Escuela Técnica Superior de Ingeniería Informática

Máster Universitario en Inteligencia Artificial Avanzada:Fundamentos, Métodos y Aplicaciones.

Trabajo Fin de Máster

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos.

Autor: Juan Gª-Rabadán Gascón

Director: Dr. Dº Enrique Amigó Cabrera

Febrero 2022

Tabla de Contenidos

RESUMEN: **1**
Palabras clave: Ontología lingüística, conceptos, términos, Minería de textos, Ley de Zipf, polisemia **1**
..... **2**
1 Introducción **5**
1.1 Motivación *5*
1.2 Planteamiento del problema y objetivos *9*
1.2.1 Elección del corpus y normalización del mismo 9
1.2.2 Extracción de términos 10
1.2.3 Paso de términos a conceptos 10
1.2.4 Relaciones entre conceptos: 11
1.2.5 Reglas de asociación entre conceptos: 12
2 Estado de la cuestión **13**
2.1 Leyes de Zipf y Mandelbrot *13*
2.2 Ontologías *17*
2.3 Ontologías lingüísticas *19*
2.3.1 Relaciones semánticas en ontologías lingüísticas 21
2.4 WordNet como ontología lingüística *25*
2.4.1 Otras ontologías linguísticas relacionadas con WordNet, EuroWordnet 27
2.4.2 SENSUS 28
2.5 Aprendizaje automático de ontologías *29*
2.6 Obtención de conceptos *30*
2.7 Extracción de términos *32*
2.7.1 Extracción de términos con patrones léxicos 32
2.7.2 Extracción de términos con patrones sintácticos 33
2.7.3 Extracción de términos basados en “colocaciones” 35
2.8 Aprendizaje de relaciones no taxonómicas *37*
2.8.1 Reglas de Asociación 38
2.8.2 Reglas de Asociación en Minería de Textos 40
2.8.3 Valoración de las Reglas de Asociación en Minería de Textos 42
2.9 Conclusiones al estado de la cuestión *44*
3 Desarrollo del trabajo **47**
3.1 Disposición de una ontología lingüística general *47*
3.2 Elección de la temática sobre la que proyectar la ontología lingüística general. Un hito para desambiguar la polisemia *48*
3.3 Formación y normalización del corpus representativo de la temática *49*
3.3.1 Normalización del corpus 50
3.4 Configuración de Términos. Naturaleza y número de los mismos *52*
3.4.1 Tripleta contextual 53
3.4.2 N-gramas etiquetados 53
3.5 Extracción de los conceptos mas significativos de un dominio mediante un corpus *54*
3.5.1 Paso de *términos* a *conceptos*. De *tripleta contextual* a *tripleta conceptual* *54*
3.5.2 Fijación del número *p* de tripletas-conceptuales más significativas 61
3.5.3 N-gramas etiquetados para capturar términos y convertirlos en conceptos 74
3.5.4 Fijación del número *j* de n-gramas etiquetados más significativos 76
3.5.5 Valores de umbral, número de tripletas distintos(p), longitud del n-grama(n) y número de n-gramas(j). Relaciones entre ellos 82

3.6 Relaciones entre conceptos(tripletas conceptuales) y una Ontología Lingüística General. Elaboración de las ontologías lingüísticas ligeras temáticas con forma de diccionario.....89
3.6.1 Relación de sinónimos. Elaboración del diccionario de sinónimos.....91
3.6.2 Relación de cohiponimos-hiperónimos. Elaboración del diccionario de hipónimos 95
4 Reglas de asociación con conceptos(tripletas conceptuales).....99
4.1 Utilidad y novedad de las reglas extraídas con tripletas conceptuales.....101
4.2 Fijación del soporte y la confianza en una Matriz de transacciones Frases versus Tripletas Conceptuales.....106
4.3 Tipos de reglas de asociación:.....108
4.4 Paso de reglas complejas a reglas simples.....109
4.5 Relación de coocurrencia. Elaboración de diccionarios a partir de Reglas de Asociación complejas.....111
5 Revisión, corrección, supresión y ampliación de términos por parte del experto.....113
5.1 Aportando valores(sinónimos, cohipónimos) a claves extraídas por el sistema.....114
5.2 Aportaciones por parte del experto.....115
5.3 Ampliación y revocación de las tripletas extraidas.....116
5.4 Revisando el diccionario de Reglas de Asociación.....117
5.5 Expresiones fijas: locuciones, modismos y otras.....118
6 Aplicaciones mas comunes de las ontologías lingüísticas ligeras con forma de diccionario.....121
6.1 Ampliación.....127
6.2 Sustitucion.....130
6.3 Posibles usos combinados de ampliación y sustitución con diversos diccionarios.....132
6.4 Ampliación con el diccionario de Reglas de Asociación.....136
7 Caso de uso: Elaboración de ontologías lingüísticas ligeras de temática religiosa judeo-cristiana.....141
7.1 Corpus, temática y normalización.....141
7.2 Extracción de términos.....142
7.2.1 Extracción de la lista de tripletas contextuales actuando como términos.....142
7.2.2 Extracción de los j n-gramas etiquetados más frecuentes actuando como términos. 143
7.3 Extracción de conceptos a partir de los términos.....145
7.3.1 Extracción de conceptos(tripletas conceptuales) a partir de tripletas contextuales(términos).....145
7.3.2 Extracción de conceptos(tripletas conceptuales a partir de j n-gramas etiquetados 147
7.3.3 Fusión de los conceptos extraídos mediante los términos del corpus y los conceptos extraídos de los j n-gramas etiquetados.....148
7.4 Generación de la ontología lingüistica ligera de sinónimos.....151
7.5 Generación de la ontología lingüística ligera de hipónimos.....152
7.6 Reglas de asociación con conceptos(tripletas conceptuales).....152
7.6.1 Fijación del soporte y la confianza:.....154
7.6.2 Comentarios a las RA extraídas.....155
7.6.3 Ontología lingüística ligera generada con RA para la temática religiosa judea cristiana :.....158
7.7 Intervención del experto.....158
7.8 Ilustración de la desambiguación de la polisemia.....159
7.8.1 Elección de una temática lo suficientemente concreta.....160
7.8.2 Desambiguando con la etiqueta sintáctica.....161

7.8.3 Desambiguando con el contexto y con p.....161
8 Conclusiones y ampliaciones futuras.....163
8.1 Conclusiones.....163
8.2 Ampliaciones futuras.....*166*

Índice de tablas

Tabla 1: Relaciones-semánticas.....21
Tabla 2: Matriz Transacciones Valor.....37
Tabla 3: Desarrollo del trabajo.....47
Tabla 4: Shakespeare: Umbral, j(codo), n(n-grama).....84
Tabla 5: Austen: Umbral, j(codo), n(n-grama).....86
Tabla 6: Chesterton: Umbral, j(codo), n(n-grama).....87
Tabla 7: Matriz transacciones(frases en forma de lista de tripletas conceptuales) versus atributos(tripletas conceptuales).....100
Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos.....122
Tabla 9: Ampliar y/o sustituir con RA.....123

Índice de ilustraciones

Ilustración 1: Particularizacion Ontología General a Ontologia Temática:usos....8
Ilustración 2: Zipf versus Mandelbrot....15
Ilustración 3: Espectro de ontologías según semántica....19
Ilustración 4: "Pipeline" ontologías lingüísticas....20
Ilustración 5: Formas, polisemia Adaptado de Konchady,1986....25
Ilustración 6: Arquitectura EuroWordnet: Figura tomada de [Gómez-Pérez,2004]....27
Ilustración 7: Diccionario de una temática versus muchos textos de dicha temática....60
Ilustración 8: Fijación de p. Desambiguación con p y etiqueta sintáctica.Ejemplo calar....64
Ilustración 9: Textos de Chesterton....67
Ilustración 10: Chesterton: Conceptos, p, términos, umbrales....68
Ilustración 11: Chesterton: Ideal Zipf....69
Ilustración 12: Textos de Shakespeare....70
Ilustración 13: Shakespeare: Conceptos, p, términos, umbrales....71
Ilustración 14: Ideal. Shakespeare Zipf....71
Ilustración 15: Textos de Austen....72
Ilustración 16: Austen Conceptos, p, términos, umbrales....73
Ilustración 17: Austen: Ideal Zipf....73
Ilustración 18: N-grama etiquetado: 1 Término, n conceptos....76
Ilustración 19: Ejemplo de 4 4-gramas etiquetados mas frecuentes en los textos de Shakespeare 77
Ilustración 20: Shakespeare con 4-gramas....79
Ilustración 21: Codo con j = 8. Shakespeare con 4-gramas etiquetados....80
Ilustración 22: Shakespeare. Frecuencia de n-gramas etiquetados versus Rango. Codo j....84
Ilustración 23: Shakespeare. Frecuencia de n-gramas etiquetados versus Rango. Codo j....84
Ilustración 24: Austen. Frecuencia de n-gramas etiquetados versus Rango. Codo j....85
Ilustración 25: Chesterton. Frecuencia de n-gramas etiquetados versus Rango. Codo j....87
Ilustración 26: Esquema de la elaboración de la lista de tripletas conceptuales mas significativas 90
Ilustración 27: Elaboración del diccionario de sinónimos....92
Ilustración 28: Dimensiones del diccionario de sinónimos....93
Ilustración 29: Dimensiones del diccionario de cohipónimos hiperónimos....96
Ilustración 30: Elaboración del diccionario de cohipónimos....97
Ilustración 31: De lo abstracto a lo concreto:términos, conceptos, relaciones, etc....98
Ilustración 32: Sentido de las RA....104
Ilustración 33: Vocabulario del texto y diccionario....126
Ilustración 34: Biblia: 10 tripletas contextuales....143
Ilustración 35: Codo n-gramas etiquetados Biblia....143
Ilustración 36: Biblia 14 4-gramas etiquetados como términos....144
Ilustración 37: Biblia: 70 tripletas conceptuales mas frecuentes....146
Ilustración 38: Biblia 24 conceptos extraídos de los 14 4gramas....147
Ilustración 39: T2 / T1 Biblia Conceptos aportados por los 14 4 -gramas....148
Ilustración 40: Sinónimos desambiguados de Jesus....151
Ilustración 41: Sinónimos desambiguados de book....151
Ilustración 42: Cohipónimos de "heaven"....152
Ilustración 43: Matriz de transacciones frases normalizadas versus Conceptos, Biblia....153
Ilustración 44: Reglas de asociación Biblia: Confianza = 0.5, Soporte = 0.001....155

Ilustración 45: Biblia: OLL(diccioniario) con RA.....158
Ilustración 46: Synsets de "hand".....160
Ilustración 47: Tripletas Conceptuales con "hand".....161
Ilustración 48: Lema "son" en Wordnet y en la temática bíblica.....162

Índice de algoritmos

Texto 1: Algoritmo de búsqueda de un concepto - tripleta conceptual - a partir de una tripleta contextual....56
Texto 2: Algoritmo de generación de tripletas conceptuales desde n-gramas etiquetados....81
Texto 3: Algoritmo de elaboración del diccionario de sinónimos....91
Texto 4: Algoritmo de elaboración del diccionario de cohipónimos....95
Texto 5: Algoritmo de elaboración del diccionario con RA....111
Texto 6: Algoritmo de reasignación de sinónimos....118
Texto 7: Algoritmo de ampliación con diccionario....129
Texto 8: Algoritmo de sustitución con diccionario....132
Texto 9: Algoritmo de sustitución con sinónimos y ampliación con hiperónimos....137

RESUMEN:

En el trabajo que se presenta se pretende, como objetivo principal, el desarrollo, y posterior síntesis, de un método propuesto por el ponente que, a partir de un corpus temático bien escogido, una ontología lingüística general y otros recursos permita elaborar, para una determinada temática, unas ontologías lingüísticas ligeras representativas de dicha temática particularizando la Ontología Lingüística General. En pos de dicho objetivo se alcanzaron hitos constitutivos parciales como la extracción de los términos del corpus con una determinada estructura y su posterior paso a conceptos con otra determinada estructura; para ello se precisó de la ayuda de dicha ontología lingüística general así como del paradigma de las Reglas de Asociación. La elección de una y otra estructura fue encaminada a facilitar la selección adecuada de los conceptos mas representativos, asi como la desambiguación de la polisemia. Cada una de estas ontologías lingüísticas ligeras representa una relación. En particular, en este trabajo se sintetizaron relaciones de sinonimia, de hiperonimia y de coocurrencia.

Palabras clave: Ontología lingüística, conceptos, términos, Minería de textos, Ley de Zipf, polisemia

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos

ABSTRACT:

The main objective of the academic work presented here is the development, and subsequent synthesis, of a method proposed by the author that, from a well-chosen thematic corpus, a general linguistic ontology and other resources allows to elaborate, for a certain theme, some light linguistic ontologies representative of said theme, particularizing the General Linguistic Ontology. In pursuit of this objective, partial constitutive milestones were reached, such as the extraction of the terms from the corpus with a certain structure and their subsequent transition to concepts with another certain structure; For this, the help of said general linguistic ontology was needed, as well as the paradigm of the Rules of Association. The choice of one or the other structure was aimed at facilitating the adequate selection of the most representative concepts, as well as the disambiguation of polysemy. Each of these lightweight linguistic ontologies represents a relation. In particular, in this academic work synonymy, hyperonymy and co-occurrence relations were synthesized.

Keywords: Linguistic ontology, concepts, terms, text mining, Zipf's Law, polysemy

1 Introducción.

1.1 Motivación.

La generalización en el uso de Internet ha provocado que la cantidad de información a almacenar y procesar crezca de manera vertiginosa. Esta información proviene de muy diversas fuentes y es de muy diversa naturaleza. Por tanto, para que entes heterogéneos, en particular humanos y máquinas, puedan, usando dicha información, comunicarse entre sí con el fin último de compartir conocimiento ha surgido el concepto de *Ontología.* Una ontología puede definirse, en una primera instancia, como "una especificación formal y explícita de una conceptualización compartida". Así, en un determinado dominio y -quizá para una determina tarea-, una ontología dirá que conceptos son importantes para ese dominio y cuales no, yendo más allá, dirá que relaciones pueden establecerse entre dichos conceptos. Tradicionalmente, la forma en que se creaban esas ontologías partía de un experto humano en determinado dominio del que se pretendía extraer los conceptos y las relaciones entre los mismos; esto acarreaba una serie de problemas no menores, entre ellos: sesgos personales, lentitud en la generación de la ontología, inconsistencias, incapacidad de abordar ontologías grandes, falta de metodología reproducible, etc.

Por éstas y otras razones durante los últimos años se ha intentado automatizar la construcción de ontologías, utilizando el texto escrito como fuente de información. Esto no es extraño, teniendo en cuenta que el lenguaje natural es la vía que más se ha utilizado para representar el conocimiento (la Web, en su inmensa mayoría, contiene la información en forma de texto). Este formato resulta algo menos atractivo que otros como el sonido, las imágenes y el

video, pero es, sin lugar a duda, el principal medio de comunicación entre seres humanos en la actualidad[Schreibman, 2015]. Esta naturaleza textual, en gran medida, determinará como aprender la ontología en cuestión.

Por tanto, un primer hito en este trabajo será proveer los textos, en forma de corpus, adecuados para cada temática con el objetivo de que siendo estos textos ajustados a dicha temática puedan proveer, primero, los términos convenientes para cada temática y luego, sobre estos, los conceptos pertinentes. No obstante, para enlazar automáticamente los términos con sus conceptos adecuados habrá de usarse una Ontología Lingüística General y también a partir de ésta se extraerán las relaciones que en este caso serán de sinonimia y de hiponimia-hiperonimia. Esta provisión, claro, habrá de poder automatizarse reduciendo, por consiguiente, al mínimo la intervención del experto humano. Sin embargo, este enfoque no está exento de problemas que se heredan o provienen de la propia esencia de las lenguas naturales y de los textos que las representan. En particular, se abordará la polisemia y la homonimia. Así, en el caso que se presenta podría darse que dicha polisemia se manifiesta en que el mismo término se puede enlazar con dos o más conceptos dejando en manos de la entidad que pretende comprender los textos la desambiguación de la polisemia. Así pues, otra motivación -la principal-, será poder generar a partir de un corpus temático y una ontología lingüística general una ontología lingüística ligera particular que contenga los conceptos representativos de dicha temática, provea de desambiguación semántica e igualmente provea de las relaciones de sinonimia e hiponimia-hiperonimia.

Asi pues en el presente trabajo, dada una temática concreta, se pretende **estudiar la** viabilidad para sintetizar un método que permita elaborar, mediante procedimientos computacionales, una serie de ontologías lingüísticas ligeras para dicha temática que habitualmente tendrán forma de diccionario. Cada una de estas ontologías ligeras representará una relación. Estas ontologías, en ocasiones, serán relaciones taxonómicas como el diccionario de cohipónimos que convergen en su hiperónimo común; en otras no taxonómicas como un diccionario de sinónimos que convergen en su representante canónico; así como un diccionario de simples relaciones de coocurrencia basado en el paradigma de las Reglas de Asociación.

Para tal empeño se estudiarán diversas técnicas de extracción de los términos fundamentales de la temática en cuestión representada por un corpus significativo. seleccionando aquellas mas aptas o buscando la mejora del desempeño general haciendo concurrir varias de ellas. A continuación, se presentarán métodos para transformar esos términos en conceptos del dominio. En pos de esto se buscará la representación adecuada de dichos conceptos, se explorará la distribución estadística que dichos conceptos siguen en el corpus esperando que se acerque a una Ley de Zipf-Mandelbrot y mediante el uso de una Ontología Lingüística General y otros medios externos se sintetizarán las relaciones entre dichos conceptos. Debe entenderse que la elección de unas u otras representaciones de términos y conceptos determinarán las aplicaciones y la utilidad de las ontologías lingüísticas ligeras extraídas.

Se persigue este objetivo con la finalidad de que dichos diccionarios para un dominio determinado puedan ser, en muy diversas tareas de minería de textos, recuperación de la información, documentación o programación del lenguaje natural, reutilizados como una herramienta mas ágil y con menos consumo de recursos de lo que sería el uso directo de una ontología lingüística general. De igual forma, se persigue que la existencia de esos diccionarios "ad hoc" resuelva algunos problemas de ambigüedad inherentes a la lengua natural que no serían resueltos por el uso directo de una Ontología Lingüística General. Además, estos diccionarios deben resultar útiles, de tamaño controlado y reutilizables en muy diversas tareas de dicho dominio. Es decir, dado un dominio para el que se han elaborado dichos diccionarios se sustituirá el costoso uso de la ontología lingüística general (ver *Ilustración 1: Particularizacion Ontología General a Ontologia Temática:usos*) por el uso, más ágil y pronto, de los diccionarios previamente elaborados para tal dominio. (parte derecha de la misma ilustración)

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos

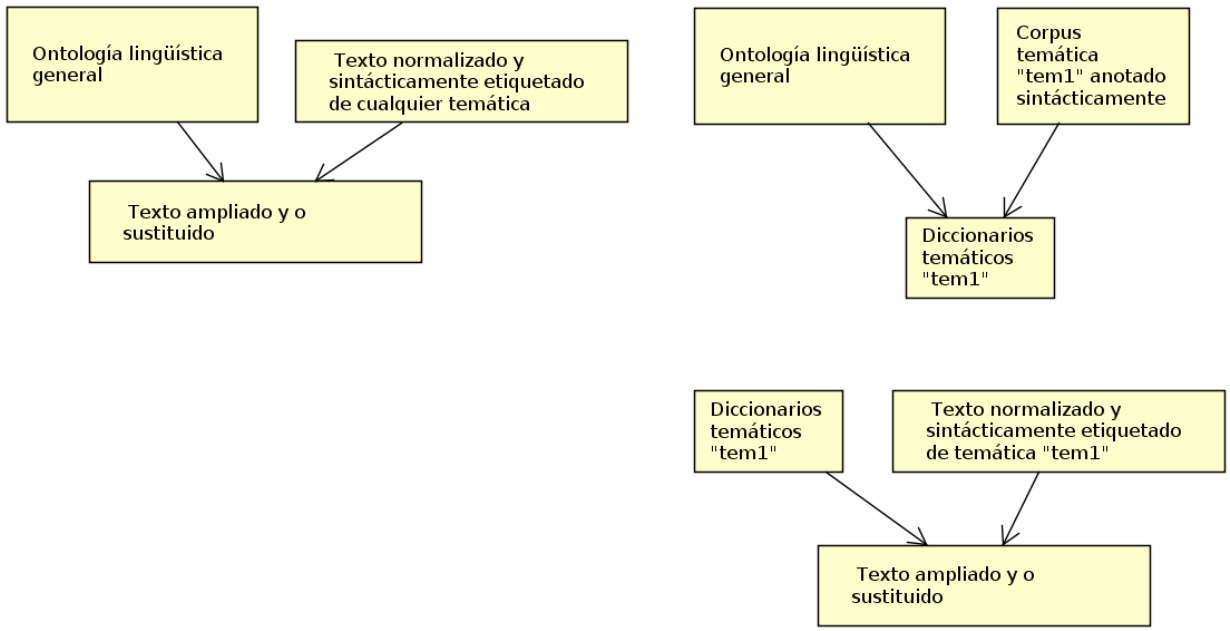


Ilustración 1: Particularizacion Ontología General a Ontologia Temática:usos

1.2 Planteamiento del problema y objetivos.

En la *Ilustración 4: "Pipeline" ontologías lingüísticas* de la página 20 se representará el método general para aprender, dado un corpus representativo, una ontología lingüística general. En el epígrafe Leyes de Zipf y Mandelbrot. de la página 13 se comentará cuál es la naturaleza matemática de esta ley y sus implicaciones conceptuales. Dichas leyes, y el comportamiento de las mismas, serán fundamentales en la fijación de los parámetros del método a investigar. En el presente trabajo siguiendo la *Ilustración 4: "Pipeline" ontologías lingüísticas* se proponen, investigan y estudian medios concretos para sintetizar un método que permita la generación de ontologías lingüísticas ligeras temáticas con especial incidencia en que dichas ontologías sean una herramienta útil en la desambiguación de la polisemia. Este método partirá de un corpus temático bien escogido y una Ontología Lingüística General. El estudio de la viabilidad del método propuesto, sus hitos constitutivos -que coinciden con los subepígrafes del presente epígrafe *1.2 Planteamiento del problema y objetivos.-,* la fijación de parámetros óptimos del método y su posterior implementación; así como el estudio de sus aplicaciones es el objeto de esta investigación .

1.2.1 Elección del corpus y normalización del mismo.

El método propuesto descansa, en primer lugar, en la capacidad de escoger correctamente un corpus adecuado a la temática en cuestión que se quiere representar en forma de ontologías lingüísticas ligeras. Debe tenerse en cuenta que éste es un primer hito en la desambiguación y un primer paso necesario que incidirá en la utilidad de usar la Ontología Lingüística Ligera temática, en presencia de documentos de dicha temática, en vez del simple uso de la Ontología Lingüística General. Por tanto, la primera pregunta que se ha de intentar responder es:

" ¿ Cómo debe formarse el corpus del que aprender para una determinada temática ? ¿ Qué características debe tener ?"

Una vez fijado este corpus debe investigarse que normalizaciones son adecuadas y cuales no para el objetivo final de este trabajo. Cabe adelantar que las normalizaciones finalmente

escogidas deben facilitar la significación estadística de las palabras en concurrencia y facilitar su posterior transformación en términos y conceptos. Así pues, otra pregunta a investigar es:

"¿ Qué procesos de normalización del corpus son mas adecuados para el trabajo en cuestión ?"

1.2.2 Extracción de términos

Una vez escogido y normalizado el corpus habrá de estudiarse cuál es la mejor configuración posible para un *término*. En definitiva: qué se considerará *término* en este trabajo concreto. Una vez hecho esto, entre otras cuestiones, habrá de contestarse las siguientes preguntas:

"¿Cuál es la distribución estadística en el corpus de entrenamiento de dichos términos ?"

"¿ Sigue, dicha distribución, una ley de Zipf-Mandelbrot ?"

"¿Cuántos de estos términos son importantes y como influye en la elaboración de las ontologías lingüísticas ligeras cuántos de estos términos se consideran ?

"¿ Facilita la configuración escogida la desambiguación de la polisemia ?"

"¿ Qué métodos son más adecuados para la extracción de los términos más representativos del corpus ? "

1.2.3 Paso de términos a conceptos

Una vez representado y normalizado el Corpus en una determinada estructura de datos que albergue los *términos* mas importantes surgen otro hitos importantísimos en el presente trabajo. Si en el epígrafe anterior se dijo que habrá de investigarse qué se considerará

termino ahora, de forma natural, surge otro investigación clave: qué se considerará *concepto* en este investigación y como pasar de *término* a *concepto*.

Una vez dicho esto cabe reformular las preguntas del epígrafe anterior.

"¿Cuál es la distribución estadística de dichos conceptos ?"

"¿ Sigue, dicha distribución, una ley de Zipf-Mandelbrot y por tanto pocos conceptos representan muchos términos?"

"¿ Facilita la configuración escogida y el número de conceptos finalmente escogidos la desambiguación de la polisemia ?"

"¿ Qué métodos son más adecuados para pasar de los términos del corpus a los conceptos del mismo ? "

"¿Cuántos de estos conceptos son importantes y como influye en la elaboración de las ontologías linguísticas ligeras cuántos de estos conceptos se consideran ? Otro hito en la fijación de parámetros"

No obstante, hay que decir que en el proceso de paso de *término* a *concepto* debe investigarse el *contexto* de los términos del Corpus y como la elección de unos u otros contextos pueden facilitar el paso de *término* a *concepto*

1.2.4 Relaciones entre conceptos:

Puesto que, llegados a este punto, se habrá estudiado y, probablemente, elegido una representación adecuada para los conceptos y el número de los mismos el siguiente hito de investigación es el estudio de las posibles relaciones semánticas entre dichos conceptos.

"¿Cuáles son estas relaciones ?"

"¿ Cómo se extraen ? "

"¿ Cuáles de estas relaciones ayudarán en la síntesis de las ontologías lingüísticas ligeras a desarrollar ? "

"¿Cómo ayuda la Ontología lingüística general, en dicha síntesis ?

"¿Cómo ayudan estas relaciones en la desambiguación de la polisemia ?"

1.2.5 Reglas de asociación entre conceptos:

Como se dejará constancia en el Estado de la Cuestión el paradigma de las RA entre diversos entes es bien conocido y está muy estudiado. No obstante, en este trabajo que, entre otras cosas, pretende extraer los conceptos fundamentales de una temática basada en un Corpus con ayuda, en ocasiones, de una Ontología Linguísita General la indagación en como las **RA entre conceptos** se comporta se hace imprescindible. Por tanto se pretenderá resolver cuestiones como éstas:

"*¿ Con los corpus escogidos, la normalización dada y tal o cual representación de los conceptos se pueden extraer reglas de asociación estadísticamente significativas ? "*

"¿Aportan estas RA a las ontologías lingüísticas ligeras nuevo vocabulario o simplemente aportan relaciones entre los conceptos ya extraídos?"

"¿Estas relaciones son distintas de las ya extraídas por otros métodos ?"

2 Estado de la cuestión.

2.1 Leyes de Zipf y Mandelbrot.

En 1949 [Zipf, 1949] hizo notar mediante experimentación que los hablantes de una determinada lengua tendían a usar pocas palabras (< 20% en inglés) en muchas ocasiones. Justificó semejante afirmación en lo que Zipf llamó "principio del mínimo esfuerzo": es decir, los hablantes, por tal principio, tienden a usar frecuentemente palabras que resultan fáciles de escribir, pronunciar o recordar. Naturalmente, esto conlleva que otras muchas palabras, que no tienen dichas características, tienden a ser usadas en muy pocas ocasiones, incluso dándose, en determinadas circunstancias, cierta sobreabundancia de palabras cuya frecuencia de uso es 1 (*hapax legomena*). Por tanto, se consigna el hecho de que los hablantes buscan comunicar el máximo de información con el mínimo de palabras. Zipf observó que ordenando las palabras de un determinado corpus representativo en orden decreciente por su frecuencia de aparición puede establecerse una relación entre dicha frecuencia(f) y el rango(r) que dicha palabra ocupa en tal ordenación; r tomará valores entre 1 y el número de palabras distintas usadas en dicho corpus(vocabulario del mismo V, $r \in 1..|V|$). Así, manifestó que el producto $f * r$ permanece aproximadamente constante (C). Es decir, el producto de la frecuencia de una palabra por su rango es muy parecido al producto de otra palabra por su rango. Puede entenderse que esto da una estructura estadística, si se interpretan las palabras como su ente atómico, a una determinada lengua y además se ha podido observar que esto sucede en gran parte de las lenguas

naturales. Es útil, en determinados ámbitos, añadir una interpretación probabilística a la Ley de Zipf. Para autores como [Erar,2002] la ley de Zipf es la probabilidad de que una palabra ocurra con cierta frecuencia en un documento, lo que introduce cierto grado de azar en la elección de unas palabras en detrimento de otras.

Zipf también observó una relación entre la frecuencia de aparición de las palabras y el número de significados de éstas. Si dichas palabras se ordenan, de nuevo, de forma decreciente por su aparición en un corpus representativo de un lengua se tiene que:

$$numeroDeSignificados = \frac{1}{\sqrt{(r)}}$$

Esto coincide con la intuición y con el principio del mínimo esfuerzo pues el hablante intentará usar el mínimo número de palabras y para ello es útil usar aquellas que más significados tienen, con esto se deja al interlocutor que escucha, o que lee, la tarea de elegir el significado adecuado a la interlocución.

No obstante, la Ley de Zipf $f(r) = \frac{C}{r}$ es una ley estrictamente lineal y sin embargo en la experimentación con corpus representativos se ha observado que dicha ley rige especialmente en las zonas centrales del intervalo $1..|V|$ y se aleja en las zonas iniciales de dicho rango y en las zonas finales cercanas a $|V|$. Haciendo determinadas correcciones en la fórmula inicial de dicha Ley de Zipf puede conseguirse una mayor aproximación al ideal lineal para todos los valores de r. Así, se presenta la siguiente expresión:

$f(r) = \frac{C}{(m+r)^B}$ donde $B \in 1..2$ y $m \in 0..100$ Ley de Mandelbrot

Esta expresión fue derivada por B. Mandelbrot[Mandelbrot,1953]. Debe consignarse que la Ley de Zipf es un caso particular de la ley de Mandelbrot haciendo $m = 0$ y $B = 1$ Ajustando estos parámetros puede conseguirse un gran acercamiento a una curva ideal lineal para todos los valores de r. En general, m es más importante para los valores bajos de r y B pesa mucho más para los valores medios-altos de r. Se observa en *Ilustración 2: Zipf versus*

Mandelbrot como la curva azul se aparta de la curva naranja manteniéndose cuasi horizontal hasta un determinado valor de r (donde r empieza a pesar más que m) y a partir de éste coincide con la curva naranja(Zipf) puesto que $B= 1$. La curva verde se aleja cada vez más de la curva naranja puesto que $B=2$.

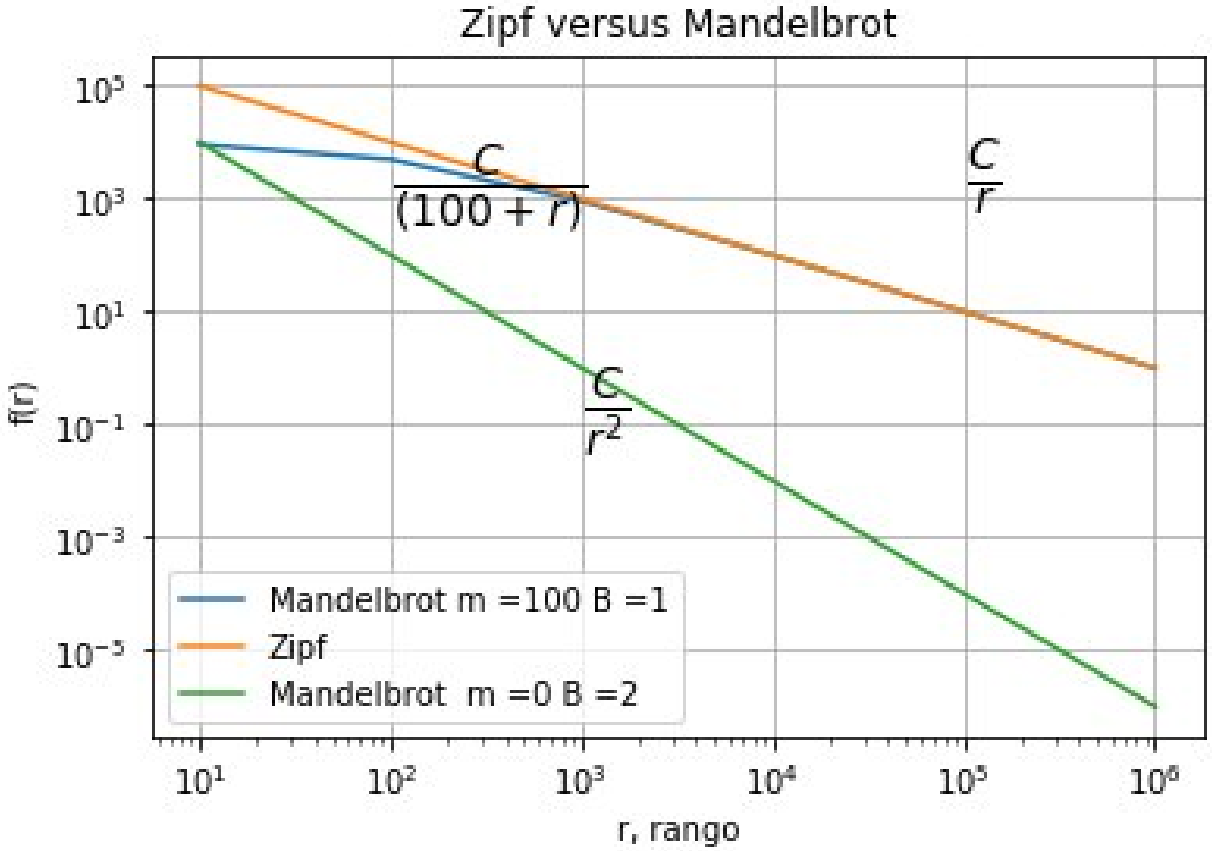


Ilustración 2: Zipf versus Mandelbrot

El parámetro *B* puede usarse para detectar un sucesivo empobrecimiento de la lengua(en cuanto a vocabulario) sí dicho parámetro se acerca a su límite superior. Subsiguientemente, si B es cercano a uno el número de *palabras únicas* será grande ampliándose el vocabulario usado y tendiendo a la erudición. En general, para que rija la Ley de Zipf deben tomarse textos con una determinada longitud que [Pierce,2012] fija para el idioma inglés en 120.000 palabras.

Hasta aquí se ha hecho una presentación teórica y matemática de las leyes de Zipf-Mandelbrot y se ha expuesto como rigen tomando como unidades lingüísticas las palabras de un

determinado texto o corpus. Sin embargo, a partir de aquí cabe hacerse dos preguntas fundamentales entreveradas entre sí:

- ¿Para que unidades lingüísticas rige la ley de Zipf ?

- ¿Es útil la ley de Zipf-Mandelbrot para extraer los conceptos fundamentales de un texto ?
  - ✔ Mediante la extracción de palabras con un alto contenido semántico.
  - ✔ Representación y extracción de dichos conceptos por determinadas unidades lingüísticas.

En un primer acercamiento a las respuestas de estos interrogantes ya en 1928 [Condon,1928] expuso que :

> *"la frecuencia de las palabras en un texto seguiría una ley cuantitativa de utilidad disminuyente muy similar a la ley de Weber-Fechner en psicología; por tanto, la frecuencia de uso de una palabra mediría el efecto de su utilidad en la transmisión de ideas entre los individuos."*

Yendo mas allá Urbizagástegui y Restrepo [Urb,2011] exploran la eliminación de ruido en la extracción de palabras con fuerte contenido semántico para ello proponen la eliminación de *stopwords* y el uso del *Índice de Goffman*. Asimismo, prueban a lematizar el corpus concluyendo que para la extracción de palabras claves de un texto lo más adecuado es la eliminación de *stopwords* y la *lematización* y que las unidades lingüísticas así cribadas(palabras lematizadas sin stopwords) siguen distribuyéndose según la Ley de Zipf-Mandelbrot con lo que pueden extraerse las mas frecuentes como representantes de un texto. Haciendo con dichas unidades lingüísticas -ejerciendo de conceptos-, así obtenidas un lenguaje de intermediación entre dicho texto y quien pretenda hacer uso de él.

Reforzando lo anterior [Corral,2015] usando como espacio de experimentación 10 novelas largas escritas en cuatro idiomas con diferentes ontogenias(español, inglés, francés y finés) comparó la vigencia de la ley de Zipf-Mandelbrot para cada una de estas novelas

representadas para todos los casos como sucesión de palabras y como sucesión de lemas(tras lematizar cada una de las novelas). Observó que en todos los casos la ley de Zipf-Mandelbrot rige tanto para palabras como para lemas, si bien el parámetro *B* sufría leves variaciones entre una y otra configuración. En gran parte de los casos *B* era mayor para la versión lematizada que para la versión con palabras.

2.2 Ontologías.

Una ontología, en el campo de la informática y ciencias relacionadas es, ante todo, un sistema para organizar el conocimiento en un dominio concreto. Dicho conocimiento vendrá sustentado por los conceptos en los que se sustancia dicho conocimiento y en las relaciones, de diverso tipo, entre dichos conceptos. Por tanto, antes de proseguir, se ha de dar una definición operativa de lo que en el campo de las ontologías informáticas se considera un concepto: Así, citando y traduciendo a [Corcho,2000] se dirá :

> *Un concepto es una entidad de la que se puede decir algo y por lo tanto puede ser la descripción de una tarea, función, acción, estrategia, proceso de razonamiento, etc.*

En torno a los conceptos de un dominio y sus relaciones girarán casi todas las definiciones generales de ontología en el campo de la informática. De hecho [Gruber, 1993] sostiene que una ontología es:

> *Una especificación formal de una conceptualización compartida*

mientras que [Guarino,1998] define la ontología como:

> *Un producto de ingeniería consistente en un vocabulario específico usado para describir una realidad más un conjunto de asunciones relacionadas con el significado del vocabulario.*

Ese vocabulario, en una realidad concreta, determinará la representación de los conceptos de interés. De igual forma hablar de ingeniería permite atisbar que determinadas

ontologías de un determinado tipo pueden ser sintetizadas mediante una metodología formal automatizable total o parcialmente. Pueden traerse a colación muchas definiciones más de ontología pero todas incidirán en los siguientes parámetros:

1. Conceptualización. Es decir, el primer paso en la génesis de una ontología será fijar los conceptos interesantes de tal ontología para tal dominio y para tal empeño. De hecho, con conceptualización se quiere decir que una ontología es un modelo abstracto acerca de cómo las personas piensan sobre las cosas del mundo, restringido a un dominio o área determinada y sobre ésta a una tarea particular; fijando para ello que conceptos se consideran imprescindibles en dicho dominio y que conceptos, aquellos que no se registran, se consideran prescindibles. Debe decirse que para un mismo dominio y sobre éste diversas tareas puede haber distintas ontologías.
2. Relaciones entre estos conceptos. Se hablará de diversas relaciones, taxonómicas o no, entre los conceptos.
3. Formal. Los conceptos y las relaciones entre estos deben ser expresados, con mayor o menor complejidad en un lenguaje formal de representación del conocimiento.
4. Compartición. Las ontologías deben poder ser compartidas y cooperar entre máquinas o sistemas heterogéneos y entre estos y los humanos.

Pueden darse muchas clasificaciones de las ontologías según muy diversos parámetros o criterios, sin embargo conviene remitirse al concepto de "espectro de las ontologías" [Lassila,2001] donde las ontologías son clasificadas en función de su menor o mayor complejidad semántica. Así, según la Ilustración 3 *Espectro de ontologías según semántica* las ontologías se clasifican en una línea conceptual ascendente desde semánticas muy simples - en realidad, muy simplificadas expresadas con constructos taxonómicos muy simples- a semánticas muy complejas – en realidad, muy elaboradas expresadas con lenguajes lógicos conceptualmente complejos-. Debe decirse que un mismo dominio puede ser expresado con diversas ontologías mas o menos apropiadas para unas determinadas tareas.

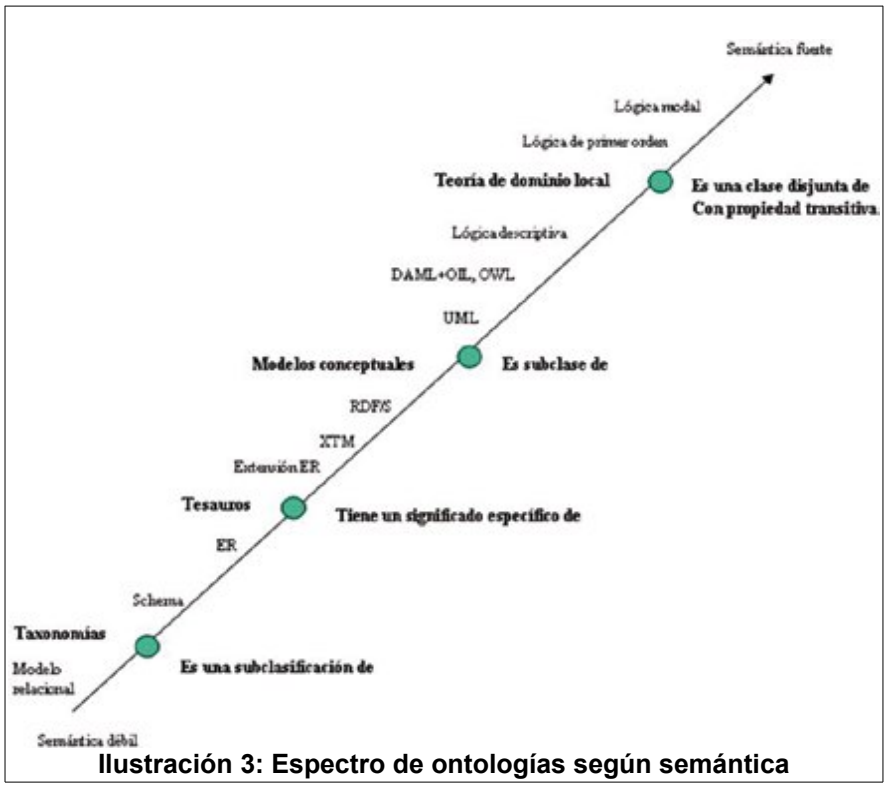


Ilustración 3: Espectro de ontologías según semántica

2.3 Ontologías lingüísticas.

Lo que hará distinguible las ontologías lingüísticas – en adelante OL- de otro tipo de ontologías será, obviamente, el tipo de conceptos con los que trata y el tipo de relaciones que emergen entre dichos conceptos. Así, se tiene que la fuente de la que se extraerán las ontologías lingüísticas serán los textos. Estos textos podrán ser representados en función de términos lingüísticos, estos términos se generarán a partir de las palabras -o tokens normalizados generados sobre éstas- de los textos y de estos términos emergerán los conceptos a partir de los cuales, mediante relaciones, crear la ontología. Un término se define, precisamente, como una expresión o conjunto de palabras -quizá, tokens normalizados -, que determinan unívocamente un concepto del dominio. Según [Cherfi,2002]

> *Un término consiste en una o más palabras, consideradas juntas como una construcción sintáctica homogénea y atómica. Este término tiene sentido solo en el contexto en el que es usado. Este contexto será considerado como el dominio de aplicación y este término denotará un objeto - abstracto o concreto- en dicho dominio.*

Las relaciones entre estos conceptos, principalmente, serán semánticas y, por lo general, la misma ontología tendrá diversos grafos semánticos representando diversas relaciones. Debe decirse que de un mismo conjunto de textos pueden extraerse diversos conjuntos de términos, según la naturaleza final de estos, y de este conjunto de términos, según el método usado para extraer los conceptos, pueden extraerse distintos conjuntos de conceptos.

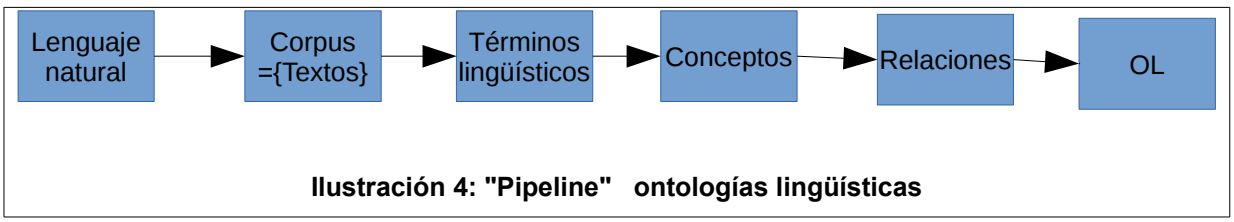


Ilustración 4: "Pipeline" ontologías lingüísticas

2.3.1 Relaciones semánticas en ontologías lingüísticas.

Si se dijo que una de las razones de ser de las ontologías era la de formalizar y taxonomizar, si es posible, los objetos – conceptos- de un dominio y las relaciones entre estos; en una ontología lingüística las relaciones, que se dan entre conceptos, tendrán fuerte contenido semántico y sobre éste descansará la proyección de la ontología lingüística general a campos concretos de aplicación.

Las relaciones semánticas entre los conceptos de la ontología lingüística tendrán sentido – o no-, en función de la etiqueta sintáctica que estos conceptos tengan.

Etiqueta sintáctica/Relación	Sinonimia	Implicación	Hiperonimia/ Hiponimia	Holonimia Meronimia
Verbos	Sí Comer-Minchar	Sí Vivir-Respirar	Sí Percibir-Ver	No
Sustantivos	Sí Ordenador-Computador	No	Sí Humano-Mujer	Sí Cuerpo-Cabeza
Adjetivos	Sí Agraciado-Guapo	No	No	No
Adverbios	Sí Aquí-Acá	No	No	No

Tabla 1: Relaciones-semánticas

Debe quedar claro que los ejemplos que aparecen en la Tabla 1: Relaciones-semánticas de la página 21 son meramente ilustrativos pues las relaciones en las ontologías lingüísticas pueden ser entre conceptos por ejemplo: hiperónimos/hipónimos o entre palabras por ejemplo: todos los lemas de un determinado concepto(sinónimos).

- ***Relaciones de sinonimia(Igualdad de significado):***

Todo hablante de una lengua natural, en presencia de dos palabras de dicha lengua, sabe si éstas tienen - o no -, igual o parecido significado. El caso más claro es el de la *sinonimia denotativa;* es decir dos o más palabras se usan para referirse, denotar, a una misma entidad. Así, en palabras de [Moreno,2000]

> *Dos palabras son sinónimos denotativos (X;Y) si pertenecen a la misma clase gramatical y cualquier oración O1 que contenga X tendrá las mismas condiciones veritativas que cualquier otra oración O2 igual a O1 en todo menos en el hecho de que donde O2 tiene Y, O1 tiene X.*

"Juan entró en el excusado"
"Juan entró en el servicio"

Estas frases, continua Moreno,

> *presentan las mismas condiciones veritativas, de modo que si la primera es verdadera, también lo es la segunda y que si la segunda es verdadera también lo es la primera*

Se aprecia que para hablar de sinonimia denotativa se abstraen los aspectos referenciales o connotativos pues, connotativa, referencial y estilísticamente, existen diferencias de significado, *excusado* es mas elegante mientras que *servicio* es más funcional, aséptico y tiene mayor urbanidad.

Habitualmente, dado un concepto reflejado en una ontología lingüística todas las palabras que denotan dicho concepto serán sinónimas entre sí -miembros de un mismo conjunto-. Ahora bien, no es un concepto lo que se encontrará habitualmente en un texto, sino

una palabra, quizá una palabra sintácticamente etiquetada en un contexto circundante donde dicha palabra aparece.

Según se disponga de una información u otra se podrá recuperar el concepto mas cercano a esa palabra con algún criterio y devolver todas las palabra sinónimas de ese concepto -sus lemas-. Es decir, la relación de sinonimia en una ontología lingüística viene dada por todos los lemas pertenecientes a un concepto dado. Ver *Ilustración 5: Formas, polisemia Adaptado de Konchady,1986* de la página 25 .Cabe decir que esto tiene sentido para las cuatro categorías sintácticas principales (verbos,sustantivos, adjetivos y adverbios).

Se incide, de nuevo y ahora mas orientado a las Ontologías Lingüísticas, que son muy pocas las palabras perfectamente sinónimas -sinónimas absolutas-, frecuentemente existe un matiz, un detalle o un criterio estilístico que marca diferencias. Sin embargo éstas quedan, en ese determinado nivel de abstracción, ocultas por la ontología general lingüística y así, sin más, recuperado un concepto todos sus lemas serán considerados sinónimos. Podría caber una última revocación de lo dicho si finalmente un experto humano – o un nivel de decisión superior - decide eliminar un determinado lema que considere no pertinente para un determinado dominio. De igual forma cabría añadir algún lema a criterio del experto y, de nuevo, en función del dominio o de un ámbito temporal o geográfico.

- ***Relaciones de hiperonimia/hiponimia***

Ambas relaciones en la ontología lingüística se producen entre conceptos, no entre palabras o lemas.

Son relaciones inclusivas transitivas, no reflexivas y antisimétricas determinando, por tanto, una taxonomía entre conceptos. Debe entenderse como una relación clase/miembro donde los miembros(cohipónimos) son independientes entre sí y la clase(hiperónimo) es un colección o acumulación de miembros. Como consecuencia, puede hablarse de una operación de síntesis. El hiperónimo es el elemento que domina a todos sus hipónimos(cohipónimos) que son, por tanto, los elementos dominados.

Pruebas para determinar la condición de hipónimo/hiperónimo, pueden ser:

Hipo *es un elemento de* ***Hiper*** donde **Hipo** e **Hiper** son conceptos e **Hiper** es un nombre colectivo extensional. (cuchillo, tenedor, y cuchara son elementos de cubertería).

Hipo *es un tipo/clase de* ***Hiper*** En este caso los hiperónimos son nombres que designan una característica común de sus hipónimos.

Un concepto será hiperónimo frente a sus hipónimos (cohipónimos) si estos, conservando la semántica de su hiperónimo lo especializan, matizan o añaden nueva información. El hiperónimo generaliza a todos sus cohipónimos, mientras que estos especializan a su hiperónimo (o hiperónimos).

Así se establece un grafo dirigido acíclico(GDA) entre los conceptos de la ontología lingüística, donde un enlace dirigido de un concepto A sobre un concepto B implica que A es un hiperónimo inmediato de B y éste un hipónimo inmediato de A.

Esta disposición en GDA permite hablar de hiperónimos/hipónimos inmediatos o de grado n sí entre los conceptos A y B hay n enlaces. Esto facilitará, si resulta útil, recuperar dado un concepto sus hiperónimos/hipónimos hasta un determinado nivel.

Debe decirse que los conceptos de hiperonimia e hiponimia solo son conceptualmente aceptables entre verbos o entre sustantivos.

Como se introdujo en el epígrafe de sinonimia en el uso habitual de estas relaciones, se partirá de una palabra, no de un concepto, por eso primero debe recuperarse en función de la etiqueta sintáctica y del contexto el concepto mas cercano y a partir de él todos los lemas de sus hiperónimos o hipónimos.

2.4 WordNet como ontología lingüística.[1]

Citando y traduciendo a [Miller,1990]

> *Wordnet es una base de datos lexicográfica en línea diseñada para ser usada bajo el control de un programa de ordenador. Sustantivos, verbos, adjetivos y adverbios están organizados en conjuntos de sinónimos, cada uno representando un concepto. Estos conceptos están enlazados según diversas relaciones semánticas.*

En WordNet, una *wordform* – forma-, está representada por una cadena de caracteres ASCII, y un sentido, o significado, está representado por el conjunto de (uno o más) sinónimos que tienen dicho significado. WordNet contiene más de 118.000 palabras diferentes - formas-, y más de 90.000 sentidos – significados-, de palabras diferentes. Aproximadamente el 17% de las palabras en WordNet son polisémicas mientras que aproximadamente el 40% tiene uno o más sinónimos.

Table 1
Illustrating the Concept of a Lexical Matrix:
F_1 and F_2 are synonyms; F_2 is polysemous

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				⋱	
M_m					$E_{m,n}$

Ilustración 5: Formas, polisemia Adaptado de Konchady,1986

1 https://wordnet.princeton.edu/

En Wordnet están representadas diversos tipos de relaciones; bien entre palabras (sinonimia, antonimia) bien entre conceptos – synsets- (hiperonimia/hiponimia). Ver tabla1 Relaciones-semánticas. Además, gran parte de los *synsets* tienen glosas de definición así como ejemplos representativos de uso. Estas dos últimas circunstancias serán especialmente útiles para, junto con el etiquetado sintáctico(sustantivos, verbos, adverbios y adjetivos), intentar resolver la ambigüedad inherente a la polisemia o a las palabras homógrafas.

*2.4.1 Otras ontologías linguísticas relacionadas con WordNet, E*uroWordnet[2].

EuroWordnet [Vossen,1998] pretende extender los conceptos de Wordnet a varias lenguas europeas conservando el espíritu original. Además, en aras de ser de utilidad en tareas de traducción automática, establece relaciones entre los mismos conceptos -synsets- expresados en diversas lenguas para ello necesita un nivel semántico superior que han de respetar todas las lenguas que pretendan adscribirse a Eurowordnet conservando sus particularidades regionales en cada Wordnet local.

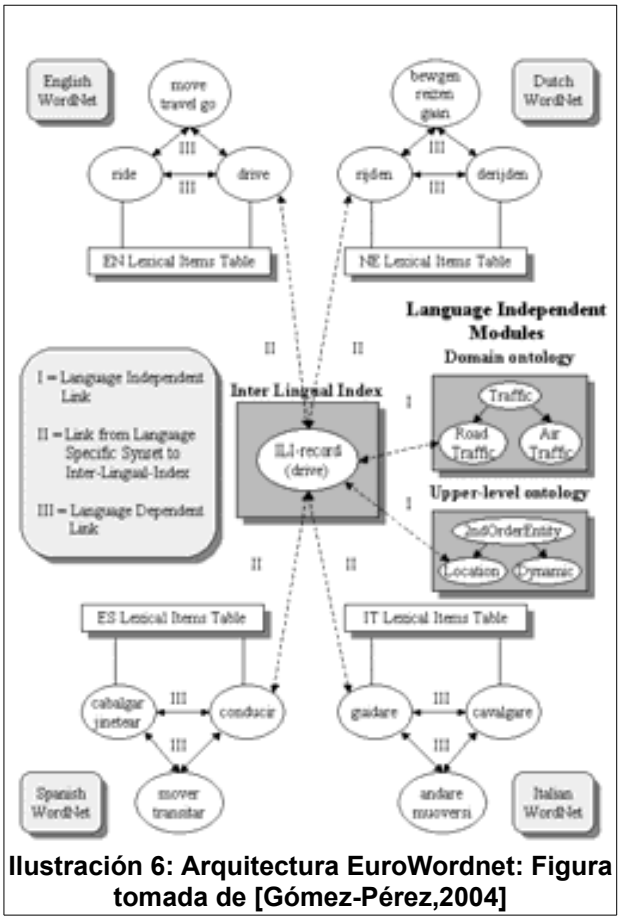


Ilustración 6: Arquitectura EuroWordnet: Figura tomada de [Gómez-Pérez,2004]

2 https://archive.illc.uva.nl/EuroWordNet/

2.4.2 SENSUS[3]

La ontología Sensus -anteriormente conocida como ontología Pangloss -, es una ontología "integrada" disponible gratuitamente producida por el Information Sciences Institute (ISI), California. Es el resultado de integrar:

- El modelo superior PENMAN
- La ontología ONTOS
- Las categorías semánticas LDOCE para sustantivos
- WordNet
- Diccionario bilingüe español-inglés Harper-Collins

Los niveles superiores de la ontología - denominados Ontology Base (OB) -, constan de aproximadamente 400 términos que representan distinciones generalizadas necesarias para los módulos de procesamiento lingüístico (analizador, generador). El OB es el resultado de fusionar manualmente el modelo superior de PENMAN con ONTOS. La región media de la ontología consta de unos 50.000 conceptos de WordNet. Se llevó a cabo una fusión automática de WordNet y LDOCE con verificación manual y el resultado de esta fusión, dada la fusión anterior de OB y WordNet, es una ontología vinculada a un rico léxico del inglés. Una fusión final con el Diccionario bilingüe español-inglés Harper-Collins vincula las palabras en español con la ontología (uno de los objetivos del trabajo es apoyar la traducción automática español-inglés).

Hay pocos detalles de la estructura de la ontología o de las entradas individuales accesibles de forma pública. La fuente electrónica de la ontología consta de varios archivos de definición de palabras y conceptos.

Es importante entender que tanto en SENSUS como en EuroWordNet y en otras muchas OL se ha reutilizado, total o parcialmente, el conocimiento condensado en WordNet lo que favorece la idea de que muchas OL pueden ser elaboradas tomando varias fuentes preexistentes y otras OL. De igual forma muchas OL pueden ser particularizadas para un dominio concreto haciendo de éstas un recurso mas ágil y recuperable en dicho dominio.

3 http://www.isi.edu/natural-language/

2.5 Aprendizaje automático de ontologías.

Si se parte del hecho de que gran parte del conocimiento humano está registrado de manera textual y que, a su vez mediante varios métodos pero principalmente mediante el acceso a la Web, dicho conocimiento es fácilmente recuperable y además, de forma empírica, se constata que la información textual registrada crece – y crecerá - mucho más rápidamente que el personal humano capaz de procesarla[Gantz,2010] se hace inevitable la exploración de vías capaces de procesar automáticamente esa ingente cantidad de información textual y, en particular, su síntesis en forma de ontología. A partir de aquellos hechos y de esta necesidad compendiando conocimientos provenientes de campos diversos relacionados como la extracción de información, el aprendizaje automático, el procesamiento del lenguaje natural, minería de textos, estadística, etc. con nuevas disciplinas *ad hoc* ha surgido el paradigma del aprendizaje automático de ontologías *ontology learning* [Maedche,2001]. En particular, se ha abordado el aprendizaje de ontologías a partir de textos[Wong, 2012], pues estos, se incide, son la fuente humana habitual de registrar el conocimiento.

En dicho aprendizaje automático de ontologías pueden darse una serie de pasos que, a modo orientativo, suponen un primer intento de arquitectura de un sistema de aprendizaje de ontologías. Básicamente esto ya se presentó en la figura Ilustración 4: "Pipeline" ontologías lingüísticas de la página 20 pero conviene comentarla y ampliarla.

- Formación de un corpus o selección de textos representativos de un determinado dominio. Qué, efectivamente, estos textos sean estadísticamente significativos y representativos de la temática será crucial para que una ontología de dicho dominio sea considerada apta y sobre todo será de gran ayuda en pos de desambiguar polisemias y homonimias.
- Extracción de los términos fundamentales de dicho corpus.
- Paso de dichos términos a conceptos del dominio.
- Extracción de relaciones entre los conceptos..
- Ampliación de conceptos y relaciones con otros recursos previos ya existentes.

2.6 Obtención de conceptos.

Una vez extraídos, por los medios que se comentarán más adelante, los términos más significativos del dominio en cuestión a través de un corpus representativo, independientemente de la naturaleza y de la cantidad de los mismos, se hace necesario el paso de dichos términos a conceptos del dominio. La conexión entre el lenguaje natural y la idea de concepto es estudiada por la semiótica[Eco,1977], a través de la relación entre el signo y el concepto. Esta conexión, de hecho, es la que permite asegurar que es posible obtener conceptos a partir de textos ya que, según esto, vienen representados por las palabras – que son la parte constituyente de los términos- que contienen.

Se puede afirmar, con carácter general, que en cualquier método de obtención de conceptos a partir de un corpus textual se distinguen dos etapas:

- Reconocimiento de los términos del texto. En particular determinando, por métodos lingüísticos, heurísticos, frecuentistas u otros, cuantos de estos términos se convertirán potencialmente en conceptos.

- Generación o determinación de los conceptos a partir del conjunto de términos finalmente determinados.

Debe entenderse que la calidad final de las ontologías lingüísticas ligeras sintetizadas dependerá, en gran medida, de la correcta elección de los conceptos y del número apropiado de los mismos. El caso mas sencillo es cuando un término puede asignarse únicamente a un solo concepto y cuando un concepto puede asignarse, igualmente, a un único término. Fenómenos lingüísticos habituales como la polisemia y como la homonimia - en entornos textuales, términos homógrafos -, dificultan dicha tarea, así como los sinónimos que pueden expresar todos ellos el mismo concepto; recuérdese la figura de la página 27 Formas, polisemia Adaptado de Konchady,1986. Así, es probable encontrar que varios términos determinen el mismo concepto. Varios métodos se apoyan, para intentar determinar el significado de los términos y desambiguarlo, en el uso de ontologías lingüísticas generales como las ya citadas Wordnet y Eurowordnet. Trabajos que usan esta metodología son: [Wimmer,2013] o [Yarowsky,1992].

Otros métodos clásicos de desambiguación de significados vienen compilados en [Witschel,2004]

En general, gran parte de los métodos de desambiguación de significados usan una funcion de cercanía[Sánchez,2012] entre el contexto del término a desambiguar y una glosa o ejemplo de uso tomados de una Ontología Lingüística General de cada uno de los significados -conceptos -, en contienda; devolviendo, obviamente, el más cercano. Entre éstas la más usada, convertida ya en un estándar de facto, es la función de Lesk[Lesk, 1986] que básicamente toma la cardinalidad de la intersección entre los tokens del contexto y los de la glosa o algún ejemplo representativo de uso. Se han estudiado muchas variantes y adaptaciones del algoritmo original que conservando la esencia del algoritmo juegan con ampliar o variar lo que se considera contexto del término, asi como lo que se considera glosa; de igual forma también se han probado diversas opciones para medir la *cercanía* entre contexto y glosa. Un ejemplo muy usado es Lesk adaptado / extendido: [Banerjee,2002]

2.7 Extracción de términos.

En *2.3 Ontologías lingüísticas.* de la página 19 se dieron unas explicaciones formales y una definición del concepto de *término.* No obstante, en este apartado se expondrá el estado actual de la cuestión en cuanto a la extracción de términos en textos mas allá de que en unos casos la idea concreta de término pueda variar con la idea concreta de término en otros casos. Se pretende, pues, abstraer la idea concreta y cuando esto no sea posible se mencionará expresamente. Básicamente bastará con considerar un término como una lista de tokens lingüísticos atómicos – de uno o más elementos -, en un determinado texto tras – o no – una determinada normalización. Así, pueden considerarse en un determinado texto un término de un solo token como "Castilla" o una lista de tokens como "vinos tintos de Valdepeñas" o esta misma frase tras normalizarla eliminado stopwords y lematizando "vino tinto Valdepeñas". En casi todos los métodos habituales se forma de manera automática una lista de expresiones candidatas a término. De esta lista un nivel de decisión superior - automático o un experto humano – selecciona cuales y cuantas de estas expresiones son finalmente términos; para ello suele medirse cual es la potencialidad de esta expresión para convertirse en un concepto [Kageura,1996] y [Nakagawa,2004] entre muchos otros, usan este enfoque.

2.7.1 Extracción de términos con patrones léxicos.

En aquellos dominios - medicina, química, farmacopea etc.-, que posean una jerga muy concreta y cuyas reglas léxicas de formación de las palabras de dicha jerga sean rígidas una parte importante de los términos seguirán un muy determinado patrón léxico. Esto, obviamente, implica que lo que, con este método concreto, se extraerán son términos de un solo token pero a la vez facilita que los métodos -habitualmente expresión regular o autómata finito - que capturarán dichos términos sean conceptualmente sencillos.

Desde luego, este método con ser utilísimo en estos dominios que se han citado deben, en gran parte de las ocasiones, ser complementados con otros. De igual forma, el método adolece de generalidad y la expresión regular o expresiones regulares – u otros métodos de extracción -, a utilizar deben ser preparadas, ex profeso, para cada dominio.

Fukuda en [Fukuda,1998] utiliza este método en textos biológicos para identificar proteinas. [Pecina,2010] propone extraer colocaciones con métodos léxicos.

2.7.2 Extracción de términos con patrones sintácticos.

Para la extracción de términos más complejos puede acudirse a análisis menos superficiales, en particular aquellos que buscan patrones sobre el etiquetado sintáctico de los textos objeto de estudio. En tal empeño ha de tenerse cuidado con el tipo de normalización al que se somete a los textos, por ejemplo: no suele ser útil eliminar *stopwords*[4] pues precisamente muchas de estas *stopwords* son los tokens que proveen de estructura sintáctica a los textos. Sin embargo, si suele ser útil, en pos de extraer términos estadísticamente significativos, lematizar los tokens, como tal lematización, el resultado de este proceso son lemas, *palabras de diccionario,* a las que puede asignárseles un etiqueta sintáctica. Es igualmente útil cribar signos ortográficos.

Barker en [Barker,2000] y Witten en [Witten,2005] proponen considerar como expresiones candidatas a términos todas las frases nominales. Sarkar en [Sarkar, 2016] aporta una sencilla expresión regular sobre las etiquetas sintácticas para recuperar las frases nominales más sencillas en el idioma inglés:

{<DT>? <JJ>* <NN.*>+}

Cabe comentar este patrón pues codifica que se pueden recuperar todas las frases nominales que tengan uno o ningún determinante seguidos de 0 o más adjetivos seguidos de 1 o más sustantivos de diversos tipos (útil si no se ha lematizado previamente). Es interesante comentar que este patrón encontraría términos formados por un único sustantivo.

Desde luego, ajustar el patrón para cada dominio particular puede hacer mucho más efectivo el método. Se apunta que este método suele aportar muchas expresiones candidatas por

4 Lista de *stopwords* en inglés http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

lo que será necesario, para el buen desempeño del mismo, acompañarlo por métodos de ponderación de las expresiones candidatas a términos. En general, estos métodos asignan una valoración a cada una de las frases nominales extraídas y se quedan con las *n* mas altamente puntuadas. Ahora sí, como paso previo, es necesario eliminar las *stopwords* puesto que éstas, por su propia naturaleza, son frecuentes y uniformemente distribuidas a lo largo de todas las frases nomimales. En pos de la valoración de cada una de las frases nominales se tratará de buscar fórmulas que tengan en cuenta criterios locales -aquellos que computan sobre la frase concreta en cuestión- y criterios globales – por oposición, aquellos que tienen en cuenta criterios computados a partir de todas las frases nominales-, con esto se persigue puntuar mas alto aquellos términos más específicos y no puntuar alto aquellas frases que simplemente repitiesen muchas veces un token habitual. Es fácil prever que en cuanto a los primeros un factor fundamental será la frecuencia de aparición de tal término en tal frase(TF) f_{ij} quizá, con alguna normalización; en cuanto a los segundos Salton[Salton,2003] propone varios criterios para computar la aportacion global, entre estos el más usado es IDF . Así el valor asignado a cada token i en la frase nominal j vendrá dado por la expresión matemática:

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{(k + df_i)}\right)$$

w_{ij} es el peso de un token determinado i en una frase nominal j

tf_{ij} es el número de ocasiones en que el token i aparece en la frase nominal j

df_i es el número de frases nominales en donde aparece el token i . Toma valores entre 1..N

N es el número total de frases nominales

k previene de apariciones de divisiones por cero. Un valor muy pequeño

Así, el peso total de una frase nominal j $W_j = \sum_{i=1}^{|vocabulario|} w_{ij}$

2.7.3 Extracción de términos basados en "colocaciones".

En palabras de [Bustos, s.f.][5]:

> *Una colocación es una combinación estable de palabras que se emplea de manera preferente, en lugar de otras también posibles, para referirse a un determinado objeto o estado de cosas de la realidad extralingüística. Se trata de combinaciones como vino tinto, pronunciar un discurso, asquerosamente rico o fracasar estrepitosamente, que a cualquier hablante nativo le resultan conocidas, pero que no son en modo alguno evidentes.*

Asimismo, Bustos continua diciendo:

> *Las colocaciones se refieren a un fenómeno que se puede constatar en cualquier lengua: de todas las combinaciones de palabras que en principio son posibles para referirse a una realidad dada, en la práctica únicamente se utilizan unas pocas o, incluso, una sola.*

De estas palabras se desprende que, integrando criterios frecuentistas que serán muy útiles en la computación de las lenguas naturales, una *colocación* puede caracterizarse como una secuencia o grupo de palabras -tokens- que tienden a ocurrir con mucha frecuencia, de modo que esta frecuencia es mayor de lo que sería una simple ocurrencia aleatoria o casual.

La idea que sustenta el paso de colocación a término y de éste a concepto es la alta cohesión sintáctica entre las palabras de la colocación y que el significado de la colocación emerge de la aparición conjunta de dichas palabras y no solo de la agregación de los significados particulares de cada una de las palabras.

Por tanto, a la hora de extraer automáticamente las colocaciones formadas por n palabras o tokens $w_1, w_2 \ldots w_n$ en un texto determinado tendrán especial relevancia las expresiones matemáticas de probabilidad conjunta puesta en relación con el producto de las probabilidades individuales. En el supuesto de independencia condicional entre las palabras de la presunta *colocación*

5 https://blog.lengua-e.com/2010/que-son-las-colocaciones/

$$P(w_1, w_{2,} \ldots w_n) = \prod_{i=1}^{n} P(w_i)$$

En este caso $w_{1,} w_2, w_n$ no sería una colocación . Sin embargo si

$P(w_1, w_{2,} \ldots w_n) \gg \prod_{i=1}^{n} P(w_i)$ habría fundadas sospechas de que $w_{1,} w_{2,} w_n$ es una colocación o, en su defecto, una expresión fija: locución, modismo o refrán. En cualquier caso significa que su aparición conjunta no es debida al azar. Son varias las expresiones matemáticas que se utilizan en la literatura para detectar potenciales colocaciones. Entre éstas cabe destacar:

"pointwise mutual information"[Church,1990] definida como:

$$pmi(x, y) = \log \frac{p(x, y)}{p(x) * p(y)}$$ x é y son palabras o tokens

Cuanto más grande sea $pmi(x, y)$ mas sospechas habrá de que se está en presencia de una colocación (x,y). Por tanto *pmi* puede ser utilizado para extraer las n colocaciones mejor puntuadas.

Son muchos los autores que sostienen que en una colocación, para que sea tal, sus palabras constituyentes son insustituibles por otras palabras. Esto será matizado por el autor de este trabajo. Por ejemplo nadie duda de que en un entorno oleícola se puede considerar una colocación el término: "aceite oliva virgen extra" pero bien podría serlo la expresión "aceite oliva virgen superior" o en un entorno vinícola la terna "vino tinto tempranillo" bien podría ser extraida como colocación, con igual semántica para el término "vino tinto cencibel". El autor de este trabajo pretende mostrar que, en determinados contextos, existen colocaciones que no son *expresiones fijas*

Autores como Sarkar [Sarkar, 2016] han propuesto considerar los *n-gramas* más frecuentes como los candidatos idóneos a colocación. El paso de *n-grama* a *término* se sostiene en la idea de que los n-gramas frecuentes lo son, precisamente, por su condición de *términos* del texto. Sobre estas ideas otros autores como [Smadja,1991] proponen usar la etiquetación sintáctica de los tokens del n-grama para conceder a dichos n-gramas mas verosimilitud como colocación en unos contextos u otros y pudiendo recuperar determinados n-gramas en detrimento de otros. Desde luego, conviene a la hora de finalmente escoger los n-gramas mas frecuentes

tratar de forma particular las *locuciones* o *expresiones idiomáticas*[6] (noche toledana, pata de palo) identificándolos quizá mediante comparación con un diccionario de expresiones idiomáticas con la misma normalización que el texto del que se extraen los n-gramas o mediante la intervención de un experto.

2.8 Aprendizaje de relaciones no taxonómicas.

	Atributo-1	Atributo-2	…….	Atributo-n
T1	True	False	…….	Talse
T2	False	False	……..	True
. . . Tn	. . . True	. . . False	…….. ……. ……. True	. . . False

Tabla 2: Matriz Transacciones Valor

En general, el aprendizaje de relaciones no taxonómicas (meronimia, sinonimia, antonimia, etc) es más complejo que el de relaciones taxonómicas porque aquéllas dependen fuertemente de la lengua natural en cuestión y dentro de éstas del dominio específico. Se han alcanzado, no obstante, ciertos hitos. Por ejemplo, en el caso de la meronimia intentando aprender patrones léxico-sintácticos, obviamente para cada idioma. En el caso del idioma inglés puede consultarse el trabajo de [Berland,1999]

En el caso de la sinonimia y con ayuda de una fuente externa como WordNet dado un par (lema, etiqueta-sintáctica) y un contexto en el texto donde ese par aparece (lema, etiqueta-sintáctica,contexto) se puede a partir de ambos y con ayuda de WordNet derivar un concepto (lema, etiqueta-sintáctica,concepto) y de este concepto extraer todos su sinónimos. Semejante idea, simple conceptualmente, se usará con profusión en este trabajo.

6 Se considera expresión idiomática o locución aquella expresión cuya semántica no puede ser extraída total o parcialmente del significado de sus constituyentes. Así, *noche toledana* significa una noche de calor extremo o *pata de palo* es una prótesis ortopédica.

También, en [Perkins,2010] puede verse un sencillo código para reemplazar una palabra antecedida por una negación por su antónimo con ayuda de WordNet.

Sin embargo, como se expone a continuación, en la extracción o aprendizaje de relaciones no taxonómicas el paradigma más usado en gran parte de las ocasiones es el de las Reglas de Asociación.

2.8.1 Reglas de Asociación.

Las reglas de asociación [Agrawal,1993] -en adelante RA- son un método que pretende descubrir patrones de *coocurrencia* en grandes bases de datos especialmente dispuestas a tal efecto. Así, deben disponerse dichas bases de datos en forma de transacciones donde cada una de éstas registran si tal o cual atributo ha tenido lugar en la transacción en estudio dando naturaleza booleana a la matriz generada. En realidad, se persiguen asociaciones no ya entre atributos sino entre los posibles valores de dichos atributos. Ver *Tabla 2: Matriz Transacciones Valor.*

Se buscan pues, cuales subconjuntos de atributos-valor tienden a *coocurrir* frecuentemente en un conjunto de transacciones y para cada subconjunto frecuente de k atributos-valor que reglas de implicación ($2^k - 2$) hay entre sus subconjuntos y que fuerza tiene dicha implicación. Este método, per se, no dice nada de la naturaleza de dichas *coocurrencias* y de sus reglas, tampoco si existe relación taxonómica entre los antecedentes y consecuentes de las reglas, menos aun debe asumirse necesariamente una relación de causalidad; sólo dice que tal coocurrencia se da y además puede medirse la fuerza de la inferencia de dicha *coocurrencia* en cada regla.

Para ello, el método primero busca qué conjuntos de valores de los atributo ocurren frecuentemente pretendiendo con ello descartar asociaciones debidas únicamente al azar, en este sentido es muy utilizado el algoritmo *Apriori AQ* [Michalski, 1983] o el algoritmo *Aclose* o *Close Algorithm*. [Pasquier,1999]

Después con cada uno de estos conjuntos frecuentes de valores busca las reglas que mas fuerza tienen aportando con ello, además, un sentido a la implicación.

Formalizando lo dicho hasta ahora se tiene que:

$T=\{t_i|i=1..n\}$ Es un conjunto de n transacciones cada una formada por un conjunto de elementos(valores de los atributos en cuestión).

$t_i=\{a_{(i,j)}|j\in 1...m_i, a_{(i,j)}\in C\}$ y cada elemento $a_{(i,j)}$ pertenece a un conjunto de valores de atributos C.

El algoritmo tiene como objetivo determinar reglas de asociación del tipo: $X_k \Rightarrow Y_k$

Donde:

X_k es el antecedente de la regla, $X_k \subset C$

Y_k es el consecuente de la regla, $Y_k \subset C$

Además se cumple $X_k \cap Y_k = \varnothing$

Para seleccionar los conjuntos frecuentes se usa el concepto de *soporte*. Así, sea X_k un conjunto de k atributos-valores y sea n la cardinalidad del conjunto de transacciones considerados.

Entonces se llamará $soporte(X_k)=\frac{|\{t_i|X_k\subseteq t_i \wedge t_i\in T\}|}{n}$

Se debe fijar un umbral mínimo que todos los conjuntos candidatos deben superar para seguir siendo considerados en el algoritmo. Este umbral mínimo debe ser escogido con especial cuidado para cada una de las posibles aplicaciones en las que se utilicen las reglas de asociación.

El concepto de soporte para un conjunto de atributos-valores puede ser extendido para el caso de una regla de asociación $Y_k \rightarrow Z_k$ sin más que considerar $X_k = Y_k \cup Z_k$ de lo que se desprende que el soporte para $Y_k \rightarrow Z_k$ es igual al soporte de $Z_k \rightarrow Y_k$

Ahora toca medir la fuerza de la implicación encontrada en la regla $X_k \Rightarrow Y_k$ Existen varias medidas para tal empeño pero la más utilizada es la confianza que puede ser definida así:

$$confianza(X_x \rightarrow Y_k) = \frac{(soporte(X_k, Y_k))}{(soporte(X_k))}$$

Esto puede interpretarse como que cuanto mas alta es la confianza más probable es para Y_k estar presente en una transacción que contiene a X_k lo que no deja de ser una estimación de la probabilidad condicionada de Y_k cuando X_k está presente.

Debe decirse que la confianza es un candidato, existen muchos otros, firme para la criba de las reglas más interesantes de aquellas extraídas por un algoritmo; haciendo que solo pasen la criba aquellas reglas en las que su nivel de confianza supere un umbral previamente fijado. De nuevo, este umbral deberá ser escogido con especial cuidado para cada una de las tareas en las que las reglas de asociación puedan ser utilizadas.

Debe tenerse muy en cuenta el gran número de reglas de asociación que pueden extraerse de un conjunto de transacciones con k pares atributo-valor. Así, [Tan,2006] fija el número de reglas extraíbles en:

$$R = 3^k - 2^{(k+1)} + 1$$

Expresión ésta de evidente crecimiento exponencial con lo que se hará necesario, en gran parte de las aplicaciones, métodos, analíticos o heurísticos, que permitan determinar la utilidad de las reglas o cuan novedoso es el conocimiento que dichas reglas aportan para considerar exclusivamente un número limitado de las mismas.

2.8.2 Reglas de Asociación en Minería de Textos

En el epígrafe anterior se ha hecho una presentación del concepto de Regla de Asociación junto con algunas referencias bibliográficas. Se comentan ahora algunos de los campos donde dicho concepto se ha aplicado con evidente éxito:

- Análisis de la cesta de la compra donde se analizan patrones de compra conjunta.
- Recuperación conjunta de información. Por ejemplo: sitios webs que suelen visitarse conjuntamente dentro de un lapso con objeto de situar dichas páginas en un mismo servidor.
- Análisis de textos. Este trabajo se centrará en dicha área. EART (Extract Association Rules from Text) [Mahgoub,2008]

Una vez establecido el campo en donde se querrán aplicar las RA lo primero será escoger el corpus significativo de dicho campo y someter a cada uno de los documentos de dicho corpus al preprocesamiento pertinente(lematización/estemización, cribado de *stopwords, tokenizacion)* a continuación deberá fijarse que se considerará como transacción y que objetos serán tomados como pares atributo-valor. En el caso de la Minería de Textos pueden considerarse como transacciones las frases, los párrafos o incluso - dado un corpus - los documentos. En cuanto a los pares atributos-valor estos deben ser los términos lingüísticos o los conceptos adecuados a la tarea en concreto. Estos términos o conceptos pueden derivarse de las palabras del texto, tokens, n_gramas u otros. Por ejemplo, [Mahgoub,2008] y [Bhujade,2011] buscan establecer coocurrencias frecuentes entre palabras claves del texto previamente señaladas como tales.

Debe decirse que la cuidada selección de ambas nociones permitirá atenuar la tendencia que en la Minería de Textos tiene la matriz transacción/atributos-valor en devenir como dispersa.

Puesto que ahora se está en un contexto lingüístico algunos autores han considerado que, a partir de dicho contexto, se podían hacer esfuerzos para intentar ir más alla en la identificación automática de las relaciones entre los términos asociados, trascendiendo la ya asumida relación de *coocurrencia* frecuente. Kavalec y Svátek [Kavalec,2005] presentan un método para asignar automáticamente un nombre a estas relaciones, identificando el tipo de la misma. Este método se utilizó en el sistema OntoLearn [Velardi,2005]. No obstante, sigue siendo una relación de tipo no taxonómica, de tal suerte que no puede establecerse una jerarquía conceptual entre los términos del consecuente y del antecedente ni entre los de éste y los de aquel. Básicamente, este

método pretende extraer de los verbos más frecuentes encontrados en una determinada ventana alrededor de la asociación de términos considerada el tipo de relación, etiquetando dicha relación con el verbo. En principio esta ventana será un subconjunto de la frase o de la transacción considerada. Entre los verbos más frecuentes se escogerá como nombre de la relación aquel que maximice la expresión:

$$MAX_v\, AE(c_1 \wedge c_2|v) = \frac{(P(c_1 \wedge c_2|v))}{(P(c_1|v)P(c_2|v))}$$

Donde:

AE "above expectations"

$v \in V$ (verbos candidatos frecuentes)

$c_1 c_2$ son los conceptos

$P(c_i|v)$ es la frecuencia con la que el concepto c_i se encuentra en la misma transacción que el verbo v

$P(c_1 c_2|v)$ es la frecuencia condicionada con la que los conceptos c_1 y c_2 se encuentran en la misma transacción que el verbo v calculándose de la siguiente forma:

$$P(c_1 \wedge c_2|v) = \frac{|\{t_i|v, c_1, c_2 \in t_i\}|}{|\{t_i|v \in t_i\}|}$$

2.8.3 Valoración de las Reglas de Asociación en Minería de Textos.

Durante la generación algorítmica de las RA se ha abordado el crecimiento exponencial de las mismas podando el árbol de generación de los conjuntos frecuentes basándose en el siguiente principio:

> *"(Principio Apriori): Si un conjunto de pares atributos-valor es frecuente entonces todos sus subconjuntos deben serlo"[Tan,2006]*

Este principio permite en la propia desenvoltura del algoritmo no considerar conjuntos de atributos-valor poco prometedores y con ello no considerar las reglas que de ellos podrían derivarse. Sin embargo, lo que se persigue ahora es dar un valor de utilidad o medir el conocimiento nuevo aportado por las reglas ya generadas permitiendo, por tanto, establecer una relación de orden entre todas las reglas.

Así, puede utilizarse una fuente externa a las RA -entre otras, WordNet o EuroWordNet- que considere los conceptos asociados a las mismas y que tenga dispuestos entre dichos conceptos las relaciones semánticas apropiadas entre estos pudiéndose medir – según diversos criterios semánticos- cuán lejanos – o cercanos - están estos. Para ello es necesario poder acomodar el tipo de conceptos que manejen las RA y el tipo de conceptos que maneje la fuente externa. Dicha fuente externa, habitualmente, tendrá forma de ontología lingüística en el campo de la Minería de Textos y la Programación del Lenguaje Natural.

Por ejemplo [Basu,2001], propone un nuevo método de estimación del grado de novedad de reglas descubiertas por métodos de Data Mining usando la base de conocimiento léxico WordNet [Ferllbaum,2010] que contiene miles de conceptos enlazados semánticamente por relaciones tales como antónimos, hiperónimos, homónimos, etc. En este modelo el grado de novedad de una regla generada se calcula utilizando la distancia semántica, según varios conceptos semánticos, que existe entre los conceptos del antecedente y los conceptos del consecuente de la regla.

Esta distancia se basa en el conocimiento estructural que aporta la jerarquía de conceptos en WordNet. El principio es que cuanto mayor sea la distancia entre los términos de la regla, mayor es el grado de novedad que ésta posee, debido a que la relación semántica entre ellos no es muy frecuente o poco común. Por ejemplo, en el contexto de informes de venta en un supermercado, una regla del tipo “Cerveza → Pañales” podría interpretarse como “las personas que compran cervezas también compran pañales”, este patrón puede considerase más novedoso que el entregado por “Cerveza → Patatas-fritas”, debido a que la distancia semántica que existe en el primer caso en WordNet es mayor que en el segundo. Además, dado que los conceptos

"cerveza" y "patatas fritas" corresponden a alimentos que están semánticamente más cercanos en WordNet se considera que el conocimiento que pueden aportar no es muy relevante, debido a la corta distancia semántica que los une.

La viabilidad de este sistema se ha probado comparando puntuaciones asignadas por dicho sistema y puntuaciones asignadas por comités de expertos humanos. No obstante, deben consignarse algunas debilidades del sistema, éstas son:

- Cómo asignar distancias cuando entran en juego conceptos no recogidos por la fuente externa de conocimiento.
- Depencia de la existencia o no de una ontología adecuada.

2.9 Conclusiones al estado de la cuestión.

Se ha presentado la Ley de Zipf que, según se ha visto, rige para muchas unidades lingüísticas. Esto abre una vía para estudiar si tal Ley se cumple y, en tal caso, en que condiciones para determinadas unidades lingüísticas aun sin estudiar. Estas unidades lingüísticas serán la base para la formación de los términos a extraer. De igual forma se ha presentado el concepto general de Ontología y, en particular el concepto de Ontología Lingüística. A continuación se ha comentado el estado actual del paradigma de *aprendizaje de ontologias lingüísticas a partir de textos* exponiendo como, de forma automática, se forman los términos de los que se sintetizarán los conceptos de dicha ontología. A partir de estos conceptos en el campo de la lingüística se ha hablado de las posibles relaciones entre estos, ora taxonómicas, ora no taxonómicas y como, hasta la fecha, se han extraído con unas metodologías u otras. Debe entenderse que el hecho de que los términos o conceptos extraídos sigan o no la Ley de Zipf marcará si es más o menos viable la extracciòn automática de una ontología lingüística y en ese caso la metodología a seguir. También se ha introducido como puede combatirse la polisemia de los términos mediante la ayuda externa de una ontología lingüística general y el contexto en que dicho término aparece.

3 Desarrollo del trabajo.

El desarrollo del trabajo vendrá pautado por la siguiente tabla:

1º	Disposición de una ontología lingüística general
2º	Elección de la temática sobre la que proyectar la ontología lingüística general. Un hito para desambiguar la polisemia.
3º	Formación y normalización del corpus representativo de la temática
4º	Extracción de los ítems(términos, conceptos) mas significativos del corpus
5º	Elaboración de los diccionarios (Ont. Ling. Ligeras)de sinóminos e hipónimos-hiperónimos
6º	Elaboración del diccionario de Reglas de Asociación

Tabla 3: Desarrollo del trabajo

3.1 Disposición de una ontología lingüística general.

En el Estado de la Cuestión se dio una visión de conjunto sobre el estado actual del paradigma de ontología lingüística haciendo hincapié en los conceptos de la misma y en las relaciones semánticas que entre dichos conceptos se establecen. Es evidente que antes de proceder al trabajo hay que tener claro de que ontologías lingüísticas generales se dispone –en particular, para cada lengua natural- y sobre cada una de éstas constatar que servicios, mediante un determinado interface, ofrece.

Debe, sobre todo, quedar claro que relaciones semánticas ofrece dicha ontología y cotejarlas con el tipo de ontologías lingüísticas ligeras, habitualmente en forma de diccionario, que se pretende elaborar. También hay que tener en cuenta si la extracción o elaboración de dichas relaciones es más o menos compleja. Se recuerda que la ontología lingüística general será el recurso fundamental para dado un concepto extraído del corpus poder asociarle otros conceptos, en particular, sinónimos y cohipónimos.

3.2 Elección de la temática sobre la que proyectar la ontología lingüística general. Un hito para desambiguar la polisemia.

Puesto que el objetivo final de este trabajo es tener un método para sintetizar una serie de ontologías lingüísticas ligeras –en particular, de sinónimos e hiperónimos– que, para una determinada temática, especialicen la ontología lingüística general la elección y fijación de la misma es uno de los primeros pasos en el trabajo. Se debe, antes de nada, calibrar si para tal o cual temática es rentable la elaboración de dichas ontologías lingüísticas ligeras temáticas con forma de diccionarios. Si la temática es demasiado amplia no merece la pena la elaboración de los diccionarios. En este caso deberá usarse la ontología lingüística general que ya habrá sido entrenada y elaborada en función de un corpus - igualmente - general. Con el uso de dicha ontología general no se resolverían problemas como la polisemia o los homógrafos puesto que una palabra homógrafa como *hoz*[7] aparecería con parecida frecuencia en el corpus general tanto en su acepción como apero agrícola o en su acepción como accidente geográfico, teniendo, por tanto, que ser desambiguada por programación para cada aparición de hoz en el texto a procesar.

En el caso de que la temática fuese lo suficientemente particular y, en el ejemplo de *hoz*, se dispusiese de sendos corpus temáticos disjuntos(ámbito geológico, ámbito agrícola) en el corpus geológico y en el diccionario basado en él los sinónimos de *hoz* serían cañón, desfiladero, angostura etc. en cambio en el corpus agrícola el sinónimo de *hoz* sería guadaña. De igual forma los cohipónimos, de haberlos, de *hoz* en un ámbito u otro serían igualmente disjuntos. Por lo que la desambiguación vendría ya recogida en el diccionario de sinónimos o

7 https://dle.rae.es/hoz

hipónimos y no habría de hacerse por programación para cada vez que apareciese el término *hoz,* simplemente se recuperarían del diccionario ad hoc. Es decir, dada una temática lo suficientemente particular y un diccionario entrenado en un corpus de dicha temática el propio diccionario -ontología lingüística ligera- desambiguaría el significado de una palabra polisémica u homógrafa.

Asimismo, si no se va a hacer uso de los diccionarios en varias ocasiones en un determinado lapso para esa determina temática tampoco podría merecer la pena tal como se representa en la *Ilustración 7: Diccionario de una temática versus muchos textos de dicha temática de la página 60* Es decir, rige el principio de elaborar una sola vez y utilizar muchas.

3.3 Formación y normalización del corpus representativo de la temática.

Una vez fijada la temática deberá buscarse un corpus representativo de la misma de tal modo que se puedan extraer, primero, los términos y luego los conceptos más significativos de dicha temática. Debe, dicho corpus, poder anotarse sintácticamente y normalizarse -en particular, lematizarse -. No debe importar demasiado que el corpus sea grande, siempre y cuando esto aporte mayor significación estadística, porque como se vió en *Ilustración 7: Diccionario de una temática versus muchos textos de dicha temática* el diccionario se elabora una vez y se usa tantísimas otras.

El éxito en la elaboración de las ontologías lingüisticas ligeras propuesto descansa, en gran parte, en que el corpus de entrenamiento esté bien escogido y delimitado -circunstancia ésta, en nuestros días, muy común en todas la tareas de AI- . Si el corpus no está bien escogido y delimitado a una tarea concreta el diccionario elaborado no presentará especiales ventajas frente al uso de la ontología lingüística general.

Una vez hecho esto se debe fijar un método de extracción de *términos* y *conceptos* y el número de estos que serán extraídos.

Puede decirse que:

La elaboración de las ontologías lingüísticas ligeras -en adelante, OLL- será útil y viable si la temática es lo suficientemente particular y sobre dicha temática es posible elaborar un corpus representativo de donde extraer los términos y conceptos mas significativos hacia las que hacer converger los diccionarios con que se representarán las OLL. Asimismo, si se dispone de una -- o unas -- ontologías lingüísticas generales de las que extraer las relaciones semánticas y finalmente se va a hacer uso de dichos diccionarios en reiteradas ocasiones.

3.3.1 Normalización del corpus.

El proceso de normalización del Corpus temático debe estar encaminado a facilitar la significación estadística de los términos a extraer y a facilitar la posterior elaboración de las ontologías lingüísticas ligeras temáticas, con especial incidencia, en la desambiguación de la polisemia.

En primer lugar, el cribado de signos ortográficos se hace necesario pues estos son numéricamente importantes pero lo son poco desde el punto de vista semántico ya que de un signo ortográfico no pueden extraerse relaciones semánticas, ni sinónimos ni hiperónimos ni tiene utilidad, al menos a efectos de este trabajo, la búsqueda de patrones de coocurrencia entre signos ortográficos.

En segundo lugar, asumiendo que deben cribarse las *palabras vacías stopwords[Luhn,1958],* puesto que éstas, por definición, carecen de significado y, por tanto, no podrán dar origen a términos ni a conceptos fijar que será, para este trabajo, considerado *stopwords* es tarea fundamental. El ponente de este trabajo dudó en utilizar un conjunto de *stopwords* general para cada idioma o si, por el contrario, considerar un conjunto de *stopwords* para cada temática particular. Tras la experimentación se tomó la decisión de usar un conjunto de stopwords general porque esto evita no considerar en la elaboración de las ontologías lingüísticas ligeras palabras que aun siendo muy frecuentes en una determinada temática no lo son en otras pervirtiendo el objetivo de extraer los términos más frecuentes de una determinada

temática. Así pues, el uso de un conjunto de stopwords general que cribe palabras frecuentes pero carentes de significado en todos las temáticas, muy en especial *function words,* es la elección tomada.

Ahora, en tercer lugar, cabría preguntarse que medios para facilitar la significación estadística de los términos del corpus contraponer e investigar. Surgen, de inmediato, dos opciones: lematización versus *estemización*.

Si se considera la *estemización* el resto del trabajo quedaría muy limitado pues sabido es que una palabra de diccionario sometida a la *estemización* podría dar como resultado una raíz que no estuviese contemplada en el diccionario canónico ni por ende en una ontología lingüística general. Esto haría imposible asignar a dicha raíz una etiqueta sintáctica y por tanto no se podría considerar ningún tipo de desambiguación basada en la etiqueta sintáctica haciendo imposible, además, extraer por los métodos propuestos sinónimos o hiperónimos. Solo cabría buscar patrones de coocurrencia. Debe decirse, sin embargo, que el proceso de *estemización* es varios órdenes de magnitud mas rápido que el de lematización.

Si como alternativa -será la que finalmente se tome en este trabajo -, se considera la lematización, asumiendo que dicho proceso, como se acaba de decir, es mucho mas lento pero que, sin embargo, da como salida una palabra de diccionario esto hará que se cuente con con una serie de ventajas:

- El proceso de lematización haciendo converger determinadas clases de palabras en su representante facilitará la significaciòn estadística de los términos y, a partir de estos, de los conceptos de una determinada temática: Un ejemplo resultará muy ilustrativo: todos los tiempos verbales convergerán en el infinitivo.
- Todos los tokens del corpus, al ser palabras de diccionario, podrán ser sintácticamente etiquetados. Esto ya facilitará cierto grado de desambiguación de la polisemia. No es lo mismo el rápido de un rio (sustantivo) que un humano rápido(adjetivo) y por tanto darán origen a dos conceptos distintos.

- De igual forma podrán buscarse, con la ayuda de la ontología lingüística general sinónimos e hiperónimos.

Así pues, como hito en este trabajo, el corpus habrá quedado representado en una lista dual de lemas con sus etiquetas sintácticas en donde ninguno de estos lemas es una *stopword* o un signo ortográfico.

3.4 Configuración de Términos. Naturaleza y número de los mismos.

En *2.3 Ontologías lingüísticas.* se ha dado una definición canónica de *término* y en *1.2 Extracción de términos* se han dado las preguntas que debe responder esta investigación. La configuración final de lo que, para este trabajo, se considere *término*, debe facilitar, entre otras cosas, los siguientes objetivos:

1. Paso de dichos términos del corpus a conceptos de las Ontologías Lingüísticas Ligeras.
2. Desambiguación de la polisemia. Para ello se propone que formen parte de la configuración final de los *términos* etiquetas sintácticas y contexto. Este contexto será o bien la frase del corpus donde aparece el *término* o bien todo un determinado n-grama etiquetado.
3. Búsqueda de las relaciones semánticas a reflejar en las Ontologías Lingüísticas Ligeras.

Así pues se proponen, para este trabajo, dos configuraciones concretas para *término: Tripleta contextual* y *n-grama etiquetado*.

3.4.1 Tripleta contextual

Se define como *tripleta contextual* (lema,tag,contexto) una terna donde el primer elemento es un lema, el segundo una etiqueta sintáctica[8] y el tercero, un contexto; bien una frase donde dicho lema aparece en el corpus de entrenamiento o bien un *n-grama* donde aparece el lema en cuestión. Obviamente el tercer elemento ha debido facilitar la asignación de etiqueta sintáctica al primer elemento. Si se parte de un corpus que tras su normalización expuesta en *3.3.1 Normalización del corpus.* su lista dual [(lema,tag)..] tiene *m* elementos el número de tripletas contextuales con una frase como contexto de dicho corpus será m^9. Se considerarán los *m* términos en forma de tripleta contextual puesto que antes de su conversión en conceptos -tripleta conceptual -, no se cribarán ni podrá considerarse una distribución de frecuencias sobre los mismos, puesto que el hecho de que el tercer elemento sea una frase o un n-grama hace que casi todos las tripletas contextuales aparezcan con frecuencia uno.

3.4.2 N-gramas etiquetados

Se define un *n-grama etiquetado* como una estructura de n elementos, donde cada elemento será un par (lema,tag).

Así, se tendrá que un n-grama será [($lema_1$,tag_1), ($lema_2$,tag_2), … ,($lema_n$,tag_n)]

El paso fundamental a considerar aquí es que los *n-gramas* más frecuentes del corpus son considerados *colocaciones*, éstas son consideradas *términos* y de estos términos, por los medios que se propondrán en el epígrafe siguiente, se extraerán conceptos a considerar en las Ontologías Lingüísticas Ligeras.

Emergen aquí dos problemas de exploración de valores para parámetros:

1. Investigar el valor de *n.* Teniendo en cuenta que el n-grama debe ser a la vez capaz de capturar los términos formados por varias palabras y ser además elemento de desambiguación. Se apunta que si n es grande habrá más contexto de desambiguación pero habrá mas dificultad en que los n-gramas sean estadísticamente significativos.

8 En este trabajo solo se considerarán adjetivos(a), sustantivos(n), verbos(v) y adverbios(r)

9Podría haber excepciones muy determinadas a esto. Por ejemplo, en letanías, donde exactamente la misma frase se repite muchas veces. No obstante, eso, en esencia, no cambia nada de lo expuesto.

2. Cuántos de los *j* n-gramas etiquetados mas frecuentes han de ser considerados términos. Para ello ha de considerarse la distribución de frecuencias de dichos n-gramas.

En este caso se han de cribar los *j* n-gramas más frecuentes antes de que estos sean utilizados para generar conceptos.

3.5 Extracción de los conceptos mas significativos de un dominio mediante un corpus.

En el apartado *2.4 WordNet como ontología lingüística.* se ha dado, citando a [Miller,1990] una definición de lo que en *Wordnet* se considera un *synset*. Puede, por generalización y a efectos de este trabajo, asumirse tal definición para caracterizar lo que en una Ontología Lingüística General se considerará como *concepto.* En particular, este *concepto* estará formado por un conjunto de lemas sinónimos, una glosa de definición y un conjunto de ejemplos de uso de dicho concepto. De igual forma, se ha dado en *3.4 Configuración de Términos. Naturaleza y número de los mismos.* lo que, en este trabajo, se considerará *término*. Ahora se aborda, siguiendo *1.3 Paso de términos a conceptos* como a partir de esos *términos* extraer los conceptos del corpus y que forma y número tendrán dichos conceptos.

3.5.1 Paso de términos a conceptos. De tripleta contextual a tripleta conceptual

Ya ha quedado fijado lo que significa *concepto* en una ontología lingüística general ahora se debe fijar lo que, para este trabajo, se considerará un *concepto*, el número significativo de estos y su estructura.

En cuanto a su estructura un *concepto* se asimilará a una *tripleta conceptual* y está será definida por una tripleta (lema, tag, concepto). A efectos prácticos el tercer elemento de esta tripleta será conceptualmente equivalente a la noción de *synset* en *Wordnet*.

Debe exponerse un método para pasar de una *tripleta contextual* a una *tripleta conceptual.*

Por tanto, provistos ya:

- En primer lugar, con un término que será expresado en forma de tripleta contextual (lema, etiqueta_sintáctica, contexto circundante)
- En segundo lugar, con una determinada Ontología Lingüística General que provee de conceptos candidatos – aquellos que tienen el mismo lema y la misma etiqueta sintáctica que el término -, de los que se extraerán la glosa o ejemplos de uso de cada uno de los conceptos candidatos
- En tercer lugar con una función de cercanía con la que comparar cada uno de los conceptos candidatos con el contexto y, con ello, asignar al término el concepto mas cercano.

  En la página *56* en el *Algoritmo de búsqueda de un concepto - tripleta conceptual - a partir de una tripleta contextual* se propone un algoritmo para pasar de tripleta contextual a tripleta conceptual.

En este caso, junto al corpus, se usan como medios externos para la asignación de concepto a un término, una ontología lingüística general de las que ya se hablado en este documento y una función de cercanía. Entre éstas la más usada, convertida ya en un estándar de facto, es la función de Lesk[Lesk, 1986] que básicamente toma la cardinalidad de la intersección entre los tokens del contexto y los de la glosa o algún ejemplo representativo de uso.

```
“”””
Termino =(lema,tag,contexto)  una tripleta contextual
OL  habitualmente WordNet,EuroWordNet
funciónDeCercanía habitualmente Lesk
Permite pasar de un término, tripleta contextual
a un concepto  tripleta conceptual (lema, tag, concepto)
“”””



Funcion  desambiguar(lema, tag, contexto,funciónDeCercanía, OL)
        conceptosCandidatos = conceptos(OL, palabra, tag)   #wn.synsets(palabra, tag)
        distancia = funcionDeCercanía(conceptosCandidatos[0].glosa, contexto)
        solución = conceptosCandidatos[0]
        for cC in conceptosCandidatos[1:]
                d = funciónDeCercanía(cC.glosa, contexto)
                if d <  distancia:
                        distancia = d
                        solucion = cC


        devolver(palabra, tag, solucion) ## Se devuelve una tripleta conceptual
```

Texto 1: Algoritmo de búsqueda de un concepto - tripleta conceptual - a partir de una tripleta contextual

Una vez fijada la tripleta conceptual como la estructura apta para representar, en este trabajo, los conceptos el objetivo es encontrar los *p* conceptos en forma de tripleta conceptual(lema, etiqueta sintáctica, concepto) más significativos[10] en cada corpus; *p* podrá ser fijado en función del tamaño deseado para el diccionario buscado e igualmente teniendo en cuenta que será otro hito para desambiguar la polisemia - ver la ilustración *8 Fijación de p. Desambiguación con p y etiqueta sintáctica.Ejemplo calar* -. Y para cada una de estas tripletas

10 No siempre tienen que ser los mas frecuentes, pero en gran parte de los casos la frecuencia de los ítems marcará su condición de significativo

conceptuales, mediante la ontología lingüística general, intervención del experto u otros métodos complementarios, derivar sus sinónimos y sus hipónimos hasta un determinado nivel -habitualmente, se escogerán sus hipónimos inmediatos-. En el primero de los casos el lema -primer componente de la tripleta -, será el representante canónico de todos sus sinónimos y en el segundo el concepto -tercer miembro de la tripleta - será el hiperónimo común para todos los cohipónimos derivados.

Para que los diccionarios sintetizados sean, finalmente, útiles debe quedar claro que los *p* conceptos extraídos deben serlo, además de por el lema en sí, en función de la etiqueta sintáctica de éste y en función de su contexto circundante. Elementos estos, capitales en diversos hitos de desambiguación.

Se adjunta un ejemplo ilustrativo en español, sobre el término *calar*[11]. Dicho término puede ser, entre otras acepciones:

1. Adjetivo. Sinónimo de calizo
2. Sustantivo. Lugar en el que abunda la tierra caliza.
3. Verbo. Acepción 9 en el Diccionario de la Real Academia Española (DRAE). Conocer las intenciones de alguien.
4. Verbo. Acepción 16 en el DRAE. Dicho de un material u objeto. Permitir que un líquido pase a través de él.

La primera pregunta que hay que hacer es: ¿ es el mismo concepto *calar* como adjetivo que como verbo o que como sustantivo ? Evidentemente no y el etiquetado sintáctico así como, el contexto permitirán discernir que calar como adjetivo y calar como sustantivo son conceptos diferentes y, por tanto, que deben contabilizarse cada cual por un lado. Así, se tendrá dos conceptos distintos:

(calar, a, concepto1) y (calar,n, concepto2) donde concepto2 es distinto de concepto1

Puesto que los sinónimos e hipónimos a extraer de concepto1 y concepto2 son distintos

11 https://dle.rae.es/calar

esto dará lugar a entradas distintas en el diccionario --si es que finalmente ambos conceptos están entre los *p* más significativos, ver epígrafe *3.5.2 Fijación del número p de tripletas-conceptuales más significativas.*--. Además esas entradas tendrán sinónimos distintos:

dictSinonimos[calizo][a] = calar (a de adjetivo)
dictSinonimos[pedregal][n] = calar (n de sustantivo(nombre))

Tanto mas en el diccionario de hiperónimos y hipónimos donde calar, puesto que es adjetivo (*ver Tabla 1: Relaciones-semánticas),* no tendrá entrada. Debe decirse que en este caso la presencia de etiquetas sintácticas distintas permite desambiguar la polisemia. Así pues la presencia de etiquetas sintácticas se muestra como un posible elemento de desambiguación de la polisemia.

Pero, ¿ qué pasa cuando dos conceptos coinciden en la forma de su lema y en su etiqueta sintáctica, pero difieren en el significado, verbigracia en el ejemplo de calar como permear, horadar y de calar como conocer, desenmascarar ? En el mismo redactado de la pregunta, implícitamente, se ofrecen ya dos conjuntos de sinónimos disjuntos en función de que el significado sea uno u otro.
¿ Cómo puede, en este caso, abordarse la *desambiguación* y elegir uno u otro conjunto de sinónimos o hiperónimos/hipónimos*?* La respuesta es mediante el contexto en el que la palabra aparece -por ejemplo, la frase del texto en que la palabra figura o si la palabra forma parte de un n-grama tomando éste como contexto- comparando dicho contexto con las definiciones -glosa - o ejemplos de uso presentes --para cada concepto candidato-- en la ontología lingüística general. Para ello provistos de:

- Una ontología lingüística general con conceptos compuestos cada uno de ellos de:
  - Un conjunto de sinónimos
  - Definición del concepto mediante una glosa.
  - Ejemplos de uso de dicho concepto en la lengua natural.

- Un contexto circundante del concepto a desambiguar. Por ejemplo, la frase o el n-grama en que aparece el ítem.

- Un algoritmo de cercanía entre el contexto circundante y la definición o ejemplos de los conceptos(tripletas-conceptuales) candidatos. Entre estos algoritmos de cercanía ha tomado gran relevancia el Algoritmo de Lesk [Lesk, 1986] que es el que se ha usado en todas las pruebas. No obstante puede considerarse la experimentación con cualquier otro algoritmo de cercanía.

Se somete a un cotejo para cada uno de los conceptos candidatos (aquellos que tienen la misma forma y la misma etiqueta sintáctica extraídos de la ontología) con el contexto en donde aparece el ítem en cuestión y se escoge el concepto más cercano según el algoritmo de cercanía provisto. De este concepto finalmente seleccionado se escogen sus sinónimos o sus hipónimos con los que se formará el diccionario. *(Ver Texto 1: Algoritmo de búsqueda de un concepto - tripleta conceptual - a partir de una tripleta contextual de la página 56)*

Así, si en el ejemplo que se está tratando *calar* aparece en un contexto circundante mas propio de un ámbito psicológico, personal o social el algoritmo de desambiguación debe ser capaz de encontrar que se está en la acepción 9 y devolver sinónimos como *desenmascarar, conocer* para forma el diccionario de sinónimos.

En este párrafo se ve claro que el corpus escogido para formar los *p -ver epígrafe 3.5.2-* conceptos mas significativos debe ser de una temática u otra. Si se va a trabajar en un ámbito de relaciones psicológicas o personales y el corpus está formado por textos de dicha temática el diccionario extraído tendrá como sinónimos de *calar: desenmascarar, conocer.* Sin embargo, si se va a trabajar en una temática geológica y el corpus es de dicha temática los sinónimos de *calar* serán *horadar, permear.* Con esto se tendrá dos diccionarios ya elaborados y siempre que se quiera trabajar en el ámbito psicológico se usará el diccionario psicológico para tantos textos psicológicos como procedan y otro tanto para trabajar con el diccionario elaborado para ámbitos geológicos; sin tener que usar la ontología lingüística general con el excesivo consumo de recursos que eso conllevaría. *(Ver Ilustración 7: Diccionario de una temática versus muchos textos de dicha temática de la página 60)*

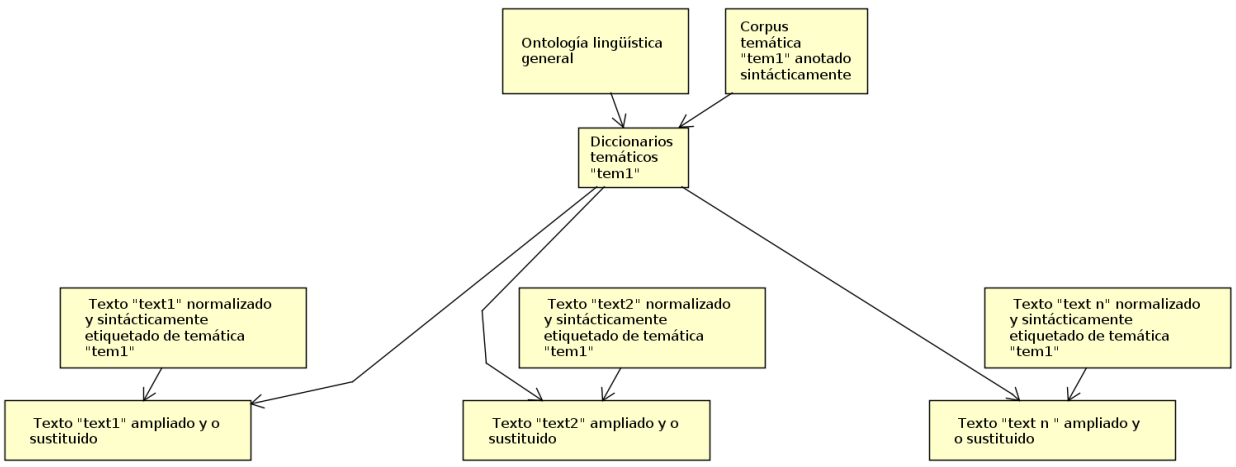


Ilustración 7: Diccionario de una temática versus muchos textos de dicha temática

Se observa como con este proceder se ha pretendido abordar, resolviéndolo razonablemente bien, dos casos de ambigüedad presentes en la lengua natural:

1. Polisemia: cuando una forma -significante- presenta dos o mas significados; es decir varios significados coinciden en un mismo lexema. Será el contexto puesto en comparativa con la ontología lingüística el que permita recoger en el diccionario unos u otros sinónimos o hipónimos. El ejemplo de *calar* en su condición de verbo ilustra lo dicho.

2. Homónimos. Varios significados convergen en un mismo significante. Este trabajo siempre se refiere a su caso particular de homógrafos. La misma forma de proceder es apta para resolver los homógrafos. Puesto que al algoritmo le dará igual que los conceptos en contienda provengan de la misma fuente etimológica(polisemia) o no(homónimos). Debe decirse que en una parte significativa de los casos las palabras homógrafas tienen etiquetas sintácticas distintas: Vela(1ª persona de presente de indicativo del verbo velar frente a vela(cirio comúnmente de parafina envolviendo una mecha) y esto junto con la lematización ya provee de un escalón de desambiguación.

En definitiva; una vez que ha quedado claro que la misma palabra -normalizada a lema- puede aparecer en varios conceptos; que dichos conceptos se representarán por la tripleta

conceptual(palabra, tag, contexto) y que la tripleta (lemaX, tagT, contextoA) es distinta de la tripleta (lemaX, tagT, contextoB) emerge la tarea de a cada tripleta asignarle una importancia. Ordenar en función de esa importancia las tripletas y escoger las *p* mejor puntuadas.

3.5.2 Fijación del número p de tripletas-conceptuales más significativas.

Una vez representado el corpus normalizado por una lista de conceptos en forma de tripletas conceptuales fijar cuántas de estas tripletas -conceptos - distintos *(p)* serán utilizadas en la elaboración de los diccionarios es la tarea inmediata. Para que *p* - y por tanto el diccionario derivado- sea pequeño se buscará representar los conceptos mas frecuentes en detrimento de los menos frecuentes. Debe entenderse que estos conceptos mas frecuentes son, además, representativos de la temática del texto a procesar. Cabe volver a señalar, llegados a este punto, cuan importante es el paso previo de eliminación de *stopwords* durante el proceso de normalización pues ha eliminado términos frecuentes -gran parte de ellos *function words*- pero poco o nada significativos desde el punto de vista semántico y de la temática en cuestión.

Para fijar este *p* se utilizará, como se dijo, un criterio frecuentista. Es decir: la elección de los *p* conceptos estará fuertemente basada en la distribución de frecuencias que estos tengan en la lista de conceptos. Por tanto se pasará de una lista de tripletas-contextuales(longitud m) a una lista dual de elementos donde cada elemento será un par formado por una tripleta-conceptual y su frecuencia de aparición en el corpus(longitud k). Esta lista dual estará ordenada por la frecuencia de manera decreciente. Así, se escogerán los *p* primeros elementos (los mas frecuentes) de la lista dual asumiendo que estos serán los más significativos. Pero, ¿cómo fijar este *p* ? Para responder a esto primero se ha de fijar un *umbral*; dicho umbral marcará -expresado en tantos por uno- cuantos elementos de la lista de tripletas-contextuales se quieren representar (por ejemplo un umbral de 0,5 significa que se quiere recuperar el 50% de los elementos de la lista de términos en forma de tripleta contextual). Debe entenderse, y además es lo que se busca, que puede haber muchos elementos en la lista de tripletas-contextuales que convergerán en un solo elemento en la lista de tripletas conceptuales (los más frecuentes en la lista de tripletas conceptuales) y estos son, precisamente, los que se quieren recuperar. Así:

Sean las listas siguientes :

- *lista de (ítems) t. contextuales* ← *[term1, term2,.... , term-m]*

  (lista de longitud m)

- lista dual ← [*(itemx, f1), (itemy, f2),.. (itemk, fk)]*

  (lista de longitud k), k <= m)

  - *f1, f2, ... fk Frecuencia de aparición del item k*

Donde: *fi>=fj si i<j*

$\forall i, j \in 1..k$ é $itemx..itemk \in conceptos(k\ conceptos\ distintos)$

y $\sum_{i=1}^{k} f_i = m$

item-k es una tripleta (lema, etiqueta-sintáctica, concepto) y *fi* la frecuencia del item i en la lista de ítems.

Entonces se fijará el *p* mas pequeño tal que: $\sum_{i=1}^{p} f_i \geq umbral \times m$

Obviamente con esto se busca que los *p* ítems – tripletas-conceptuales- mas frecuentes sean los que finalmente se recojan, como valores, en el diccionario y así, éste sea pequeño y representativo de la temática a procesar. En la medida en que p sea pequeño y, aun así, represente muchos términos se estará en el camino correcto. Al recuperar los *p* tripletas conceptuales mas frecuentes se podrá basar el diccionario a elaborar en que para cada uno de las *p* tripletas los lemas del concepto de la tripleta ,extraídos de la ontología lingüística general y sinónimos entre si, como claves del diccionario convergerán en el diccionario de sinónimos en el lema de la tripleta -como valor del diccionario -ver epígrafe *3.6.1 Relación de sinónimos. Elaboración del diccionario de sinónimos.* -. Del mismo modo todos los lemas de cada uno de los hipónimos del concepto -extraidas de la ontología lingüistica general a partir del tercer componente de la tripleta - convergerán en el diccionario de hiperónimos a su hiperónimo común. Los lemas del concepto de cada tripleta y su etiqueta sintáctica serán las claves del diccionario a sintetizar y el lema del ítem -primer componente de la tripleta- será el valor de dicho diccionario -ver epígrafe *3.6.2 Relación de cohiponimos-hiperónimos. Elaboración del diccionario de hipónimos* -.

Todo este proceso, trayendo, de nuevo, a colación el ejemplo del término en español "calar" queda recogido en la página *64* con la *Ilustración 8.* Se ve como la elección de *p* así como la etiqueta sintáctica permite desambiguar. No obstante, todo ello es posible porque la temática - y el corpus ad hoc de la misma - permite recuperar términos orográficos y rehusar los términos psicológicos.

Se examinan primero algunos casos particulares que permitirán esclarecer los conceptos:

- Sea *fi* constante para todos los items *(f)*. Es decir la frecuencia para todos los ítems es la misma. En este caso se buscará el *p* más pequeño tal que:

$$p \times f \geq umbral \times m$$

  - Si f = 1 $\rightarrow$ $p = umbral \times m$ Es decir si se quiere representar como umbral el 50% de los casos habrá que elegir el 50 % de los items(conceptos) de la lista de items
  - Si f es 2 $\rightarrow$ $p = (umbral \times m)/2$ Es decir si se quiere representar como umbral el 50% de los casos habrá que elegir el 25 % de los items(conceptos) de la lista de items si f = 2.

  Queda claro que habrá $\binom{m/f}{p}$ posibilidades distintas – para *f* constante - igualmente (poco) significativas de elegir las *p* tripletas conceptuales(conceptos). No hay, pues, -una vez fijado *p* – manera informada de elegir que *p* tripletas conceptuales son las mas significativas pues todas lo son en parecida magnitud.

Se observa que en este caso no se puede intuir un criterio claro sobre que *p* items elegir y además *p* no es pequeño.

No obstante, si se supone que la distribución de las tripletas conceptuales(ítems) sigue una determinada Ley de Zipf-Mandelbrot como ocurre en muchos ámbitos y con muchos elementos de la lingüística computacional se tiene que para pocas tripletas conceptuales*(p*) se cubren muchos términos del corpus a procesar y además el criterio para elegir que *p* tripletas usar, ahora sí, está claro: las *p* más frecuentes tal que la suma de éstas cubran un determinado umbral de términos.

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos

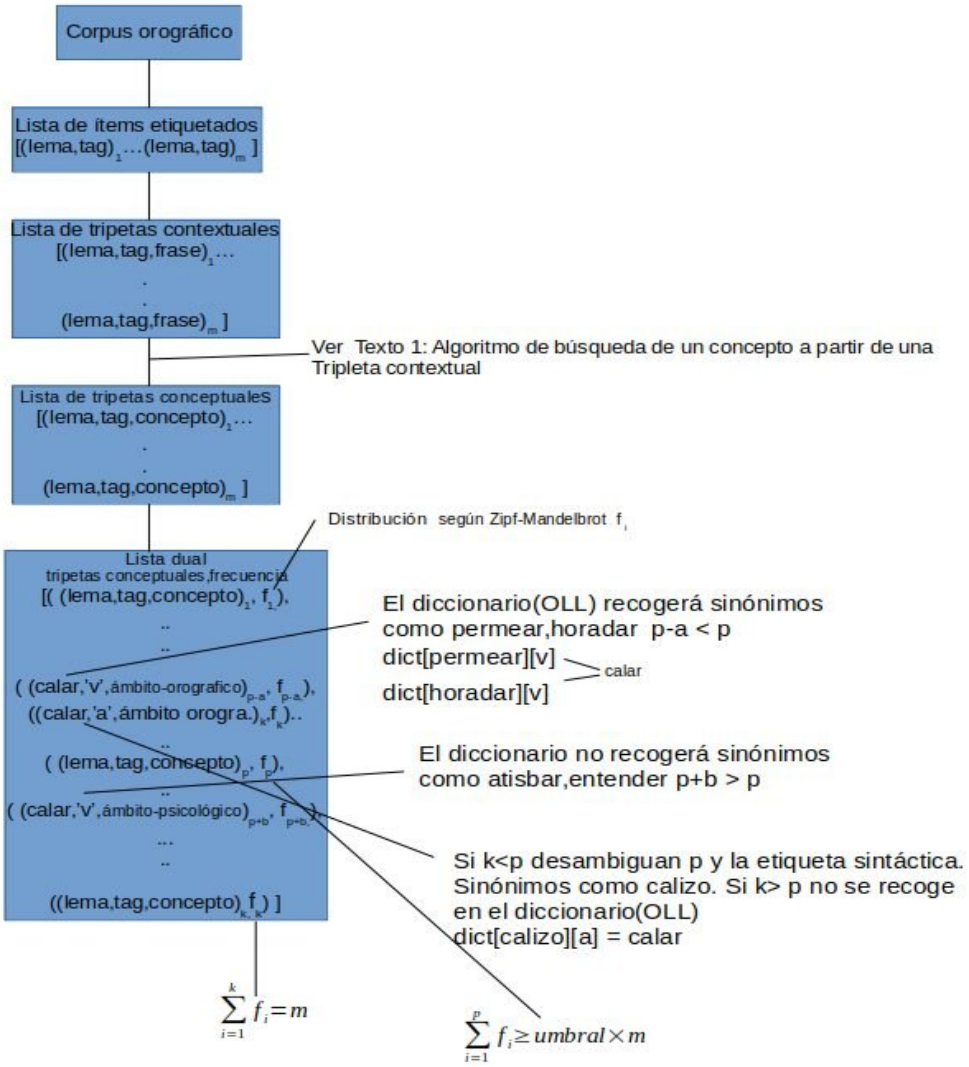


Ilustración 8: Fijación de p. Desambiguación con p y etiqueta sintáctica.Ejemplo calar

G. K. Zipf (ver 1.1 Leyes de Zipf y Mandelbrot.)comprobó empíricamente que por el principio del mínimo esfuerzo los hablantes tienden a usar un pequeño porcentaje de las palabras muy frecuentemente y, por tanto, el resto de las palabras se usan muy poco esto lo reflejó en su libro [Zipf, 1949] "Human behavior and the principle of least effort: An introduction to human ecology". Además comprobó que, ordenadas las palabras de forma decreciente según su frecuencia el producto $frecuencia \times ranking$ permanece constante o con muy poca variabilidad en torno a un valor central. A esto además ayuda el hecho de que bastantes palabras son polisémicas u homónimas con lo que aumenta su frecuencia. **El reto está ahora en ver si dado el texto o corpus temático tras ser normalizado, lematizado y representado en forma de tripletas-conceptuales dichas tripletas siguen una Ley de Zipf o están cerca de seguirla**. Teniendo ,además, en cuenta que :

- La eliminación de *stopwords* en el proceso de normalización criba muchas palabras frecuentes.
- Muchas palabras en dicho proceso de normalización convergen en sus formas canónicas o lemas lo que facilita la presencia de términos canónicos frecuentes.
- La desaparición de polisemia u homonimia puesto que ahora se tendrá que una palabra homómima o polisémica se bifurca en varias tripletas-conceptuales tras el etiquetado sintáctico y la asignación de conceptos lo que facilita que la frecuencia baje. Además, como ya se ha comentado, en un corpus temático aparecerá preeminentemente un significado frente a otros o directamente algunos significados no aparecerán.

Para comprobar que dicha suposición es asumible se decidió probar en varios corpora y para cada uno de ellos(tras normalizarlos) se experimentó de la siguiente forma:

- Se creo la lista de tripletas conceptuales y sobre ésta su distribución de frecuencias.
- Tomando como entrada el valor numérico de la tripleta mas frecuente se generó la distribución ideal siguiendo una ley de Zipf.
- Se trazan las curvas de ambas distribuciones para ver hasta que punto se ajusta la lista de tripletas conceptuales ordenada por su frecuencia a una distribución de Zipf.

- Finalmente para varios umbrales $(25\%, 50\%, 75\%, 100\%)$ se tabulan los valores de *p* y el porcentaje de términos distintos del texto normalizado que abarca. A la vista de estos datos podrá asumirse – o no - que las *p* primeras tripletas conceptuales(conceptos) serán la entrada de los diccionarios a elaborar pues cubren con un valor bajo de *p (conceptos)* un determinado *umbral* de las tripletas-contextuales(*términos)*. Se incluye el umbral 100% como elemento de control pues obviamente para cubrir todos los ítems - tripletas conceptuales - se han de abarcar todos los términos distintos aunque estos tengan solo frecuencia 1.

A) Experimentos para las siguientes obras de Chesterton:

En el siguiente gráfico *Ilustración 9: Textos de Chesterton* comparativo entre los textos de Chesterton[12] como corpus y una ley de Zipf que toma como entrada la primera frecuencia de los textos de Chesterton se observa que para un valor bajo de *p* (p conceptos) se recogen gran parte de los términos de dichos textos

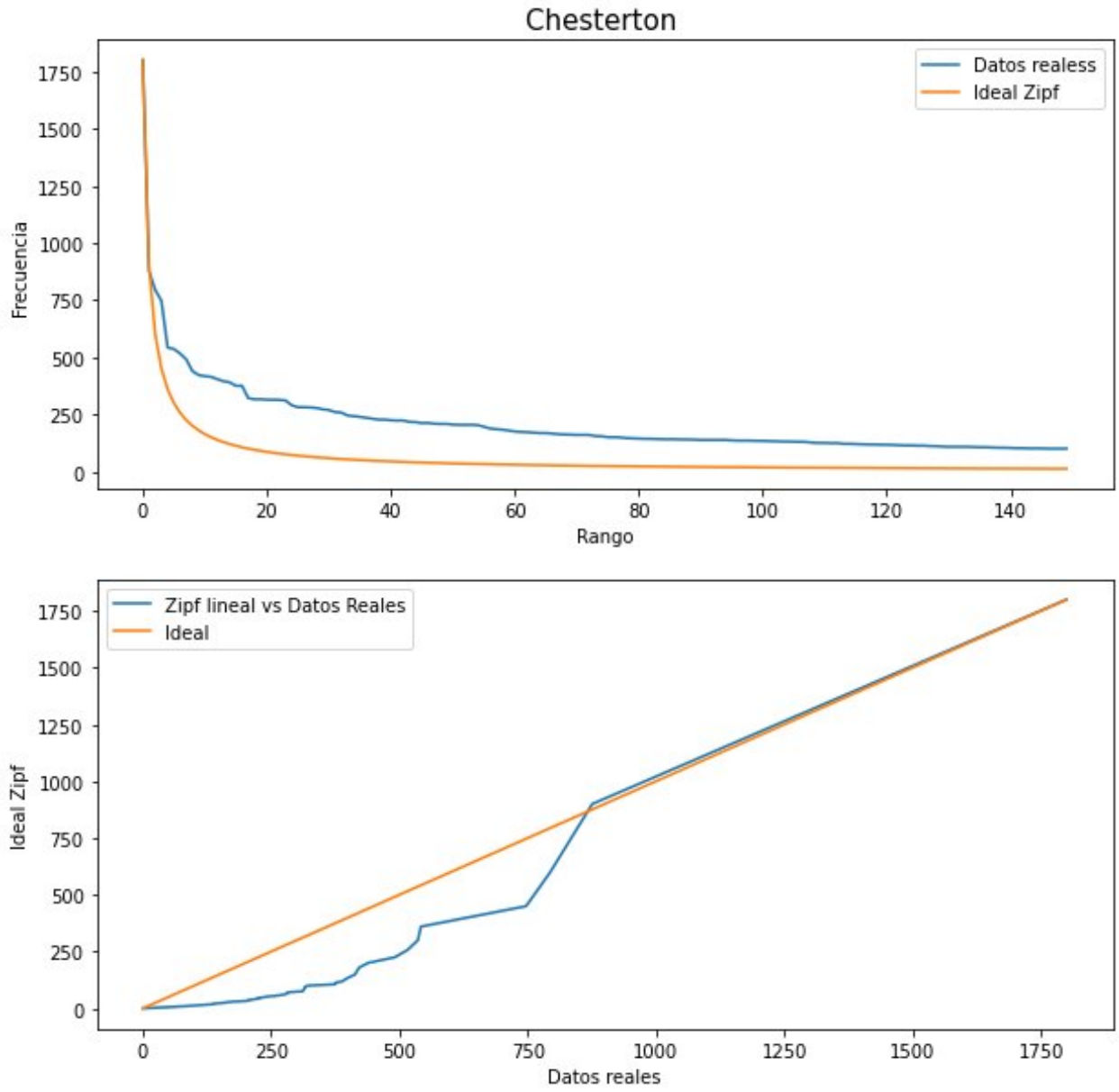


Ilustración 9: Textos de Chesterton

12 'chesterton-ball.txt', 'chesterton-brown.txt', 'chesterton-thursday.txt'] en /nltk_data/corpora/gutenberg"

Este corpus, una vez normalizado y convertido en lista de tripletas contextuales tiene m =103477 términos de los cuales 15291 son conceptos distintos, si f_i es la frecuencia de cada uno de los conceptos se tiene que $\sum_{i=1}^{15291} f_i = 103477$

Se obtienen para p y distintos umbrales los siguientes resultados:

Chesterton	Conceptos	Términos	
	Tri. Conceptuales	Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	97,00	25869,25	0,63 %
0,50	521,00	51738,50	3,41 %
0,75	2342,00	77607,75	15,32 %
1,00	15291,00	103477,00	100,00 %

Ilustración 10: Chesterton: Conceptos, p, términos, umbrales

Es decir con $p = 521$ conceptos(tripletas conceptuales) distintos quedarían cubiertos, al menos, el 50% de los términos del corpus de Chesterton, esto supone $521/15291 = 0.03407$.

$\sum_{i=1}^{521} f_i \geq 103477 \times 0{,}5 \qquad \nexists j < 521$ tal que $\sum_{i=1}^{j} f_i \geq 103477 \times 0{,}5$

La interpretación de los resultados para el resto de umbrales y el resto de corpora es similar.

Pueden compararse con los resultados que se obtendrían para una ley de Zipf ideal tomando como frecuencia inicial la frecuencia del concepto mas frecuente de la lista de conceptos de Chesterton.

Ideal Zipf Chesterton	Conceptos Tri. Conceptuales	Términos Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	7,00	25869,25	0,05 %
0,50	93,00	51738,50	0,61 %
0,75	1190,00	77607,75	7,78 %
1,00	15291,00	103477,00	100,00 %

Ilustración 11: Chesterton: Ideal Zipf

$$\sum_{i=1}^{93} f_i \geq 103477 \times 0{,}5 \quad \nexists j < 93 \quad \sum_{i=1}^{j} f_i \geq 103477 \times 0{,}5$$

B) Experimento con un corpus con obras de Shakespeare

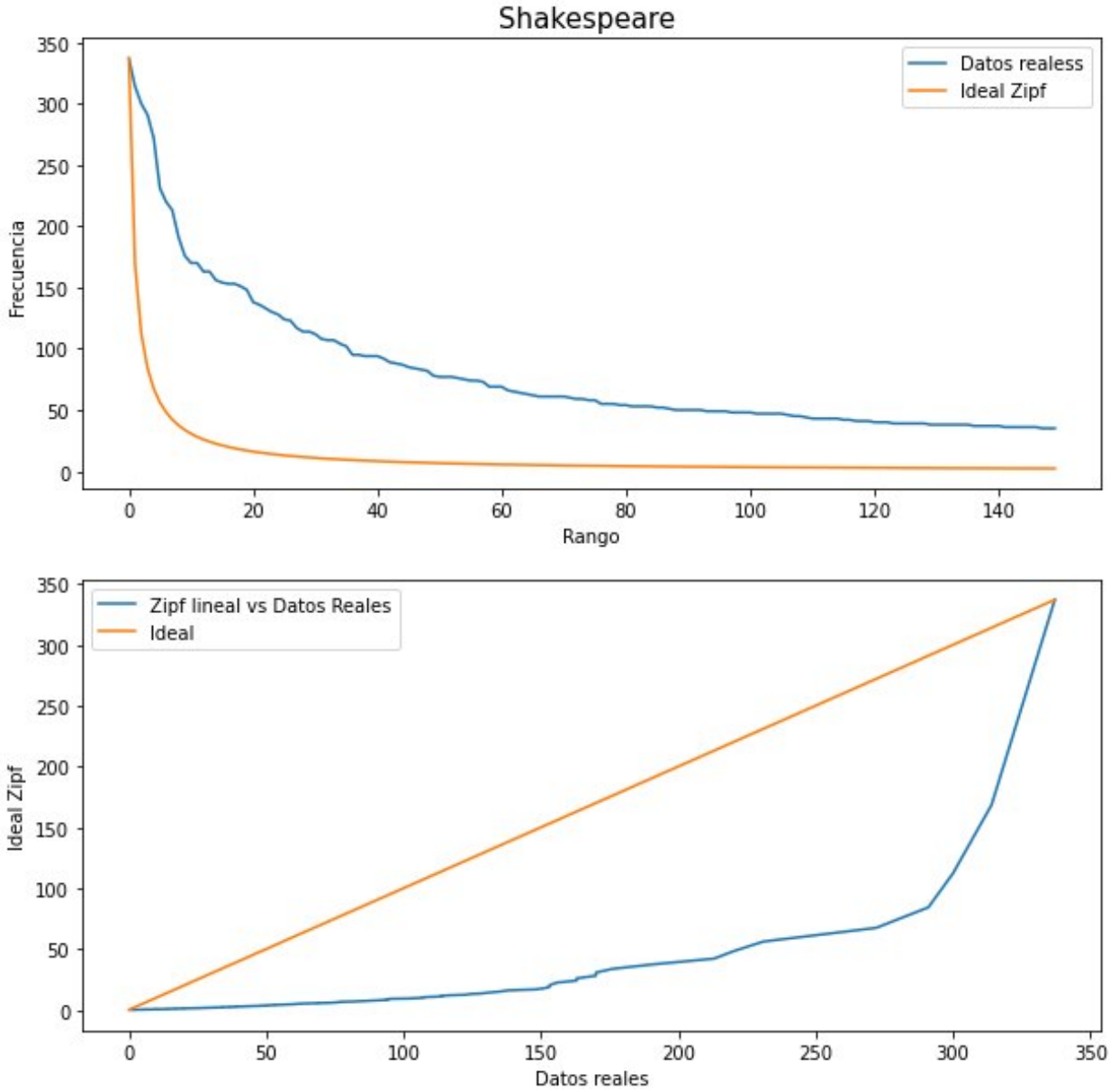


Ilustración 12: Textos de Shakespeare

En el siguiente gráfico comparativo entre los textos de Shakespeare[13] y una ley de Zipf que toma como entrada la primera frecuencia de los textos de Shakespeare se observa que para un valor bajo de *p* (conceptos) se recogen gran parte de los términos de dichos textos. Aunque se ha

13 ['shakespeare-caesar.txt', 'shakespeare-hamlet.txt', 'shakespeare-macbeth.txt']

de decir que en este caso la curva Shakespeare se aleja bastante mas de la curva ideal que en el caso Chesterton. No obstante aun se puede elegir un valor de *p* bajo para cubrir gran parte de los términos.

Este corpus, una vez normalizado y convertido en lista de ítems tiene 37133 items de los cuales 9304 son distintos: Se obtienen para *p* y distintos umbrales los siguientes resultados:

Shakespeare	Conceptos	Términos	
	Tri. Conceptuales	Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	82,00	9283,25	0,88 %
0,50	456,00	18566,50	4,90 %
0,75	2050,00	27849,75	22,03 %
1,00	9304,00	37133,00	100,00 %

Ilustración 13: Shakespeare: Conceptos, p, términos, umbrales

Pueden compararse con los resultados que se obtendrían para una ley de Zipf ideal tomando como frecuencia inicial la frecuencia del concepto mas frecuente de la lista de conceptos de Shakespeare.

Ideal Zipf	Conceptos	Términos	
Shakespeare	Tri. Conceptuales	Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	6,00	9283,25	0,06 %
0,50	72,00	18566,50	0,77 %
0,75	820,00	27849,75	8,81 %
1,00	9304,00	37133,00	100,00 %

Ilustración 14: Ideal. Shakespeare Zipf

C) Experimentos para las siguientes obras de Jane Austen:

En el siguiente gráfico comparativo entre los textos de Austen[14] y una ley de Zipf que toma como entrada la primera frecuencia de los textos de Austen se observa, de nuevo, que para un valor bajo de *p* se recogen gran parte de los items de dichos textos.

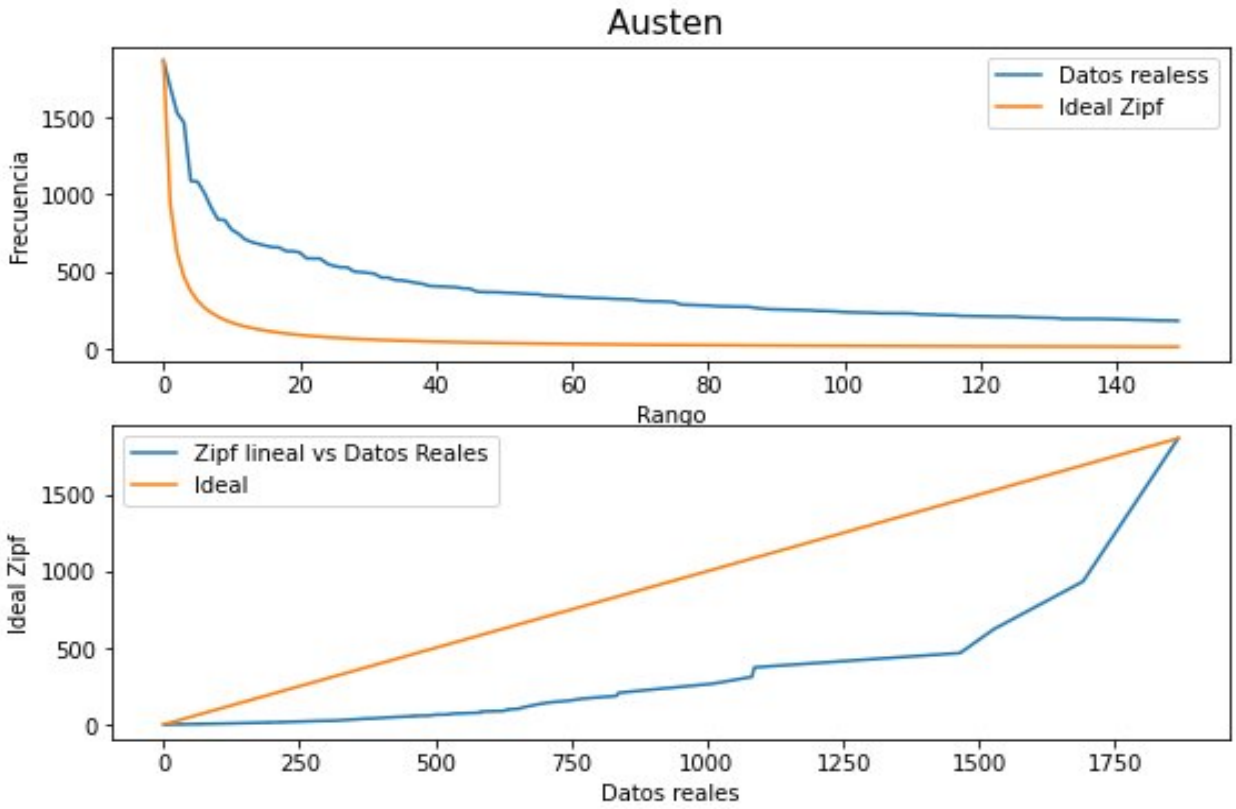


Ilustración 15: Textos de Austen

14 Éstas son las obras de Jane Austen procesadas como Corpus: ['austen-emma.txt', 'austen-persuasion.txt', 'austen-sense.txt']

Austen	Conceptos	Términos	
	Tri. Conceptuales	Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	76,00	41361,75	0,57 %
0,50	344,00	82723,50	2,57 %
0,75	1346,00	124085,25	10,04 %
1,00	13410,00	165447,00	100,00 %

Ilustración 16: Austen Conceptos, p, términos, umbrales

Pueden compararse con los resultados que se obtendrían para una ley de Zipf ideal tomando como frecuencia inicial la frecuencia del concepto mas frecuente de la lista de conceptos de Shakespeare.

Ideal Zipf Austen	Conceptos	Términos	
	Tri. Conceptuales	Tri. Contextuales	% de conceptos
umbral	p	Umbral * m	usados
0,25	6,00	41361,75	0,04 %
0,50	72,00	82723,50	0,54 %
0,75	820,00	124085,25	6,11 %
1,00	13410,00	165447,00	100,00 %

Ilustración 17: Austen: Ideal Zipf

Se concluye, según esta experimentación – contestanto a algunas de las preguntas de los epígrafes Extracción de términos 1.2.2,Paso de términos a conceptos 1.2.3 , que la distribución de frecuencias de un corpus temático representado por sus tripletas conceptuales es asimilable a una Ley de Zipf-Mandelbrot hasta el punto de que un valor bajo de *p* permite recuperar los conceptos fundamentales de dicha temática.

3.5.3 N-gramas etiquetados para capturar términos y convertirlos en conceptos.

Recogiendo alguna de las preguntas del epígrafe *1.1* muy en particular a la de *" qué se considerará término en este trabajo"* en este epígrafe se presenta la noción de *n-grama etiquetado* como *término*

Hasta ahora, según se ha visto, se ha usado un criterio estrictamente frecuentista haciendo la suposición de que las tripletas conceptuales mas frecuentes serán, automáticamente, también las mas significativas para recuperar las ideas principales de la temática que se pretende recoger en las ontologías lingüísticas ligeras (con forma de diccionario). Puesto que esto no siempre es así: -- piénsese en una expresión como *aceite de oliva virgen extra* que, con un criterio estrictamente frecuentista, podría dejar de recuperar algunos de sus componentes (virgen, extra) siendo toda la expresión en sí un término muy significativo --, se explorarán otros métodos para extraer los términos fundamentales de un corpus de determinada temática.

Se desarrolla esto considerando un corpus, y un posterior ejemplo, de temática oleícola. Es evidente que en dicha temática el 4-grama tras ser normalizado y etiquetado *((aceite,n) (oliva,n) (virgen,a) (extra,a))* es una de las ideas principales y significativas de un corpus oleícola -en definitiva, un término- y, por tanto, aparecerá frecuentemente en el texto normalizado y convertido en n-gramas etiquetados. Es decir, tal n-grama es un término fundamental de la temática. Por tanto, toda ontología lingüística ligera basada en dicha temática debería contener los sinónimos y cohipónimos de todos y cada una de las tripletas conceptuales extraídas del 4-grama tomando, además, dicho 4-grama como contexto de desambiguación. Éste es un buen momento para examinar por primera vez la *Ilustración 19 de la página 77*

'Tal n-grama será, se incide de nuevo, frecuente tras la conversión del corpus normalizado en n-gramas etiquetados y si el método de elaboración del diccionario como se apuntará en el método *Texto 2 Algoritmo de generación de tripletas conceptuales desde n-gramas etiquetados* de la página 81 contempla todos los items de los *j 4-gramas* más frecuentes el diccionario sintetizado contemplará una entrada:

dict[puro][a] = virgen.

Sin embargo, si solo se hubiese contemplado la elaboración del diccionario con los *p* items(conceptos) mas frecuentes es probable que el concepto (virgen, a, contexto-oleicola) no hubiese sido recuperado entre los *p* conceptos más frecuentes pues dicho concepto no aparecerá o, si lo hace, aparecerá con poca frecuencia fuera del contexto (aceite de oliva virgen extra) en el resto del texto. Sí es, por contra, muy probable que conceptos como (aceite, a, contexto-oleícola) si hubiesen sido recogidos entre las *p* tripletas conceptuales mas frecuentes. Probablemente conceptos como (extra, a, contexto oleícola) también deberían haber sido recuperados del n-grama y no de las tripletas con su sentido de *superior*.

dict[superior][a] = extra

Por tanto, si se usa en un texto nuevo oleícola el diccionario así sintetizado, frases como "aceite de oliva puro superior" y "aceite de oliva virgen extra" serán tomadas como frases con una gran cercanía semántica.

Debe comentarse que en este caso un término (el n-grama etiquetado frecuente) puede dar lugar a n conceptos(n tripletas conceptuales para cada una de las n duplas (lemas, tag) del que se compone el n-grama todas ella desambiguadas para formar cada uno de los n conceptos mediante todo el n-grama. En *Ilustración 18: N-grama etiquetado: 1 Término, n conceptos* de la página 76 se ilustra todo lo dicho

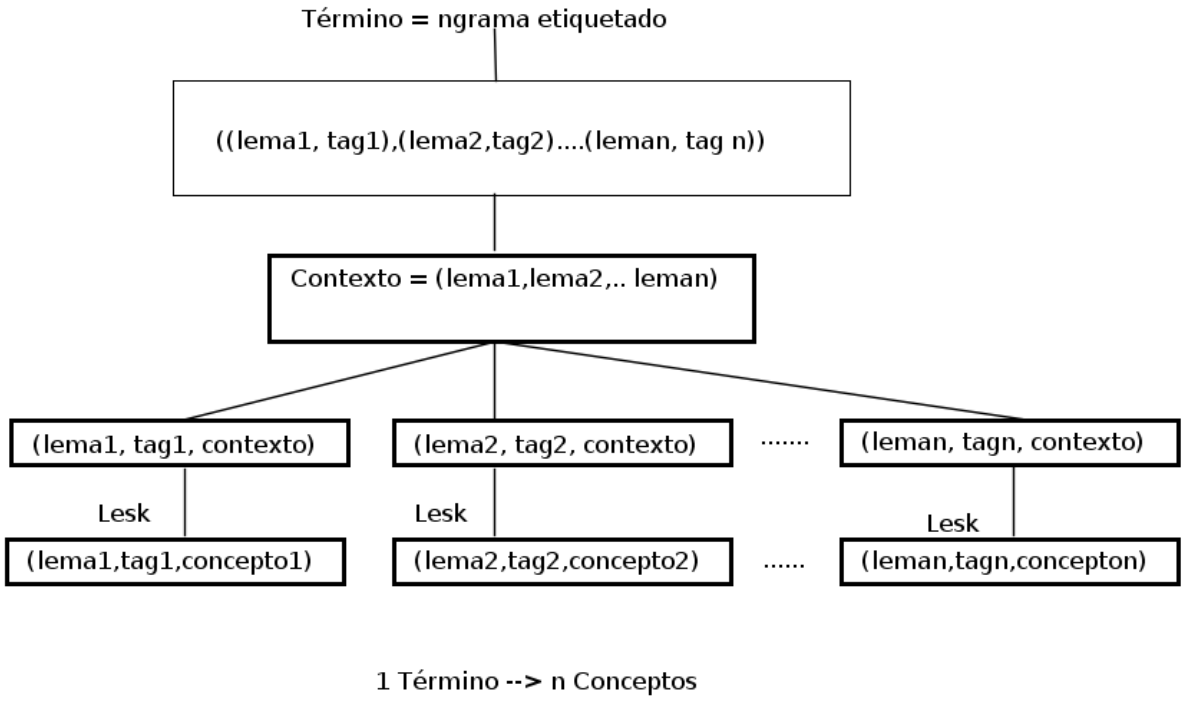


Ilustración 18: N-grama etiquetado: 1 Término, n conceptos

3.5.4 Fijación del número j de n-gramas etiquetados más significativos.

Se ha comentado – y se ha ilustrado, recuérdese el ejemplo oleícola - la necesidad de complementar - incluso sustituir- la idea de recuperar los *p* tripletas-conceptuales más frecuentes como primer paso para la elaboración de las ontologías lingüísticas ligeras con forma de diccionarios. Se introdujo, a esos efectos, la posibilidad de encontrar los *j* n-gramas etiquetados más frecuentes - cada uno de estos *j* n-grama es un término importante de la temática- de un texto o corpus con la intención de que estos aporten, con el modo que se verá, tripletas conceptuales importantes para representar la temática.

En este epígrafe partiendo de la idea de los n-gramas etiquetados mas importantes se explora la hipótesis de que los *j* n-gramas sintácticamente etiquetados distintos más frecuentes[15] -una vez normalizado el texto- deban, de una manera u otra, ser recuperados y usados para la elaboración del diccionario con el fin de que ítems -tripletas conceptuales - que no sean muy

15 Aquí si se asume que los n-gramas mas frecuentes sean los más significativos.

frecuentes pero sí muy significativos -pues se han extraído de un n-grama frecuente- sean considerados en la síntesis de las ontologías lingüística ligeras con forma de diccionario. Antes de eso, debe dejarse sentado que la recuperación de los *j n-gramas* etiquetados más importantes -por frecuentes- es, per se, un hito en este trabajo pues fijados *j* y *n* se habrán encontrado los *j* términos de *n* lemas más importantes del corpus.

```
(((('enter', 'n'), ('three', 'n'), ('witch', 'n'), ('1', 'n')), 3),
 ((('thunder', 'n'), ('enter', 'n'), ('three', 'n'),      ('witch', 'n')), 3),
 ((('trouble', 'n'), ('fire', 'n'), ('burne', 'n'), ('cauldron', 'n')), 3),
 ((('fire', 'n'), ('burne', 'n'), ('cauldron', 'n'), ('bubble', 'n')), 3),
```

Ilustración 19: Ejemplo de 4 4-gramas etiquetados mas frecuentes en los textos de Shakespeare

Si se parte de un texto que tras haber sido normalizado a una lista de *m'* ítems -donde cada ítem tiene una palabra normalizada y su etiqueta sintáctica- y se quiere formar una lista de *n*-gramas normalizados y etiquetados ésta sería de longitud:

$$m = m' - n + 1$$

Tras haber – si finalmente procede - fijado este *j d*e cada uno de sus n-gramas se extraerían las tripletas(palabra, etiqueta, concepto) con las que elaborar el diccionario o con las que complementar un diccionario parcial y previamente establecido. Para ello se usará de cada n-grama cada uno de sus pares (palabra, etiqueta) y como contexto circundante para extraer su concepto adecuado todo el *n*-grama(ver *Ilustración 18: N-grama etiquetado: 1 Término, n conceptos)*. Esto da una primera pista de que este *n* (el tamaño del n-grama) debe ser escogido con cuidado pues, como se verá, si *n* crece habrá mas contexto y la desambiguación será mas acertada pero los n-gramas serán menos frecuentes y ,quizá, con ello se alejarán de seguir una Ley de Zipf-Mandelbrot o simplemente se dificultará la extracción de los n-gramas mas significativos deviniendo en un caso extremo en que no hay n-gramas -para ese *n*- más significativos que otros.

Debe, pues, buscarse un *n* que genere el suficiente contexto para desambiguar adecuadamente y a la vez permita aislar los n-gramas mas significativos.

Se contemplan varias posibilidades que se estudiarán empíricamente usando, para ello, los mismos corpora que en el epígrafe anterior:

1. Si la distribución de frecuencias de los n-gramas etiquetados sigue una Ley de Zipf-Mandelbrot con los *j* n-gramas etiquetados más frecuentes en vez de con los *p* ítems más frecuentes(ver epígrafe anterior en la página 61) se podría elaborar un diccionario de tamaño controlado con la ventaja frente al epígrafe anterior de que lemas -per se- no muy frecuentes pero si miembros de n-gramas frecuentes y por tanto significativos serán contemplados y otros frecuentes pero poco significativos serán desechados. En este caso se sustituye totalmente - y no se amplía - el método del epígrafe anterior*(3.5.2 Fijación del número p de tripletas-conceptuales más significativas.).* Para fijar *j* se determinará éste siguiendo el método del epígrafe anterior: es decir el *j* más pequeño que haga que se cubra un determinado umbral de los n-gramas.

$$\sum_{i=1}^{j} f_i \geq umbral \times m$$

(m, en este caso, es el n.º de n-gramas etiquetados totales en el que se ha transformado el texto normalizado)

Con esta expresión debe tenerse en cuenta que el número final de tripletas conceptuales utilizadas en la síntesis del diccionario será $|tripletas| \leqslant n \times j$ pues puede haber, y de hecho así sucede, muchas tripletas que aparezcan en varios n-gramas.

Si lo que se quiere es controlar el tamaño del diccionario a sintetizar usando el nº de tripletas la inecuación a utilizar sería:

$$\sum_{i=1}^{j} f_i \geq \frac{umbral}{n} \times m$$

Si se observa la *Ilustración 20: Shakespeare con 4-gramas* se ve que para cubrir el $x\%$ de los ítems totales es necesario recolectar prácticamente los $x\%$ de los ítems distintos puesto que la frecuencia es muy parecida y cercana a 1 para todos los n-gramas etiquetados. Rige, pues, lo que se ha expresado en el apartado *3.5.2 Fijación*

del número p de tripletas-conceptuales más significativas. para ítems con frecuencia constante solo que en este caso en vez de con ítems con n-gramas etiquetados. En dicha ilustración - *Ilustración 20: Shakespeare con 4-gramas* - se expresa en las cuatro primeras líneas que prácticamente todos los n-gramas etiquetados tienen frecuencia 1. En la *Ilustración 21: Codo con j = 8. Shakespeare con 4-gramas etiquetados* se compara una distribución Zipf ideal y la distribución de los 4-gramas etiquetados en el corpora de Shakespeare; comparando la una y la otra se observa la lejanía entre ambas. Resultados similares se han obtenido para 3-gramas, 4-gramas y 5-gramas en el resto de los corpora. Luego queda descartada la alternativa de usar esta opción basada en la Ley Zipf-Mandelbrot sin perjuicio de que otras aportaciones venideras para usar, en exclusiva, los n-gramas etiquetados como elemento atómico para sintetizar el diccionario deban ser exploradas.

```
-------------------------------------------------------
Con 9214    términos distintos se cubren al menos el 25.0% de los items totales. El 24.86% de los items distintos.
Con 18497   términos distintos se cubren al menos el 50.0% de los items totales. El 49.91% de los items distintos.
Con 27779   términos distintos se cubren al menos el 75.0% de los items totales. El 74.95% de los items distintos.
Con 37061   términos distintos se cubren al menos el 100 % de los items totales. El 100.0% de los items distintos.
Con 9       términos distintos se cubren al menos el 25.0% de los items totales. El 0.02 % de los items distintos.
Con 144     términos distintos se cubren al menos el 50.0% de los items totales. El 0.39 % de los items distintos.
Con 2312    términos distintos se cubren al menos el 75.0% de los items totales. El 6.24 % de los items distintos.
Con 37061   términos distintos se cubren al menos el 100 % de los items totales. El 100.0% de los items distintos.
```

Ilustración 20: Shakespeare con 4-gramas

2. En el caso de que la distribución de los n-gramas etiquetados no se acerque a una Ley de Zipf (como sucede en todos los corpora explorados) siempre pueden escogerse los *j* n-gramas más frecuentes y extraer para cada uno de esos *n-gramas* sus *n* tripletas conceptuales. Sin embargo, si la variabilidad de las frecuencias de los n-gramas etiquetados es pequeña este método deviene en, prácticamente, escoger una combinacion de *j* n-gramas al azar de las $\binom{m}{j}$ posibles combinaciones donde *m* es el número de n-gramas etiquetados distintos; con lo que se deja en manos del azar que los n-gramas escogidos sean significativos, por tanto el método adolece de un criterio informado para escoger *j*. Sin embargo, las pruebas realizadas en los diversos corpora permiten comprobar visualmente que la gráfica rank-frecuencia de estos n-gramas etiquetados presentan un codo muy acusado (este codo marca los n-gramas significativos – por

frecuentes -- de los que no lo son ver *Ilustración 21: Codo con j = 8. Shakespeare con 4-gramas etiquetados)* y, por tanto, dónde se presenta este codo es un buen candidato como *j*. Por tanto,

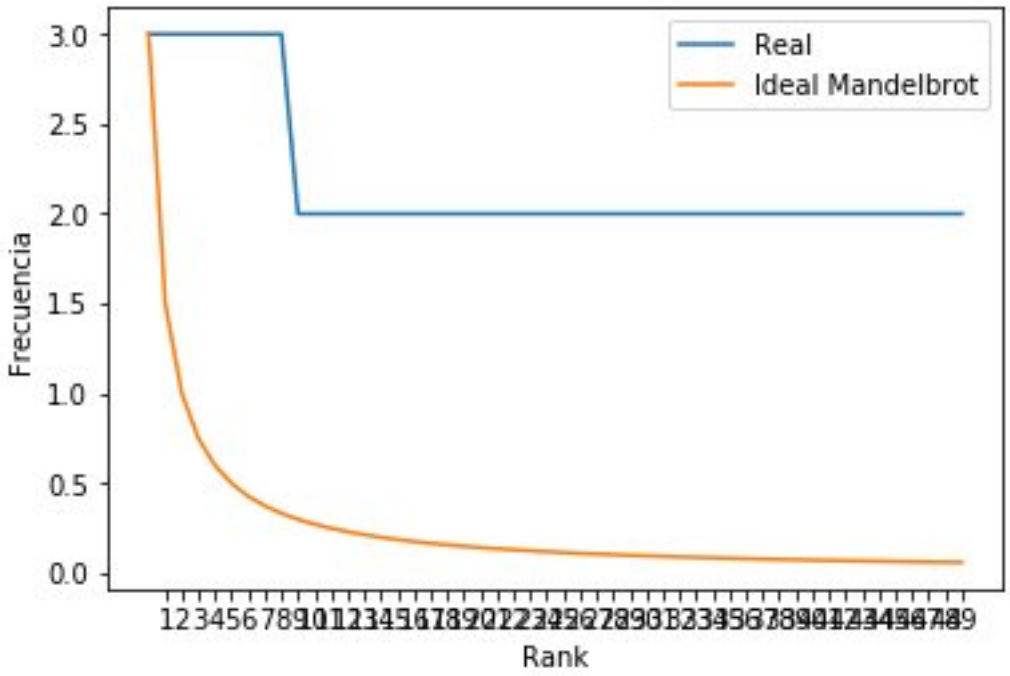


Ilustración 21: Codo con j = 8. Shakespeare con 4-gramas etiquetados

parece necesario abordar la elaboración final del diccionario conservando el método del epígrafe anterior *(3.5.2 Fijación del número p de tripletas-conceptuales más significativas.)* y combinándolo con lo expuesto en el epígrafe presente. Para ello se propone usar la unión de las tripletas conceptuales extraídas con el método *3.5.2 Fijación del número p de tripletas-conceptuales más significativas.* con las tripletas extraídas de *j* n-gramas (asignando a j el valor que hace codo en la gráfica ver *Ilustración 23: Shakespeare. Frecuencia de n-gramas etiquetados versus Rango. Codo j)* haciendo así que el diccionario sintetizado contenga algunos *conceptos* que a pesar de no ser especialmente frecuentes si son significativos pues forman parte de n-gramas frecuentes. Así:

Sea:

T1 Conjunto de p tripletas conceptuales extraídas por el método del epígrafe *3.5.2 Fijación del número p de tripletas-conceptuales más significativas.* fijado un determinado umbral. Donde $|T1| = p$

T2 Conjunto de tripletas extraídas de los j n-gramas mas frecuentes. Donde

$$|T2| \leqslant n \times j$$

Devolver $T1 \cup T2$ Donde $|T1 \cup T2| \leq |T2| + p$

Si $|T2 \setminus T1|$ es grande (cercana a $|T2|$) el método habrá aportado muchas tripletas significativas pero poco frecuentes extraídas de los n-gramas

Si $|T1 \cup T2|$ es cercana a p el método no habrá aportado apenas tripletas significativas pero poco frecuentes.

Independientemente del método usado y del número de n-gramas extraídos*(j)*. Una vez extraídos estos habrá de usarse el algoritmo del *Texto 2: Algoritmo de generación de tripletas conceptuales desde n-gramas etiquetados* para extraer las tripletas conceptuales de los j n-gramas

```
Para cada n-grama de (j) N-gramas:
    contexto = n-grama
    Para cada (lema, etiqueta) de n-grama
        concepto ←desambiguar(lema,etiqueta,contexto)
        solucion ←solucion union (lema,etiqueta, concepto)

devolver solucion
```

Texto 2: Algoritmo de generación de tripletas conceptuales desde n-gramas etiquetados

A la vista del algoritmo puede darse un tope máximo del número de tripletas conceptuales que -en el mejor de los casos, todas las tripletas visitadas son distintas-, se extraerán:

$$|tripletas| \leqslant n \times j$$

El hecho de que cada tripleta tenga como su tercera dimensión un concepto ayuda a que el resultado final no esté demasiado lejano de su tope $(n\,x\,j)$ pues tripletas con su palabra y etiqueta sintáctica iguales pueden, debido a su contexto(el n-grama en este caso), pertenecer a conceptos diferentes.

3.5.5 Valores de umbral, número de tripletas distintos(p), longitud del n-grama(n) y número de n-gramas(j). Relaciones entre ellos.

Llegados hasta aquí mediante la exposición razonada y experimentación en diversos corpora se ha establecido que la mejor manera conocida de hacer un diccionario útil, de tamaño controlado y representativo es mediante la concurrencia y posterior interacción entre la extracción, dado un ***umbral***, de los ***p*** tripletas conceptuales más frecuentes y la extracción de los ***j*** ***n***-gramas más significativos, extrayendo de estos, sus tripletas conceptuales. Qué tal método, basado en semejante concurrencia, sea, finalmente, útil dependerá de que la elección de los valores de dichos parámetros y la relación entre los mismos sea la adecuada.

Para dar unas pautas en semejante empeño tomando los mismos corpora en los que se ha basado toda la experimentación anterior se decidió ampliar la misma haciendo entrar en juego la extracción de los *j* *n*-gramas mas significativos y viendo si la presencia de estos n-gramas(términos) y la posterior extracción de sus tripletas conceptuales generaba mejores diccionarios. Esto puede medirse viendo la cardinalidad del conjunto $T2 \setminus T1$ y cuan significativas son las tripletas finales de dicho conjunto.

Así pues, la experimentación, dados los corpora de Shakespeare, Chesterton y Austen consistirá en:

- Jugar con los valores *n* del n-grama con 3,4,5. Son valores razonables pues valores como 1 y 2 estarían muy cercanos conceptualmente a la simple extracción de ítems y n-gramas significativos de valores mayores que 5 no se dan, menos aun en la lengua inglesa que tiende a frases cortas y teniendo en cuenta, además, que el proceso de normalización ha cribado muchas de las *function words*.
- Para fijar los valores de *j* se grafica la frecuencia de aparición de los n-gramas etiquetados para cada valor de n frente a su rango y visualmente se extrae el valor de *j*

que hace codo en la gráfica. Desde luego cualquier otro método sobrevenido que pueda encontrarse para, dado *n*, fijar *j* deberá ser considerado como ampliación de este trabajo.

- Una vez fijado umbral y dado que $p=f(umbral,m)$ esto dará - como se expuso en las consideraciones finales del epígrafe *3.5.2* -, el conjunto $T1$ de las p tripletas conceptuales más frecuentes. De igual manera mediante la elección de un n para los n-gramas y un j se extraerán $|T2|\leq n\times j$ tripletas. Ver según los valores del par $\langle umbral,p\rangle$ y del par $\langle n,j\rangle$ cuantas tripletas tiene el conjunto[16] $T2/T1$ y cuan significativas son estas tripletas no presentes en $T1$ y sí en $T2$ así como cual es el tamaño final del diccionario dará idea fidedigna de cuales deben ser los valores de *umbral, p, n* y *j*. Obviamente se incluye el umbral = 100% con efectos de control pues en este caso el método de los n-gramas etiquetados no puede aportar niguna tripleta que no haya ya capturado el método de las tripletas pues éste, para dicho umbral, contará con todos las tripletas distintos del texto.

16 / Diferencia de conjuntos

A) Corpora Shakespeare

Sea la gráfica siguiente utilizada en la fijación de j según n = 3,4,5:

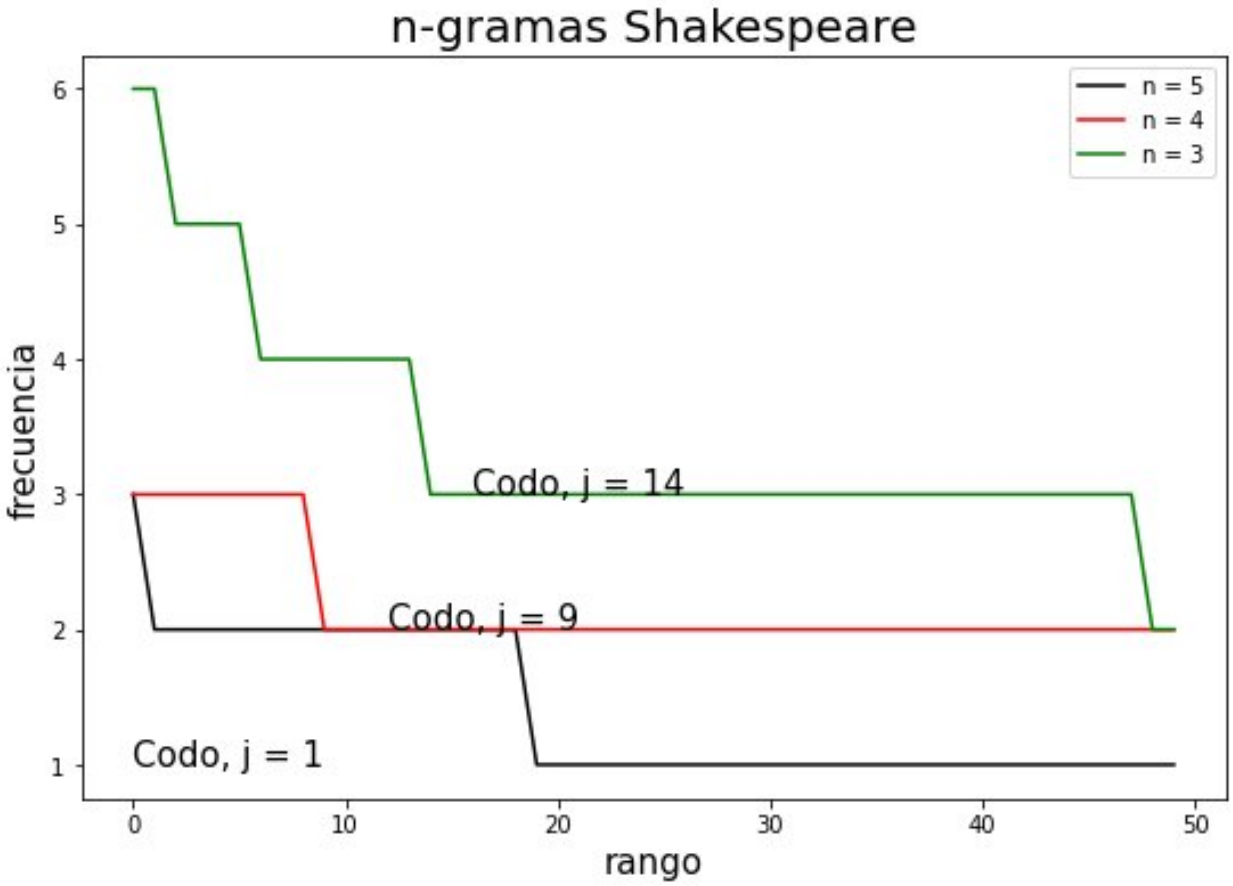


Ilustración 22: Shakespeare. Frecuencia de n-gramas etiquetados versus Rango. Codo j

SHAKESPEARE		n-gramas = 3	j= 14	n-gramas=4	j=9	ngramas= 5	j=1
25%	82/9304	|T2| = 20 ; |T2-T1|=11		|T2| =28 ; |T2-T1|=22		|T2| = 5 ; |T2-T1|=5	
50%	456/9304	|T2| = 20 ; |T2-T1|=2		|T2| = 28 ; |T2-T1|=12		|T2| = 5 ; |T2-T1|=3	
75%	2050/9304	|T2| =20 ; |T2-T1|=0		|T2| = 28 ; |T2-T1|=0		|T2| = 5 ; |T2-T1|=0	
100%	9304/9304	|T2| =20 ; |T2-T1|=0		|T2| = 28 ; |T2-T1|=0		|T2| = 5 ; |T2-T1|=0	

Tabla 4: Shakespeare: Umbral, j(codo), n(n-grama)

|T2| Número de tripletas extraídas desde los n-gramas etiquetados

|T2-T1| Número de tripletas aportadas por los n-gramas(aquellas que no estaban ya contempladas en T1).

B) Corpora de Austen.

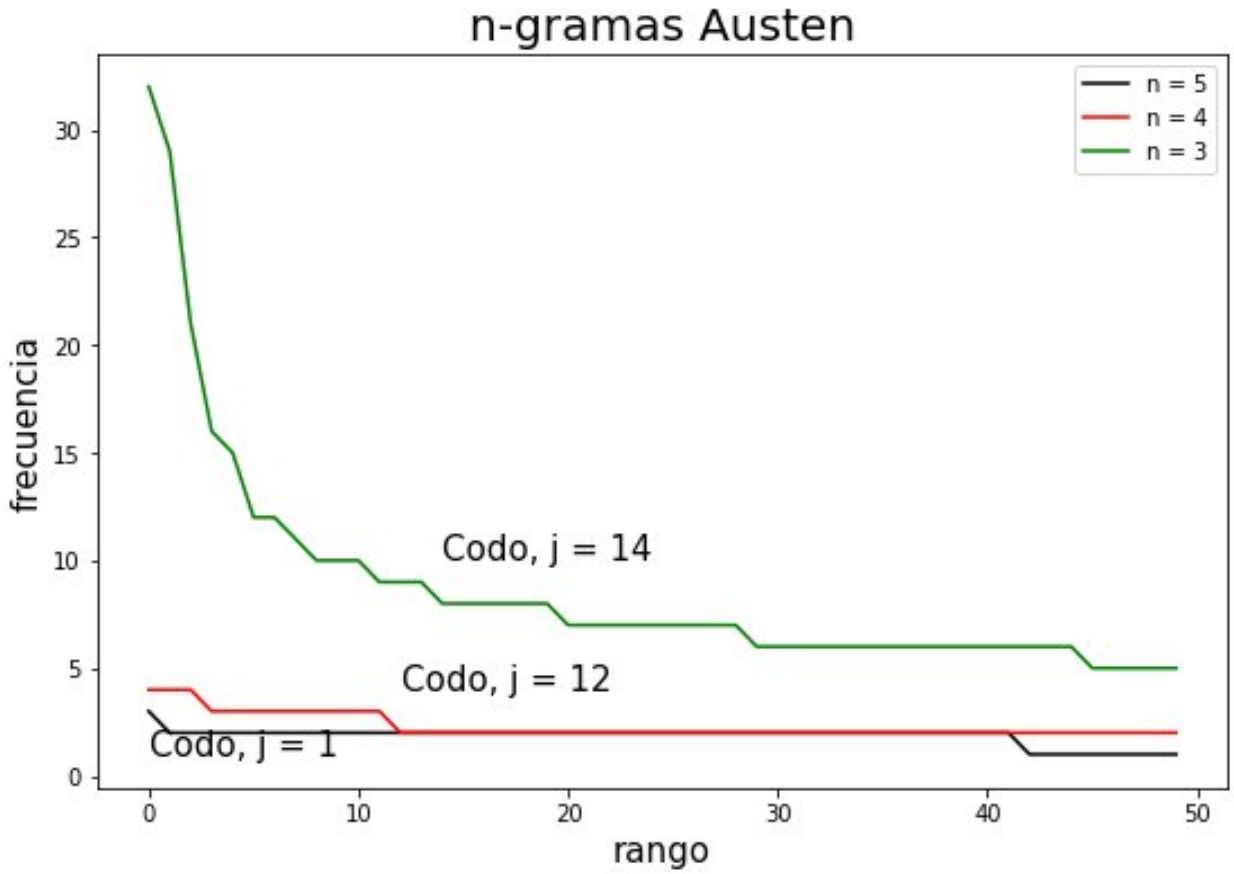


Ilustración 24: Austen. Frecuencia de n-gramas etiquetados versus Rango. Codo j

AUSTEN		n-gramas = 3	j= 14	n-gramas= 4	j=12	n-gramas= 5	j= 1
25%	76/13410	|T2| = 28 |T2-T1|=14		|T2| =27 |T2-T1|=16		|T2| = 4 ; |T2-T1|=4	
50%	344/13410	|T2| = 28 |T2-T1|=2		|T2| =27 |T2-T1|=6		|T2| = 4 ; |T2-T1|=3	
75%	1346/13410	|T2| = 28 |T2-T1|=0		|T2| =27 |T2-T1|=2		|T2| = 4 ; |T2-T1|=2	
100%	13410/13410	|T2| = 28 |T2-T1|=0		|T2| =27 |T2-T1|=0		|T2| = 4 ; |T2-T1|=0	

Tabla 5: Austen: Umbral, j(codo), n(n-grama)

C) Corpora de Chesterton

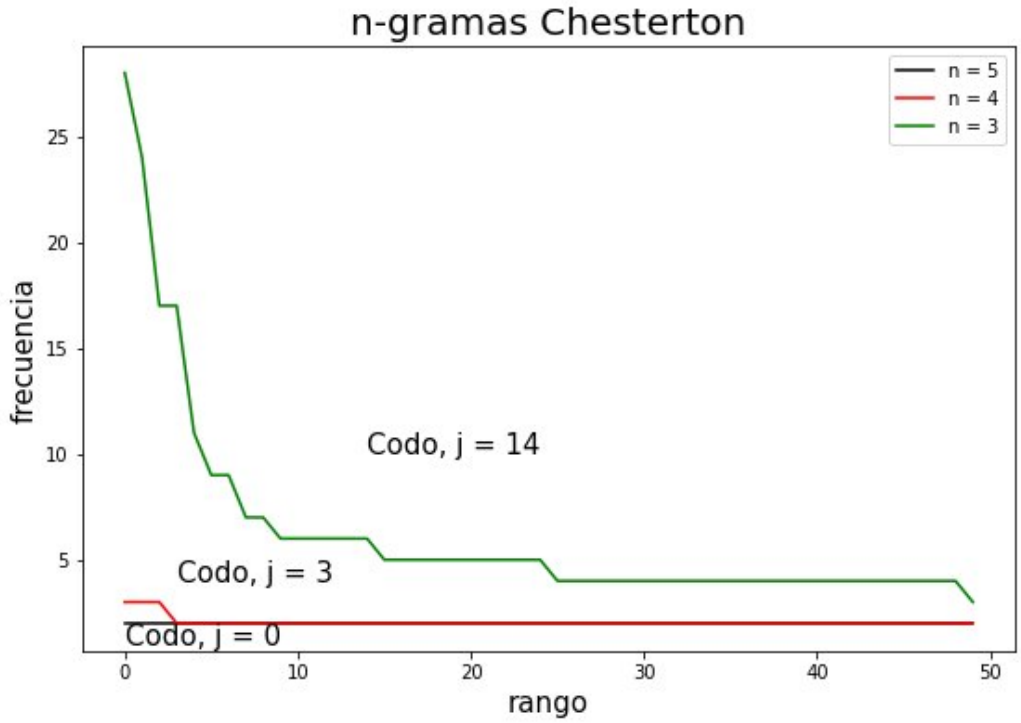


Ilustración 25: Chesterton. Frecuencia de n-gramas etiquetados versus Rango. Codo j

CHESTERTON		n-gramas = 3	j= 14	n-gramas= 4	j=3	n-gramas= 5	j= 0
25%	97/15291	|T2| = 30 ; |T2-T1| =14		|T2| =12 ; |T2-T1|=7		|T2| = 0 ; |T2-T1|=0	
50%	521/15291	|T2| = 30 ; |T2-T1| =5		|T2| = 12 ; |T2-T1| =4		|T2| = 0 ; |T2-T1|=0	
75%	2342/15291	|T2| =30 ; |T2-T1| =0		|T2| = 12 ; |T2-T1| =1		|T2| = 0 ; |T2-T1|=0	
100%	15291/15291	|T2| =30 ; |T2-T1| =0		|T2| = 12 ; |T2-T1| =0		|T2| = 0 ; |T2-T1|=0	

Tabla 6: Chesterton: Umbral, j(codo), n(n-grama)

A la vista de la experimentación con los Corpora de Shakespeare , Austen y Chesterton se concluye que:

- Para n = 5 resulta complejo fijar *j* y no hay 5-gramas que se muestren significativos pues no hay 5-gramas etiquetados que sean especialmente frecuentes frente a los demás por lo tanto $n=5$ pierde peso como valor útil en un método general. No obstante si aparece algún corpora donde existe un 5-grama etiquetado especialmente frecuente siempre se pueden incluir las tripletas extraidas de dicho 5-grama.
- Para umbral >= 75% el método de los n-gramas etiquetados no aporta niguna tripleta nueva al diccionario. Esto permitiría no considerar para dicho valor de umbral el método de los n-gramas etiquetados; aun así, como envés de la moneda, esto significará que se han contemplado en el diccionario ítems que aun siendo frecuentes no son significativos.
- Para valores de un umbral = 50 % y n = 4 se recuperan tripletas significativas que no había recuperado el método de los ítems mas frecuentes. Por tanto la lista de tripletas final debe ser la unión de las tripletas aportadas por el método estrictamente frecuentista y por las tripletas aportadas por el método de los 4-gramas.

Por ejemplo en el caso de Shakespeare, estos son las tripletas aportadas, donde se encuentran palabras que parecen habituales en la obra de Shakespeare como *witch* o *beware y* que no habían sido aportadas mediante los ítems

- Tripletas aportadas : {('burne', 'n', 'No hay synset'), **('witch', 'n', Synset('witch.n.02'))**, ('primus', 'n', Synset('primus_stove.n.01')), ('reynol', 'n', 'No hay synset'), ('secunda', 'n', 'No hay synset'), ('ides', 'n', Synset('ides.n.01')), ('prima', 'n', Synset('prima.n.01')), ('actus', 'n', 'No hay synset'), ('scoena', 'n', 'No hay synset'), ('cauldron', 'n', Synset('caldron.n.01')), ('bubble', 'n', Synset('house_of_cards.n.01')), **('beware', 'v', Synset('beware.v.01'))**}

En el caso de Austen se tienen tripletas recuperadas que son tan significativos como :

Tripletas aportadas : {('lay', 'v', Synset('lay.v.04')), ('reign', 'v', Synset('reign.v.01')), ('alone', 'r', Synset('entirely.r.02')), ('sir', 'v', 'No hay synset'), ('lovely', 'r', 'No hay synset'), ('bed', 'n', Synset('seam.n.03'))}

3.6 Relaciones entre conceptos(tripletas conceptuales) y una Ontología Lingüística General. Elaboración de las ontologías lingüísticas ligeras temáticas con forma de diccionario.

Ahora se parte de una lista de tripletas conceptuales(lema, etiqueta-sintáctica, concepto) sintetizada, en la propuesta del ponente, mediante los métodos presentados en el epígrafe *3.5* cuyo proceso puede verse graficado en la *Ilustración 26* . Asimismo, se cuenta con una Ontología Linguística General de la que extraer información. Con el concurso de uno y otro recurso se extraerán las ontologías lingüísticas ligeras donde, en principio, cada una representará una relación que, en su estado final, tendrá forma de diccionario. Estas relaciones se darán bien entre los componentes de las tripletas conceptuales y los componentes de la Ontología Lingüística General apuntados por el tercer elemento de dicha tripleta conceptual en concurso o bien entre las propias tripletas conceptuales como en el caso de las RA. Del tercer término *concepto* de la tripleta conceptual se pueden recuperar todos sus lemas e igualmente se pueden recuperar todos sus conceptos hipónimos -cohipónimos-, en ambos casos desde una ontología lingüística general. Debe reforzarse, antes de proseguir, la idea de que, en ambos casos, los valores del diccionario serán los lemas más representativas del texto que actuarán como cánones hacía los que colapsarán las claves de dicho diccionario. Precisamente la presencia como tercer elemento de la tripleta conceptual de un determinado concepto permitirá desambiguar la polisemia. Pues no se recuperán los mismos sinónimos ni los mismos conceptos hipónimos para la tripleta *(hoz, n, concepto-orográfico)* que para *(hoz, n, concepto-apero-agrícola)*:

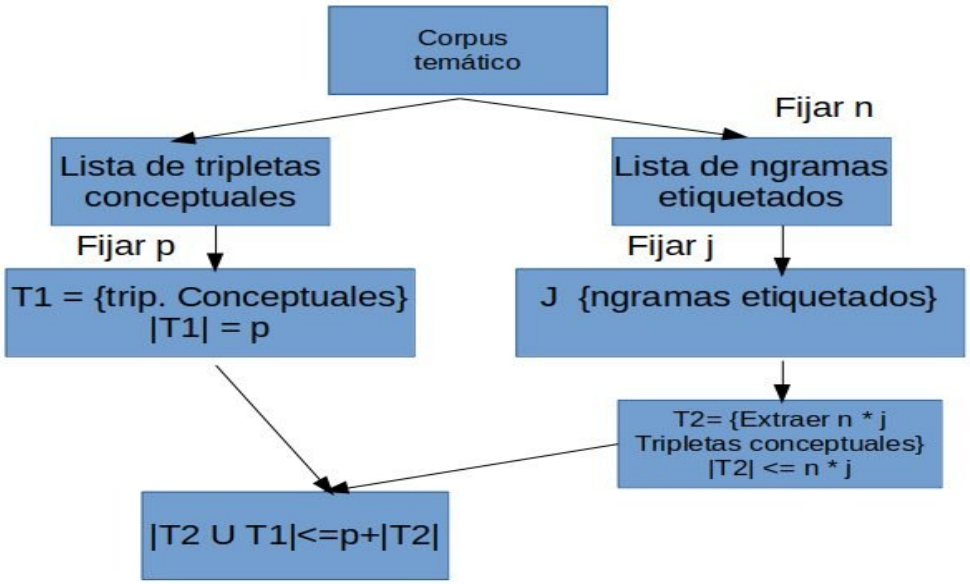


Ilustración 26: Esquema de la elaboración de la lista de tripletas conceptuales mas significativas

3.6.1 Relación de sinónimos. Elaboración del diccionario de sinónimos.

En este caso cada valor del diccionario será un representante común de todos sus sinónimos y todas las claves (palabra, etiqueta_sintáctica) convergerán en dicho representante canónico. La elaboración del diccionario se muestra en forma de algoritmo en *Texto 3: Algoritmo de elaboración del diccionario de sinónimos* y se representa en la *Ilustración 27: Elaboración del diccionario de sinónimos*. Se garantiza por la propia construcción del algoritmo que al final del mismo:

$$dict.claves \cap dict.valores = \emptyset$$

```
dict = {} , OL = OntologíaLingüística general  escogida
Para cada lema, etiqueta, concepto  de listaDeTripletas  Conceptuales
        sinonimos = extraerLemas(concepto,OL)
        dict1 = {}
        Para cada sin de sinonimos:
                if sin not in dict.claves
                        dict1[sin][etiqueta] = lema
        dict.añadir(dict1)
devolver dict
```

Texto 3: Algoritmo de elaboración del diccionario de sinónimos

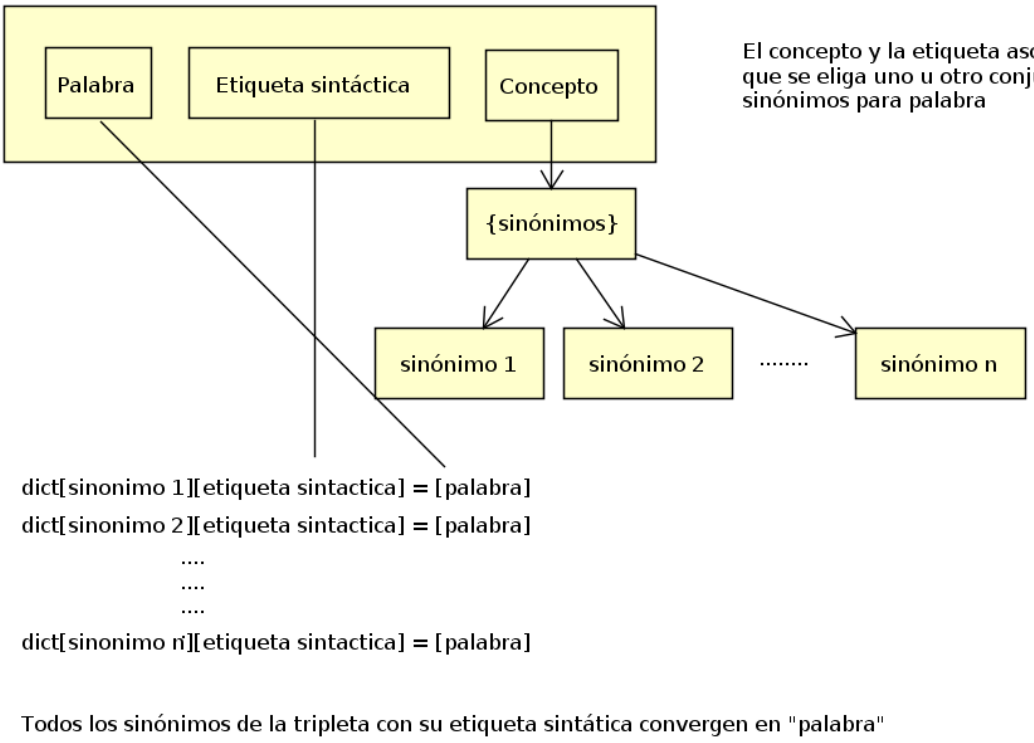


Ilustración 27: Elaboración del diccionario de sinónimos

Puede darse una fórmula de las dimensiones que tendrá el diccionario de sinónimos sintetizado:Elaboración del diccionario de hiperónimos-hipónimos.

Sean:

- nT el número de elementos de la lista de tripletas conceptuales con concepto asignado.

  (Se excluyen tripletas (lema, tag, 'NoHaySynset')

- nlC el número medio de lemas por concepto. Es decir cuantas palabras tiene, en promedio, cada concepto reflejado en la ontología lingüística.
- nValores el número de elementos distintos que tendrá el diccionario = nT. El número de sinónimos hacia los que convergerán los textos. Valores muy frecuentes en el corpus de entrenamiento. $len(set(dictSinonimos.values()))$ en un lenguaje funcional tipo Python
- nEDict Número de entradas del diccionario. $len(dictSinonimos.keys())$ en Python

$nEDict = nT \times nlC$ si

$\forall i, j \in conceptos\ de\ la\ lista\ de\ tripletas\ lemas(i) \cap lemas(j) = \emptyset$

en otro caso $nEDict \leq nT \times nlC$

Ilustración 28: Dimensiones del diccionario de sinónimos

3.6.2 Relación de cohiponimos-hiperónimos. Elaboración del diccionario de hipónimos

Se parte, igualmente, de la lista de tripletas conceptuales. De cada concepto de cada tripleta si viene acompañado por una etiqueta sintáctica verbo o sustantivo se extraerán de la ontología lingüística general sus conceptos hipónimos --recuérdese que se dijo que era importante elegir una ontología que permitiese fácilmente extraer tales relaciones semánticas--. De cada uno de estos hipónimos se extraen sus lemas y la unión de estos lemas junto con sus etiquetas sintácticas serán las claves del diccionario. La palabra de la tripleta será el valor de dicho diccionario; el hiperónimo común de todos sus cohipónimos. Es decir, todos los cohipónimos convergerán en su hiperónimo. En el algoritmo *Texto 4: Algoritmo de elaboración del diccionario de cohipónimos* y en la Ilustración *30 Elaboración del diccionario de cohipónimos* se desarrollan estas ideas.

```
Dict = {}, OL = OntologíaLingüística General escogida
Para cada lema, etiqueta, concepto  de listaDeTripletas
        si etiqueta en {verbo, sustantivo}
                conceptosHipònimos = extraerHipónimos(concepto, OL)
                dict1 = {}
                Para cada cp de conceptosHiponimos:
                        lemas =  extraerLemas(cp, OL)
                                Para cada l de lemas
                                        if l not in dict.claves
                                                dict1[l][etiqueta] = lema
                dict.añadir(dict1)
devolver dict
```

Texto 4: Algoritmo de elaboración del diccionario de cohipónimos

Puede darse una fórmula estimativa del n.º de entradas que tendrá el diccionario de hipónimos sintetizado:

Sean:

- nT el número de elementos de la lista de tripletas. (Excluidas (lema, tag, noHaySynset))
- nch el número medio de conceptos hipónimos por concepto. Es decir cuantos conceptos hipónimos tiene, en promedio, cada concepto reflejado en la ontología lingüística.
- nlC el número medio de lemas por concepto. Es decir cuantas palabras tiene, en promedio, cada concepto reflejado en la ontología lingüística.
- nValores el número de elementos distintos que tendrá el diccionario = nT. El número de hiperónimos hacía los que convergerán los textos. Valores muy frecuentes en el corpus de entrenamiento. $len(set(dictHiperonimos.values()))$ en Python
- nEDict Número de entradas del diccionario.

Si

Si se asume que las etiquetas sintácticas de la lista de tripletas siguen la misma distribución que las etiquetas sintácticas en el idioma de trabajo y en dicha idioma el porcentaje de sustantivos y verbos es *psv* expresados en tantos por uno entonces se puede afinar la expresión $nEDict = psv \times nT \times nlC \times nch$

Ilustración 29: Dimensiones del diccionario de cohipónimos hiperónimos

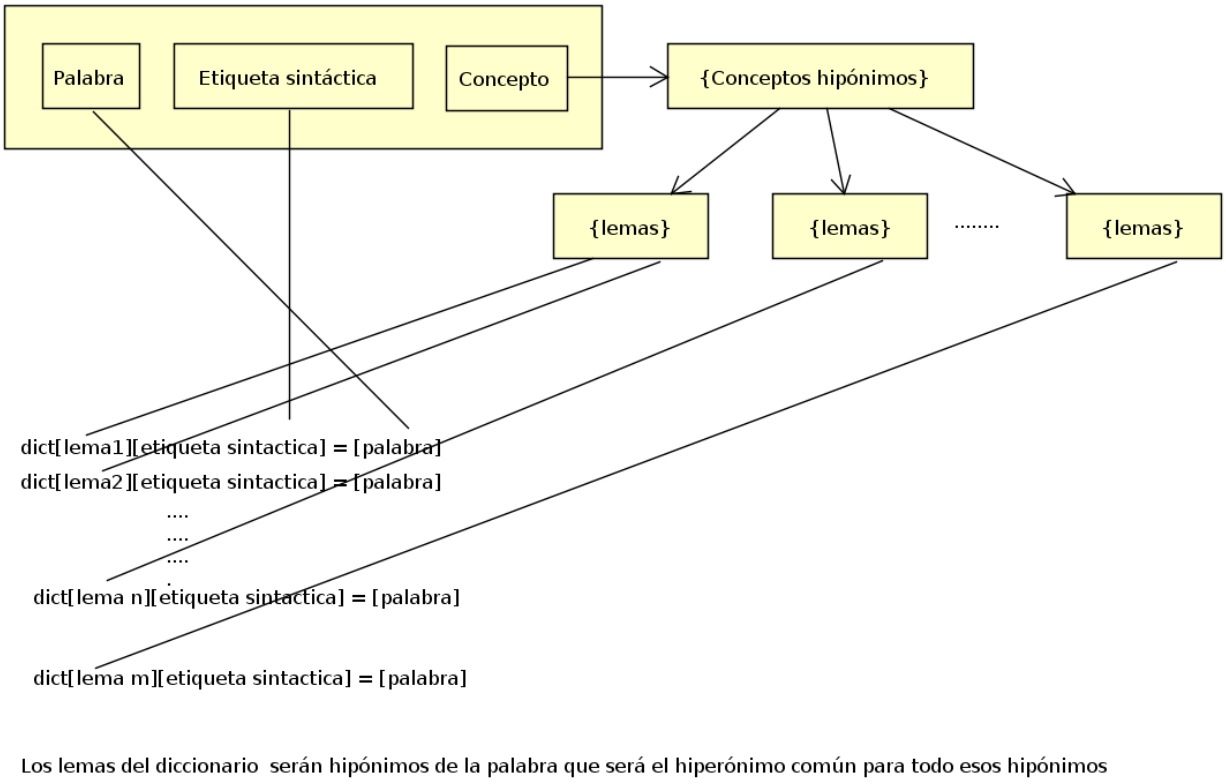


Ilustración 30: Elaboración del diccionario de cohipónimos

En la figura *De lo abstracto a lo concreto:términos, conceptos, relaciones, etc Ilustracion 31* y antes de considerar las RA entre conceptos, que merecen ellas solas un capítulo aparte, se señalan como las nociones generales planteadas en el capítulo 2 han sido resueltas en el capítulo 3.

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos

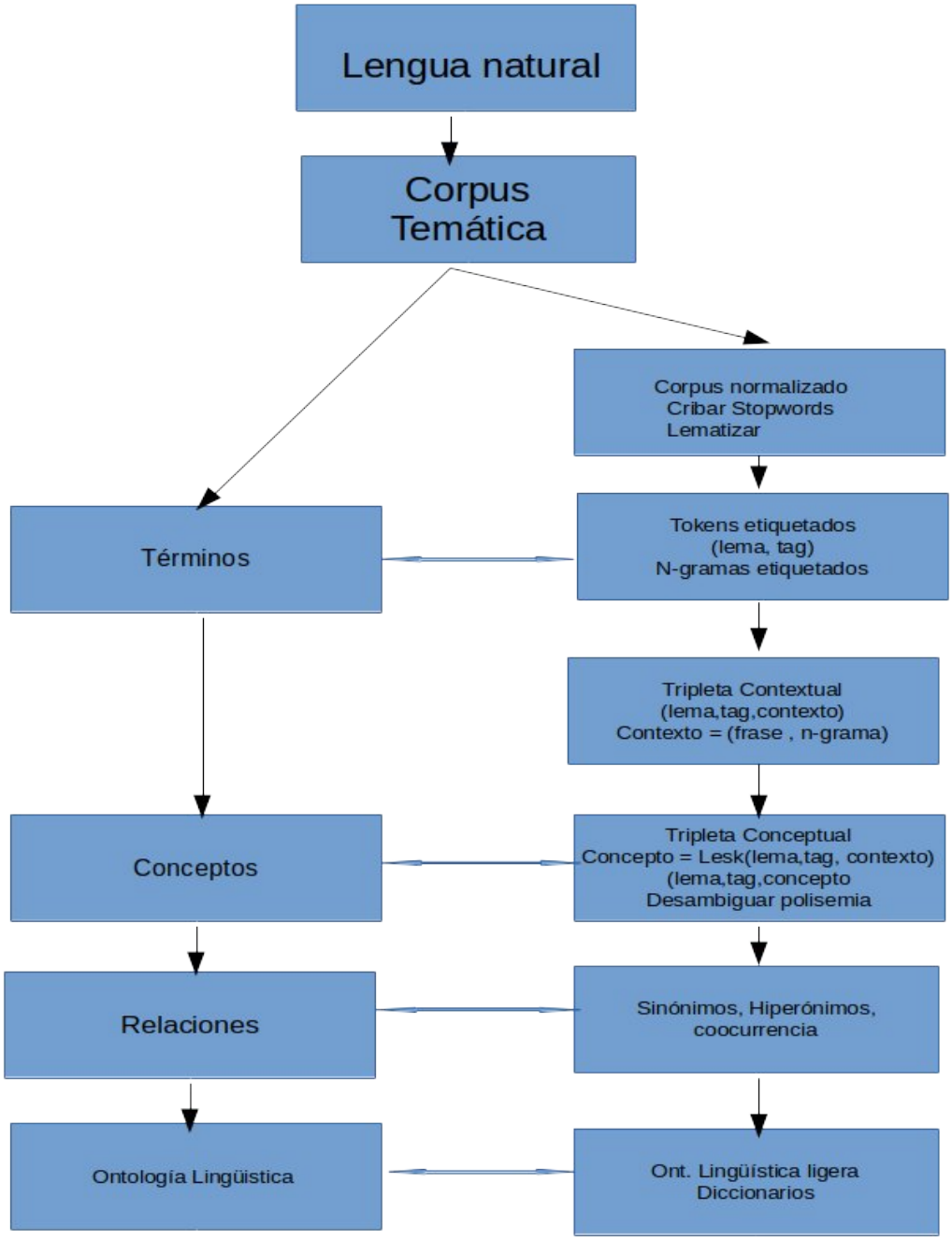


Ilustración 31: De lo abstracto a lo concreto:términos, conceptos, relaciones, etc

4 Reglas de asociación con conceptos(tripletas conceptuales).

En el estado de la cuestión, como corresponde, se ha hecho una semblanza general del paradigma de las reglas de asociación – en adelante RA-. Asimismo, se han comentado las particularidades que dicho paradigma tiene para poder ser usado en la Minería de Textos. Muy en particular, se introdujo el concepto de matriz de transacciones ver Tabla 2 *Matriz Transacciones Valor* y se comentó que dicha matriz es el punto de partida ineludible de toda tarea de RA. Además, en este trabajo se han dado los métodos para extraer los conceptos en forma de *tripleta conceptual* mas importantes de un determinado dominio representado por un corpus significativo de dicho dominio. De igual forma, de cualquier corpus pueden extraerse las frases constitutivas del mismo, un vez normalizadas éstas a su forma de lista con tripletas conceptuales. Así pues, surge de forma natural la necesidad de explorar el concepto de Regla de Asociación en Minería de Textos cuando la matriz se forme por un determinado número de transacciones -las frases en su forma normalizada de lista de tripletas conceptuales -, y los atributos sean un determinado subconjunto de las tripletas conceptuales extraídas del dominio en estudio. Ver *Tabla 7 Matriz transacciones(frases en forma de lista de tripletas conceptuales) versus atributos(tripletas conceptuales)*

	Tripleta Conceptual 1	Tripleta Conceptual 2	Tripleta Conceptual 3	……..	Tripleta Conceptual m
Frase normalizada 1	True	True	False	…….	True
Frase normalizada 2	False	True	False	…….	True
. . . . Frase normalizada n	True	True	True	……..	True

Tabla 7: Matriz transacciones(frases en forma de lista de tripletas conceptuales) versus atributos(tripletas conceptuales)

Así, las potenciales RA extraídas tendrán la forma de:

$$[tripleta-conceptual] \rightarrow [tripleta-conceptual]$$

Con esta introducción ya puede intuirse que se presenta otro nuevo hito en el problema de la polisemia/homonimia, puesto que si se extrae una regla como: *hoz* → *cañón* en un dominio orográfico ya habrá quedado consignado que la coocurrencia se produce entre los conceptos con ámbito orográfico de hoz y de cañón y no, por ejemplo, entre las acepciones de apero agrícola y arma de calibre grueso, ya que lo que se ha extraído, asociándolos, no son lemas sino conceptos – tripletas conceptuales- . Así, lo verdaderamente extraído habrá sido[17]:

$$[(hoz,'n',ámbito-orográfico)] \rightarrow [(cañón,'n',ámbito-orográfico)]$$

17 Se recuerda, por si no se ha leido el resto del trabajo, que el ámbito orográfico, tipicamente en forma de synset si se usa Wordnet, habrá sido extraido mediante una ontología lingüística general.

Se introduce aquí – se desarrollará más adelante - el hecho de conservar la etiqueta sintáctica de las tripletas conceptuales en la regla de asociación lo que permitará dar a ésta unos usos u otros. Muy en particular a la hora de trasladar la RA encontrada a la Ontología Lingüística Ligera -en adelante OLL- a sintetizar.

4.1 Utilidad y novedad de las reglas extraídas con tripletas conceptuales.

En gran parte de las tareas con Reglas de Asociación, el número de reglas extraídas suele ser elevado y, por tanto, es muy útil valorar cada una de estas reglas para poder cribar, en función de dicha valoración, cuantas de aquellas reglas finalmente se considerarán, en el caso que nos ocupa en la elaboración de Ontologías Lingüísticas Ligeras. Son muchos los criterios que se siguen, gran parte de ellos *-ver 2.8.3 Valoración de las Reglas de Asociación en Minería de Textos. en la página 42-,* tienden a apreciar mucho aquellas reglas que descubren asociaciones no obvias; es más: esta circunstancia está en la génesis de las propias RA. Sin embargo, en este caso, puesto que lo que se pretende es elaborar una ontología lingüística ligera que relacione conceptos, las relaciones obvias deben considerarse y registrarse en la ontología ligera a desarrollar con parecida valoración que las reglas no obvias. Además, debe decirse, que la naturaleza dispersa de la matriz de tripletas conceptuales hace, antes bien al contrario, que el número de reglas extraídas no sea grande y por tanto, no sea imprescindible, en gran parte de las ocasiones, cribar las reglas.

Puesto que el método de las RA, muy a menudo, se usará en conjunción con los métodos para extraer, de la ontología lingüística general, sinónimos y cohipónimos de los lemas de las tripletas conceptuales mas frecuentes las reglas mas apreciadas -más útiles, finalmente- serán aquellas que recojan otro tipo de relaciones.

Un ejemplo aclarará lo que se busca: Sea un dominio oleícola. En dicho dominio será fácil encontrar que el conjunto { aceite, oliva} es frecuente -aparecen ambos términos en un número significativo de las frases del corpus-. Aceite y oliva no son sinónimos ni tienen relación de hipónimo/hiperónimo, sin embargo, son términos claramente relacionados, si se pasa de

dichos términos a sus conceptos oportunos para el dominio en cuestión se tendrá que se puede extraer la siguiente regla.

$$\{(oliva, 'n', ámbito-oleícola)\} \rightarrow \{(aceite, 'n', ámbito-oleícola)\}$$

Se dejan de considerar posibles asociaciones en un ambiente textil de la tripleta (oliva, 'a', ámbito-textil) como tonalidad apagada del color verde.

Si dicha relación se recoge en una ontologia lingüística ligera, por ejemplo en forma de entrada de un diccionario de la siguiente forma: *dict[oliva][n] = aceite* y aparecen en un texto nuevo dos frases, con sus respectivas normalizadas, en principio ortogonales, pero muy relacionadas semánticamente, como:

Juan elabora aceite → Juan elaborar aceite

Maria cultiva olivas → Maria cultivar oliva

Ampliando[18] la segunda con el diccionario de RA, tras normalizar se tendrá:

María cultivar oliva aceite.

Ahora dichas frases que están semánticamente relacionadas ya no son ortogonales, sino que, con algún grado de similitud, reflejan la cercanía semántica. Piénsese, para mas incidencia, en el caso en que *cultivar* no se hubiese extraído como hipónimo de *elaborar*.

También son muy útiles, en este trabajo, las RA cuando han aparecido en el corpus tripletas contextuales especialmente frecuentes pero a las que no ha sido posible -por no existir en la ontología lingüística general u otras circunstancias, asignarle un concepto[19]- por tanto estas

18 Se justificará por qué se amplía con los diccionario de RA y no se sustituye. Básicamente porque la sustitución no conserva la verdad ni la coherencia semántica entre conceptos que pueden tener, para empezar, distinta etiqueta sintáctica.

19 A efectos prácticos esto significa que no hay un synset en WordNet que contenga el par (lema, etiqueta sintáctica) encontrado en el corpus de entrenamiento. Estas tripletas contextuales serán, si procede, convertidas en tripletas conceptuales del siguiente modo: (lema, tag, "nohayconcepto")

tripletas no estarán reflejadas en las ontologías ligeras desarrolladas de sinónimos e hiperónimos/ hipónimos. Empíricamente se ha constatado que habitualmente suelen ser descubiertas relaciones entre *entidades nombradas* que no tienen entrada en la Ontología Lingüística General y el concepto al que pertenecen puesto que estas *entidades nombradas* son una particularización material del concepto con el que se relacionan. Por seguir con los ejemplos agrícolas: es fácil entender que el conjunto {valdepeñas,vino } será frecuente en un dominio vitivinícola y que en ese dominio aparecerá finalmente esa relación, mientras que en un dominio de circulares de la UNED aparecerá fácilmente el conjunto frecuente {valdepeñas, ciudad}[20] sirva este último ejemplo para ilustrar varias circunstancias:

- Según un dominio u otro, se consigue desambiguar el significado resolviendo la polisemia, cosa ésta que se lleva persiguiendo en todo este trabajo. En un caso hablamos de un vino, en el otro de una pequeña *agrociudad* con Centro Asociado de la UNED.
- En el caso de que se pueda asociar una *entidad nombrada* con el concepto al que pertenece como instancia se habrá descubierto una relación taxonómica pues Valdepeñas *es un* tipo de vino y Valdepeñas *es un* pueblo o pequeña *agrociudad*.
- Las RA pueden conseguir relacionar dos conceptos que difícilmente tendrán entrada en una Ontología Lingüística General y con ello reflejarlo en una Ontología Ligera. En el caso del conjunto frecuente {valdepeñas, UNED}. Un nombre propio y unas siglas no suelen estar recogidas en una Ontología Lingüística general.
- Supóngase ahora un dominio, como podría ser el de las circulares de la UNED donde son frecuentes los conjunto {valdepeñas ,UNED} y { valdepeñas, ciudad} ¿ qué sentido tendrán las relaciones extraidas a → b o b→ a? Si se ha de maximizar la confidencia de cada una de las dos reglas extraidas de cada conjunto frecuente serán Ciudad → Valdepeñas y Uned → Valdepeñas

20 En Valdepeñas está físicamente ubicado el CA de la UNED de Ciudad Real.

Valdepeñas	Ciudad
Valdepeñas	Ciudad
Valdepeñas	Ciudad
Valdepeñas	Ciudad
Valdepeñas	UNED
Valdepeñas	UNED
Valdepeñas	UNED

Ilustración 32: Sentido de las RA

porque, tomando como ilustración el primer de los casos y según el cuadro anterior se tendrá que la confianza para Ciudad → Valdepeñas es

$$\frac{|\{ciudad, Valdepeñas\}|}{|\{ciudad\}|} = 4/4 = 1$$

Mientras que para Valdepeñas → ciudad es

$$\frac{|\{ciudad, Valdepeñas\}|}{|\{valdepeñas\}|} = 4/7 < 1$$

Con igual razonamiento tendrá mayor confianza la regla UNED → valdepeñas

Otro caso en el que las RA se mostrarán muy útiles, ampliando otros ontologías lingüísticas ligeras, es en aquellas relaciones que se producen entre conceptos con etiquetas sintácticas diferentes; debe recordarse según la *Tabla 1: Relaciones-semánticas de la página 21* que las relaciones de sinonimia se producen entre lemas con la misma etiqueta sintáctica y que las relaciones de hiponimia/hiperonimia se producen solo entre verbos o entre sustantivos. Esta utilidad viene determinada porque las RA son capaces de extraer relaciones entre lemas con etiquetas sintácticas diferentes[21].

Esto permitirá encontrar relaciones muy interesantes, por ejemplo, entre verbos y sustantivos, en especial cuando éste actúa de objeto directo de aquel . De nuevo, se encuentra aquí una muestra más de la importancia de elegir y circunscribir bien el dominio y, sobre éste, el corpus de entrenamiento: supóngase que el corpus de entrenamiento versa sobre comida vegana.

21 Si se considera que las cuatro categorías sintácticas fundamentales son (verbos, sustantivos, adjetivos, adverbios) y recordando que en este trabajo las asociaciones son entre tripletas conceptuales que, como tales, conservan la etiqueta sintáctica de las RA

En este caso se encontrarán conjuntos frecuentes como { (comer, ‘v’) ,{verdura, ‘n’) } y conjuntos como { (comer, ‘v’) ,{carne, ‘n’) } no serán frecuentes. Por tanto, se registrará en la ontología lingüística ligera derivada para ese dominio concreto una asociación extraida del conjunto { (comer, ‘v’) ,{verdura, ‘n’) } en detrimento de potenciales asociaciones extraidas del conjunto { (comer, ‘v’) ,{carne, ‘n’) } que al no ser frecuente no generará asociaciones. Con lo que en futuro usos de la ontología lingüística ligera *vegana* no se asociará la carne con la comida. Si el corpus de entrenamiento versara sobre comida, en general, podrían darse ambos conjuntos frecuentes { (comer, ‘v’) ,{verdura, ‘n’) } { (comer, ‘v’) ,{carne, ‘n’) }

Puede hacerse, a partir de las RA considerando las etiquetas sintácticas, una digresión en la temática de este trabajo dejando constancia para posterior estudio de lo útiles que pueden ser las RA que asocian sustantivos o verbos con adjetivos. En principio, cabe desaconsejar registrar esas asociaciones en la ontología lingüística ligera a desarrollar puesto que ampliar -menos aun, sustituir – en el sentido del epígrafe de la página *121 Aplicaciones mas comunes de las ontologías lingüísticas ligeras con forma de diccionario.* con un adjetivo una frase puede dar lugar a cercanías semánticas espurias. Sin embargo, se debe dejar apuntado lo útiles que las RA con tripletas conceptuales pueden ser extrayendo opiniones si por ejemplo se descubren reglas como:

$$\{ (\text{aceite,n,synset1}),(\text{martos}^{22},\text{n,synset2}) \} \rightarrow \{ (\text{bueno,a,synset3}) \}$$

$$\{ (\text{aceite,n,synset1}),(\text{colza ,n,synset4}) \} \rightarrow \{ (\text{malo,a,synset5}) \}$$

Por tanto, una vez dicho esto, en este trabajo se cribarán las reglas del tipo:

$$\{tripleta-conceptual\} \rightarrow (lema,'a',concepto)$$

También, en este trabajo se cribarán las reglas donde su consecuente o ascendente sea de cardinalidad uno y su etiqueta sintáctica sea un adverbio. Pues no tiene mucho sentido ampliar con un adverbio un conjunto de sustantivos o verbos o mezcla de ambos. Sin embargo si el consecuente de la regla o el antecedente tiene cardinalidad mayor que uno y uno de sus tripletas conceptuales es un adverbio conviene considerar si se ha de mantener la regla encontrada.

22 https://www.martos.es/ Agrociudad conocida mundialmente por la calidad de sus aceites

No obstante, deben señalarse, las debilidades de este método; como tal cabe mencionar que las RA no aportan vocabulario nuevo a la ontología lingüística ligera a desarrollar sino que solo pueden encontrar relaciones de coocurrencia entre conceptos existentes en el corpus de entrenamiento. Asimismo, debe decirse, desde el punto de vista más computacional que conceptual, la gran cantidad de recursos que consume el método por lo que cabe considerar si en todas las ocasiones merece la pena desplegarlo. En este último caso puede darse como indicación la mayor o menor presencia de tripletas frecuentes a las que no haya podido asignársele un concepto.

4.2 Fijación del soporte y la confianza en una Matriz de transacciones Frases versus Tripletas Conceptuales.

Si se recuerda que el soporte, *min-support,* en las RA determina un umbral mínimo que deben cumplir los conjuntos frecuentes para posteriormente de dichos conjuntos extraer las Reglas de Asociación; en el caso concreto que nos ocupa, antes de ir a la parte nuclear, deben hacerse varias consideraciones:

- La Matriz de Transacciones frases versus tripletas-conceptuales -*ver Tabla 7*- tendrá una naturaleza dispersa, sobre todo si se consideran todas las tripletas conceptuales. Recuérdese que se ha asumido que la distribución de frecuencias de las tripletas-conceptuales en el corpus de entrenamiento seguirá una Ley de Zipf-Mandelbrot con lo que habrá pocas tripletas que tengan frecuencias altas y muchas otras que tendrán frecuencias muy bajas. Esto rige aun en el caso de las tripletas frecuentes puesto que considerada su frecuencia en relación con el número de frases aquella resultará pequeña.
- El número de apariciones de cada tripleta conceptual en la Matriz de Transacciones será menor o igual que la frecuencia de aparición de dicha tripleta en la lista de tripletas conceptuales. Si se asume que una tripleta conceptual aparecerá a lo sumo una vez en cada frase se tiene que el número de veces que aparece determinada tripleta conceptual en la matriz es igual a la frecuencia de dicha tripleta en la lista de tripletas.

Asumiendo estas dos consideraciones pueden darse unas pautas para fijar el *soporte*: Si se quiere que las asociaciones extraídas por el algoritmo sean solo entre las p tripletas más frecuentes y se fija n como el número de frases del corpus bastaría con fijar el soporte así:

$$soporte = \frac{f_p}{n} \qquad f_p \text{ es la frecuencia de la p-más-frecuente tripleta}$$

Si se han recuperado con anterioridad las *n-gramas* más frecuentes del corpus -entendidos estos *n-gramas* como listas de *n* tripletas conceptuales frecuentes-, pueden usarse para mediante la fijación de un determinado soporte intentar extraer asociaciones entre las tripletas de dichos n-gramas. Para ello se recuperan los j más frecuentes - en este caso se vió que no rige la Ley de Zipf y se empleó el método del codo- Ver epígrafe *3.5.5 Fijación del número j de n-gramas etiquetados más significativos.*

En cuanto a la fijación de la *confianza* de las RA no se ha conseguido extraer un criterio taxativo, pero si pueden darse unas pautas para ir bajando el nivel de confianza hasta que dichas pautas con determinados umbrales, o un subconjunto de las mismas, se alcancen. Así, :

- Aparecen reglas que reflejan asociaciones entre tripletas conceptuales a las que no ha podido asignársele un concepto -synset, si se usa WN-.
- Han aparecido reglas que relacionan conceptos con distinta etiqueta sintáctica.
- El número de reglas es manejable.

4.3 Tipos de reglas de asociación:

Se clasificarán las RA extraídas en cuatro tipo de reglas, se enumeran en grado creciente de complejidad.

- Reglas simples. Tanto antecedente como consecuente están formados por una única tripleta-conceptual.

  $Tripleta-Conceptual \rightarrow Tripleta-Conceptual$

- Regla compleja *n_a_1*. El antecedente está formado por *n* Tripletas-conceptuales y el consecuente por una única Tripleta-conceptual.

  $\{Tripleta-Conceptual\,1,\ldots.,Tripleta-Conceptualn\} \rightarrow Tripleta-Conceptual$

- Regla compleja *1_a_n*. El antecedente está formado por 1 Tripleta-conceptual y el consecuente por *n* Tripletas-conceptuales.

  $Tripleta-Conceptual \rightarrow \{Tripleta-Conceptual\,1,\ldots.,Tripleta-Conceptualn\}$

- Regla compleja *n_a_m*. El antecedente está formado por *n* Tripletas-conceptuales y el consecuente por m Tripletas-conceptuales.

$\{Tripleta-Conceptual\,1,\ldots.,Tripleta-Conceptualn\} \rightarrow \{Tripleta-Conceptual\,1,\ldots.,Tripleta-Conceptualm\}$

Esta clasificación es operativa para, al menos, dos circunstancias:

1. Intentar, llegado el caso reducir su complejidad para transformar una regla compleja, según algún criterio, en un conjunto de reglas simples.
2. Según, un tipo u otro de regla trasladarla de una forma u otra a la Ontología Lingüística Ligera, típicamente con forma de diccionario.

No obstante debe comentarse que las reglas complejas por la propia naturaleza dispersa de la *Matriz de Transacciones frases versus tripletas-conceptuales* asi como por, en gran parte de los casos, la longitud pequeña de las frases normalizadas, sobre todo en algunas lenguas como el inglés, tienden a aparecer en pocas ocasiones.

4.4 Paso de reglas complejas a reglas simples.

En este caso se asume que las RA pueden ser consideradas como expresiones de la lógica de proposiciones. Asimismo, para facilitar la exposición de los cálculos se asumirá que *n y m* son iguales a 2 con lo que las reglas serán representadas con las letras a,b,c y d.

- Paso de reglas *n_a_1* a regla simple: Sea, por simplificación la regla a,b → c.

  $a \wedge b \rightarrow c ; \neg(a \wedge b) \vee c ; \neg a \vee \neg b \vee c ; \neg a \vee c \vee \neg b \vee c ; a \rightarrow c \vee b \rightarrow c$

- Paso de reglas 1_a_n a regla simple: Sea, por simplificación la regla $a \rightarrow b, c$.

  $a \rightarrow b \wedge c ; \neg a \vee (b \wedge c) ; \neg a \vee b \wedge \neg a \vee c ; a \rightarrow c \wedge b \rightarrow c$

- Paso de reglas n_a_m: Sea por simplificación la regla a,b → c,d

  $a \wedge b \rightarrow c \wedge d \equiv (a \rightarrow c \vee b \rightarrow c) \wedge (a \rightarrow d \vee b \rightarrow d)$

Pueden existir circunstancias en que una regla compleja deba ser reducida a una única regla simple. Estas circunstancias pueden ser impuestas por la naturaleza de la estructura de datos que finalmente albergará la Ontología Lingüística Ligera o por circunstancias de espacio de dicha estructura. En ese caso se plantea la siguiente reducción:

- Para el caso de reglas *n_a_1*. Se acaba de ver que:

  $a \wedge b \rightarrow c \equiv a \rightarrow c \vee b \rightarrow c$

  En este caso para escoger como regla simple final una de las dos reglas simples en disyunción se puede seleccionar aquella que maximice un determinado criterio numérico, por ejemplo la confianza *(confidence)*

  $MAX_{(r \in R)}(confidence(a \rightarrow c), confidence(b \rightarrow c))$ $R = \{a \rightarrow c,\ b \rightarrow c\}$

- Para el caso de reglas 1_a_n. Se acaba de ver que:
  $a \wedge b \rightarrow c \equiv a \rightarrow c \wedge b \rightarrow c$

  En este caso para escoger como regla simple final una de las dos reglas simples en conjución se puede seleccionar aquella que minimice un determinado criterio numérico, por ejemplo la confianza *(confidence)*

$$MIN_{(r\in R)}\left(condifence\left(a\rightarrow c\right),confidence\left(b\rightarrow c\right)\right) \qquad R = \{a\longrightarrow c,\ b\longrightarrow c\ \}$$

- Para el caso de reglas n_a_m. Se acaba de ver que:
  $a\wedge b\rightarrow c\wedge d\equiv\left(a\rightarrow c\vee b\rightarrow c\right)\wedge\left(a\rightarrow d\vee b\rightarrow d\right)$

  En este caso para escoger como regla simple final una de las cuatro reglas simples se procede aplicando

  $$MIN_{(r\in R)}\left(MAX\left(confidence\left(a\rightarrow c\right),confidence\left(b\rightarrow c\right)\right)MAX\left(condifence\left(a\rightarrow d\right),confidence\left(b\rightarrow d\right)\right)\right)$$

  $R = \{a\longrightarrow c,\ b\longrightarrow c\ ,a\longrightarrow d,\ b\longrightarrow d\}$

4.5 Relación de coocurrencia. Elaboración de diccionarios a partir de Reglas de Asociación complejas.

```
R    conjunto de reglas
R =  ordenar(R, de menor a mayor confidencia)
R = cribarSegunTags( R )   ##Eliminar aquellas con el consecuente adjetivos
                           ## Eliminar aquellas con adverbios en cons o ante de
                           ##  cardinalidad 1
Si pasar Reglas Complejas a Simples:
      R = SimplificarReglas(R)
dict  Diccionario dict = {}
Para cada r  perteneciente a     R
      si r  es regla simple  r: a → b :
            dict[a.lema, a.tag ] = b.lema
      sino  si r es regla  n_a_ 1  r: a1,a2,...an →b1
            dict[a1.lema,a1.tag],  ….    ,[an.lema,an.tag] = b1.lema
      sino si r es regla 1_a_n  r: a → b1,b2, ...bn
            dict[a.lema,a.tag] = [b1.lema, ….. , bn.lema]
      sino si r es regla  n_a_  r: a1,a2 ... an  → b1,b2,… bm
            dict[a1.lema,a1.tag]  ,  …..  ,  [an.lema,an.tag]=  [b1.lema,  …..  ,
bm.lema]
devolver dict
```

Texto 5: Algoritmo de elaboración del diccionario con RA

5 Revisión, corrección, supresión y ampliación de términos por parte del experto.

Con lo dicho hasta ahora se podría haber generado ya diccionarios temáticos muy útiles. Sin embargo, la intervención, como paso final, de un experto o expertos humanos[23], en la forma que se expondrá a continuación, mejorará la eficacia y la utilidad de los diccionarios generados. Es éste un caso particular que muestra la realidad empírica, al menos hasta la fecha y en gran parte de los casos, que indica que la Inteligencia Artificial cuando más efectiva se muestra es en colaboración con el ser humano y con una última posibilidad por parte de éste de revocación de lo inducido, deducido o aprendido por parte de aquélla.

23 Se habla aquí de expertos humanos. No obstante cabe la posibilidad de que el experto bien pudiera ser un nivel de abstracción superior automatizado al empleado en la elaboración del diccionario hasta la fecha.

5.1 Aportando valores(sinónimos, cohipónimos) a claves extraídas por el sistema.

Existen varios escenarios en donde el uso de la ontología lingüística general (OLG) no bastará para extraer los conceptos y los términos que las sustentan de un corpus de entrenamiento. Uno de estos casos es cuando hay un par (palabra,etiqueta-sintáctica) muy frecuente en el corpus pero al que no se le ha podido adjudicar por los medios computacionales habituales un concepto adecuado para terminar de formar la tripleta (palabra, etiqueta-sintáctica, concepto)[24]. Un caso particular es cuando el par frecuente es un neologismo y dado que la formalización de la ontología lingüística general – en adelante OLG- precede al uso habitual de este neologismo no se le puede, en consecuencia, asignar un concepto. Caso parecido sucederá con extranjerismos que pueden ser habituales en determinadas temáticas y con términos procedentes de jergas, argots o jerigonzas que no aparecerán en la OLG.

Por tanto un humano experto en la temática en juego deberá elaborar -o ampliar- diccionarios ad-hoc para solventar esta *zona oscura* que no se cubre con la ontología lingüística general. Así, para todos aquellos pares frecuentes extraídos a los que no se ha podido, mediante la OLG, asignar el concepto el experto humano deberá asignarles sus sinónimos y, si procede, sus cohipónimos. Obviamente estos diccionarios – o ampliaciones de los mismos- deben contemplarse para cada temática. En este caso particular el experto suple lo que debería haber suplido la Ontología Lingüística General.

Ejemplo: en una temática vitícola es posible que el par (airén, sustantivo) sea muy frecuente en un corpus de dicha temática y que los medios computacionales -simplemente detectando lo frecuente que es- lo señalen como tal, sin embargo una OLG será incapaz de asignarle un concepto por ser una palabra propia de un argot agrario vitícola. En ese momento entraría en juego el experto vitícola o el ingeniero agrónomo que podría indicar que los sinónimos de airén son: valdepeñera, lairén, forcallat, layren etc. Por tanto, como se decía, el experto ha de completar -aportando sinónimos-, lo que el sistema ha iniciado detectando

24 Se recuerda que es de este concepto -synset, si se usa Wordnet- de donde han de extraerse sinónimos y cohipónimos.

frecuencias altas. Cosa parecida habría de hacerse si el experto aporta cohipónimos que convergieran en *airén* como su hiperónimo.

Debe mencionarse que las RA paliarán, en parte, la gestión de las tripletas a las que, mediante la OLG no se le ha podido asignar un concepto. Si estas tripletas son frecuentes las RA encontrarán asociaciones con otros conceptos del corpus. Si bien esto no aporta vocabulario nuevo, sino relaciones de coocurrencia entre conceptos del corpus que ayudarán en la desambiguación de la polisemia.

5.2 Aportaciones por parte del experto.

Cabe la posibilidad de que el experto aporte, según su propio conocimiento de la temática en cuestión sus entradas a los diccionarios. Aquí, el experto complementa lo que debería haber sido provisto por el corpus temático. No obstante, debe tenerse en cuenta el tamaño del diccionario extraído por los métodos computaciones para que lo aportado por el experto basado en su propio conocimiento no supere un determinado porcentaje de lo aportado por los métodos computacionales. Si es el experto el que aporta una parte no desdeñable del conocimiento cabría preguntarse si los métodos computacionales usados son aptos o si la temática elegida está lo suficientemente bien delimitada o el corpus es el adecuado.

En concreto lo que se está diciendo aquí es que un experto en una temática puede introducir uno – o unos – conceptos en el diccionario temático no extraídas por los métodos computacionales. Esta idea tendrá que ser reflejada en los diccionarios apropiados y para cada uno de estos se habrá de implementar las entradas pertinentes extrayendo de este concepto su lema y los sinónimos de éste así como sus cohipónimos. Cabe aquí la posibilidad de que el concepto que el experto ha considerado importante, pero no así el sistema, si que esté reflejado en la OLG y sea de ésta de donde se pueden extraer sus sinónimos o cohipónimos.

5.3 Ampliación y revocación de las tripletas extraidas.

Otra posibilidad de intervención por parte del experto es mediante la inspección de cada una de los valores finalmente extraídos por los métodos computacionales, ora ampliando con algún valor, ora revocando alguno de los valores asignados por el sistema. En el caso de la ampliación bien pueden seguirse los aspectos del epígrafe *5.2 Aportando valores(sinónimos, cohipónimos) a claves extraídas por el sistema.* Es decir: prestando atención a neologismos, localismos, extranjerismos o eufemismos que serán añadidos como valor a una determinada clave del diccionario.

Un ejemplo de extranjerismo: Supóngase que en una temática hípica ha surgido la tripleta (jinete, n, contexto-hípico) como una de las ideas preeminentes del corpus temático en cuestión. Una vez aislado el concepto de la OLG se extraen los sinónimos *caballista* y *centauro*. Tras esto las entradas en el diccionario serían

dictSin[caballista][n] = jinete

dictSin[centauro][n] = jinete

Sin embargo el experto en hípica puede tener el conocimiento de que el término *jockey* -un anglicismo y, en cierto modo también, un neologismo- se usa ampliamente en la literatura hípica y aunque no se ha extraido de la OLG, debe, pues, decidir insertar una nueva entrada en el diccionario que no ha sido recogida por la OLG

dictSin[jockey][n] = jinete

Debe decirse que, de nuevo, aquí el experto puede ser humano u otro nivel superior de abstracción por ejemplo otra Ontologia Lingüística como un diccionario de extranjerismos.

Otro ejemplo palmario de esta circunstancia surge con los eufemismos. Por ejemplo, el verbo *armonizar*[25] nadie en un contexto general lo consideraría sinónimo de subir o elevar, de hecho en el DRAE no tienen entrada esas acepciones. Sin embargo, en periodos electorales y

25 https://dle.rae.es/armonizar?m=form

hablando de temática fiscal o impositiva es justo el significado que tiene en gran parte de los casos. Si, de una forma u otra, no se contemplasen estos significados restringidos la idea fundamental en un corpus impositivo podría no ser extraída.

De parecida forma el experto podría revocar valores del diccionario que, aun siendo recogidos por la OLG, han quedado desfasados, obsoletos o pueden, en la temática, que nos ocupa resultar ofensivos para determinado colectivo.

5.4 Revisando el diccionario de Reglas de Asociación.

Se ha comentado que el diccionario aprendido a partir de las RA entre tripletas conceptuales es especialmente útil para sintetizar relaciones no taxonómicas y otro tipo de relaciones que no son las cubiertas hasta ahora tales como las de sinonimia y las de hiperónimos/cohipónimos. De igual modo, se ha comentado que estas relaciones se pueden producir entre conceptos con distintas etiquetas sintácticas. Se comentará que el diccionario aprendido debe ser, únicamente, usado para ampliar la representación de nuevas frases pues resulta imposible con la generalidad de las relaciones aprendidas garantizar el mantenimiento de la verdad y la cohesión semántica -ver *Tabla 9: Ampliar y/o sustituir con RA* - . Sin embargo, nada impide que se encuentren relaciones entre tripletas conceptuales con la misma etiqueta sintáctica y que además, entre ellas tengan relación de sinonimia o de hiperonimia. Si el experto o un nivel de abstraccion superior detecta esto, dichas reglas pueden ser movidas al diccionario pertinente de sinónimos y esto facilitaría, si procede según lo correcto en dichos diccionarios no solo la posibilidad de ampliar sino de sustituir *(ver Capítulo 6 Sustitucion.)* Así pues, el experto o un algoritmo pueden llevar a cabo esta tarea *Texto 6: Algoritmo de reasignación de sinónimos:*

R: conjunto de reglas de coocurrencia simples extraidas.

r = a → b

Puede entenderse que existe un concepto(synset) que tiene á a ý b como lemas

entonces se considerán sinónimos a y b

Para :

borrar dictRA[a][tag] = b

insertar dictSin[a][tag] = b

devolver dictRA, dictSin

Texto 6: Algoritmo de reasignación de sinónimos

5.5 Expresiones fijas: locuciones, modismos y otras.

Se llamará *expresión fija* en fraseología a aquellas unidades léxicas con significado propio distinto del significado individual de las palabras que lo forman. Entre estas *expresiones fijas* se considerarán las *locuciones, modismos, refranes* y otros. Una definición de *locución* en palabras de [Casares, 1950] aclarará los conceptos.

> *Llamaremos en adelante locución a la "combinación estable de dos o más términos, que funciona como elemento oracional y cuyo sentido unitario consabido no se justifica, sin más, como una suma del significado normal de los componentes". Noche oscura no es locución porque nos limitamos a añadir al concepto ordinario de "noche" el también corriente de "oscuridad" mediante un calificativo. Noche toledana sí es locución, porque el hecho de conectar la "noche" con "Toledo" no justifica que con ambos vocablos se designe "una noche en la que no es posible dormir"*

A esta definición formal debe añadírsele una característica fundamental y es que la locución, sin pérdida de significado, puede ser sustituida por una palabra. La etiqueta sintáctica de dicha palabra permitirá clasificar las locuciones en adjetivas, sustantivas, verbales y otras.

Por ejemplo: una verdad *como un templo* puede ser sustituida por una verdad *palmaria o* verdad *incontestable*. Se estaría, pues, ante una locución adjetiva. Abundando en esto se puede traer a colación un ejemplo de locución nominal: *brazo de gitano* por *pastel.* Tanto así para gran parte de las etiquetas sintácticas.

Los modismos, aun siendo expresiones fijas, no suelen admitir su sustitución por una palabra además, pueden conjugarse: *no veo tres en un burro, no ves tres en un burro* y tienen un fuerte sentido figurado.

Los refranes pueden considerarse modismos con un fuerte contenido moralizante o pedagógico y, sin embargo, no suele tener sentido conjugarlos.

En el ámbito de este trabajo, y de este epígrafe, lo importante es tener en cuenta hasta que punto las diversas clases de *expresiones fijas* son clases cerradas o no. Así como evitar que estas *expresiones fijas* puedan generar tripletas conceptuales espurias que alimenten las ontologías lingüísticas ligeras a generar. Desde luego si se encuentra, como ya se ha dicho, que un n-grama frecuente es *aceite oliva virgen extra* es adecuado buscar los sinónimos de extra como superior. Sin embargo si se encuentra de forma frecuente el refrán *a Dios rogando y con el mazo dando* no tiene sentido almacenar los sinónimos o cohipónimos de *Dios* o de *mazo* porque nadie va a decir *al Creador rogando y con el marro golpeando.*

Llegados hasta aquí lo importante es separar las *expresiones fijas* de los n-gramas frecuentes que posteriormente generarán entradas en los diccionarios, obviamente para que aquéllas no generen entradas en los diccionarios. Para ello se presentan, según la clase de *expresión fija* unos protocolos:

1. Locuciones.

   Se considera clase cerrada y por tanto suceptible de ser compendiada en un diccionario. Así, en el proceso de normalización del texto y mediante comparativa con el diccionario se puede sustituir la locución (clave del diccionario) por su palabra equivalente(valor del diccionario). Esto es fácilmente automatizable.

   El experto humano puede intervenir si encuentra, a su juicio, alguna locución no contemplada en el diccionario de locuciones. Y en cualquier caso no dejando que en los n-gramas frecuentes finales se cuele ninguna locución.

2. Modismos.

   Puesto que , en gran parte de los casos, no hay palabra equivalente a un modismo no se puede obrar como en el epígrafe anterior.

   Se puede cruzar el conjunto de n-gramas frecuentes con el diccionario de modismos para eliminar del conjunto de n-gramas frecuentes los modismos.

   El experto puede revisar el conjunto de n-gramas frecuentes para eliminar de él lo que considere modismo. Esta intervención es importante por la propiedad de los modismos de ser conjugables y por tanto la plausibilidad de que alguna forma conjugada no esté en el diccionario de modismos.

3. Refranes.

   Se debe operar igual que en el caso de los modismos. Si bien es más fácilmente automatizable puesto que los refranes, como se ha dicho, no se conjugan y pueden considerarse una clase cerrada.

6 Aplicaciones mas comunes de las ontologías lingüísticas ligeras con forma de diccionario.

Una vez elaboradas las ontologías lingüísticas ligeras, con forma de diccionarios, pertinentes de una determinada temática y ante la presencia de un nuevo texto -de dicha temática- normalizado, lematizado y sintácticamente etiquetado cabe concretar que usos más comunes -aplicaciones-, sobre dicho texto, pueden darse a los diccionarios.

Surge, así, en primera instancia la propuesta de hacer evolucionar el texto con ayuda de dichos diccionarios - que en este trabajo serán los de sinónimos, hiperónimos y RA-. Se escrutan las posibilidades de para aquellos términos - en este capítulo 6 término ha de entenderse como un par (lema,etiqueta sintáctica) que será la clave del diccionario en cuestión -, del texto que aparezcan como claves en el diccionario ser sustituidos o ampliados por los valores del diccionario correspondientes. En la tabla *Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos* se exponen, sobre cinco principales características, que implicaciones tiene una determinada evolución -ampliación, sustitución - de un texto con un determinado diccionario – sinónimos, hiperónimos o RA-. Para recordar la estructura de los diccionarios

sintetizados debe revisitarse el epígrafe 3.6 *Relaciones entre conceptos(tripletas conceptuales) y una Ontología Lingüística General. Elaboración de las ontologías lingüísticas ligeras temáticas con forma de diccionario.*

	DICCIONARIO DE SINÓNIMOS	***DICCIONARIO DE HIPERÓNIMOS***
AMPLIAR	CONSERVAN LA VERDAD NO HAY PÉRDIDA DE INFORMACIÓN HAY REDUNDANCIA ESPACIO $2\times\lvert longitudFrase\rvert$ VOCABULARIO $voc' \subset voc \cup dictSin.valores$	CONSERVAN LA VERDAD NO HAY PÉRDIDA DE INFORMACIÓN NO HAY REDUNDANCIA ESPACIO $2\times\lvert v\rvert+2\times\lvert n\rvert+\lvert a\rvert+\lvert r\rvert$ VOCABULARIO $voc' \subset voc \cup dictHip.valores$
SUSTITUIR	CONSERVAN LA VERDAD NO HAY PÉRDIDA DE INFORMACIÓN NO HAY REDUNDANCIA ESPACIO Permanece igual VOCABULARIO $voc' \subset voc \setminus dictSin.claves \cup dictSin.valore$	CONSERVAN LA VERDAD HAY PÉRDIDA DE INFORMACIÓN NO HAY REDUNDANCIA ESPACIO Permanece igual VOCABULARIO $voc' \subset voc \setminus dictHip.claves \cup dictHip.valores$

Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos

	DICCIONARIO DE REGLAS DE ASOCIACION
AMPLIAR	CONSERVAN LA VERDAD NO HAY PÉRDIDA DE INFORMACIÓN HAY REDUNDANCIA O NO: En función de la naturaleza oculta de la asociación descubierta ESPACIO $espacio \leqslant 2 \times \lvert longitudFrase \rvert$ En general el espacio se incrementa poco porque pocas son las entradas del diccionario de RA VOCABULARIO: No se amplia vocabulario, pues éste no proviene de una fuente externa, sino del propio corpus de entrenamiento $voc' = voc$
SUSTITUIR	**NO** CONSERVAN LA VERDAD Por tanto no se considera la sustitución con el diccionario de r. de asociación

Tabla 9: Ampliar y/o sustituir con RA

Así, con respecto a los cinco puntos que tiene cada una de las posibles combinaciones se tiene que:

1. CONSERVAR LA VERDAD. Con ello se hace referencia a que la introducción de sustituciones o ampliaciones de términos no implique contradicción alguna en la frase a evolucionar ni altere su contenido semántico. En el único caso en que no puede garantizarse esto es sustituyendo con el diccionario de RA. Por tanto se descarta ya esta posibilidad y se cierra esta vía -sustituir con el diccionario de RA-.

   El siguiente ejemplo en un corpus católico, ilustra lo dicho para las RA:

   Sea la entrada en el diccionario de RA *dictRA[cristo][n] = maría* extraída de una asociación *cristo → maría*

   Sea la frase y su normalizada *Cristo es redentor → cristo redentor.*

Sea la frase anterior sustituida mediante el diccionario de RA: *maria redentor*

Cualquiera que conozca la teodicea católica verá que la frase sustituida no conserva la verdad y es incluso herética.

2. PÉRDIDA DE INFORMACIÓN. En el único caso en el que hay pérdida es el caso de sustitución con hiperónimos pues al sustituir un hipónimo por su hiperónimo, generalizando, se pierden detalles en la información transmitida. Si se sustituye el término *italiano* por su hiperónimo *europeo* se conserva la verdad pero se podría pensar que tal europeo es de, por ejemplo, Alemania.
3. REDUNDANCIA. Se alude a cuando se representa y se almacena la misma información varias veces. En el caso de *Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos* sucede al ampliar con sinónimos pues se representa el mismo contenido semántico varias veces con distintas palabras. No obstante, esta situación puede ser útil al considerar la representación matricial de los textos pues tal ampliación puede hacer variar los pesos de los términos de dicha matriz. Más aun en la ampliación con RA pues, en este caso, no hay vocabulario proveniente de la Ontología Lingüística General sino del corpus de entrenamiento. Dando, por tanto, mas importancia a los términos relevantes del corpus.
4. ESPACIO. Tras hacer evolucionar – ampliando o sustituyendo - la frase; ésta, si se ha ampliado, ocupará más espacio. Las expresiones consignadas en la tablas *Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos y Tabla 9: Ampliar y/o sustituir con RA* deben ser tomadas como límites superiores cuando todos los términos de la frase a evolucionar pertenecen a las claves del diccionario en cuestión. De igual forma debe decirse que en el caso de la ampliación con el diccionario de hiperónimos los términos *v,n,a,r* aluden respectivamente a los verbos, sustantivos, adjetivos y adverbios como categorías sintácticas que se han usado en todo este trabajo. En el caso de ampliar con hiperónimos y en las condiciones ya citadas de límite superior se doblarán sólo los verbos y los sustantivos de la frase en cuestión pues, conceptualmente hablando, solo esas categorías sintácticas tienen hiperónimos.

5. VOCABULARIO. Una vez que se ha hecho evolucionar un texto el vocabulario del mismo habrá, en consecuencia, variado[26]. Esto es muy importante pues, por ejemplo, en una ulterior representación del texto en una matriz términos-documentos será el nuevo vocabulario el que determine la dimensión de dicha matriz. Así, en el caso de las sustituciones la cardinalidad del nuevo diccionario disminuirá, sin embargo en el caso de las ampliaciones la cardinalidad del diccionario aumentará. Para mayor abundamiento en el vocabulario ver los epígrafes *6.1 y 6.2*

En el siguiente cuadro se ponen en relación algunos parámetros del texto a procesar – en particular su vocabulario - con parámetros de los diccionarios a ser usados en dicho proceso.

26 Existen circunstancias plausibles pero poco probables en que esto no es asi. Por ejemplo si ningún término pertenece al conjunto de las claves del diccionario en uso

Sea *dict* un determinado diccionario con la estructura de *Ilustración 27: Elaboración del diccionario de sinónimos.* Y *voc* el vocabulario de un texto a procesar(evolucionar). Se sigue la terminología de *Ilustración 28: Dimensiones del diccionario de sinónimos*

- Por construcción
- $|dict.valores| = nT$ (nT número de Tripletas Conceptuales)
- $|dict.claves| = nT \times nlC$ (diccionario de sinónimos)
- $|dict.claves| = nT \times nlC \times nch$ (diccionario de hiperónimos)
- $voc = \{(palabra, etiquetaSintáctica)\}$ Vocabulario etiquetado del texto a procesar
-

(Un término evolucionable es un término(par) que puede evolucionar bien mediante sustitución bien mediante ampliación)

- $términosEvolucionables \subset dict.claves$
- $términosNoEvolucionables \subset dict.valores \cup otrosTérminos$
- $otrosTerminos \not\subset dict.valores \wedge otrosTerminos \not\subset dict.claves$
- $|voc| \leq |dict.claves| + |dict.valores| + |otrosTérminos|$

Ilustración 33: Vocabulario del texto y diccionario

A la vista del cuadro *Ilustración 33* pueden hacerse consideraciones así como dar medidas de cuan apropiado será procesar tal o cual texto con los diccionarios provenientes de tal o cual corpus de la misma temática. Así, si se tiene un texto(con su correspondiente vocabulario) e i diccionarios provenientes de diversos corpus y se ha de elegir el mejor diccionario la elección vendrá dada por el diccionario(i) que maximice la expresión :

$$MAX_i[\alpha \times |voc \cap dict.claves| + \beta \times |voc \cap dict.valores|]$$

$$\alpha + \beta = 1$$

Los valores concretos de α y β pueden ajustarse según el tipo de diccionario que se pretenda examinar. Por ejemplo, en el caso de un diccionario de RA tenderán a ser mas parecidos en torno a 0.5 pues tanto dict.claves como dict.values son términos del corpus de entrenamiento. En el caso de diccionario de hipónimos y sinónimos α debe ser más grande que β.

A continuación en los epígrafes 6.2 y 6.1 se comentan las consecuencias que tienen en el vocabulario del texto la ampliación y sustitución con uno y otro diccionario; poniendo énfasis en las cardinalidades del nuevo vocabulario puesto en referencia con la cardinalidad de las claves del diccionario *dict.claves* y con la cardinalidad de los valores del diccionario *dict.valores*.

6.1 Ampliación.

Los diccionarios se usarán para, mediante ampliación, expandir el número de símbolos que se tomarán para representar un texto. Debe decirse que con el diccionario de RA no se da tal expansión pues los valores de dicho diccionario pertenecen al corpus de entrenamiento y no a una fuente externa al mismo como es la Ontología Lingüística General.

En este caso se amplía – o se deja como está - el vocabulario del texto original *voc*. En ambos casos se pretende que los conceptos mas representativos estén, mediante ampliación, muy presentes en el texto resultante[27] pero dejando presente el término original que *llama* al término representativo. Es decir claves y valores del diccionario estarán, en una medida u otra, presentes en el texto ampliado. El vocabulario resultante será etiquetado como *voc'*

27 Esto, por ejemplo, facilitaría la agrupación de frases semánticamente similares en clústers.

- Si $voc = dict.claves$

  Debe decirse que está posibilidad es muy poco plausible -imposible si el texto tiene adjetivos o adverbios y el diccionario es el de hiperónimos-, pero conviene consignarla en el análisis como límite de las posibilidades $voc \subset dict.claves \; y \; dict.claves \subset voc$

  Se observa que todo los términos se amplían y todos los términos de dict.valores aparecerán.

  - $voc' = dict.claves \cup dict.valores$
    - $|voc'| = nT + nT \times nlC$ Diccionario de sinónimos
    - $|voc'| = nT + nT \times nlC \times nch$ Diccionario de hiperónimos

- Si $voc \neq dict.claves \wedge voc \cap dict.claves = \emptyset$

  No hay evolución(ampliación en este caso) $voc' = voc$

  - $voc \neq dict.valores$ No tiene sentido usar este diccionario con este texto.
  - $voc \cap dict.valores \neq 0$ Sobreajuste del diccionario al texto. Relación estrecha entre ambos pero no hay margen para la generalización. Puede deberse a que se ha escogido un umbral demasiado alto. Si no se revisa éste, diccionario inútil.

- Si $voc \neq dict.claves \wedge voc \cap dict.claves \neq \emptyset$

  Habrá ampliación, en concreto $|voc \cap dict.claves|$ términos

  $voc' \subset voc \cup dict.valores$

  $|voc'| \leq |voc| + |dict.valores| \qquad |voc'| \leq |voc| + nT$

  Si además $voc \cap dict.valores \neq \emptyset$ el diccionario es apto para hacer evolucionar el texto pues comparten temática.

  - Si además $voc \subset dict.Claves$ Entonces, claro, se ampliarán $|voc|$ términos y $|voc'| = 2 \times |voc|$

    Además esto implica que no hay en *voc* elementos de *dict.valores*

  - Si además $dict.Claves \subset voc$

$voc' = voc \cup dict.valores$ Se amplían todos los términos y aparecerán todos los

términos de dict.valores.

- Si $voc \cap dict.valores = \emptyset$

$$|voc'| = |voc| + |dict.values|$$

- Si $voc \cap dict.valores \neq \emptyset$

$$|voc'| = |voc| + |dict.valores| - |voc \cap dict.valores|$$

Un algoritmo de ampliación será:

```
Funcion AMPLIARAUX(texto, diccionario)
    ampliación = [ ]
    for palabra, etiqueta en texto:
        if diccionario[palabra][etiqueta]:
            ampliacion.append((diccionario[palabra][etiqueta],etiqueta))
    devolver ampliacion

Funcion AMPLIAR(texto, diccionario)
    devolver Concatenar(texto, AMPLIARAUX(texto,diccionario))
```

Texto 7: Algoritmo de ampliación con diccionario

6.2 Sustitucion.

La sustitución implica que estos diccionarios se podrán usar como una herramienta para hacer converger unos símbolos en otros símbolos canónicos y, por tanto, reducir el número de símbolos que se usan para representar un texto y hacer converger a éste hacía las ideas principales – los valores del diccionario-. Así, en un texto virgen etiquetado aquellos pares (palabra, etiqueta) que aparezcan como una de las claves de un diccionario serán sustituidas por dict[palabra][etiqueta]; el representante de todos los sinónimos de un mismo concepto o el hiperónimo de todos sus cohipónimos. No se contempla el uso del diccionario de RA para sustituir.

Así, si *voc* es el vocabulario del texto normalizado y etiquetado y *voc'* el vocabulario del texto procesado y teniendo presente el cuadro Ilustración 33: Vocabulario del texto y diccionario y siendo dict el diccionario en cuestión . Mediante la sustitución se tendrá qué:

- Si $voc = dict.claves$
  Todos los términos se sustituyen $voc' = dict.valores$ y $|voc'| = nT$

- Si $voc \neq dict.claves \wedge voc \cap dict.claves = \emptyset$
  No hay ninguna sustitución $voc' = voc$
  - Y además $voc \cap dict.valores = \emptyset$
    Texto y Diccionario de temáticas diferentes no tiene sentido usar éste para procesar aquel.
  - Y además $voc \cap dict.valores \neq \emptyset$
    En la medida en que $|voc \cap dict.valores|$ se acerque a $|dict.valores|$ la temática del texto y del corpus de donde provenga el diccionario serán la misma indicando un sobreajuste del texto y del diccionario. Conviene, pues, inspeccionar la posibilidad de bajar el *umbral* para que aparezcan términos de dict.claves y haya sustitución.

- Si $voc \neq dict.claves \wedge voc \cap dict.claves \neq \emptyset$

  $voc' \subset dict.valores \cup otrosTerminos$ pues los términos pertenecientes a dict.claves serán sustituidos.

  $|voc'| \leq nT + |otrosTerminos|$

  - Si además $dict.claves \subset voc$ entonces:

    Tras la sustitución todo el conjunto $dict.valores$ aparecerá en voc' y se tendrá que:

    $voc' = dict.valores \cup otrosTerminos$ y $|voc'| = nT + |otrosTerminos|$

  - Si además $voc \subset dict.claves \rightarrow voc' \subset dict.valores$ Puesto que todos los términos se sustituyen $|voc'| \leq nT$

Teniendo en cuenta que los diccionarios pueden ser de sinónimos o hipónimos.

En el caso de los diccionarios de sinónimos, se tendrá:

$$|voc'| \leq |voc \setminus dicSinonimos.claves| + |dictSinonimos.valores|$$

$$|voc'| \leq (|voc| - nT\ x\ nlC) + nT$$

En el caso de usar un diccionario de hiperónimos los cálculos son similares. El cálculo del n.º de valores del diccionario de cohipónimos es

$|dictHiperonimos.claves| = nEDict = psv \times nT \times nlC \times nch$ por tanto

$$|voc'| \leq (|voc| - nT \times nlC \times nch \times psv) + nT$$

A la vista de las expresiones queda claro que se produce una reducción de $|voc'|$ frente a $|voc|$ sí $voc \cap dictSinonimos.claves \neq \emptyset$ pues, en el caso del diccionario de sinónimos se suprimen $nT\ x\ nlC$ términos y se añaden solo *nT.* Piénsese que esto reduce, por ejemplo, una posible representación del texto en un matriz término-documento y además se consigue colapsar el documento hacia sus ideas mas representativas.

Un posible algoritmo de sustitución será:

```
Funcion  SUSTITUIR(texto, diccionario)
        nuevoTexto = []
        for  palabra, etiqueta in  texto:
                if   diccionario[palabra][etiqueta]:
                        palabra, etiqueta  = diccionario[palabra][etiqueta], etiqueta
        nuevoTexto.append(palabra, etiqueta)
        devolver  nuevoTexto
```

Texto 8: Algoritmo de sustitución con diccionario

6.3 Posibles usos combinados de ampliación y sustitución con diversos diccionarios.

En los epígrafes *6.1 Ampliación. y 6.2 Sustitucion.* se han dado los algoritmos para, respectivamente, ampliar y sustituir con un diccionario provisto un texto; de igual forma se ha comentado la incidencia que esto tendrá en el vocabulario evolucionado de dicho texto. Siguiendo la *Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos y Tabla 9: Ampliar y/o sustituir con RA* se consigna que existe la posibilidad de combinar, como sumo cinco evoluciones con a lo sumo tres diccionarios, para afinar, según conveniencia, el texto resultante.

Así, se tendrían $\sum_{(i=1)}^{5} \binom{5}{i}$ posibilidades a considerar si se quieren probar todas las posibles combinaciones de:

1. Ampliar con sinónimos.
2. Ampliar con hiperónimos.
3. Sustituir con sinónimos.
4. Sustituir con hiperónimos.
5. Ampliar con diccionario de Reglas de Asociación

No obstante y atendiendo a la *Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos* se puede restringir el espacio de búsqueda a aquellos operadores que no implican aspectos negativos como:

- Presencia de redundancia.
- Pérdida de información.

Con lo que ahora se tendría como posibles combinaciones a considerar solo:

A) Sustituir con sinónimos.

B) Ampliar con hiperónimos.

C) Sustituir con sinónimos y ampliar con hiperónimos.

A estas tres posibles combinaciones se añadirá, por defecto, la de ampliar con el diccionario de RA porque éste, según se ha visto, suele ser pequeño y además facilita la intervención de aquellos lemas a los que no haya podido asignárseles un concepto.

Con esto se tendrán una herramientas que servirán, entre otras muchas cosas, para acercar, de diversos modos, frases semánticamente similares pero que son cuasi ortogonales léxicamente hablando. Además, si se restringe la evolución a los operadores A) B) o C) no se introducirá ruido[28] en el texto resultante.

Un ejemplo trivial servirá para ilustrar lo dicho y para dar una primera idea del alcance del trabajo. Supóngase que en un determinado corpus representativo del mundo del motor se ha extraído que la palabra *coche*, en su condición de sustantivo, y en un determinado contexto es la más frecuente de dicho texto y a continuación en dicho contexto automovilístico se encuentra que el verbo *conducir* es muy frecuente. Mediante la ontología lingüística general escogida y, quizá la revisión del experto, se encuentra que para la tripleta (coche, sustantivo, concepto) sus sinónimos son *automóvil* y *carro*[29] y sus hipónimos *berlina* y *utilitario*; de igual forma se concluye que *manejar*, en el contexto automovilístico es un sinónimo de *conducir* (conducir, 'v', ámbito automovilístico). Además, mediante las Reglas de Asociación se encuentra que *Renault* (Renault, 'n', sin contexto) y *coche* coocurren en un número significativo de frases del corpus

28 Pérdida de información o redundancia de la misma

29 El hecho de que el corpus de entrenamiento sea automovilístico permite que coche y carro sean tomados como sinónimos y no como dos palabras "disjuntas" si se estuviese en un ámbito ganadero donde carro lo sería, por ejemplo tirado por una yunta.

de entrenamiento. Es importante señalar que el sistema debe encontrar los sinónimos adecuados al dominio y el contexto en cuestión y excluir los inapropiados; de nada sirve que el sistema detectara *conducir* como sinónimo de *portarse o de proceder de una forma u otra*[30]. De hecho, ésta es una de las razones que llevan a la elaboración de estos diccionarios en detrimento del uso de la ontología general.

En una notación poco formal pero cercana a muchos lenguajes de programación se tendrá que en un dominio automovilístico:

dictSinonimos[automóvil][n] = coche
dictSinonimos[carro][n] = coche

dictSinonimos[manejar][v] = conducir
~~dictSinonimos[portar][v] = conducir~~ Inapropiado en este dominio.

dictHiperonimo[berlina][n] = coche
dictHiperonimo[utilitario][n] = coche

dictRA[renault][n] = coche

Sean las frases y sus normalizadas:

- Juan maneja un carro. → manejar carro
- Juan conducirá un coche. → conducir coche

En este caso, una vez normalizadas ambas frases -eliminando la *stopword un y* usando el infinitivo-, se tendría que ambas frases son semánticamente idénticas pero léxicamente ortogonales. Si se usa el diccionario de sinónimos y se sustituyen los sinónimos -claves del diccionario- por su forma canónica - utilizando el operador A) - se habría conseguido dar para ambas frases la misma representación y por tanto hacerlas coincidir también léxicamente.

30 Octava acepción en el DRAE

- conducir coche
- conducir coche

Se observa además que no se introduce redundancia ni hay pérdida de información, manteniendo además el mismo espacio y reduciendo el vocabulario.

Sean las frases y sus normalizadas:

- Juan maneja una berlina. → manejar berlina
- Pepa conduce un utilitario. → conducir utilitario

Estas frases son semánticamente muy similares pero no idénticas. Tal cual el proceso de normalización las deja se muestran, como en el caso anterior, léxicamente ortogonales. Para este caso un uso combinado del diccionario de sinónimos sustituyendo *manejar* por *conducir* y del diccionario de hiperónimos sustituyendo tanto *berlina* como *utilitario* por *coche* dejarían las frases normalizadas así

- conducir coche
- conducir coche

Efectivamente se consigue forzar la cercanía semántica en ambos casos Juan y Pepa conducen coche pero se ha perdido que lo que conducen no es exactamente el mismo artefacto. Por tanto, con esta combinación de operadores -sustituir con sinónimos y sustituir con hiperónimos- se pierde información aunque se conserva el espacio y se reduce el vocabulario. Para subsanar esto se propone usar el diccionario de hiperónimos, en vez de sustituyendo, ampliando con lo que se tendría ahora:

- conducir berlina coche
- conducir utilitario coche

Las frases son ahora mucho mas cercanas pero no exactamente iguales con lo que se gana en cercanía semántica pero se preservan las diferencias. No hay ni redundancia ni pérdida de

información, aunque aumenta el espacio, si bien aunque el vocabulario cambia no aumenta su cardinalidad.

6.4 Ampliación con el diccionario de Reglas de Asociación.

Finalmente se consigna un ejemplo ampliando con el diccionario de reglas de asociación: Sean las frases y sus normalizadas:

Juan tiene un Renault → tener renault
Juan conduce un coche → conducir coche

Según los diccionarios de sinónimos e hiponimos que se han considerado en este epígrafe ni ampliando ni sustituyendo se conseguiría dotar a las frases de la cercanía que, sin duda, semánticamente tienen, si se asume que *Renault* es una marca determinada de coche. Puesto que Renault es un nombre propio difícilmente ninguna ontología lingüística general tendría referencia alguna a dicho nombre. Sin embargo como se ha extraido la regla:
renault → coche y se ha trasladado al diccionario de reglas de asociacion como:

dictRA[renault] = coche

Ampliando se tendría que:

Juan tiene un renault → tener renault coche
Juan conduce un coche → conducir coche

Con lo que las frases,al menos, dejan de ser ortogonales léxicamente hablando.

Así, viendo la utilidad de hacer evolucionar un texto ampliando con hiperónimos y sustituyendo con sinónimos puede darse un algoritmo haciendo uso de los ya consignados en *Texto 7: Algoritmo de ampliación con diccionario y en Texto 8: Algoritmo de sustitución con diccionario*

```
Funcion  SUSTITUIRYAMPLIAR(texto, dicSinonimos, dicHiperonimos)

    COBEGIN   #artificio notacional para indicar concurrencia
            texto1 = SUSTITUIR(texto, dicSinonimos)
            texto2 = AMPLIARAUX(texto,dicHiperonimos
    COEND
    devolver  texto1 concatenado texto2
```

Texto 9: Algoritmo de sustitución con sinónimos y ampliación con hiperónimos

Llegados hasta aquí conviene hacer una reflexión que fije, según todo lo antedicho, la utilidad de estos diccionarios frente al uso directo de la ontología lingüística:

- Solo los términos -sus lemas- verdaderamente importantes para un dominio serán, si procede, sustituidas o ampliadas con los diccionarios. Esto hará el vocabulario mas restringido o mas representativo de un dominio en cuestión frente al uso directo de la ontología que sustituiría o ampliaría gran parte de las palabras en el texto.

- Recuperar de un diccionario "ad hoc" una determinada palabra consumirá, en varios órdenes de magnitud, menos recursos que la recuperación, desambiguación y extracción directa de sinónimos o hiperónimos de una ontología lingüística general. En este sentido puede decirse que el diccionario será, en cierto modo, una *caché* de la ontología lingüística general.

- Existe, de forma intrínseca, en el diccionario una resolución de la polisemia/homonimia.

7 Caso de uso: Elaboración de ontologías lingüísticas ligeras de temática religiosa judeo-cristiana.

7.1 Corpus, temática y normalización.

Se utilizará en este Caso de Uso un Corpus bíblico compuesto por la Biblia del Rey Jacobo (King James Bible). Una pequeña semblanza en palabras de [López, 2011] es:

> *La Biblia del Rey Jacobo fue un gran esfuerzo colectivo de traducción llevado a cabo entre 1604 y 1611 por 47 estudiosos nombrados por el rey Jacobo VI de Escocia y I de Inglaterra. Los traductores se dividieron en seis comités, dos en Westminster, dos en Cambridge y dos en Oxford; el trabajo de cada traductor fue revisado por los demás miembros del grupo, luego el trabajo de cada grupo por los demás grupos y finalmente la obra completa por dos miembros de cada comité. El resultado final prima, pues, la tradición y el consenso. Además, las instrucciones explícitas del rey fueron utilizar siempre*

> *que se pudiera las opciones de una versión anterior, la Biblia de los Obispos (1568). Esa traducción era una revisión de la Gran Biblia (1539), revisión a su vez de la Biblia de Matthew (1537), fruto de la revisión de las Biblias de Coverdale (1535) y Tyndale (1525, 1530). La Biblia del Rey Jacobo fue la tercera de las Biblias inglesas «autorizadas», después de la Gran Biblia (1539) y la Biblia de los Obispos (1568).*

Puede revisarse la versión digital utilizada en :

bible-kjv.txt

A las tareas habituales de normalización ya consignadas se ha añadido la eliminación numérica de los versículos.

7.2 Extracción de términos.

7.2.1 Extracción de la lista de tripletas contextuales actuando como términos.

Tras la fijación de la temática, la selección del corpus representativo de la misma y la normalización de dicho corpus debe presentarse como primer hito la representación del mismo como lista de Tripletas contextuales según la definición presentada en *3.4.1 Tripleta contextual*

El resultado de dicho proceso da como resultado una lista de 374.761 tripletas contextuales que quedan recogidas para quién quiera examinarlas en:

- bibTripletasContextuales.txt
- bibTripletasContextuales.pickle

Se muestran, aquí, a modo de ejemplo, 10 de ellas, sin ningún criterio de preferencia:

```
('king', 'n', ['king', 'james', 'bible'])
('james', 'n', ['king', 'james', 'bible'])
('bible', 'a', ['king', 'james', 'bible'])
('old', 'a', ['old', 'testament', 'king', 'james', 'bible'])
('testament', 'n', ['old', 'testament', 'king', 'james', 'bible'])
('king', 'n', ['old', 'testament', 'king', 'james', 'bible'])
('james', 'n', ['old', 'testament', 'king', 'james', 'bible'])
('bible', 'a', ['old', 'testament', 'king', 'james', 'bible'])
('first', 'a', ['first', 'book', 'moses', 'call', 'genesis'])
('book', 'n', ['first', 'book', 'moses', 'call', 'genesis'])
('moses', 'n', ['first', 'book', 'moses', 'call', 'genesis'])
```

Ilustración 34: Biblia: 10 tripletas contextuales

7.2.2 Extracción de los j n-gramas etiquetados más frecuentes actuando como términos.

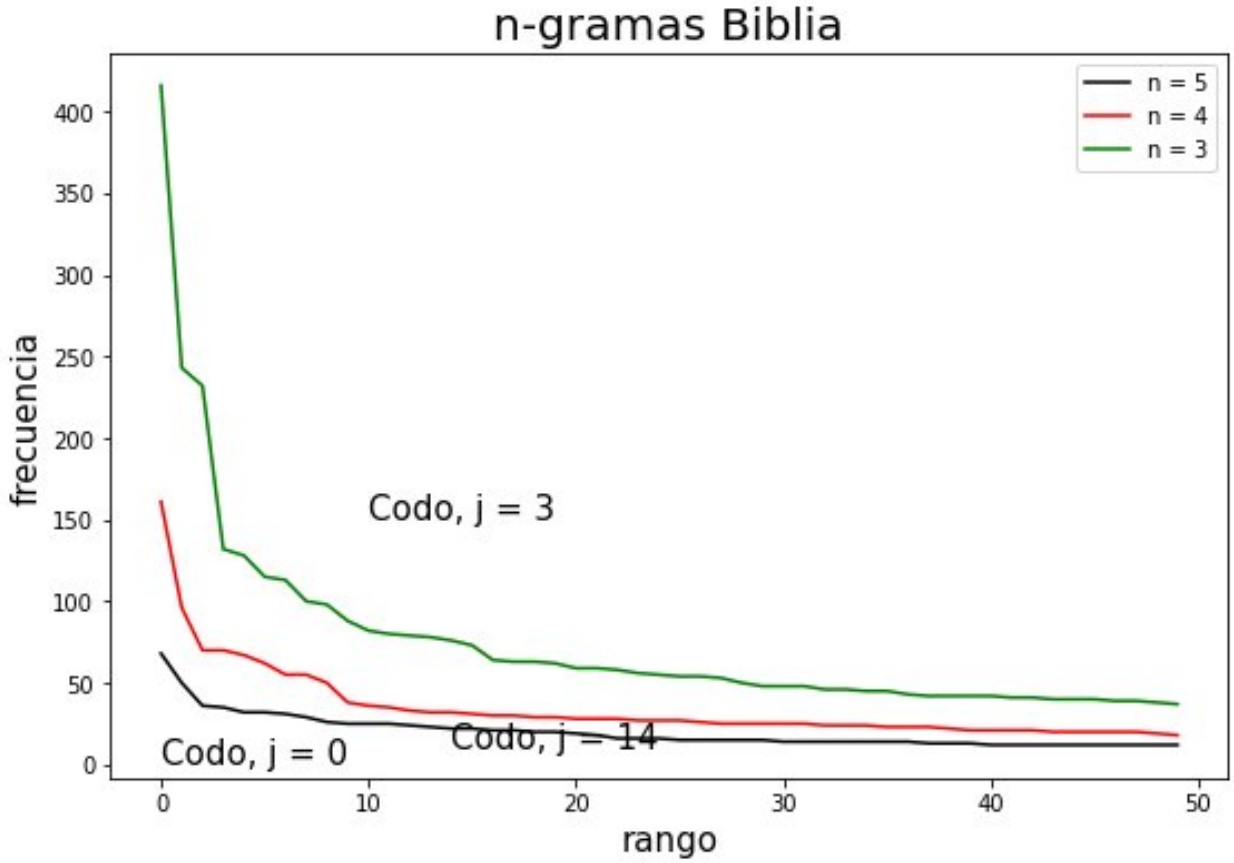


Ilustración 35: Codo n-gramas etiquetados Biblia

En el grueso del trabajo mediante exposición teórica y posterior comprobación empírica ver *Fijación del número j de n-gramas etiquetados más significativos. 3.5.4 s*e ha justificado la

elección de *n = 4* como el parámetro adecuado para equilibrar, a la vez, la significación estadística de los 4-gramas etiquetados extraídos que, se recuerda, funcionarán como *términos* y la capacidad de estos de proveer un contexto de desambigüación lo suficientemente amplio. De parecida forma, se introdujo el "*método del codo"* para la fijación del número *j* de 4-gramas que funcionarán como *términos*. En la *Ilustración 35: Codo n-gramas etiquetados Biblia* se justifica la elección de *j = 14* como el parámetro adecuado para la extracción en el ámbito bíblico de los *j n-gramas* (14, 4-gramas)etiquetados mas frecuentes. Aunque ha de decirse que se ha escogido un *j* mayor del *j* que estrictamente hace el mayor codo. Puede esto considerarse con una primera intervención del experto.

Puede verse en:

- bib4gramasEtiquetados.pickle
- bib4gramasEtiquetados.txt

la lista ordenada de 4-gramas etiquetados con su correspondiente frecuencia.

Se aporta, para mayor comodidad los 14 n-gramas etiquetados ordenados por su frecuencia. Debe, en otra intervención del experto u otra instancia de abstracción superior examinarse si existen locuciones, modismos o refranes en la lista de las 14 4-gramas; si es así deberían desecharse como términos de los que extraer conceptos. Se aprecia que no es el caso y por tanto como un nuevo hito en este trabajo se dejan fijadas los 14 4-gramas como términos.

```
[((('thus', 'r'), ('saith', 'v'), ('lord', 'n'), ('god', 'n')), 161),
 ((('lord', 'n'), ('spake', 'n'), ('unto', 'a'), ('moses', 'n')), 96),
 ((('spake', 'n'), ('unto', 'a'), ('moses', 'n'), ('say', 'v')), 70),
 ((('thus', 'r'), ('saith', 'v'), ('lord', 'n'), ('host', 'n')), 70),
 ((('say', 'v'), ('thus', 'r'), ('saith', 'v'), ('lord', 'n')), 67),
 ((('therefore', 'r'), ('thus', 'r'), ('saith', 'v'), ('lord', 'n')), 62),
 ((('lord', 'n'), ('say', 'v'), ('unto', 'a'), ('moses', 'n')), 55),
 ((('word', 'n'), ('lord', 'n'), ('come', 'v'), ('unto', 'r')), 55),
 ((('lord', 'n'), ('come', 'v'), ('unto', 'r'), ('say', 'v')), 50),
 ((('unto', 'r'), ('say', 'v'), ('son', 'n'), ('man', 'n')), 38),
 ((('come', 'v'), ('unto', 'r'), ('say', 'v'), ('son', 'n')), 36),
 ((('write', 'v'), ('book', 'n'), ('chronicle', 'n'), ('king', 'n')), 35),
 ((('year', 'n'), ('old', 'a'), ('begin', 'v'), ('reign', 'v')), 33),
 ((('unto', 'a'), ('moses', 'n'), ('say', 'v'), ('speak', 'n')), 32)]
```

Ilustración 36: Biblia 14 4-gramas etiquetados como términos

7.3 Extracción de conceptos a partir de los términos.

7.3.1 Extracción de conceptos(tripletas conceptuales) a partir de tripletas contextuales(términos)

En el apartado *Extracción de la lista de tripletas contextuales actuando como términos. 7.2.1* se ha presentado como hito la expresión del Corpus temático (la Biblia en este apéndice) en forma de lista de tripletas contextuales; ahora se persigue el hito de mediante el *Algoritmo de búsqueda de un concepto - tripleta conceptual - a partir de una tripleta contextual 1* expresar esa lista en forma de tripletas conceptuales. Esta lista -reforzando lo expuesto en este trabajo - , ahora sí, agrupando por el tercer componente de la tripleta conceptual - concepto, synset - tendrá pocos elementos especialmente frecuentes que formarán una distribución de frecuencias cercana a una ley de Zipf lo que significa que con pocos conceptos se representarán gran parte de los términos del corpus. Si se fija como umbral representar el 50% de los términos *(374761 tripletas contextuales)* se tiene que para $p = 213$ en la lista:

$$\sum_{i=1}^{p} f_i \geq 374761 * 0{,}5 \quad y \quad \nexists p' < p \, tal \, que \sum_{i=1}^{(p')} f_i > 374761 * 0{,}5$$

Por lo tanto se confirma lo presentado en este trabajo pues con 213 conceptos ($p= 213$) pueden representarse $374761 \times 0{,}5 = 187381$ términos. Es muy aclaratorio consignar que hay 18233 tripletas conceptuales distintas(conceptos) y por ende:

$$\sum_{i=1}^{18233} f_i = 374761$$

Se presenta en forma de fichero una lista dual de tripletas conceptuales con su frecuencia tal como se consignó en el apartado *3.5.2 Fijación del número p de tripletas-conceptuales más significativas.*

biblTConcFrecBiblia.pickle

A continuación, para mayor facilidad de lectura y de modo ilustrativo, se consignan las 70 primeras tripletas (ver *Ilustración 37*)conceptuales ordenadas por su frecuencia.

Se hace notar, se desarrollará más adelante, que existe un elevado número de tripletas a las que no ha podido asignársele un concepto -synset -. Tras la semblanza hecha de la Biblia del Rey Jacobo se puede adivinar que aparecerán como particularidad en este Caso de Uso gran presencia de arcaísmos.

```
[(('shall', 'v', 'No hay synset'), 9838), (('lord', 'n', "Synset('overlord.n.01')"), 7808),
(('unto', 'r', 'No hay synset'), 5574), (('say', 'v', "Synset('suppose.v.01')"), 5475),
(('thy', 'a', 'No hay synset'), 3741), (('come', 'v', "Synset('occur.v.02')"), 3727),
(('son', 'n', "Synset('son.n.02')"), 3440), (('thou', 'n', "Synset('thousand.n.01')"), 3063),
(('thee', 'n', 'No hay synset'), 2972), (('god', 'n', "Synset('idol.n.01')"), 2913),
(('unto', 'a', 'No hay synset'), 2708), (('upon', 'r', 'No hay synset'), 2662),
(('day', 'n', "Synset('sidereal_day.n.01')"), 2609), (('go', 'v', "Synset('rifle.v.02')"), 2535),
(('king', 'n', "Synset('king.n.09')"), 2520), (('people', 'n', "Synset('multitude.n.03')"), 2136),
(('man', 'n', "Synset('world.n.08')"), 2108), (('israel', 'n', "Synset('israel.n.02')"), 2080),
(('one', 'n', "Synset('one.n.02')"), 1986), (('child', 'n', "Synset('child.n.04')"), 1946),
(('house', 'n', "Synset('house.n.09')"), 1909), (('hand', 'n', "Synset('handwriting.n.01')"), 1907),
(('also', 'r', "Synset('besides.r.02')"), 1769), (('ye', 'n', 'No hay synset'), 1728),
(('take', 'v', "Synset('take.v.24')"), 1722), (('make', 'v', "Synset('make.v.15')"), 1704),
(('thing', 'n', "Synset('thing.n.12')"), 1630), (('give', 'v', "Synset('render.v.04')"), 1538),
(('father', 'n', "Synset('founder.n.02')"), 1459), (('us', 'r', 'No hay synset'), 1451),
(('even', 'r', "Synset('even.r.04')"), 1388), (('men', 'n', "Synset('world.n.08')"), 1354),
(('city', 'n', "Synset('city.n.03')"), 1258), (('hath', 'n', 'No hay synset'), 1257),
(('thou', 'a', 'No hay synset'), 1237), (('every', 'r', 'No hay synset'), 1236),
(('word', 'n', "Synset('word.n.07')"), 1229), (('shalt', 'n', 'No hay synset'), 1214),
(('let', 'v', "Synset('permit.v.01')"), 1189), (('land', 'n', "Synset('land.n.01')"), 1184),
(('ye', 'v', 'No hay synset'), 1164), (('know', 'v', "Synset('know.v.09')"), 1107),
(('great', 'a', "Synset('great.s.01')"), 1060), (('therefore', 'r', "Synset('therefore.r.01')"), 1047),
(('thou', 'v', 'No hay synset'), 1041), (('may', 'v', 'No hay synset'), 1027),
(('saith', 'v', 'No hay synset'), 999), (('name', 'n', "Synset('name.n.06')"), 969),
(('hath', 'v', 'No hay synset'), 950), (('jesus', 'n', "Synset('jesus.n.01')"), 939),
(('among', 'r', 'No hay synset'), 916), (('offering', 'n', "Synset('offer.n.01')"), 904),
(('hast', 'n', 'No hay synset'), 893), (('brother', 'n', "Synset('buddy.n.01')"), 886),
(('god', 'n', "Synset('god.n.03')"), 869), (('neither', 'r', 'No hay synset'), 859),
(('year', 'n', "Synset('year.n.02')"), 846), (('hear', 'v', "Synset('hear.v.03')"), 845),
(('way', 'n', "Synset('way.n.11')"), 839), (('servant', 'n', "Synset('servant.n.01')"), 835),
(('put', 'v', "Synset('put.v.01')"), 825), (('heart', 'n', "Synset('kernel.n.03')"), 811),
(('place', 'n', "Synset('topographic_point.n.01')"), 806), (('two', 'n', "Synset('deuce.n.04')"), 798),
(('accord', 'v', "Synset('harmonize.v.01')"), 797), (('bring', 'v', "Synset('bring.v.05')"), 790),
(('thus', 'r', "Synset('thus.r.02')"), 735), (('moses', 'n', "Synset('moses.n.02')"), 734),
(('earth', 'n', "Synset('ground.n.09')"), 699), (('david', 'n', "Synset('david.n.03')"), 696)]
```

Ilustración 37: Biblia: 70 tripletas conceptuales mas frecuentes

7.3.2 Extracción de conceptos(tripletas conceptuales a partir de j n-gramas etiquetados

En el apartado *7.2.2 Extracción de los j n-gramas etiquetados más frecuentes actuando como términos.* se ha sintetizado la lista de los *14 4gramas* más significativos. Ahora, mediante el algoritmo *2 Algoritmo de generación de tripletas conceptuales desde n-gramas etiquetados* se pretende extraer la lista de los conceptos asociados a dicha lista. Es interesante repasar la ilustración *N-grama etiquetado: 1 Término, n conceptos* para, de un vistazo, recordar el proceso. Se consigna que el número de conceptos(tripletas conceptuales) extraídas será menor de

$$14 \times 4 = 56$$

El resultado del proceso puede verse en el fichero: *tcp4por14setBiblia.pickle*

No obstante, dado que finalmente solo se han extraído 24 conceptos se presentan en este informe.

```
[('speak', 'n', 'No hay synset'),
 ('say', 'v', "Synset('suppose.v.01')"),
 ('therefore', 'r', "Synset('therefore.r.01')"),
 ('man', 'n', "Synset('world.n.08')"),
 ('reign', 'v', "Synset('reign.v.01')"),
 ('god', 'n', "Synset('idol.n.01')"),
 ('begin', 'v', "Synset('begin.v.10')"),
 ('word', 'n', "Synset('word.n.07')"),
 ('son', 'n', "Synset('son.n.02')"),
 ('lord', 'n', "Synset('overlord.n.01')"),
 ('unto', 'a', 'No hay synset'),
 ('moses', 'n', "Synset('moses.n.02')"),
 ('spake', 'n', 'No hay synset'),
 ('saith', 'v', 'No hay synset'),
 ('write', 'v', "Synset('write.v.10')"),
 ('come', 'v', "Synset('occur.v.02')"),
 ('king', 'n', "Synset('king.n.09')"),
 ('unto', 'r', 'No hay synset'),
 ('thus', 'r', "Synset('thus.r.02')"),
 ('book', 'n', "Synset('script.n.01')"),
 ('chronicle', 'n', "Synset('history.n.02')"),
 ('host', 'n', "Synset('master_of_ceremonies.n.01')"),
 ('year', 'n', "Synset('year.n.02')"),
 ('old', 'a', "Synset('old.a.02')")]
```

Ilustración 38: Biblia 24 conceptos extraídos de los 14 4gramas

7.3.3 Fusión de los conceptos extraídos mediante los términos del corpus y los conceptos extraídos de los j n-gramas etiquetados.

En los dos epígrafes anteriores se han sintetizado, respectivamente, dos listas de tripletas conceptuales. Una proveniente de la lista de términos(tripletas contextuales) a la que retomando la notación de *3.5.4 y 3.5.5* se llamará T1 y la otra proveniente de la lista de n-gramas etiquetados a la que se llamará T2. La lista final de la que extraer las ontologías lingüísticas ligeras será la formada, según se consignó en la presentación teórica y experimental de este trabajo, por[31]: $T1 \cup (T2 \setminus T1)$ Antes de proseguir convendría repasar la Ilustracion 26 : *Esquema de la elaboración de la lista de tripletas conceptuales mas significativas*

Si se observa $T2 \setminus T1$ se cae en la cuenta de que son muy pocos los conceptos recuperados frente a los que ya había recuperado T1. Por tanto, no aporta, numéricamente hablando, gran cosa(5 conceptos). *(VER Ilustración 39: T2 / T1 Biblia Conceptos aportados por los 14 4 -gramas*

```
T2 -T1
Out[15]:
{('begin', 'v', "Synset('begin.v.10')"),
 ('book', 'n', "Synset('bible.n.01')"),
 ('chronicle', 'n', "Synset('history.n.02')"),
 ('spake', 'n', 'No hay synset'),
 ('speak', 'n', 'No hay synset')}
```

Ilustración 39: T2 / T1 Biblia Conceptos aportados por los 14 4 -gramas

31 El ponente de este trabajo sabe perfectamente que $T1 \cup (T2 \setminus T1) = T2 \cup T1$ pero prefiere la expresión registrada por su mayor poder aclaratorio de lo que se pretende en este trabajo

No obstante si se observa ('book', 'n', "Synset('bible.n.01')") se debe caer en la cuenta de varias circunstancias que ilustran la parte mollar de este trabajo. Para ello primero se muestran todos los posibles *synsets*(conceptos) del lema *'book'*

Synset('book.n.01') a written work or composition that has been published (printed on pages bound together) ['book'] ['I am reading a good book on economics']

Synset('book.n.02') physical objects consisting of a number of pages bound together ['book', 'volume'] ['he used a large book as a doorstop']

Synset('record.n.05') a compilation of the known facts regarding something or someone ['record', 'record_book', 'book'] ["Al Smith used to say, `Let's look at the record'", 'his name is in all the record books']

Synset('script.n.01') a written version of a play or other dramatic composition; used in preparing for a performance ['script', 'book', 'playscript'] []

Synset('ledger.n.01') a record in which commercial accounts are recorded ['ledger', 'leger', 'account_book', 'book_of_account', 'book'] ['they got a subpoena to examine our books']

Synset('book.n.06') a collection of playing cards satisfying the rules of a card game ['book'] []

Synset('book.n.07') a collection of rules or prescribed standards on the basis of which decisions are made ['book', 'rule_book'] ['they run things by the book around here']

Synset('koran.n.01') the sacred writings of Islam revealed by God to the prophet Muhammad during his life at Mecca and Medina ['Koran', 'Quran', "al-Qur'an", 'Book'] []

Synset('bible.n.01') the sacred writings of the Christian religions ['Bible', 'Christian_Bible', 'Book', 'Good_Book', 'Holy_Scripture', 'Holy_Writ', 'Scripture', 'Word_of_God', 'Word'] ['he went to carry the Word to the heathen']

Synset('book.n.10') a major division of a long written composition ['book'] ['the book of Isaiah']

Synset('book.n.11') a number of sheets (ticket or stamps etc.) bound together on one edge ['book'] ['he bought a book of stamps']

Synset('book.v.01') engage for a performance ['book'] ['Her agent had booked her for several concerts in Tokyo']

Synset('reserve.v.04') arrange for and reserve (something for someone else) in advance ['reserve', 'hold', 'book'] ['reserve me a seat on a flight', 'The agent booked tickets to the show for the whole family', "please hold a table at Maxim's"]

Synset('book.v.03') record a charge in a police register ['book'] ['The policeman booked her when she tried to solicit a man']

Synset('book.v.04') register in a hotel booker ['book'] []

En primer lugar, mediante la desambiguación basada en el contexto del n-grama, debe consignarse que se ha recuperado el concepto adecuado al contexto bíblico:

- Primero mediante el etiquetado sintáctico, puesto que (book,'v') remite al acto de reservar alojamiento en ámbitos hoteleros y, desde luego, no procede en un ambiente bíblico. Así pues no generará en las Ontologías Linguísticas Ligeras ni sinónimos ni cohipónimos.

- En segundo lugar de todos las posibles semánticas para (‘book’, ‘n’) mediante el contexto provisto por el n-grama ha recuperado el concepto adecuado.

  Synset('bible.n.01') the sacred writings of the Christian religions ['Bible', 'Christian_Bible', 'Book', 'Good_Book', 'Holy_Scripture', 'Holy_Writ', 'Scripture', 'Word_of_God', 'Word'] ['he went to carry the Word to the heathen']

- En tercer lugar una vez recuperado el concepto adecuado debe observarse cuan importantes son los sinónimos aportados (*ver Ilustración 41 Sinónimos desambiguados de book)* y como de no ser por la contribución de dichos conceptos aportados por los n-gramas etiquetados un concepto particularmente importante en la temática en juego no habría sido tenido en cuenta.

Llegados hasta aquí, se ha alcanzado un hito fundamental: tener generada la lista de conceptos(tripletas conceptuales) con las que, a partir de dicha lista y con la ayuda de una ontología general externa – Wordnet en este caso de uso-, y, quizás del experto, generar las ontologías lingüísticas ligeras de sinónimos y cohipónimos. Esta lista que tendrá 218 elementos distintos, según se ha ido justificando y contruyendo puede examinarse y recuperarse de:

lTCpBibliaFinal.pickle

No obstante, debe decirse que de esas 218 tripletas conceptuales a 46 de los mismas no se les ha podido asignar un concepto (synset). Por tanto, se generarán las ontologías lingüisticas ligeras con forma de diccionarios de sinónimos y cohipónimos a partir de una lista de 172 tripletas conceptuales.

7.4 Generación de la ontología lingüistica ligera de sinónimos.

A partir de las 172 tripletas conceptuales y siguiendo el *3 Algoritmo de elaboración del diccionario de sinónimos* se ha generado, con forma de diccionario, una ontología lingüística ligera con 416 entradas que puede recuperarse del fichero:

'dictSinBib.pickle'

La idea fundamental del algoritmo puede recordarse viendo *Ilustración 27 Elaboración del diccionario de sinónimos:*

De modo ilustrativo y abundando en el ejemplo de *(book, n)* se consigna como ha quedado resuelto esto en el diccionario generado, así como la figura de *Jesús* tan importante en la temática que nos ocupa. Debe observarse como recupera los sinónimos adecuados y no los impropios para la temática en cuestión: Las claves del diccionario son los sinónimos de los valores del mismo ("book" y "jesus") (*ver Ilustración 41: Sinónimos desambiguados de book Sinónimos desambiguados de Jesus)*

```
(('bible', 'n'), 'book'),
(('christian_bible', 'n'), 'book'),
(('book', 'n'), 'book'),
(('good_book', 'n'), 'book'),
(('holy_scripture', 'n'), 'book'),
(('holy_writ', 'n'), 'book'),
(('scripture', 'n'), 'book'),
(('word_of_god', 'n'), 'book'),
```

Ilustración 41: Sinónimos desambiguados de book

```
(('jesus_of_nazareth', 'n'), 'jesus'),
(('the_nazarene', 'n'), 'jesus'),
(('jesus_christ', 'n'), 'jesus'),
(('savior', 'n'), 'jesus'),
(('saviour', 'n'), 'jesus'),
(('good_shepherd', 'n'), 'jesus'),
(('redeemer', 'n'), 'jesus'),
(('deliverer', 'n'), 'jesus'),
```

Ilustración 40: Sinónimos desambiguados de Jesus

7.5 Generación de la ontología lingüística ligera de hipónimos.

A partir de las 172 tripletas conceptuales y siguiendo el algoritmo *4 Algoritmo de elaboración del diccionario de cohipónimos* se ha generado, con forma de diccionario, una ontología lingüística ligera con 919 entradas que puede recuperarse del fichero:

'dictHipBib.pickle'

La idea fundamental del algoritmo puede recordarse viendo *Ilustración 30: Elaboración del diccionario de cohipónimos*.

Se ilustra el trabajo con los cohipónimos obtenidos para el concepto "heaven" dado que es especialmente didáctico ver cuantos términos bíblicos recupera:

```
(('heavenly_city', 'n'), 'heaven'),
(('garden_of_eden', 'n'), 'heaven'),
(('elysium', 'n'), 'heaven'),
(('eden', 'n'), 'heaven'),
(('city_of_god', 'n'), 'heaven'),
(('bosom_of_abraham', 'n'), 'heaven'),
(('promised_land', 'n'), 'heaven'),
(('celestial_city', 'n'), 'heaven'),
(("abraham's_bosom", 'n'), 'heaven'),
(('paradise', 'n'), 'heaven'),
(('elysian_fields', 'n'), 'heaven'),
(('valhalla', 'n'), 'heaven'),
(('walhalla', 'n'), 'heaven'),
(('holy_city', 'n'), 'heaven'),
```

Ilustración 42: Cohipónimos de "heaven"

7.6 Reglas de asociación con conceptos(tripletas conceptuales).

Para tal empeño según, de modo general, se ha visto en el capitulo 4 *Reglas de asociación con conceptos(tripletas conceptuales).* debe formarse la matriz de transacciones ver *Tabla 7: Matriz transacciones(frases en forma de lista de tripletas conceptuales) versus*

atributos(tripletas conceptuales) de la que aprender las RA. Esta matriz de transacciones puede recuperarse del fichero:

bibMatrizTransaccionesTCp.pickle

Se presenta, aquí, una imagen que da cuenta de su estructura y dimensiones:

Out[4]:

	aaron/a/No hay synset	...	zuzims/n/No hay synset
0	0	...	0
1	0	...	0
2	0	...	0
3	0	...	0
4	0	...	0
...	...	...	...
30098	0	...	0
30099	0	...	0
30100	0	...	0
30101	0	...	0
30102	0	...	0

[30103 rows x 18233 columns]

Ilustración 43: Matriz de transacciones frases normalizadas versus Conceptos, Biblia

Como se ve la matriz consta de 30.103 filas que coincide con el número de frases del Corpus y 18.233 columnas que coincide con el número de conceptos distintos extraídos.

7.6.1 Fijación del soporte y la confianza:

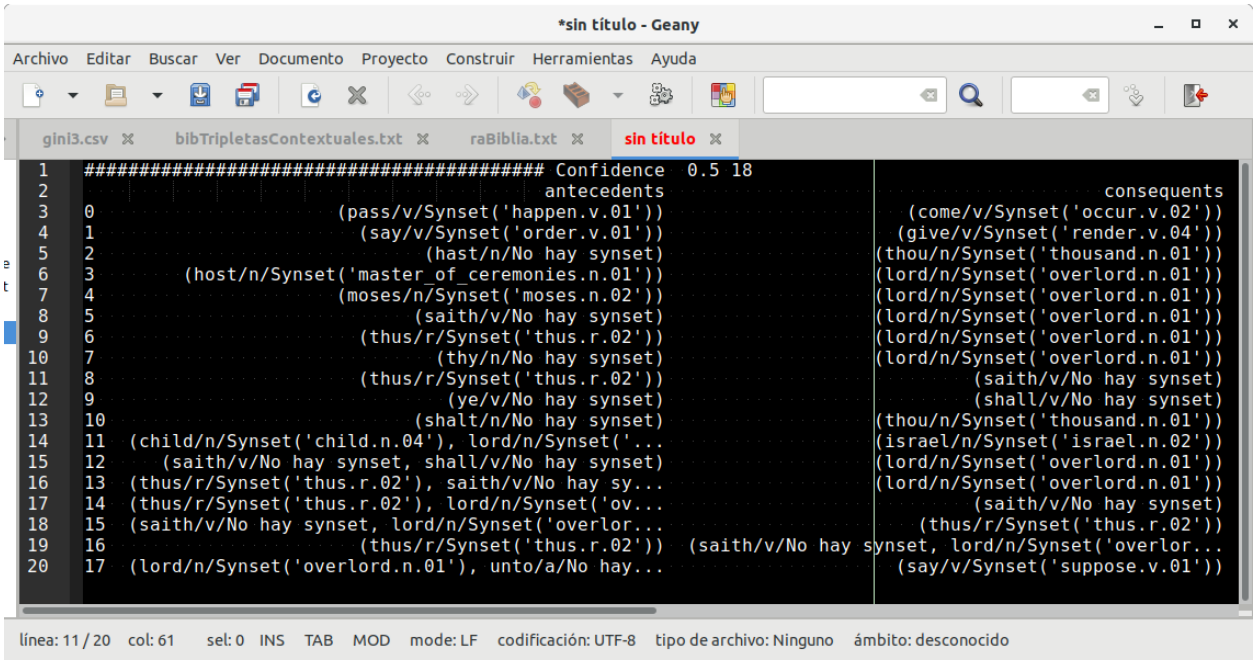
En la sección *4.2 Fijación del soporte y la confianza en una Matriz de transacciones Frases versus Tripletas Conceptuales.* se ha dado una expresión para fijar el soporte. En el caso que nos ocupa el número de transacciones(nº de frases del Corpus) es $n = 30103$ y para $p = 213$ $f_p = 279$. Así pues:

$$soporte = \frac{f_p}{n} \rightarrow \frac{279}{30103} = 0{,}00926 \approx 0{,}001$$

Para diversos niveles de confianza se obtienen el siguiente n.º de reglas:

0,8	5
0,5	18
0,3	82
0,15	213

Equilibrando que el nivel de confianza sea lo suficientemente alto como para no aportar relaciones espurias y que el el número de reglas permita aportar relaciones no obvias se fija la *confianza* a un nivel de 0,5. Además de que como se ve en la ilustración 44 la *Reglas de asociación Biblia: Confianza = 0.5, Soporte = 0.001* dicho nivel de confianza recupera relaciones que contienen tripletas conceptuales a las que no ha podido asignarles un concepto determinado(reglas 2,5,9 entre otras) y reglas que asocian conceptos con distinta etiqueta sintáctica(regla 5)




```
    ######################################## Confidence  0.5 18
                                        antecedents                                                    consequents
0                          (pass/v/Synset('happen.v.01'))                        (come/v/Synset('occur.v.02'))
1                           (say/v/Synset('order.v.01'))                        (give/v/Synset('render.v.04'))
2                                (hast/n/No hay synset)                      (thou/n/Synset('thousand.n.01'))
3       (host/n/Synset('master_of_ceremonies.n.01'))                      (lord/n/Synset('overlord.n.01'))
4                          (moses/n/Synset('moses.n.02'))                      (lord/n/Synset('overlord.n.01'))
5                                (saith/v/No hay synset)                       (lord/n/Synset('overlord.n.01'))
6                           (thus/r/Synset('thus.r.02'))                       (lord/n/Synset('overlord.n.01'))
7                                 (thy/n/No hay synset)                       (lord/n/Synset('overlord.n.01'))
8                           (thus/r/Synset('thus.r.02'))                              (saith/v/No hay synset)
9                                  (ye/v/No hay synset)                               (shall/v/No hay synset)
10                               (shalt/n/No hay synset)                      (thou/n/Synset('thousand.n.01'))
11  (child/n/Synset('child.n.04'), lord/n/Synset('...                      (israel/n/Synset('israel.n.02'))
12     (saith/v/No hay synset, shall/v/No hay synset)                      (lord/n/Synset('overlord.n.01'))
13  (thus/r/Synset('thus.r.02'), saith/v/No hay sy...                      (lord/n/Synset('overlord.n.01'))
14  (thus/r/Synset('thus.r.02'), lord/n/Synset('ov...                              (saith/v/No hay synset)
15  (saith/v/No hay synset, lord/n/Synset('overlor...                         (thus/r/Synset('thus.r.02'))
16                          (thus/r/Synset('thus.r.02'))  (saith/v/No hay synset, lord/n/Synset('overlor...
17  (lord/n/Synset('overlord.n.01'), unto/a/No hay...                          (say/v/Synset('suppose.v.01'))
```

Ilustración 44: Reglas de asociación Biblia: Confianza = 0.5, Soporte = 0.001

7.6.2 Comentarios a las RA extraídas.

En primer lugar han de cribarse las reglas según lo expuesto en *4.1 Utilidad y novedad de las reglas extraídas con tripletas conceptuales*. Por ello, las reglas 6, 8,15,16 se criban puesto que tienen un único adverbio. En el antecedente la 6,8 y 16 y en el consecuente la 15. Por tanto, no se considerarán para la generación de Ontología Lingüística Ligera alguna.

En segundo lugar, se comentan las reglas 5 *saith,n → lord,n* y 7 *thy,n → lord,n* porque dan cuenta de que se han extraído reglas con un patrón muy determinado con una relación semántica más profunda que las simples asociaciones de sinonimia e hiperonimia. Esta profundidad viene dada por varias circunstancias, se pormenorizan algunas:

1. Relacionan lemas con distinta etiqueta sintáctica.
2. Relaciona una tripleta conceptual a la que no ha podido asignársele un concepto con otra tripleta conceptual a la que sí.
3. *Saith* y *Thy* son arcaísmos en inglés y por tanto sería difícil extraer relaciones semánticas de estos términos con Ontologías Lingüísticas Generales. En este caso, las RA han encontrado una relación muy útil y difícilmente extraíble por otros paradigmas. Cabe

pensar que es *Lord (Dios, Ser supremo)* quien habla usando arcaísmos no en tanto como tales sino como lenguaje mayestático.

En tercer lugar se comenta la regla 0 entre verbos: Se aporta la regla y los *synsets* a los que pertenecen ambas tripletas de los que se recupera sus lemas y su glosa de definición.

(pass/v/Synset('happen.v.01')) --> (come/v/Synset('occur.v.02'))

Synset('happen.v.01') ['happen', 'hap', 'go_on', 'pass_off', 'occur', 'pass', 'fall_out', 'come_about', 'take_place'] come to pass

Synset('occur.v.02') ['occur', 'come'] come to one's mind; suggest itself

Éste es un claro ejemplo de desambigüación acertada de la polisemia. Entre todos los posibles synsets de (pass, v) y todos los posibles de (come, v) se ha extraido la asociación pertinente entre ellos para la temática bíblica. Obsérvese la definición(glosa) de *Synset('happen.v.01')* donde aparece "come" y como la intersección del conjunto de lemas de uno y otro concepto es no vacía. Sin embargo, tal relación no hubiera sido extraida con la relación de sinonimia ni de hipónimos-hiperónimos.

En cuarto lugar, las reglas 2 y 10 reflejan, de nuevo, relaciones entre arcaísmos. De forma indirecta permiten acercar los términos *shalt* y *hast* puesto que al ser ambos claves del diccionario de RA ampliarán las frases donde aparezcan con el término *thou* aportando cercanía semántica entre ambas frases. Ver *Ilustración 45: Biblia: OLL(diccioniario) con RA*

En quinto lugar las reglas 3 y 4 (*moses → lord* y *host → lord*) sin necesidad de conocer profundamente la teodicea judeo-cristiana aportan relaciones que sin ser perfectamente sinónimas si indican formas concurrentes de referirse al ser supremo o deidad máxima de la religión judeo-cristiana(moses,host, lord). Es más interesante de lo que en principio cabría suponer reflejar esto pues al tratarse de un corpus judeo-cristiano no se han extraido asociaciones como (buda → lord) o (alá → lord). De igual modo ayudan en la desambiguación de la polisemia pues remiten al ámbito religioso y no a la condición de título nobiliario de *lord.*

En sexto lugar se comentan las reglas 13,14,15,16, son reglas, obviamente, que proceden del conjunto frecuente { thus, saith, lord}. Las reglas 15, 16 ya fueron cribadas y no se incide más sobre ello. Sin embargo, las reglas 13,14 son, per se, interesantes pues permiten ilustrar como en este caso no es apropiado reducir estas reglas complejas a reglas simples en el modo expuesto en *4.4 Paso de reglas complejas a reglas simples.* sino que conviene reflejar en la Ontología Lingüística Ligera que la presencia de dos de las tres palabras del conjunto frecuente aunque una de ellas sea adverbio permite pensar en la tercera. Este razonamiento se hace extensivo a la regla 11 que relaciona a los "niños de Israel" (una forma de llamar al pueblo de Israel o los Israelitas solo en ambientes bíblicos o judeo-cristianos) con el término *Lord* en su acepción de deidad judeo-cristiana. Así si aparecen juntos, en la misma frase, los términos *child* y *lord* remiten, en esta temática, a i*srael*

Cabe, finalmente, comentar como las RA han permitido paliar, en parte, la debilidad consistente en no poder, a partir de la Ontología Lingüística General, asignar un synset a un par (lema, tag) frecuente. Se recuerda que estas tripletas sin synset no generarán entradas en los diccionario de sinónimos ni en los de cohipónimos. Así, consignando que de la lista de 218 tripletas conceptuales a partir de las cuales se construyen las ontologías linguísticas ligeras, 46 de las mismas no tienen concepto -synset - asociado es interesante notar como 6 de esos 46 han sido reflejados en la OLL en forma de diccionario construido con las reglas de asociación (*hast, saith, thy, ye, shalt, unto*).

7.6.3 Ontología lingüística ligera generada con RA para la temática religiosa judea cristiana :

Una vez extraídas las reglas de asociación solo queda, mediante el *Texto 5: Algoritmo de elaboración del diccionario con RA* elaborar la ontología lingüística ligera con forma de diccionario ver *Ilustración 45: Biblia: OLL(diccioniario) con RA*. Puede verse el fichero:

dictBibliaRA.py

```
dictBibliaRA = {
      ('pass','v') : 'come',
      ('say','v') : 'give',
      (('child', 'n' ),('lord','n')) : 'israel',
      ('hast','n'):'thou',
      ('host', 'n'): 'lord',
      ('moses', 'n'): 'lord',
      ('saith', 'n'): 'lord',
      ('thy', 'n'): 'lord',
      ('ye', 'v'): 'shall',
      ('shalt', 'n'): 'thou',
      (('saith', 'v' ),('shall','n')) : 'lord',
      (('thus', 'r' ),('saith','v')) : 'lord',
      (('thus', 'r' ),('lord','n')) : 'saith',
      (('lord', 'n' ),('unto','a')) : 'say',
       }
```

Ilustración 45: Biblia: OLL(diccioniario) con RA

7.7 Intervención del experto.

La temática escogida para este capítulo no es un tema fácil, presenta particularidades -como la presencia frecuente de arcaismos-, que puede hacer especialmente apropiada la intervención de un experto. En este caso, un exégeta. No obstante, debe decirse que con los medios exclusivamente automáticos expuestos en los epígrafes anteriores y sin perjuicio de que pudieran ajustarse algunos parámetros se han obtenido Ontologías Lingüisticas Ligeras adecuadas y de un tamaño asumible.

La presencia frecuente de arcaísmos en este caso puede sugerir la necesidad de usar, además de la ontología lingüistica general, otra ontología externa preexistente como un diccionario de arcaísmos que, habitualmente, hará corresponder un arcaísmo con sinónimos de la lengua actual.

Allá donde no llegué este diccionario puede intervenir el exégeta aportando sinónimos y cohipónimos a las tripletas incompletas extraídas automáticamente.

Pueden verse las 46 tripletas sin concepto -synset- en: *'bibTripletasSinSinset.pickle'*

En este caso es más útil hacer converger los arcaísmos en su forma estándar actual. Así, como ilustración podría ampliarse el diccionario de sinónimos de la siguiente forma:

dict[‘saith’][v] = say

dict[‘shalt’][v] = shall

dict[‘hast’][v] = have

dict[‘spake’][v] = speak

7.8 Ilustración de la desambiguación de la polisemia.

A lo largo del trabajo se ha justificado profusamente como la elaboración de las ontologías lingüísticas ligeras temáticas debe proveer a la entidad que las use de desambiguación inmediata de la polisemia eligiendo, para ello, la ontología lingüística ligera apropiada a cada temática. Son varios los hitos que se han ido presentando a tales efectos. Ahora cabe compendiar los mismos a la vez que, de forma paralela, se ilustra con lo sucedido en la elaboración de las ontologías lingüísiticas ligeras en el caso bíblico. Antes de proseguir, debe remitirse al lector a la *Ilustración 8: Fijación de p. Desambiguación con p y etiqueta sintáctica.Ejemplo calar* donde se representó gráficamente el proceso de desambiguación mediante la elección de las *p* tripletas conceptuales mas frecuentes.

7.8.1 Elección de una temática lo suficientemente concreta.

El hecho de que la temática sea judeo-cristiana (mas restringida) que una temática religiosa general hace que aparezca como concepto destacado la figura de Jesús:

('jesus', 'n', "Synset('jesus.n.01')") 939

y que otras figuras de indudable peso religioso como Allah o Buda no aparezcan no ya como conceptos sino ni siquiera como términos. Esto implica, por ejemplo que los sinónimo de ('redeemer', 'n') o ('deliverer', 'n') sean *Jesus* y no *Allah.* Evitando, pues, que Allah sea considerado como un redentor lo que en un ambiente judeo-cristiano puede ser hasta herético.

```
for n,i in enumerate(wn.synsets('hand')):
    print(n,i, i.definition(), i.lemma_names())

0 Synset('hand.n.01') the (prehensile) extremity of the superior limb ['hand', 'manus', 'mitt', 'paw']
1 Synset('hired_hand.n.01') a hired laborer on a farm or ranch ['hired_hand', 'hand', 'hired_man']
2 Synset('handwriting.n.01') something written by hand ['handwriting', 'hand', 'script']
3 Synset('hand.n.04') ability ['hand']
4 Synset('hand.n.05') a position given by its location to the side of an object ['hand']
5 Synset('hand.n.06') the cards held in a card game by a given player at any given time ['hand', 'deal']
6 Synset('hand.n.07') one of two sides of an issue ['hand']
7 Synset('hand.n.08') a rotating pointer on the face of a timepiece ['hand']
8 Synset('hand.n.09') a unit of length equal to 4 inches; used in measuring horses ['hand']
9 Synset('hand.n.10') a member of the crew of a ship ['hand']
10 Synset('bridge_player.n.01') a card player in a game of bridge ['bridge_player', 'hand']
11 Synset('hand.n.12') a round of applause to signify approval ['hand']
12 Synset('hand.n.13') terminal part of the forelimb in certain vertebrates (e.g. apes or kangaroos); -
Springfield (Mass.) Union ['hand']
13 Synset('hand.n.14') physical assistance ['hand', 'helping_hand']
14 Synset('pass.v.05') place into the hands or custody of ['pass', 'hand', 'reach', 'pass_on', 'turn_over',
'give']
15 Synset('hand.v.02') guide or conduct or usher somewhere ['hand']
```

Ilustración 46: Synsets de "hand"

('hand', 'n', "Synset('handwriting.n.01')") 1907
('hand', 'n', "Synset('hand.n.07')") 12
('hand', 'v', "Synset('pass.v.05')") 10

Ilustración 47: Tripletas Conceptuales con "hand"

7.8.2 Desambiguando con la etiqueta sintáctica.

Dado que en el caso bíblico se considerarán para generar los diccionarios los *p =213* conceptos(tripletas conceptuales) más frecuentes y a la vista de *Ilustración 47: Tripletas Conceptuales con "hand"* puede observarse como el hecho de que la frecuencia de ('hand', **'v'**, "Synset('pass.v.05')") 10) no la coloque entre las 213 tripletas conceptuales mas frecuentes hará que no se generen entrada en las OLL (tanto sinónimos como cohipónimos) y por tanto, no haya ni sinónimos ni hipónimos para *hand* como verbo. Debe entenderse que el concepto que determina ('hand', **'v'**, "Synset('pass.v.05')") no es un concepto significativo para el ambiente bíblico.

Synset('pass.v.05') place into the hands or custody of ['pass', 'hand', 'reach', 'pass_on', 'turn_over', 'give']

7.8.3 Desambiguando con el contexto y con p.

Si se observa la *Ilustración 46: Synsets de "hand"* se comprueba cuantas posibles acepciones tiene el lema “hand” y cuantas de ellas, además, son sustantivos (‘n’); en concreto 14. Si se hace lo propio con la *Ilustración 47: Tripletas Conceptuales con "hand"* se observa que dos de las dos posibles acepcion de *hand* como sustantivo han aparecido en la temática bíblica.

Es decir 12 posibles acepciones no han formado ya conceptos y por tanto es la temática bíblica la que ha desambiguado. Quedan, pues, dos acepciones que han sido recogidas en la temática bíblica. Una de ellas, aparece entre los 213 tripletas conceptuales mas frecuentes

Synset('handwriting.n.01') something written by hand ['handwriting', 'hand', 'script']

y por tanto generará entradas en los diccionarios y la otra, que no está entre las 213 tripletas conceptuales más frecuentes, no. *Synset('hand.n.07') one of two sides of an issue ['hand']*

Por tanto, $p = 213$ ha desambiguado entre estas dos posibles acepciones. Observando los synsets se toma cuenta de cuan acertada es la desambiguación pues en un entorno bíblico *hand* tomado como *manuscrito, escritura* es especialmente adecuado. Debe repararse como en un entorno general muchas de las acepciones desdeñadas serían mucho mas frecuentes y por tanto generarían entradas en los diccionarios.

Un segundo ejemplo dejará claro lo bien que ha desambiguado el sistema para la temática propuesta: Sea el lema "son" que aparece en dos *synsets* en la lista de conceptos. Estos son, además, los dos únicos conceptos que el lema *son* tiene en Wordnet. Aparecen en la posición $6 < 213$ ('son.n.02') y en la posición $1393 > 213$('son.n.01') con frecuencias respectivas 3440 y 31. Por tanto, la primera generará entradas en los diccionarios y la segunda no. *(Ver la Ilustración 48: Lema "son" en Wordnet y en la temática bíblica)*

Si se reflexiona se cae en la cuenta de que en un temática general, la acepción ('son.n.01') sería mucho mas común. No obstante, el hecho de que la acepción ('son.n.02') sea la mayoritaria solo podría suceder en un temática bíblica como ha sucedido en el caso de uso con lo que el sistema ha desambiguado correctamente. Así, puede verse como quedaría el diccionario de sinónimos con los sinónimos adecuados a la temática en la *Ilustración 48: Lema "son" en Wordnet y en la temática bíblica*.

```
6    (('son', 'n', "Synset('son.n.02')"), 3440),
 1393 (('son', 'n', "Synset('son.n.01')"), 31))]
Synset('son.n.01') a male human offspring ['son', 'boy']
Synset('son.n.02') the divine word of God; the second person in the Trinity
            (incarnate in Jesus) ['Son', 'Word', 'Logos']

dictSin('word', 'n'): son
ditSin(('logos', 'n'): son
```

Ilustración 48: Lema "son" en Wordnet y en la temática bíblica

8 Conclusiones y ampliaciones futuras.

8.1 Conclusiones.

En el capítulo 1 se han fijado una serie de hitos a alcanzar en pos de la consecución de ciertos objetivos. Estos hitos tienen una estructura jerárquica de tal manera que la consecución de determinados hitos descansa en la previa consecución de otros.

Antes de nada se sometió al corpus temático a una normalización, entre varios opciones, y tras la experimentación y la discusión teórica, y, en entre otras, se escogieron dos:

- Cribado general de *stopwords*. Se señaló y justificó la importancia de usar un solo conjunto de stopwords independientemente de la temática a tratar.
- Lematización. Se justificó su elección frente a la *stemizacion* porque eso permitió darle utilidad a los lemas extraídos mediante el uso de una ontología lingüística general. Dicha justificación puede ser considerada en trabajos similares.

Posteriormente se presentaron dos determinadas estructuras para albergar lo que en este trabajo se consideró *término*. Es decir, se han dado dos estructuras que se han demostrado muy viables para representar los términos de un texto y ayudar en la desambiguación de los mismos. Éstas son: la tripleta contextual y el n-grama etiquetado; así la tripleta contextual proveyó de un escalón de desambiguación mediante su segundo componente: la etiqueta sintáctica. Por

tanto, el corpus quedó representado mediante una lista de tripletas contextuales de longitud *m*. Sobre dicha lista no se consideró, antes de su paso a lista de tripletas conceptuales, útil hacer reflexiones sobre su distribución de frecuencias. Sobre la estructura de n-grama etiquetado se experimento para fijar el valor de n, de igual forma se experimentó para, sobre su distribución de frecuencias, dar una forma informada de escoger los *j* ngramas mas significativos. Dicho método se nominó como *"método del codo"* y puede ser tomado en cuenta para tareas similares. Este método se propuso tras descartar, mediante experimentación, que los términos expresados como n-gramas etiquetados siguieran una distribución cercana a una Ley de Zipf. Toda esa experimentación se llevó a cabo sobre tres corpora. Por tanto, puede concluirse que los n-gramas etiquetados para valores de n = 3,4,5 no siguen una Ley de Zipf en su distribución de frecuencias pero si pueden extraerse los *j* más significativos mediante el *método del codo.*

A continuación, se propusieron métodos para pasar de los términos extraídos con ambas estructuras a conceptos. Se propuso como estructura para representar los conceptos: la tripleta conceptual. Ésta se mostró muy útil pues hacía converger muchas tripletas contextuales en una sola tripleta conceptual como se observó en los tres corpora de investigación y sobre dichas tripletas conceptuales, así consideradas, se pudo examinar su distribución de frecuencias; sobre éstas se comprobó en los corpora de investigación que siguen una distribución cercana a una Ley de Zipf. Esto es: se estableció que con pocos conceptos se pueden representar muchos términos; siendo, además, estos conceptos los más significativos de la temática dada. Subsiguientemente, se dió una forma de extraer los *p conceptos* mas significativos y se vió que esta *p* era pequeña de tal suerte que servía para tener controlado el tamaño de las ontologías ligeras a desarrollar y además servía para desambiguar los términos. A los conceptos capturados de esta forma se añadieron los conceptos extraídos de los términos expresados como n-gramas etiquetados; de nuevo sobre los corpora de experimentacion se fijo que estos n-gramas proveían de conceptos interesantes cuando *n = 4 y p* se fijaba para representar, como umbral, el 50% de los términos. Proponiendo estos valores como ideales. Con esta fijación de valores se comprobó que los conceptos aportados por las n-gramas etiquetados eran pocos pero especialmente significativos para la temática en cuestión.

Así pues, se concluyó que un corpus puede ser representado por una lista de tripletas conceptuales cuya cardinalidad vendrá dada por la intersección de las p tripletas conceptuales extraídas de las *m* tripletas contextuales y las tripletas conceptuales extraídas de los *j* 4-gramas etiquetados. Además *p* se fijó para un umbral del 50% de los términos y *j* mediante *el método del codo*.

En estas condiciones se hizo intervenir a la Ontología Lingüística General y con apoyo de la misma se extrajeron a partir de las tripletas conceptuales sinónimos e hiperónimos con los que formar sendas Ontologías Lingüísticas Ligeras Temáticas se propuso, y se implementó posteriormente en el Caso de Uso, una estructura de diccionario para albergar dichas ontologías ligeras de tal suerte que cada ontología – y cada diccionario -, representó una relación semántica. Esta representación en forma de diccionario resultó especialmente apta para los usos habituales que se propusieron para las ontolologías sintetizadas, esto es, *ampliación y sustitución*. Y, en ausencia de aparición de contraindicaciones, dicha estructura en forma de diccionario puede proponerse como primera opción a albergar otras ontologías que reflejen otras relaciones semánticas.

Sin embargo, esto no fue suficiente porque la aparición en la experimentación de un número no desdeñable de tripletas conceptuales a las que mediante la Ontología Lingüística General no se les había podido asignar concepto precipitó la necesidad de encontrar medios para relacionar estas tripletas entre si y entre dichas tripletas y otras con concepto asignado. Se propuso así el paradigma de las Reglas de Asociación y se dispuso el corpus para poder investigar si se encontraban asociaciones interesantes y novedosas en el sentido de que se encontraban relaciones que no habían podido ser encontradas con anterioridad. Se propuso que las relaciones que se pretendían extraer fuesen entre tripletas conceptuales. Así se aprovechó la previa extracción de valores para p para fijar los parámetros propios de este paradigma (confianza y soporte) y se encontró que:

- Efectivamente, encontraba relaciones entre tripletas conceptuales que no tenían asignado concepto.

- Encontraba relaciones entre conceptos con distinta etiqueta sintáctica proveyendo ,por tanto, un medio para solventar las limitaciones de la tabla Tabla 1: Relaciones-semánticas.
- Encontraba, en algunas ocasiones relaciones entre más de dos conceptos, dejando para posterior discusión como reflejar estas relaciones en el diccionario.
- Se comprobó que dada una determinada relación de asociación entre conceptos dicha asociación podía ayudar en la desambiguación de los términos.

8.2 Ampliaciones futuras.

Si se asume que puede considerarse sintetizado un buen método con unos parámetros adecuados para extraer la lista de los conceptos más significativos de una determinada temática mediante un corpus propicio pueden explorarse, a partir de dicha lista de conceptos, la síntesis de otras ontologías lingüísticas ligeras a partir de otras relaciones semánticas no consideradas en este trabajo. Naturalmente, esto puede hacerse teniendo en cuenta otras relaciones contempladas en la Ontología Lingüística General, por ejemplo: la implicación *(entailment)* o ampliando el trabajo con métodos para identificar automáticamente el tipo semántico de las relaciones extraídas; especialmente apto para comenzar esta ampliación futura podría ser considerar el trabajo contemplado con las RA. Puede continuarse este trabajo propuesto examinando si para dichas relaciones los usos propuestos aquí (ampliación y sustitución) resultan apropiados, por ejemplo contemplando las características expuestas en las tablas *8 Tabla 8: Ampliar, sustituir con Diccionario de sinónimos e hiperónimos y 9 Tabla 9: Ampliar y/o sustituir con RA*. o si caben otras aplicaciones más provechosas.

Por otro parte cabe, tomando este trabajo como base y en especial su capítulo *Revisión, corrección, supresión y ampliación de términos por parte del experto.* considerar lo necesario para que la intervención del experto sea, cada vez más, hecha por niveles de abstracción superiores automatizados y no puestas en manos del experto humano. Así, la elaboración o uso de ontologías que contemplen modismos, locuciones etc y que complementen a la Ontologia Linguïstica General pueden ser motivo de estudio.

Una mayor formalización matemática del *método del codo* o incluso el estudio de métodos alternativos para extraer el número adecuado de n-gramas etiquetados como términos es también un camino a explorar.

Bibliografía y referencias.

Bibliografía

Schreibman, 2015: Schreibman, Susan, Ray Siemens, and John Unsworth, A new companion to digital humanities., 2015

Zipf, 1949: Zipf, G. K., Human behavior and the principle of least effort: An introduction to human ecology., 1949

Erar,2002: Erar, Aydin, Bibliometrics or informetrics: displaying regularity in scientific patterns by using statistical distributions, 2002

Mandelbrot,1953: Mandelbrot, Benjamin, An informational theory of the statistical structure of language.Communication theory,

Pierce,2012: , An introduction to information theory: symbols, signals and noise., 2012

Condon,1928: CONDON, EDWARD U, Statistics of vocabulary. Science, 1928

Urb,2011: URBIZAGÁSTEGUI ALVARADO, Rubén; RESTREPO ARANGO, Cristina, La ley de Zipf y el punto de transición de Goffman en la indización automática, 2011

Corral,2015: CORRAL, Álvaro; BOLEDA, Gemma; FERRER-I-CANCHO, Ramon, Zipf's law for word frequencies: Word forms versus lemmas in long texts, 2015

Corcho,2000: Corcho O. Gómez Pérez A., "Evaluating knowledge representation and reasoningcapabilities of ontology specification languages,", 2000

Gruber, 1993: Gruber, T. R, A translation approach to portable ontology specifications, 1993

Guarino,1998: Guarino, N., Formal ontology and information systems, 1998

Lassila,2001: Lassila, Ora; McGuinness, Deborah, "The role of frame-based representa-tion on the semantic web", 2001

Cherfi,2002: Cherfi, Hacène, and Yannick Toussaint, How far association rules and statistical indices helpstructure terminology. In Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering , 2002

Moreno,2000: Juan Carlos Moreno Cabrera, Curso Universitario de Lingüistica general. , 2000

Miller,1990: Miller, G. A, WordNet: An on-line lexical database. Inter-national Journal of Lexicography,

Vossen,1998: P. Vossen, EuroWordNet: A multilingual database with lexical semantic networks,

Gantz,2010: J. Gantz and D. Reinsel, The Digital Universe Decade - Are You Ready?, 2010

Maedche,2001: A. Maedche and S. Staab, Ontology Learning for the Semantic Web.IEEE Intelligent systems, vol. 16, no 2, 2001

Wong, 2012: Wong, W., Liu, W., & Bennamoun, M., Ontology learning from text: A look back and into the future, 2012

Eco,1977: Eco,Umberto, Tratado de semiótica general, 1977

Wimmer,2013: Wimmer, H., & Zhou, L, Word sense disambiguation for ontology learning., 2013

Yarowsky,1992: Yarowsky, D., Word-sense disambiguation using statistical models of Roget's categories trained on large corpora., 1992

Witschel,2004: H.F. Witschel, Terminologie-Extraktion - Möglichkeiten der Kombination statistischer und musterbasierter Verfahren, 2004

Sánchez,2012: Sánchez, D., Batet, M., Isern, D., & Valls, A., Ontology-based semantic similarity: A new feature-based approach. Expert Systems withApplications, 2012

Lesk, 1986: Lesk, M., Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, 1986

Banerjee,2002: Banerjee, Satanjeev; Pedersen, Ted , Un algoritmo de Lesk adaptado para la desambiguación del sentido de palabras usando WordNet, 2002

Kageura,1996: Kageura, K., & Umino, B, Methods of automatic term recognition: A review, 1996

Nakagawa,2004: Nakagawa, H., & Mori, T, A simple but powerful automatic term extraction method. , 2002

Fukuda,1998: K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi,, Toward information extraction: identifying protein names from biological papers, 1998

Pecina,2010: Pecina, Pavel., Lexical association measures and collocation extraction., 2010

Barker,2000: Barker, K., & Cornacchia, N., Using noun phrase heads to extract document keyphrases. In conference of the canadian society for computational studies of intelligence, 2000

Witten,2005: Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G, Kea: Practical automated keyphrase extraction., 2005

Sarkar, 2016: Sarkar D., Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data. ,

Salton,2003: Salton, G., & Harman, D., Salton, G., & Harman, D. (2003). Information retrieval. In Encyclopedia of computer science , 2003

Bustos, s.f.: Alberto Bustos, ¿ Qué son las colocaciones ?, ,

Church,1990: Church, K., & Hanks, P., Word association norms, mutual information, and lexicography., 1990

Smadja,1991: Smadja, F., From n-grams to collocations: An evaluation of Xtract. In 29th Annual Meeting of the Association for Computational Linguistics ., 1991

Berland,1999: M. Berland and E. Charniak, Finding Parts in Very Large Corpora, in 37th AnnualMeeting of the Association for Computational Linguistics on ComputationalLinguistics}„ College Park, Maryland, 1999

Perkins,2010: Perkins, J. , Python text processing with NLTK 2.0 cookbook. , 2010

Agrawal,1993: Agrawal, R., Imieliński, T., & Swami, A, Mining association rules between sets of items in large databases, 1993

Michalski, 1983: Michalski R.S., Incremental generation of VL1 hypotheses: the underlying methodology and the description of program AQ11,

Pasquier,1999: Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L, Discovering frequent closed itemsets for association rules, 1999

Tan,2006: Tan, P. N., Steinbach, M., & Kumar, V., Data mining introduction., 2006

Kavalec,2005: M. Kavalec and V. Svátek, "A Study on Automated Relation Labelling in OntologyLearning,", 2005

Velardi,2005: P. Velardi, R. Navigli, A. Cuchiarelli, and F. Neri, Evaluation of ontolearn, amethodology for automatic population of domain ontologies,, 2005

Basu,2001: Basu, S., Mooney, R. J., Pasupuleti, K. V., & Ghosh, J., Evaluating the novelty of text-mined rules using lexical knowledge. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001

Ferllbaum,2010: Fellbaum, C, WordNet. In Theory and applications of ontology: computer applications, 2010

Luhn,1958: Luhn, H. P. , The automatic creation of literature abstracts. IBM Journal of research and development, 1958

Lesk, 1986: Lesk, M., Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, 1986

Casares, 1950: Casares,Julio, Introducción a la lexicografía moderna. Madrid: CSIC, 1950

Método para la construcción automática de ontologías lingüísticas ligeras basadas en corpus temáticos