63]
[[['Event_Establish', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 63]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Organization', 0], ['Person', 0]], 62]
[[['Event', 0], ['Number', 0], ['Object', 0], ['Object', 0], ['Time_Point', 0]], 62]
[[['Person', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 61]
[[['Event', 0], ['Person', 0], ['Time_Period', 0], ['Topic', 0]], 61]

[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 60]
[[['Event', 0], ['Number', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 60]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Object', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0]], 59]
[[['Event', 0], ['Location', 0], ['Organization', 0], ['Organization', 0], ['Organization', 0], ['Person', 0]], 59]
[[['Person', 0], ['Profesion', 0], ['Property', 0], ['Time_Point', 0], ['Topic', 0]], 59]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 58]
[[['Event', 0], ['Object', 0], ['Ownership', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 58]
[[['Event', 0], ['Ownership', 0], ['Property', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 58]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 57]
[[['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 57]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 56]
[[['Event_Establish', 0], ['Organization', 0], ['Person', 0]], 56]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 55]
[[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 55]
[[['Event', 0], ['Person', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 54]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 54]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Reference', 0], ['Topic', 0]], 54]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Property', 0], ['Reference', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 53]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Related', 0], ['Topic', 0], ['Topic', 0]], 53]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Reference', 0], ['Topic', 0], ['Topic', 0]], 53]
[[['Event', 0], ['Event', 0], ['Number', 0], ['Person', 0], ['Property', 0], ['Topic', 0]], 51]
[[['Event', 0], ['Organization', 0], ['Ownership', 0], ['Topic', 0], ['Topic', 0]], 51]
[[['Event', 0], ['Object', 0], ['Person', 0], ['Profesion', 0]], 51]
[[['Event', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 50]
[[['Object', 0], ['Object', 0], ['Ownership', 0], ['Reference_Relation', 0]], 50]
[[['Object', 0], ['Ownership', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 49]

[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Time_Point', 2]], 49]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 49]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Reference_Relation', 0], ['Topic', 0]], 47]
[[['Event', 0], ['Object', 0], ['Object', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 46]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 46]
[[['Event', 0], ['Person', 0], ['Profesion', 0], ['Reference', 0], ['Topic', 0]], 46]
[[['Description', 0], ['Event', 0], ['Person', 0], ['Topic', 0]], 46]
[[['Event_Establish', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0]], 46]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Time_Point', 0]], 45]
[[['Event', 0], ['Organization', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 45]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Organization', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 44]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Topic', 0], ['Topic', 0]], 44]
[[['Event', 0], ['Location', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0]], 44]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Person', 0], ['Property', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 43]
[[['Event', 0], ['Event', 0], ['Number', 0], ['Person', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 43]
[[['Event', 0], ['Set', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 43]
[[['Event', 0], ['Object', 0], ['Object', 0], ['Person', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0]], 43]
[[['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 43]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 42]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 42]
[[['Event', 0], ['Event', 0], ['Set', 0], ['Time_Point', 0], ['Time_Point', 0]], 42]
[[['Ownership', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 41]
[[['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], 41]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Object', 0], ['Person', 0], ['Property', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 39]

[[['Location', 0], ['Ownership', 0], ['Ownership', 0], ['Property', 0], ['Property', 0], ['Topic', 0], ['Topic', 0]], 39]
[[['Event', 0], ['Location', 0], ['Location', 0], ['Object', 0], ['Topic', 0]], 39]
[[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0]], 39]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 38]
[[['Birth', 1], ['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Person', 2], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0]], 38]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 2], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 38]
[[['Event', 0], ['Organization', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 38]
[[['Birth', 0], ['Event', 1], ['Event', 0], ['Location', 1], ['Person', 0], ['Person', 0]], 38]
[[['Event', 0], ['Location', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0]], 38]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 37]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 37]
[[['Event_Establish', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0]], 37]
[[['Event_Establish', 0], ['Number', 0], ['Object', 0], ['Person', 0]], 37]
[[['Event', 0], ['Number', 0], ['Object', 0], ['Time_Point', 0], ['Time_Point', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 36]
[[['Event', 0], ['Event', 0], ['Event', 0], ['Organization', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Topic', 0], ['Topic', 0]], 36]
[[['Event', 0], ['Event', 0], ['Number', 0], ['Object', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 36]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Property', 0], ['Reference', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0]], 36]
[[['Event', 0], ['Number', 0], ['Number', 0], ['Ownership', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 36]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Ownership', 0], ['Person', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 35]
[[['Event', 0], ['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Related', 0]], 35]
[[['Event_Establish', 0], ['Person', 0], ['Property', 0], ['Topic', 0]], 35]
[[['Birth', 0], ['Event', 1], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 1]], 34]
[[['Organization', 0], ['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 34]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Relative', 0]], 34]
[[['Event', 0], ['Ownership', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 33]

[[['Event', 0], ['Event', 0], ['Location', 0], ['Location', 0], ['Person', 0], ['Person', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0]], 33]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Ownership', 0], ['Property', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 33]
[[['Organization', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 33]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 32]
[[['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Property', 0], ['Reference', 0], ['Reference_Relation', 0]], 32]
[[['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2]], 32]
[[['Event_Establish', 0], ['Ownership', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 32]
[[['Object', 0], ['Ownership', 0], ['Person', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 32]
[[['Birth', 0], ['Event', 1], ['Event', 0], ['Location', 0], ['Location', 1], ['Person', 0], ['Time_Point', 0]], 31]
[[['Event', 0], ['Person', 0], ['Profesion', 0], ['Reference', 0], ['Time_Point', 0], ['Topic', 0]], 31]
[[['Event', 0], ['Person', 0], ['Person', 0], ['Related', 0], ['Time_Point', 0], ['Topic', 0]], 31]
[[['Event_Establish', 0], ['Event_Establish', 0], ['Object', 0], ['Object', 0], ['Person', 0], ['Time_Point', 0], ['Topic', 0]], 30]
[[['Object', 0], ['Object', 0], ['Ownership', 0], ['Reference_Relation', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 30]
[[['Event', 0], ['Event', 0], ['Object', 0], ['Organization', 0], ['Topic', 0], ['Topic', 0]], 30]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Person', 0], ['Person', 0], ['Time_Period', 0]], 30]
[[['Event', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Time_Period', 0]], 30]
[[['Birth', 1], ['Event', 2], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2]], 29]
[[['Event', 0], ['Event', 0], ['Ownership', 0], ['Person', 0], ['Person', 0], ['Reference', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0]], 29]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Location', 0], ['Person', 0], ['Property', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0]], 29]
[[['Event', 0], ['Person', 0], ['Related', 0], ['Time_Point', 0], ['Time_Point', 0], ['Topic', 0], ['Topic', 0], ['Topic', 0]], 29]

[[['Event', 0], ['Location', 0], ['Person', 0], ['Profesion', 0], ['Property', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 29]
[[['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2]], 28]
[[['Event', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0]], 28]
[[['Event', 0], ['Location', 0], ['Object', 0], ['Property', 0], ['Topic', 0], ['Topic', 0]], 28]
[[['Event', 0], ['Object', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0]], 28]
[[['Number', 0], ['Object', 0], ['Object', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 28]
[[['Event', 0], ['Ownership', 0], ['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Topic', 0], ['Topic', 0]], 27]
[[['Event', 0], ['Event', 0], ['Location', 0], ['Person', 0], ['Profesion', 0], ['Time_Point', 0], ['Topic', 0]], 27]
[[['Event', 0], ['Object', 0], ['Object', 0], ['Organization', 0], ['Person', 0], ['Time_Point', 0], ['Time_Point', 0]], 27]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Time_Period', 0], ['Topic', 0]], 27]
[[['Event', 0], ['Event', 0], ['Organization', 0], ['Person', 0], ['Profesion', 0], ['Time_Period', 0]], 27]

## 9.2 Clasificación de etiquetas en detalle

En este apartado se muestran los resultados detallados de la ejecución del algoritmo de agrupación. Se ha buscado una base de conocimiento que permita mostrar cómo se clasifican las etiquetas en plantillas y la creación de agrupaciones nuevas con las etiquetas sobrantes. Las anotaciones seleccionadas han sido las correspondientes a la biografía de A. Edward Sutherland, las cuales se muestran en la tabla 9-1 clasificadas según su nivel jerárquico.

| Etiquetas nivel 0: | |
|---|---|
| Time_Point | 2 |
| Property | 2 |
| Reference | 2 |
| Relative | 7 |
| Number | 2 |
| Profesion | 6 |
| Event | 2 |
| Person | 14 |
| Ownership | 5 |
| Birth | 1 |
| Parenthesis | 1 |
| Reference_Relation | 5 |
| Object | 2 |
| Event_Establish | 1 |

| Etiquetas nivel 1: | |
|---|---|
| Death | 1 |
| Event | 1 |
| Person | 1 |
| Birth | 1 |
| Location | 1 |

| Etiquetas nivel 2: | |
|---|---|
| Time_Point | 2 |

**Tabla 9-1 Etiquetas de A. E. Sutherland**

La primera parte en la agrupación de etiquetas corresponde al uso de plantillas. Se ha buscado de entre las plantillas disponibles cuáles se pueden utilizar teniendo las etiquetas listadas en la tabla 9-1, optimizando el resultado de manera que queden el menor número de "etiquetas sobrantes" tras seleccionar las plantillas. La selección final de plantillas contiene las siguientes:

1. [[['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2]], 28]
2. [[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Relative', 0]], 34]
3. [[['Event_Establish', 0], ['Number', 0], ['Object', 0], ['Person', 0]], 37]
4. [[['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], 41]
5. [[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 55]
6. [[['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 153]
7. [[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0]], 241]

Cada línea de este listado corresponde con una plantilla, la cual está compuesta por las etiquetas que la forman (indicando su nombre y nivel jerárquico) y por un número que indica cuántas veces ha aparecido en el corpus que no tiene efectos prácticos en

la agrupación de etiquetas, puesto que solamente se utiliza en la lectura de datos para descartar aquellas que están por debajo de un cierto valor.

De las 59 anotaciones de esta base de conocimiento, después de aplicar la agrupación mediante plantillas quedan las 22 etiquetas pendientes de agrupar mostradas en la tabla 9-2, las cuales son todas de nivel 0.

| Etiquetas nivel 0: | |
|---|---|
| Reference | 2 |
| Ownership | 5 |
| Relative | 6 |
| Property | 1 |
| Number | 1 |
| Person | 4 |
| Reference_Relation | 2 |
| Profesion | 1 |

**Tabla 9-2 Pendientes tras las plantillas**

En este punto, el algoritmo procede a buscar agrupaciones nuevas entre estas etiquetas a partir de los valores de PMI. Los núcleos creados para realizar agrupaciones son:

- ['Reference_Relation', [('Profesion', '0.31'), ('Property', '0.25')], 0.28]
- ['Reference_Relation', [('Profesion', '0.31'), ('Property', '0.25')], 0.28]
- ['Profesion', [('Reference_Relation', '0.31'), ('Property', '0.20')], 0.255]
- ['Property', [('Reference_Relation', '0.25'), ('Profesion', '0.20')], 0.225]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Relative', [('Ownership', '0.18')], 0.18]
- ['Ownership', [('Relative', '0.18'), ('Reference', '0.16')], 0.17]
- ['Ownership', [('Relative', '0.18'), ('Reference', '0.16')], 0.17]
- ['Ownership', [('Relative', '0.18'), ('Reference', '0.16')], 0.17]
- ['Ownership', [('Relative', '0.18'), ('Reference', '0.16')], 0.17]
- ['Ownership', [('Relative', '0.18'), ('Reference', '0.16')], 0.17]
- ['Reference', [('Ownership', '0.16')], 0.16]
- ['Reference', [('Ownership', '0.16')], 0.16]
- ['Person', [], -1]
- ['Person', [], -1]
- ['Person', [], -1]
- ['Person', [], -1]
- ['Number', [], -1]

Es evidente que aquellas etiquetas del mismo tipo crean núcleos iguales alrededor de ellas, puesto que los valores de PMI con otras etiquetas son iguales. Solamente observando estos núcleos ya podemos deducir que a partir de Reference_Relation, Property o Ownership se pueden crear agrupaciones nuevas, puesto que son núcleos de mínimo 3 elementos.

A partir de estos núcleos se logran crear dos agrupaciones:

- [[['Reference_Relation', 0], ['Profesion', 0], ['Property', 0]], 1]
- [[['Ownership', 0], ['Relative', 0], ['Reference', 0]], 1]

Las etiquetas que siguen pendientes de agrupar son las listadas en la tabla 9-3, las cuales se analizan individualmente para buscar la agrupación en la que mejor encajen.

| Etiquetas nivel 0: | |
|---|---|
| Reference | 1 |
| Ownership | 4 |
| Relative | 5 |
| Number | 1 |
| Person | 4 |
| Reference_Relation | 1 |

**Tabla 9-3 Pendientes tras agrupaciones**

Se calcula el PMI de cada una de estas etiquetas con cada una de las agrupaciones realizadas y se ordenan los valores de mayor a menor. El primer valor de la lista es Reference_Relation con [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], para el cual se detecta que no es posible hacer encajar los elementos debido a que no hay ningún bigrama entre Reference_Relation y Birth. Entonces se procesa el segundo elemento de la lista, que es la etiqueta Relative con la agrupación [['Ownership', 0], ['Relative', 0], ['Reference', 0]], y se detecta que en este caso sí es posible colocar Relative dentro de este grupo. El algoritmo entonces añade esta etiqueta en la agrupación y la elimina del listado de etiquetas pendientes. Después, se recalcula de nuevo el valor de PMI de cada una de las etiquetas pendientes con todas las agrupaciones teniendo en cuenta esta modificación en una de ellas. El resto de clasificaciones de etiquetas dentro de agrupaciones sigue de este modo:

1. La primera relación vuelve a ser Reference_Relation y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. Se prosigue con Relative y [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0]], añadiendo la etiqueta a la agrupación.
2. La primera relación vuelve a ser Reference_Relation y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. Se añade Relative a [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0]].
3. La primera relación vuelve a ser Reference_Relation y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. Se añade Relative a [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0]].

4. La primera relación vuelve a ser Reference_Relation y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. La siguiente es de Relative con [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0]], que sí se pueden juntar.
5. La primera relación vuelve a ser Reference_Relation y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. La siguiente relación es de Reference_Relation con [['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2]], añadiendo la etiqueta a este grupo.
6. En este caso se empieza con Ownership y el grupo [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0]], pudiéndolos juntar.
7. El primer elemento ahora está formado por Number y [['Event_Establish', 0], ['Number', 0], ['Object', 0], ['Person', 0]], que sí es posible agruparlos.
8. Comenzando con Ownership y [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Ownership', 0]], se detecta que la etiqueta se puede añadir a la agrupación.
9. La primera relación es de Reference con [['Reference_Relation', 0], ['Profesion', 0], ['Property', 0]], que sí es viable.
10. En esta ocasión el primer elemento es Ownership y [['Event_Establish', 0], ['Number', 0], ['Object', 0], ['Person', 0]], añadiendo la etiqueta al grupo.
11. La primera relación es de Person con [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], pero no se encuentra un modo válido de juntarlos. Se prosigue con Ownership y [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Ownership', 0]], añadiendo la etiqueta a esta agrupación.
12. Vuelve a encontrarse como primer elemento Person y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. El siguiente de la lista actual es Person y la agrupación [['Reference_Relation', 0], ['Profesion', 0], ['Property', 0], ['Reference', 0]], las cuales se juntan.
13. La primera relación vuelve a ser elemento Person y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. Se continúa con Person y [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Ownership', 0], ['Ownership', 0]], las cuales se agrupan.
14. La primera relación vuelve a ser elemento Person y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta de nuevo. Se prosigue con Person y [['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2], ['Reference_Relation', 0]], para los que sí hay una combinación válida.
15. En esta ocasión se empieza de nuevo con Person y [['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], que se descarta otra vez. A continuación viene la relación Person y [['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Ownership', 0], ['Ownership', 0]['Person', 0]], añadiendo esta etiqueta a la agrupación.

En este punto ya hemos procesado las 16 etiquetas de nivel 0 que quedaban pendientes de agrupar en esta base de conocimiento. Las agrupaciones finales son:

- [[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Relative', 0]], 34]
- [[['Event_Establish', 0], ['Number', 0], ['Object', 0], ['Person', 0], ['Number', 0], ['Ownership', 0]], 37]
- [[['Event', 0], ['Person', 0], ['Person', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0]], 241]
- [[['Reference_Relation', 0], ['Profesion', 0], ['Property', 0], ['Reference', 0], ['Person', 0]], 1]
- [[['Ownership', 0], ['Relative', 0], ['Reference', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Relative', 0], ['Ownership', 0], ['Ownership', 0], ['Ownership', 0], ['Person', 0], ['Person', 0]], 1]
- [[['Birth', 0], ['Event', 1], ['Location', 1], ['Person', 1]], 41]
- [[['Object', 0], ['Person', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 55]
- [[['Person', 0], ['Profesion', 0], ['Reference_Relation', 0], ['Time_Point', 0]], 153]
- [[['Birth', 1], ['Death', 1], ['Parenthesis', 0], ['Person', 0], ['Profesion', 0], ['Profesion', 0], ['Property', 0], ['Reference_Relation', 0], ['Time_Point', 2], ['Time_Point', 2], ['Reference_Relation', 0], ['Person', 0]], 28]