



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Departamento de Comunicación Audiovisual, Documentación e Historia del Arte

**AMPLIACIÓN Y PERFECCIONAMIENTO DE LOS
MÉTODOS CUANTITATIVOS Y LEYES CLÁSICAS EN
RECUPERACIÓN DE LA INFORMACIÓN:
DESARROLLO DE UN SISTEMA DE INDIZACIÓN Y
SEGMENTACIÓN AUTOMÁTICA PARA TEXTOS EN
ESPAÑOL**

TESIS DOCTORAL

Presentada por:
Dña. Manuela Rodríguez Luna

Dirigida por:
Dr. D. José Llorens Sánchez

Valencia, Diciembre 2012

*A mi familia, mi marido y amigos
Porque siempre creyeron en mí.*

*Las personas grandes de espíritu e intelecto
pasan por la vida sin querer demostrar la grandeza
que esconden en su interior.
Me siento muy afortunada de haber conocido
a una de estas personas
José Llorens Sánchez.*

AGRADECIMIENTOS

Agradezco en primer lugar al director de esta Tesis Doctoral José Llorens Sánchez por todo su apoyo y dedicación, por el esfuerzo año tras año que ha podido suponer, por su paciencia, siempre acompañada de una sonrisa amable y por la gran capacidad que ha demostrado con una investigación tan novedosa e interdisciplinar.

A mi familia porque sólo ellos saben el esfuerzo y el tiempo dedicado a este trabajo de investigación.

Agradezco en especial a mi marido Marceliano Coquillat Mora, ingeniero y gran entusiasta de las matemáticas y las ciencias exactas, porque sin duda él me ayudó a comprender y vislumbrar esta ciencia tan apasionante y a aplicarla a mis investigaciones.

A mi príncipe azul, mi hijo Alejandro porque él es la razón de todo mi esfuerzo.

A todas mis amistades y gente querida por su apoyo, sus ánimos y sus buenos consejos.

RESUMEN

Se desarrolla e implementa un Sistema de Indización y Segmentación Automática para textos largos en español, contribuyendo a su categorización textual e indización automática.

Para su desarrollo, se estudian y perfeccionan los métodos cuantitativos y leyes clásicas en Recuperación de Información, como son los modelos relativos al proceso de repetición de palabras (Zipf, 1949), (Mandelbrot, 1953) y al proceso de creación de vocabulario (Heaps, 1978). Se realiza una crítica de las circunstancias de aplicación de los modelos y se estudia la estabilidad de los parámetros de manera experimental mediante recuentos en textos y sus fragmentos. Se establecen recomendaciones a priori para los valores de sus parámetros, dependiendo de las circunstancias de aplicación y del tipo de texto analizado. Se observa el comportamiento de los parámetros de las fórmulas para vislumbrar una relación directa con la tipología de texto analizado. Se propone un nuevo modelo (Log-%) para la visualización de la distribución de frecuencias de las palabras de un texto.

El objetivo final es detectar los cambios temáticos que se producen en un documento, para establecer su estructura temática y obtener la indización automática de cada una de sus partes. De este modo, se obtiene la categorización del texto o documento utilizando la enumeración de sus partes temáticas a modo de niveles o estructura arbórea.

Una vez constituidas las partes temáticas del texto en sus niveles correspondientes con los términos indizados, estos se agrupan en bloques distribuidos jerárquicamente según se desglose el documento en cuestión. El bloque inicial describe el contenido global de todo el documento con una cantidad inicial de palabras o descriptores. Seguidamente este bloque inicial se subdivide en varios bloques, los cuales corresponden a distintas partes del documento total, cada uno de estos también contiene una serie de palabras que describe el contenido y así sucesivamente hasta poder formar las divisiones necesarias y llegar a describir cada párrafo del documento en cuestión.

Los *términos* que finalmente formarán parte del mapa temático o Sistema de Indización y Segmentación Automática serán una combinación de palabras obtenidas del texto y coocurrencias de palabras que superen los umbrales adecuados. Los términos quedan colocados automáticamente en cada nivel de Segmentación utilizando similitudes entre ellos y la representación Log-% anteriormente citada.

Esta Tesis doctoral, no solo consta de una base conceptual teórica sobre indización y segmentación automática sino en la implementación y crítica de las aplicaciones informáticas que proporcionan la base para las experimentaciones de esta investigación.

ABSTRACT

It develops and implements an Automatic Indexing and Segmentation System for long Spanish texts, contributing to the categorization and automatic textual indexing.

For its development, study and improvement of quantitative methods, classic law retrieval and information such as models relating to process repetition of words (Zipf, 1949) (Mandelbrot, 1953) and the vocabulary creation process (Heaps, 1978). It is a critique of the circumstances of the application models and the study of the stability of the experimental parameters by word counts and fragments. It is to establish recommendations that are set to the priority values of its parameters, depending on circumstances of application and type of text analyzed. It observes the behaviour of the parameters formulas to discern a direct relationship to the type of text analysis. The new proposed model (log-%) is to visualize the distribution of frequencies of words of text.

The ultimate goal is to identify thematic changes that produce a document to establish its structure topic and get the automatic indexing of each of its parts. Thus, we obtain the categorization text or document using a list of its thematic parts level or as a tree structure.

Once formed the thematic parts of the text in their levels corresponding to the indexed terms, these blocks are grouped hierarchically and distributed according to the break down of the document in question. The initial block describes the overall content of the entire document with an initial amount of words or descriptors. Next this initial block is subdivided into several blocks, which correspond to different parts of the total document, each of these also contains a number of words describing the content and so on to form the necessary divisions and to reach a description of each paragraph of the document.

The terms which will ultimately form part of the thematic map or Automatic Indexing and Segmentation System will be a combination of words obtained from text co-occurrence with words that exceed the appropriate threshold. The terms are automatically placed at each level of segmentation using similarities between them and the Log-% mentioned above.

This doctoral thesis not only consists of a conceptual base theoretical indexing and automatic segmentation but implementation and review of the computer applications that provides the basis for experiments of this research.

RESUM

Es desenvolupa i implementa un Sistema d'Indexació i Segmentació Automàtica per a textos llargs en Espanyol, contribuint a la seua categorització textual i indexació automàtica.

Per al seu desenvolupament, s'estudien i perfeccionen els mètodes quantitius i lleis clàssiques en Recuperació d'informació, com són els models relatius al procés de repetició de paraules (Zipf, 1949), (Mandelbrot, 1953) i al procés de creació de vocabulari (Heaps, 1978). Es realitza una crítica de les circumstàncies d'aplicació dels models i s'estudia l'estabilitat dels paràmetres de manera experimental mitjançant recomptes en textos i els seus fragments. S'estableixen recomanacions a priori per als valors dels seus paràmetres, depenent de les circumstàncies d'aplicació i del tipus de text analitzat. S'observa el comportament dels paràmetres de les fórmules per a entreveure una relació directa amb la tipologia de text analitzat. Es proposa un nou model (Log-%) per a la visualització de la distribució de freqüències de les paraules d'un text.

L'objectiu final és detectar els canvis temàtics que es produeixen en un document, per a establir la seua estructura temàtica i obtenir la indexació automàtica de cadascuna de les seues parts. D'aquesta manera s'obté la categorització del text o document utilitzant l'enumeració de les seues parts temàtiques a mode de nivells o estructura arbòria.

Una vegada constituïdes les parts temàtiques del text en els seus nivells corresponents amb els termes indizats, estos s'agrupen en blocs distribuïts jeràrquicament segons es desglossa el document en qüestió. El bloc inicial descriu el contingut global de tot el document amb una quantitat inicial de paraules o descriptors. Seguidament aquest bloc inicial es subdivideix en diversos blocs, els quals corresponen a distintes parts del document total, cadascun d'aquests també conté una sèrie de paraules que descriuen el contingut i així successivament fins a poder formar les divisions necessàries i arribar a descriure cada paràgraf del document en qüestió.

Els termes que finalment formaran part del mapa temàtic o Sistema d'Indexació i Segmentació Automàtica seran una combinació de paraules obtingudes del text i coocurrències de paraules que superen els llindars adequats. Els termes queden col·locats automàticament en cada nivell de Segmentació utilitzant similituds entre ells i la representació Log-% anteriorment citat.

Esta Tesi doctoral no solament consta d'una base de dades conceptual teòrica sobre indexació i segmentació automàtica, sinó en la implementació i crítica de les aplicacions informàtiques que proporcionen la base per a les experimentacions d'esta investigació.

0. INTRODUCCIÓN, VISIÓN GLOBAL E HIPÓTESIS	1
1. OBJETIVOS.....	7
2. METODOLOGÍA.....	9
3. ANTECEDENTES Y ESTADO DEL CONOCIMIENTO.....	15
3.1. MODELOS DE RECUPERACIÓN DE INFORMACIÓN	17
3.1.1. <i>Modelo booleano</i>	17
3.1.2. <i>Lógica difusa</i>	19
3.1.3. <i>Clustering</i>	22
3.1.4. <i>Modelo de espacio vectorial</i>	25
3.1.4.1. Inverse Document Frequency. IDF.....	28
3.1.5. <i>Indización semántica latente</i>	31
3.1.6. <i>Redes neuronales</i>	32
3.1.7. <i>Modelo probabilístico</i>	34
3.1.8. <i>Redes bayesianas</i>	35
3.1.9. <i>Algoritmos genéticos</i>	35
3.1.10. <i>Procesamiento del lenguaje natural</i>	36
3.1.11. <i>Sistemas expertos</i>	37
3.2. ANÁLISIS DOCUMENTAL	38
3.2.1. <i>El análisis documental y sus niveles</i>	38
3.2.2. <i>Lingüística documental</i>	40
3.2.3. <i>La terminología científica y técnica</i>	40
3.2.4. <i>Orígenes de los lenguajes documentales</i>	41
3.2.5. <i>Los nuevos lenguajes de representación del conocimiento</i>	46
3.2.5.1. Ontologías	46
3.2.5.2. Folksonomías	49
3.2.5.3. Taxonomías	51
3.2.6. <i>La indización</i>	51
3.2.6.1. La indización por materias y unitérminos. El Método Taube.....	53
3.2.6.2. La indización por descriptores	54
3.2.6.3. La indización por palabras clave. Folksonomías	57
3.2.7. <i>La indización automática</i>	57
3.2.8. <i>Metadatos. Topics Maps o Mapas Conceptuales</i>	62
3.2.9. <i>Los Tesauros</i>	65
3.2.9.1. Evolución histórica de los tesauros	67
3.2.9.2. Clases de tesauros	69
3.2.9.3. Estructura interna de los tesauros	69
3.2.9.4. Los componentes fundamentales de los tesauros	70
3.2.9.5. Tipos de relación entre términos	70
3.2.9.6. Diferencias entre Tesauros y Ontologías.....	71
3.2.9.7. Diferencias entre Tesauros y Folksonomías.....	71
3.2.9.8. Diferencias entre Tesauros y Taxonomías.....	72
3.2.10. <i>La Web Semántica o Web 2.0</i>	72
3.2.10.1. La interoperabilidad en la web semántica. SKOS en el entorno Linked Open Data.....	74
4. MÉTODOS CUANTITATIVOS Y LEYES CLÁSICAS EN RECUPERACIÓN DE LA INFORMACIÓN. ESTUDIOS DESTACADOS.....	79
4.1. ESTUDIOS DESTACADOS DE HEAPS. LA LEY DE HEAPS Y SU MODELO DE CRECIMIENTO DEL VOCABULARIO RESPECTO DEL TAMAÑO DEL TEXTO	79
4.2. ESTUDIOS DESTACADOS DE ZIPF. LA LEY DE ZIPF Y SU MODELO DE FRECUENCIAS DE PALABRAS EN UN TEXTO	81
4.3. CANTIDAD DE PALABRAS SIGNIFICATIVAS EN UN DOCUMENTO	84
4.4. PODER DISCRIMINATORIO DE LAS PALABRAS	86
4.5. FÓRMULAS RELATIVAS A LA SIMILITUD ENTRE PALABRAS	86
4.5.1. <i>Fórmula de Dice</i>	86
4.5.2. <i>Fórmula de Jaccard</i>	87
4.5.3. <i>Coefficiente de Semejanza del Coseno</i>	87
4.5.4. <i>Medida de Información Mutua (Mutual Information. MI)</i>	88
4.5.5. <i>La Media Geométrica</i>	88

5.	MODELO DE CRECIMIENTO DEL VOCABULARIO. LEY DE HEAPS	89
5.1.	ESTUDIOS CUANTITATIVOS RELACIONADOS CON PALABRAS EN UN TEXTO O COLECCIÓN DE TEXTOS SIMILARES. METODOLOGÍA ESTADÍSTICA DEL ESTUDIO CUANTITATIVO	90
5.1.1.	<i>El vocabulario o número de palabras distintas utilizadas en cada texto</i>	<i>90</i>
5.1.2.	<i>Estimación del valor del vocabulario V para un texto de tamaño P palabras</i>	<i>91</i>
5.1.3.	<i>Intervalo de confianza para el promedio de valores del vocabulario en la población de fragmentos de tamaño P</i>	<i>93</i>
5.1.4.	<i>Variabilidad en el vocabulario</i>	<i>93</i>
5.1.5.	<i>Relación entre los parámetros básicos del vocabulario y el tamaño del fragmento</i>	<i>94</i>
5.2.	DEDUCCIÓN DEL MODELO DE REPRESENTACIÓN DEL VOCABULARIO RESPECTO AL TAMAÑO DEL DOCUMENTO.....	96
5.2.1.	<i>Relación entre los parámetros básicos del vocabulario y el tamaño del fragmento</i>	<i>96</i>
5.2.2.	<i>Otros problemas consecuentes a la representación $V=V(P)$ como función potencial.....</i>	<i>99</i>
5.2.3.	<i>Otras evidencias a favor de la existencia de dos tramos de características distintas</i>	<i>100</i>
5.2.4.	<i>Representación de $V=V(P)$ como dos funciones potenciales, una en cada tramo.....</i>	<i>104</i>
5.2.5.	<i>Posibilidades de la representación como función potencial</i>	<i>106</i>
5.3.	ESTUDIO DEL MODELO Y UTILIZACIÓN PRÁCTICA	113
5.3.1.	<i>Estudio de las propiedades de la función potencial como modelo de $V=V(P)$.....</i>	<i>113</i>
5.3.2.	<i>Interpretación de los coeficientes de la función potencial $V=V(P)$.....</i>	<i>119</i>
5.3.3.	<i>Determinación, a priori, de la distribución de los valores del vocabulario en fragmentos extraídos de un texto</i>	<i>122</i>
5.3.4.	<i>Breve resumen de este capítulo.</i>	<i>125</i>
5.4.	ESTUDIOS COMPLEMENTARIOS: POSIBILIDADES DE UTILIZACIÓN DEL MODELO PARA LA EXTRAPOLACIÓN	126
6.	MODELO DE DISTRIBUCIÓN DE FRECUENCIAS. LEY DE ZIPF	129
6.1.	LA LEY DE ZIPF EN RELACIÓN CON EL ESTUDIO DE PALABRAS EN LOS DOCUMENTOS. ESTUDIOS DESTACADOS	129
6.2.	INTERÉS ACTUAL POR LOS POSTULADOS DE ZIPF. BIBLIOGRAFÍA BÁSICA Y BIBLIOGRAFÍA RECIENTE.....	134
6.3.	FORMAS DE PRESENTACIÓN GRÁFICA DE LA LEY DE ZIPF/MANDELBROT Y SIGNIFICADO DE LOS VALORES DE LOS PARÁMETROS	136
6.3.1.	<i>Representación clásica.....</i>	<i>136</i>
6.3.2.	<i>Representación logarítmica.....</i>	<i>138</i>
6.3.3.	<i>Representación transformada o Espectro de frecuencias.....</i>	<i>141</i>
6.3.4.	<i>Gráficas sintéticas y significado de las fórmulas de Zipf y Zipf-Mandelbrot.....</i>	<i>144</i>
6.3.4.1.	<i>Zipf.....</i>	<i>144</i>
6.3.4.2.	<i>Mandelbrot.....</i>	<i>148</i>
6.3.4.3.	<i>Comparativa de gráficas sintéticas para Zipf-Mandelbrot con Zipf como medio de visualización de la tendencia para el exponente (e) y el sumando (Σ).....</i>	<i>155</i>
6.4.	VALORES DE LOS PARÁMETROS EN LAS FÓRMULAS DE ZIPF/MANDELBROT PARA EL AJUSTE A LAS FRECUENCIAS DE LAS PALABRAS DE UN TEXTO.....	158
6.4.1.	<i>Representación Transformada. Visualización de ejemplos.....</i>	<i>166</i>
6.5.	VARIACIÓN DE LOS VALORES DE LOS PARÁMETROS SEGÚN EL TIPO DE TEXTO, SEGÚN EL TAMAÑO DEL TEXTO Y SEGÚN EL TRATAMIENTO DADO AL TEXTO	168
6.5.1.	<i>El comportamiento del exponente (e) de Zipf en los documentos.....</i>	<i>168</i>
6.6.	AJUSTE PARCIAL DE ZIPF POR TRAMOS	192
6.7.	MODELO LOG-% DE VISUALIZACIÓN DEL EFECTO DE FALTA DE PALABRAS EN RELACIÓN A LAS PREDICHAS POR ZIPF. NUEVA FORMA DE REPRESENTACIÓN.	199
6.7.1.	<i>Agrupamiento en tramos: Modelo Log-%</i>	<i>199</i>
6.7.2.	<i>Modelo Log-% : valores en el eje de abscisas.....</i>	<i>200</i>
6.7.2.1.	<i>Procedimiento preliminar para la representación del Modelo Log-%. Agrupamiento en tramos en el rango.</i>	<i>200</i>
6.7.2.2.	<i>Procedimiento definitivo para la representación del Modelo Log-%. Agrupamiento en tramos logarítmicos en el rango.</i>	<i>202</i>
6.7.3.	<i>Modelo Log-%: valores en el eje de ordenadas</i>	<i>204</i>
6.7.4.	<i>Punto de Transición (Transition Point).....</i>	<i>210</i>
6.8.	ESTUDIOS COMPLEMENTARIOS: CÁLCULOS FÓRMULAS AJUSTADAS DE ZIPF Y MANDELBROT. PASOS 1-4.....	211
6.9.	ESTUDIOS COMPLEMENTARIOS: CÓMO AFECTA LAS PALABRAS VACÍAS Y LA EXTRACCIÓN DE RAÍCES AL VALOR DEL EXPONENTE (E) DE MANDELBROT.....	213

7.	ESTUDIOS CUANTITATIVOS RELACIONADOS CON RAÍCES. STEMMERS.....	219
7.1.	MÉTODO DE VARIEDAD DE SUCESORES	220
7.2.	ADAPTACIÓN AL ESPAÑOL DEL MÉTODO DE VARIEDAD DE SUCESORES.....	222
7.3.	MÉTODO DE SUFIJOS DE LOVINS	223
7.4.	ADAPTACIÓN AL ESPAÑOL DEL MÉTODO DE SUFIJOS DE LOVINS.....	225
7.5.	EL ALGORITMO DE PORTER.....	226
7.6.	PROCESOS DE EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN (SRI) ..	227
7.7.	PROCESOS DE EVALUACIÓN DE LOS STEMMERS	231
7.8.	ESTUDIO DE LA DISTRIBUCIÓN DE FRECUENCIAS DE ZIPF Y MANDELBROT CON LOS STEMMERS: MÉTODO DE VARIEDAD DE SUCESORES (MVS), MÉTODO DE SUFIJOS (SU) Y MÉTODO DE EXTRACCIÓN TANTO POR CIENTO (TCP-%)	233
7.9.	ESTUDIO DE LA DISTRIBUCIÓN DE FRECUENCIAS DE ZIPF AJUSTADO A LAS FRECUENCIAS PRÓXIMAS AL PT (TRANSITION POINT) CON LOS STEMMERS: MÉTODO DE VARIEDAD DE SUCESORES (MVS), MÉTODO DE SUFIJOS (SU) Y MÉTODO DE EXTRACCIÓN TANTO POR CIENTO (TCP).....	249
7.10.	EVALUACIÓN DE LOS STEMMERS: MÉTODO DE VARIEDAD DE SUCESORES (MVS), MÉTODO DE SUFIJOS (SU) Y MÉTODO DE EXTRACCIÓN TANTO POR CIENTO (TCP), PRECISIÓN-EXHAUSTIVIDAD.	260
8.	INTRODUCCIÓN A LAS PALABRAS ASOCIADAS. LA RELACIÓN SEMÁNTICA Y CONCEPTUAL DE LOS TÉRMINOS	265
8.1.	LA SIMILITUD ENTRE PALABRAS.....	265
8.2.	LAS PALABRAS ASOCIADAS	267
8.3.	COLOCACIONES (COLLOCATIONS)	267
8.4.	GRANULARIDAD.....	269
8.5.	ESTIMACIÓN DE LAS PAREJAS DE RAÍCES EN UN TEXTO	269
8.5.1.	<i>Contando parejas de raíces asociadas</i>	<i>270</i>
8.5.2.	<i>La granularidad del texto y las parejas de raíces asociadas encontradas.....</i>	<i>272</i>
8.5.3.	<i>Selección de parejas de raíces asociadas que aportan información</i>	<i>275</i>
8.6.	RELACIÓN DEL NÚMERO DE PAREJAS DE RAÍCES ASOCIADAS Y EL TAMAÑO DEL TEXTO.....	277
8.7.	EL ESPECTRO DE SIMILITUDES	280
9.	SEGMENTACIÓN AUTOMÁTICA DEL TEXTO. IDENTIFICACIÓN DE CAMBIOS TEMÁTICOS.....	285
9.1.	SEGMENTACIÓN AUTOMÁTICA DEL TEXTO (TEXT SEGMENTATION).....	285
9.2.	ESTUDIO DIACRÓNICO DEL TEXTO Y SEGMENTACIÓN DEL TEXTO.....	286
9.2.1.	<i>Granularidad de los documentos.....</i>	<i>290</i>
9.3.	NOVEDAD DE LAS PALABRAS EN UN TEXTO.....	290
9.4.	APLICACIÓN A LA SEGMENTACIÓN AUTOMÁTICA DEL TEXTO.....	293
10.	SISTEMA DE INDIZACIÓN Y SEGMENTACIÓN AUTOMÁTICA DEL TEXTO: MALLOV.....	295
10.1.	SISTEMA DE INDIZACIÓN Y SEGMENTACIÓN AUTOMÁTICA MALLOV.....	295
10.1.1.	<i>Sistema de Indización y Segmentación Automática MALLOV: Procedimiento básico...</i>	<i>299</i>
10.1.2.	<i>Sistema de Indización y Segmentación Automática MALLOV: Procedimiento completo</i>	<i>306</i>
11.	DISCUSIÓN DE RESULTADOS.....	317
12.	CONCLUSIONES	321
	BIBLIOGRAFÍA	325
	APÉNDICES	335
	APÉNDICE I: TEXTOS	335
	APÉNDICE II: BASES DE DATOS	345
	APÉNDICE III: FORMULARIOS DESTACADOS DE LAS BASES DE DATOS	347

ÍNDICE DE FIGURAS

IV

Figura 1. Método Booleano.....	18
Figura 2. Clustering. Métodos no jerárquicos	23
Figura 3. Clustering. Métodos jerárquicos	24
Figura 4. Método de Espacio Vectorial. Producto escalar de dos vectores	27
Figura 5. Red Neuronal	33
Figura 6. Ley de Heaps. Crecimiento del vocabulario respecto al tamaño de un texto	80
Figura 7. Ley de Zipf. Distribución de frecuencias de palabras	83
Figura 8. Medida de Similitud del Coseno	87
Figura 9. Interpretación de los valores de los parámetros a y b de la función potencial	119
Figura 10. Procedimiento preliminar para representación del Modelo Log-%	200
Figura 11. Procedimiento definitivo para representación del Modelo Log-%	203
Figura 12. Base de datos del Modelo Log-%	205
Figura 13. Distribución de Zipf según valor del exponente.....	214
Figura 14. Gráfico Exhaustividad-Precisión.....	228
Figura 15. Novedad de palabras en texto de Cervantes	287
Figura 16. Novedad de palabras en tramos de distintos tamaños	288
Figura 17. Novedad de palabras en tramos de distintos tamaños: 600, 1.000, 3.000, 10.000 palabras ...	288
Figura 18. Ampliación tramo final de figura 17	289
Figura 19. Novedad de palabras (medias móviles).....	289
Figura 20. Novedad de palabras en tramos solapados	290
Figura 21. Superávit de vocabulario en tramos	290
Figura 22. Tasa de aparición de novedad de palabras	291
Figura 23. Tasa de novedad de palabras (recorrido inverso)	292
Figura 24. Puntos iniciales de asunto	292
Figura 25. Puntos iniciales de asunto (en forma de valores positivos).....	293
Figura 26. Niveles en Sistema de Indización y Segmentación automática MALLOV.....	297
Figura 27. Aplicación TOPOS, formulario AnalizaZipf	347
Figura 28. Aplicación RENOS, formulario Generar Representación Numérica	347
Figura 29. Aplicación RENOS, formulario Zipf	348
Figura 30. Aplicación RENOS, formulario Triple fórmula de Heaps	348
Figura 31. Aplicación ARENA, formulario Arena.....	349
Figura 32. Aplicación COPAS, formulario Calcular Similitudes entre Términos.....	349
Figura 33. Aplicación MALLOV, formulario MALLOV	350
Figura 34. Aplicación MALLOV, formulario Sistema de Indización y Segmentación Automática	350

ÍNDICE DE TABLAS

IV

Tabla 1. Ejemplo Método Booleano.....	18
Tabla 2. Ejemplo Lógica Difusa.....	20
Tabla 3. Ejemplo Método Espacio Vectorial.....	26
Tabla 4. Operaciones y resultados del análisis documental.....	39
Tabla 5. Clases de términos que componen un tesoro	70
Tabla 6. Distribución normal del vocabulario en una población de documentos de tamaño fijo	93
Tabla 7. Muestra del tamaño de los textos en escala logarítmica.....	94
Tabla 8. Muestra del tamaño de los textos en escala logarítmica	107
Tabla 9. Cantidad de palabras predichas por la función potencial en varios textos.....	112
Tabla 10. Situaciones posibles para averiguar el vocabulario de un texto	122
Tabla 11. Indicaciones globales sobre el vocabulario de los textos	123
Tabla 12. Diferencia entre los resultados obtenidos por la fórmula recomendada y la aplicación	124
Tabla 13. Fórmulas para la extrapolación.....	128
Tabla 14. Parámetros de la distribución de Zipf-Mandelbrot	155
Tabla 15. Resultados de los parámetros fórmulas Zipf y Mandelbrot ajustadas	159
Tabla 16. Relación del exponente en la fórmula de Zipf con la tipología del texto	174
Tabla 17. Tipología y características de los textos	175
Tabla 18. Leyenda gráfico núm. 65	177
Tabla 19. Ejemplos palabras más frecuentes en textos tipología científico de resúmenes	178
Tabla 20. Ejemplo palabras más frecuentes en texto tipología científico de resúmenes	179
Tabla 21. Ejemplo palabras más frecuentes en texto tipología literario	181

Tabla 22. Resultados del procedimiento preliminar para representación del Modelo Log-%	201
Tabla 23. Porcentaje de vocabulario agrupadas en cada tramo del Modelo Log-%	201
Tabla 24. Resultados del procedimiento definitivo para representación del Modelo Log-%	203
Tabla 25. Ejemplo de las subdivisiones en 10 tramos del Modelo Log-%	211
Tabla 26. Ejemplo Stemmer método de variedad de sucesores.....	221
Tabla 27. Condiciones para la raíz del algoritmo de Porter.....	227
Tabla 28. Valores obtenidos en la distribución de frecuencias texto Cervantes con diversos Stemmers .	246
Tabla 29. Valores obtenidos en la distribución de frecuencias de Zipf ajustada al PT (Punto de Transición) texto Cervantes con diversos Stemmers	258
Tabla 30. Leyenda gráfico núm. 109.....	260
Tabla 31. Cantidad de parejas de raíces teóricas en los documentos.....	270
Tabla 32. Cantidad de parejas de raíces teóricas y pre-asociadas en varios documentos	271
Tabla 33. Cantidad de parejas de raíces teóricas, pre-asociadas y asociadas con criterio estadístico y de repetición en varios documentos	276
Tabla 34. Parámetros para la triple fórmula de Heaps.....	308
Tabla 35. Colocaciones y valor de similitud de Dice	310
Tabla 36. Términos de indización	312
Tabla 37. Términos de indización en cada nivel	314

ÍNDICE DE GRÁFICOS

V

Gráfico 1. Ejemplo Ley de Heaps	80
Gráfico 2. Ley de Zipf. Distribución de frecuencias de palabras	83
Gráfico 3. Promedio del vocabulario para el logaritmo del tamaño	95
Gráfico 4. Crecimiento del vocabulario en escala logarítmica	96
Gráfico 5. Crecimiento del vocabulario en escala logarítmica ajustada a una función lineal.....	97
Gráfico 6. Crecimiento del vocabulario en escala logarítmica ajustada a dos funciones lineales	98
Gráfico 7. Derivada numérica y derivada simbólica de un texto.....	100
Gráfico 8. Variabilidad en los textos	101
Gráfico 9. Frecuencia de las palabras y la variabilidad en los textos	102
Gráfico 10. Frecuencia de las palabras y la variabilidad en los textos en escala logarítmica.....	102
Gráfico 11. Cociente de las frecuencias de las palabras y la variabilidad en los textos	103
Gráfico 12. Incompatibilidad de la representación como una función potencial.....	104
Gráfico 13. Representación como dos funciones potenciales.....	105
Gráfico 14. Representación como dos funciones potenciales de la palabra más frecuente	106
Gráfico 15. Representación como una función potencial	107
Gráfico 16. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico	108
Gráfico 17. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico	109
Gráfico 18. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario	109
Gráfico 19. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario	110
Gráfico 20. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico	110
Gráfico 21. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico	111
Gráfico 22. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico	111
Gráfico 23. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario.....	112
Gráfico 24. Función potencial como modelo de $V=V(P)$ en 3D	114
Gráfico 25. Función potencial como modelo de $V=V(P)$	115
Gráfico 26. Representación de dos funciones potenciales.....	116
Gráfico 27. Representación de los valores del exponente para distintas funciones potenciales.....	117
Gráfico 28. Representación de los valores del exponente para distintas funciones potenciales.....	118
Gráfico 29. Coeficientes de la función potencial en varias categorías de texto.....	120
Gráfico 30. Variación de las funciones potenciales para un mismo texto	124
Gráfico 31. Distribución de frecuencias de Zipf con texto $P=4.125$	136
Gráfico 32. Distribución de frecuencias de Zipf con texto $P=428$	137
Gráfico 33. Distribución de frecuencias de Zipf con texto $P=379.948$	138
Gráfico 34. Distribución logarítmica de frecuencias de Zipf con texto $P=4.125$	139
Gráfico 35. Distribución logarítmica de frecuencias de Zipf con texto $P=428$	140
Gráfico 36. Distribución logarítmica de frecuencias de Zipf con texto $P=379.945$	140
Gráfico 37. Distribución transformada de frecuencias de Zipf con texto $P=73.173$	143

V

Gráfico 38. Distribución logarítmica transformada de frecuencias de Zipf con texto P=73.173.....	143
Gráfico 39. Distribución de frecuencias de Zipf con distintos valores del exponente (e)	145
Gráfico 40. Distribución de frecuencias de Zipf con distintos valores del exponente (e). Rangos 3-20 ..	146
Gráfico 41. Distribución de frecuencias de Zipf con distintos valores del exponente (e). Rangos 40-51	147
Gráfico 42. Distribución de frecuencias de Zipf con distintos valores del exponente (e)	148
Gráfico 43. Distribución de Mandelbrot para un mismo exponente (e) distintos Σ	149
Gráfico 44. Distribución de Mandelbrot con $\Sigma=30$ y distinto exponente (e)	150
Gráfico 45. Distribución de Mandelbrot con distinto exponente (e) rangos 13-33.....	151
Gráfico 46. Distribución de Mandelbrot con distinto exponente (e) rangos mayores	151
Gráfico 47. Tendencia de la constante (k) en la distribución de Mandelbrot	153
Gráfico 48. Tendencia de la exponente (e) en la distribución de Mandelbrot	154
Gráfico 49. Tendencia del Σ en la distribución de Mandelbrot.....	154
Gráfico 50. Gráficas sintéticas Mandelbrot. Comparativa logarítmica	156
Gráfico 51. Gráfica sintética logarítmica Zipf ajustada.....	156
Gráfico 52. Gráficas sintéticas Zipf-Mandelbrot. Comparativa logarítmica Zipf ajustada	157
Gráfico 53. Gráficas sintéticas Zipf-Mandelbrot. Comparativa Zipf ajustada.....	158
Gráfico 54. Comparación logarítmica distribuciones Zipf, Mandelbrot y Zipf ajustada.....	160
Gráfico 55. Comparación distribuciones Zipf, Mandelbrot y textos reales	161
Gráfico 56. Comparación Zipf, Mandelbrot y textos reales rangos 1000-1200.....	162
Gráfico 57. Comparación Zipf, Mandelbrot y textos reales rangos 10000-10200.....	162
Gráfico 58. Comparación Zipf, Mandelbrot y textos reales rangos 30000-30200.....	163
Gráfico 59. Comparación Zipf, Mandelbrot y textos reales rangos 48100-48300.....	164
Gráfico 60. Distribución logarítmica de Zipf con textos reales.....	165
Gráfico 61. Distribución logarítmica de Zipf con textos reales y Zipf ajustada	165
Gráfico 62. Distribución logarítmica de Zipf con textos reales y Mandelbrot ajustada	166
Gráfico 63. Representación Transformada. Distribución de Zipf con textos reales	167
Gráfico 64. Representación Transformada. Distribución de Zipf con textos reales y Zipf ajustada	167
Gráfico 65. Valor del exponente (e) de Zipf con textos literarios y científicos.....	176
Gráfico 66. Valor del exponente (e) de Zipf con textos literarios y científicos aplicado a raíces	183
Gráfico 67. Valor del exponente (e) de Zipf con textos científicos aplicado a raíces	184
Gráfico 68. Valor del exponente (e) de Zipf con textos literarios aplicado a raíces.....	184
Gráfico 69. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 1.....	187
Gráfico 70. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 2.....	188
Gráfico 71. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 3.....	189
Gráfico 72. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 4.....	190
Gráfico 73. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1 pv=386.....	191
Gráfico 74. Distribución logarítmica de Zipf con valores reales y Zipf ajustada con texto pequeño	195
Gráfico 75. Distribución logarítmica de Zipf con valores reales y Zipf ajustada con texto grande.....	195
Gráfico 76. Porcentaje de vocabulario agrupado en 10 tramos. Modelo Log-%	202
Gráfico 77. Representación Log-% con texto de P=78.719 y V=16.449	205
Gráfico 78. Representación Log-% con texto real y sintéticos.....	206
Gráfico 79. Representación Log-% con texto de P=112.466 y V=17.097	208
Gráfico 80. Representación Log-% con texto de P=244.324 y V=29.062	208
Gráfico 81. Representación Log-% con texto de P=367.570 y V=38.094	209
Gráfico 82. Representación Log-% con texto de P=450.574 y V=43.816	209
Gráfico 83. Distribución logarítmica de Mandelbrot ajustada con textos reales y pv=72	214
Gráfico 84. Distribución logarítmica de Mandelbrot ajustada con textos reales y pv=386.....	215
Gráfico 85. Distribución logarítmica de Mandelbrot ajustada con textos reales sufijos=358	216
Gráfico 86. Distribución logarítmica de Mandelbrot ajustada a tramos 2-8 con textos reales	217
Gráfico 87. Stemmer método de variedad de sucesores	221
Gráfico 88. Precisión sobre Exhaustividad de los datos ejemplo	230
Gráfico 89. Distribución de frecuencias texto Cervantes, Zipf y Mandelbrot.....	234
Gráfico 90. Distribución de frecuencias con raíces texto Cervantes, Zipf y Mandelbrot	235
Gráfico 91. Distribución de Frecuencias texto Cervantes con raíces SU 3	237
Gráfico 92. Distribución de Frecuencias texto Cervantes con raíces SU 5	238
Gráfico 93. Distribución de Frecuencias texto Cervantes con raíces MVS.....	239
Gráfico 94. Distribución de Frecuencias texto Cervantes con raíces TCP 80% 6.....	240
Gráfico 95. Distribución de Frecuencias texto Cervantes con raíces TCP 60% 5	241
Gráfico 96. Distribución de Frecuencias texto Cervantes con raíces TCP 40% 4	242
Gráfico 97. Distribución de Frecuencias texto Cervantes con raíces TCP 30% 3.....	244

Gráfico 98. Distribución de Frecuencias texto Cervantes con raíces TCP 20% 2.....	245
Gráfico 99. Distribución de Frecuencias texto Cervantes con diversos Stemmers.....	246
Gráfico 100. Distribución de Frecuencias texto Cervantes con raíces TCP 30% 3.....	248
Gráfico 101. Distribución de frecuencias texto Cervantes con raíces SU 3 y Zipf ajustado tramos 4-7 ..	250
Gráfico 102. Distribución de frecuencias texto Cervantes con raíces SU 5 y Zipf ajustado tramos 4-7 ..	251
Gráfico 103. Distribución de frecuencias texto Cervantes con raíces MVS y Zipf ajustado tramos 4-7.	252
Gráfico 104. Distribución de frecuencias texto Cervantes con raíces TCP 80% 6 y Zipf ajustado tramos 4-7.....	253
Gráfico 105. Distribución de frecuencias texto Cervantes con raíces TCP 60% 5 y Zipf ajustado tramos 4-7	254
Gráfico 106. Distribución de frecuencias texto Cervantes con raíces TCP 40% 4 y Zipf ajustado tramos 4-7.....	255
Gráfico 107. Distribución de frecuencias texto Cervantes con raíces TCP 30% 3 y Zipf ajustado tramos 4-7.....	256
Gráfico 108. Distribución de frecuencias texto Cervantes con raíces TCP 20% 2 y Zipf ajustado tramos 4-7.....	257
Gráfico 109. Valor del exponente de Zipf con distintos Stemmers aplicado a textos literarios	259
Gráfico 110. Precisión-Exhaustividad texto Cervantes con varios Stemmers	262
Gráfico 111. Precisión-Exhaustividad texto Cervantes con varios Stemmers	263
Gráfico 112. Parejas de raíces pre-asociadas según granularidad texto Azorín	272
Gráfico 113. Parejas de raíces pre-asociadas según granularidad texto Castelar	273
Gráfico 114. Parejas de raíces pre-asociadas según granularidad texto Cienti1	273
Gráfico 115. Parejas de raíces pre-asociadas según granularidad texto Legal1	274
Gráfico 116. Parejas de raíces pre-asociadas según granularidad texto Quijote.....	274
Gráfico 117. Funciones potenciales con diferentes parámetros.....	277
Gráfico 118. Parejas de raíces asociadas y función potencial obtenida texto Cienti1	278
Gráfico 119. Parejas de raíces asociadas y función potencial obtenida texto Legal1	279
Gráfico 120. Parejas de raíces asociadas y función potencial obtenida texto Quijote	279
Gráfico 121. Espectro de similitudes texto Azorín.....	281
Gráfico 122. Espectro de similitudes texto episo1	282
Gráfico 123. Representación logarítmica del espectro de similitudes texto Cervantes	283
Gráfico 124. Representación logarítmica del espectro de similitudes varios textos.....	283

0. Introducción, Visión Global e Hipótesis

La recuperación de la información y la representación de ésta con fines documentales es el objetivo primordial de muchos investigadores en la materia, debido a que la producción científica actual supera los límites de la retención y asimilación humana ya que ésta crece a una velocidad exponencial.

Nos resulta imposible en el siglo XXI analizar documentalmente toda la masa de información científica que se produce día a día a una velocidad vertiginosa sin la ayuda de la informática y las nuevas tecnologías.

Si tenemos en cuenta que desde la aparición de la World Wide Web en 1965 el incremento de la información disponible es cada vez más grande, esto incrementa nuestra responsabilidad para encontrar una solución científica eficiente y útil que atenúe esta problemática.

Asimismo, no toda la información que se genera es válida, lo que denominamos contaminación informativa. Esta situación hace que cada vez sea más difícil para los profesionales gestionarla y para los usuarios encontrar información útil. Una consecuencia del gran volumen de información es el ruido documental en los resultados de búsquedas de los usuarios.

Internet puede ser considerada como una gran base de datos a la que se puede acceder desde cualquier lugar del mundo. Por esta razón el desarrollo de mecanismos de búsqueda y de análisis documental de los documentos tiene gran relevancia en el campo de las tecnologías de la información.

Todo ello, nos plantea la posibilidad de la automatización del proceso de indización y resumen y por supuesto la recuperación de información. Superando así barreras como la indización manual, que de todos es bien conocida la problemática e inconsistencia de esta técnica sin las tecnologías adecuadas.

Toda la comunidad científica inmersa en esta cuestión investiga para conseguir aplicaciones informáticas que sean capaces de llevar a cabo las tareas documentales de un modo automatizado. Esta no es tarea fácil, porque los ordenadores no son inteligentes como las personas, pero por otro lado permiten analizar gran cantidad de datos de un modo riguroso y exacto.

Es un hecho que la existencia de nuevas aplicaciones nos permite realizar el análisis documental, la indización y por supuesto la recuperación de la información científica mediante técnicas automáticas.

Existe multitud de autores que han revolucionado el campo de la recuperación de la información con sus teorías, considerados como los pioneros de una ciencia que hoy en día está adquiriendo gran importancia debido a la explosión de la información y los medios de comunicación.

Desde los ordenadores digitales que se desarrollaron en los años 40 y principios de los 50 y desde el uso de las tarjetas perforadas hasta las primeras teorías relativas a la recuperación de información de varios autores como Calvin N. Mooers conocido por acuñar el término de recuperación de información en 1950, y autores como Gerald Salton (1968) anticipando ideas como los entornos de recuperación on-line, interfaces expertos y adentrándose en la interacción flexible entre el usuario y el sistema, el campo de la recuperación de información ha evolucionado considerablemente.

A finales de los años 40 y 50 aparecen dos autores importantísimos en la recuperación de la información que realizaron estudios basados en la frecuencia de aparición de las palabras en un documento para determinar si eran significativas para representar el contenido o las características del mismo, así pues la frecuencia de las palabras podía utilizarse para indicar el grado de significación, entre estos autores destacan Luhn (1957) y Zipf (1949).

Otros autores destacados que sentaron las bases de lo que hoy en día conocemos como recuperación de información es Heaps (1978) con su ley sobre las palabras distintas en un documento o lo que podemos llamar el vocabulario en un documento. Lotka (1926) y su ley que muestra que el número de escritores con n publicaciones en una bibliografía sigue una expresión matemática y Bradford (1948) y su ley de dispersión de la ciencia o dispersión de la bibliografía en las revistas científicas.

Considerando estos autores como los precursores de las ciencias de la recuperación de información, transcurridos los años esta ciencia ha evolucionado desde los modelos de recuperación clásicos como el modelo booleano, vectorial, probabilístico etc., hasta hoy en día con los modelos booleano extendido, modelo de conjuntos difusos, el modelo vectorial generalizado, el LSI: Latent Semantic Indexing o las redes neuronales y otros modelos extensiones del modelo probabilístico como son las redes bayesianas y las redes de inferencia bayesiana.

Sin olvidar proyectos importantes que han ayudado a que esta ciencia avance como el proyecto Cranfield I y II (1957-1966), el cual constaba de colecciones experimentales en las que se aplicaban consultas y juicios de relevancia, se aplicaban cálculos de precisión y exhaustividad para analizar los resultados y establecer comparaciones entre los modelos evaluados.

Otro de los hitos importantes ha sido la creación de las conferencias TREC (Text REtrieval Conference¹) que comenzaron en 1992 para fomentar la investigación en el campo de la recuperación de información ofreciendo grandes cantidades de colecciones y un foro para investigadores en el cual tratar sus trabajos enmarcados en un problema común.

En el siglo XXI continúa la problemática de la recuperación de información en la Web, apareciendo el deseo de recuperar, no meramente documentos, sino la respuesta a la pregunta de un usuario. Quizá contribuya al camino de la solución la utilización de metalenguajes de marcas (XML), que llevan en sí mismos ciertos metadatos con valor

¹ En <http://trec.nist.gov/>

semántico. Más aún si se hace manifiesto como en la formulación RDF o Resource Description FrameWork. El objetivo anunciado es la construcción de la llamada Web Semántica.

Visión Global

La investigación que se aborda realiza un estudio científico detallado sobre una materia con creciente importancia en las ciencias de la documentación. Este estudio científico desarrolla un prototipo de Sistema de Indización y Segmentación Automática para textos científicos de cualquier tipología en Español, utilizando técnicas de similitud entre las palabras asociadas en los textos. Además se amplían y perfeccionan los métodos cuantitativos y leyes clásicas en recuperación de la información, unas veces de modo experimental y otras de modo especulativo, con comprobación posterior, aplicándolas a varios tipos de texto, siempre largos y en Español.

Es primordial mencionar que esta Tesis doctoral hace un recorrido transversal por varios de los modelos y leyes que contribuyen a los métodos algorítmicos, tratando de encontrar mejoras mediante modificaciones o ampliaciones significativas. Todo ello trabajando siempre con textos largos y en idioma Español para dar solución a los problemas específicos que en este contexto puedan plantearse.

Aunque la mayor parte de estos métodos cuantitativos son independientes del idioma en que están escritos los textos, al menos en un análisis superficial hay algunos aspectos que son distintos de un idioma a otro. No nos referimos sólo a diferencias evidentes como la estructura gramatical morfológica de las palabras (que hace que el famoso algoritmo de Porter sea inviable en Español) sino a diferencias más sutiles en el léxico (por ejemplo palabras compuestas en alemán que en español sólo pueden expresarse con una frase), que se traducen en alteraciones en las leyes cuantitativas de la lingüística. Por tanto no es accesorio, sino esencial la acotación que se hace del objeto de estudio a sólo textos en Español.

El Sistema de Indización y Segmentación Automática del Texto denominado MALLOV categoriza cualquier texto en Español y efectúa la representación temática de éste, además obtiene los términos de indización de cada una de las partes temáticas en las que ha sido dividido. En su funcionamiento se ha buscado que haga explícitos los cálculos y resultados intermedios de cada proceso.

La Segmentación del Texto o *Text Segmentation* es una técnica utilizada en el Procesamiento del Lenguaje Natural y exclusivamente para resúmenes automáticos, aunque en esta Tesis doctoral nosotros lo utilizamos para detectar los cambios de temas en un documento. Además de esto utilizaremos esta técnica junto con otras como la similitud entre las palabras asociadas de un texto.

El sistema extrae el contenido de un documento para obtener la enumeración automática de las partes principales o estructura del texto. Esta enumeración o distribución temática se representará con una estructura arbórea jerárquica en la que se obtiene en el Nivel 0 o parte principal, los descriptores y la temática general del texto o documento. En los siguientes subniveles como el Nivel 1 se subdivide en varias partes con los términos de indización y así sucesivamente se va segmentando el contenido del texto. En las partes

temáticas obtenidas se congregan varios subdocumentos diferentes ya sean sucesivos o no del documento original, para finalmente obtener un mapa del conocimiento acerca de un documento.

Parte de la gran dificultad de esta Tesis doctoral ha radicado en el desarrollo no sólo de una base conceptual teórica sobre indización y segmentación automática sino en el desarrollo de las aplicaciones informáticas creadas exclusivamente para desarrollar las experimentaciones de nuestra investigación. Estas aplicaciones informáticas han dado lugar a un Sistema de Indización y Segmentación Automática del Texto denominado MALLOV el cual se ha desarrollado tanto conceptualmente como físicamente. Se ha considerado importante esta realización física, aunque sea a nivel de prototipo para mostrar que las hipótesis y desarrollos teóricos no son irreales.

Obviamente para nosotros hubiera resultado más sencillo realizar nuestra investigación sólo en el plano teórico, pero esto nos parecía insuficiente. Es por ello, que se han desarrollado las aplicaciones informáticas y sistemas automatizados utilizando las herramientas disponibles y un lenguaje de programación sencillo. Somos conscientes que nuestras aplicaciones se podrían perfeccionar utilizando otros lenguajes de programación como el C, pero esto no ha sido nuestro objetivo, el desarrollar nuestras aplicaciones con dicho lenguaje de programación suponía salir de la senda marcada y realizar una investigación más informática que basada en nuestro objetivo final: las técnicas documentales.

Por tanto, en último lugar, esta Tesis no solo consta de una base conceptual teórica sobre indización y segmentación automática sino en la implementación y crítica de las aplicaciones informáticas que proporcionan la base para las experimentaciones de esta investigación.

Deseamos que esta Tesis doctoral ayude y guíe a la comunidad científica interesada en encontrar soluciones a la gran cantidad de información que hoy día nos desborda y sirva de base para futuras investigaciones sobre la materia.

Hipótesis

Entender la estructura de esta tesis doctoral es primordial para la consecución final de nuestro prototipo de Modelo de Sistema de Indización y Segmentación Automática. Las diferentes técnicas cuantitativas y leyes clásicas en Recuperación de la Información desarrollada en los capítulos cinco a nueve de esta tesis doctoral convergen en el diseño y aplicación de nuestro Sistema de Indización y Segmentación Automática en el capítulo diez de esta investigación. Es decir cada capítulo independiente uno del otro finalmente confluye en el capítulo diez, por ese motivo las hipótesis van numeradas independientes referidas a cada uno de los capítulos, la nomenclatura se compone en primer lugar del número del capítulo correspondiente y en segundo lugar del número de hipótesis dentro de dicho capítulo.

Es por ello que el planteamiento de las siguientes hipótesis que han permitido desarrollar un Sistema de Indización y Segmentación Automática han sido las siguientes:

- 5.1. Contribuyen de manera efectiva al funcionamiento de los sistemas de tratamiento y recuperación de la información los metodos cuantitativos relativos al proceso de creación de vocabulario (Heaps, 1978)
- 5.2. ¿Puede utilizarse para el tratamiento de la información los parámetros que predice la Ley de Heaps o las discrepancias del texto respecto a los parámetros?
- 6.1. Contribuyen de manera efectiva al funcionamiento de los sistemas de tratamiento y recuperación de la información los metodos cuantitativos relativos al proceso de repetición de palabras (Zipf, 1949), (Mandelbrot, 1953)
- 6.2. La Ley de Zipf aplicada a un mismo texto en distintas circunstancias proporciona distinto resultado; ¿de qué depende?, ¿podemos utilizar distintas aplicaciones de la Ley de Zipf a un texto para aprovechar los distintos valores de los parámetros y obtener conclusiones al respecto?
- 6.3. Obtenidos los parámetros, ¿los datos reales del texto se ajustan a la curva predicha por la Ley de Zipf?
- 7.1. ¿Qué influencia tendrá en la degradación de la información y en el valor de los parámetros de Zipf si utilizamos un lematizador o stemmer? ¿Y si utilizamos distintos métodos de lematización, los diferentes resultados pueden servir para obtener conclusiones?
- 8.1. La construcción de pares de palabras en base a su coocurrencia, ¿puede ayudar a seleccionar una pequeña colección de términos que describan la temática de éste?
- 8.2. ¿Afectará la granularidad del texto en el número de parejas encontradas?
- 8.3. ¿De qué manera influye el tamaño total del texto en el número de parejas encontradas?
- 8.4. ¿Cómo es el espectro de similitudes de Dice (1945)?
- 9.1. Para segmentar un texto ¿Podemos explotar la fórmula de Heaps o su incumplimiento para determinar la división en fragmentos de modo que tenga sentido temático?
- 10.1. ¿Pueden reunirse todos los resultados de los capítulos cinco al nueve de esta tesis doctoral en un Sistema coherente que cumpla todas las conclusiones mencionadas y que contribuya al funcionamiento de un sistema de análisis temático de un texto largo y sus partes?

1. Objetivos

El camino recorrido en estos años de trabajo ha sido arduo y difícil, ha estado lleno de proyectos profesionales y en ocasiones de otros proyectos personales e incluso de algunos obstáculos que han supuesto paralizar la investigación. Por muy duro que haya sido el camino recorrido cuando la compañía es la adecuada se consigue llegar con éxito al final de éste y cumplir los objetivos.

La presente tesis doctoral tiene un objetivo general y una serie de objetivos específicos. El objetivo general consiste en estudiar como los métodos cuantitativos o numéricos pueden ser útiles para desarrollar un Sistema de Indización y Segmentación Automática para textos en Español. Los objetivos específicos que se detallan a continuación, han contribuido a la consecución del objetivo general. Es importantísimo aclarar que la estructura de esta tesis doctoral se compone de capítulos con estudios independientes entre sí, capítulos cinco al nueve. Y el capítulo diez es el compendio de cada uno de ellos en la aplicación del Sistema de Indización y Segmentación Automática, por ese motivo al igual que las hipótesis, los objetivos específicos van numerados independientes referidas a cada uno de los capítulos, la nomenclatura se compone en primer lugar del número del capítulo correspondiente y en segundo lugar del número de objetivo específico dentro de dicho capítulo.

- 5.1** Aplicar el modelo de crecimiento del vocabulario, con máxima precisión para obtener los detalles de cómo, cuanto y en qué circunstancias, no se ajusta exactamente a la Ley de Heaps.
- 6.1** Aplicar el modelo de distribución de frecuencias de Zipf cotejando sus resultados en distintas circunstancias y tipos de textos. Obtener especificaciones de la variación de los parámetros de Zipf
- 6.2** Cotejar la Ley de Zipf con los valores reales de frecuencias para ver si hay variaciones locales
- 7.1** Implementar varios métodos distintos de lematización de las palabras para estudiar su influencia sobre la degradación de la información y el valor de los parámetros de Zipf.
- 8.1** Implementar un sistema informático flexible, aunque no sea eficiente, que construya la colección de palabras asociadas definidas a partir de un texto y cierta granularidad de sus fragmentos y que muestre la cantidad de palabras y las posibilidades de su reducción y de su selección.
- 8.2** Utilizar este sistema para dar respuesta a las hipótesis 8.2, 8.3 y 8.4
- 9.1** Implementar un método de segmentación automática del texto que esté basado únicamente en procedimientos cuantitativos, que utilice las fórmulas obtenidas anteriormente sobre crecimiento del vocabulario (Heaps, 1978).

- 9.2** Como objetivo específico nos planteamos desarrollar un concepto de Novedad de las palabras en relación con la lectura diacrónica del texto y utilizarlo para detectar los cambios de temática.
- 10.1** Podemos diseñar un prototipo de Sistema de Indización y Segmentación Automática de la información basándonos estrictamente en métodos cuantitativos, sin utilizar los métodos léxicos o lingüísticos.

La tesis doctoral consiste en la realización de los diversos estudios que cumplen cada uno de los objetivos específicos, sus resultados convergen finalmente en el último capítulo de esta investigación, en el que se desarrolla un Sistema de Indización y Segmentación Automática para textos en Español que consigue un funcionamiento razonable sin la utilización de las potentes herramientas ya conocidas² sino únicamente con la aplicación de métodos cuantitativos.

² Véase capítulo 3

2. Metodología

La metodología a aplicar para cumplir los objetivos que se proponen en la tesis doctoral se establece en varios capítulos cada uno sobre un tema característico en recuperación de la información. En ellos se desarrollan métodos originales que están orientados a su utilización en el objetivo central de la tesis. El estudio abarca varios aspectos: se estudia el modelo de crecimiento del vocabulario aportando críticas y una nueva versión de la ley de Heaps. Se aclaran distintos aspectos relacionados con la ley de Zipf. Se implementa un método de lematización de palabras. Se estudia cuantitativamente el concepto de palabras asociadas con vistas a extraer conclusiones semánticas y se implementa una forma de segmentación del texto utilizando las fórmulas obtenidas anteriormente sobre crecimiento de vocabulario. Estos estudios desarrollados ampliamente en capítulos independientes convergen al final en la exposición del último capítulo donde se implementa el Sistema de Indización y Segmentación Automática para textos en Español.

Como el objetivo principal requiere trabajar con varios aspectos muy distintos entre sí, tal como se ha detallado en los objetivos específicos, los capítulos cinco a nueve son independientes entre sí, cada uno tiene sus objetivos específicos que de manera global puede definirse como obtener de las leyes clásicas ciertos valores numéricos o fórmulas que puedan ayudar en el análisis de un texto concreto.

Por tanto la metodología para dar respuesta a las hipótesis, cumplir los objetivos y obtener así las conclusiones finales se divide en dos vertientes principales, la metodología aplicada a la parte experimental con los documentos objeto de ensayo, es decir la descripción de hechos objetivos y la metodología aplicada a las técnicas documentales y métodos de cálculo para obtener las conclusiones mediante el desarrollo de las aplicaciones informáticas donde plasmamos la construcción de métodos. En la parte experimental se ha tenido en cuenta para las investigaciones, tanto el tamaño de los documentos como la tipología de éstos, que hemos dividido en los siguientes: histórico narrativo, literario, científico, industrial, legal y de patentes.

En primer lugar para obtener los documentos de texto de tipo literario, histórico, narrativo, se ha consultado la Biblioteca Virtual Miguel de Cervantes <http://www.cervantesvirtual.com>. En el Apéndice I se detallan las referencias bibliográficas de la edición manejada en los textos exportados de la Biblioteca Virtual:

Benito Pérez Galdós (1843-1920) y parte de su obra “*Episodios Nacionales*”

Mariano José de Larra (1809-1837) y parte de su obra “*El ideario Español*”

Marcelino Menéndez y Pelayo (1856-1912) y parte de su obra “*La ciencia española: polémicas, indicaciones y proyectos*”

Emilio Castelar (1832-1899) y varias de sus obras sobre discursos políticos y otros tipos de escritos.

Joaquín Costa (1846-1911) y varias de sus obras sobre política y otros tipos de escritos.

Miguel de Cervantes Saavedra (1547-1616) y varias de sus obras, entre ellas su obra más conocida “*El Ingenioso Hidalgo Don Quijote de la Mancha*”

Para la obtención de textos de tipo científico se ha utilizado la base de datos URBADISC, conjunto de bases de datos bibliográficas sobre urbanismo, arquitectura, medio ambiente, transporte, saneamiento, planificación de ciudades, utilización del suelo, políticas urbanas, etc. Recoge documentos publicados desde 1995 y está accesible mediante el servidor IRIS. De la base de datos URBADISC se ha exportado un texto de tipo científico mediante una consulta sobre arquitectura, construcciones, obras públicas e ingeniería civil. Posteriormente incluida en la Base de Datos general de CSIC, de donde se han obtenido también varios ejemplos. También se han exportado documentos sobre legislación de la base de datos ARANZADI. Dentro de los documentos de tipo científico hemos considerado también artículos cortos de resúmenes obtenidos de la bases datos del CSIC sobre turismo, administración, etc. E incluso hemos exportado de la base de datos del CSIC artículos médicos que se componen de referencias bibliográficas con gran cantidad de campos.

Para obtener los textos de tipo científico pero con contenido exclusivo de patentes se ha utilizado la base de datos CIBEPAT, base de datos bibliográfica producida por la OEPM (Oficina Española de Patentes y Marcas) que recoge referencias de patentes y modelos de utilidad desde 1968, patentes europeas con efectos en España desde 1986, solicitudes PCT con efectos en España desde 1989, patentes iberoamericanas desde 1966 y la quinta edición de la Clasificación Internacional de Patentes. Esta base de datos está accesible mediante el servidor IRIS e Internet.

Estos documentos exportados configuran una serie de textos con un tamaño de origen determinado, que va desde las 7.000 a las 500.000 palabras de vocabulario aproximadamente, además ha sido modificado dividiendo estos textos en partes más pequeñas para así realizar ensayos entre ellos y observar las relaciones o alteraciones sufridas en las variables de los textos objeto de estudio.

TIPOS DE TEXTOS	BASES DE DATOS ORIGEN	TEMÁTICA
LITERARIOS		
Benito Pérez Galdós (1843-1920)	BVMC ³	Literario Narrativo-Histórico
Mariano José de Larra (1809-1837)	BVMC	Literario Narrativo-Histórico
Marcelino Menéndez y Pelayo (1856-1912)	BVMC	Literario Narrativo-Histórico
Emilio Castelar (1832-1899)	BVMC	Discursos Literarios
Joaquín Costa (1846-1911)	BVMC	Discursos Literarios
Miguel de Cervantes (1547-1616)	BVMC	Literario Narrativo-Histórico
Calderón de la Barca (1600-1681)	BVMC	Poesía
Juan del Encina (1468-1529); Jorge Manrique(1440-1479); Marqués de Santillana (1398-1458); Anónimos	BVMC	Poesía
Gabriel Galán (1870-1905)	BVMC	Poesía
José M ^a de Pereda (1833-1906)	BVMC	Literario Narrativo
San Francisco de Borja (1510-1572)	BVMC	Tratados, sermones
CIENTÍFICOS		
Artículos temas científicos	URBADISC	Arquitectura, construcciones, obras públicas e ingeniería civil.
Artículos temas legales	ARANZADI	Legislación
Artículos ISO	CSIC	Resúmenes cortos, (turismo, administración, riesgo,...)
Artículos médicos	CSIC	Referencias bibliográficas incluyendo muchos campos.
Patentes	CIBEPAT	Patentes

Una vez obtenidos los documentos para efectuar nuestros experimentos se han desarrollado 6 bases de datos implementas para realizar estudios cuantitativos con los ficheros de textos y para realizar las experimentaciones y aplicar los modelos correspondientes a cada texto.

³ Biblioteca Virtual Miguel de Cervantes <http://www.cervantesvirtual.com>

BASE DE DATOS	NOMBRE DETALLADO	FUNCIONES
TOPOS	Recuentos de Palabras	<ul style="list-style-type: none"> -Cálculo de la ley de Zipf -Cálculo de la distribución transformada de Zipf -Cálculo de la ley reducida de Zipf -Ajustar fórmula de Zipf y Mandelbrot -Obtener predicción de la fórmula de Zipf y Mandelbrot -Modelo log-% en 10 tramos logarítmicos
RENOS	Representación Numérica	<ul style="list-style-type: none"> -Ajustes del exponente y coeficiente de Zipf -Cálculo de frecuencias de las palabras -Genera representación numérica de las palabras -Cálculo de vocabulario, la varianza y derivadas -Divide en partes el vocabulario -Ajusta la función potencial
ARENA-1: PIEDRAS	Aprendiendo Recuperación de Información	<ul style="list-style-type: none"> -Convierte ficheros de tipo secuencial a ficheros de acceso aleatorio -Analiza léxico -Dividir texto en documentos - Extraer palabras significativas y sus lemas o raíces -Extracción de raíces por Método de Variedad de Sucesores - Extracción de raíces por Método de Sufijos - Extracción de raíces por Método de % - Búsquedas para la recuperación de información mediante palabras, nº del documento o raíces. - Cálculos de Zipf
ARENA-2	Aprendiendo Recuperación de Información	<ul style="list-style-type: none"> - Recuperación de información: implementación del método vectorial - cálculo IDF
COPAS	Contando Palabras Asociadas	<ul style="list-style-type: none"> - Calcula las similitudes relativas de Dice de las palabras asociadas y almacena en tabla -Analiza un texto dividido en documentos -Cuenta y obtiene las palabras asociadas de un texto -Obtiene palabras asociadas de cada uno de los documentos en los que se divide un texto -Estudio de la distribución de valores de las similitudes de las palabras asociadas -Cálculo de umbrales para un número de palabras asociadas -Ajuste de los coeficientes de las fórmulas a los datos de cada texto -Distribución de frecuencias, Zipf y análogas

MALLOV	Sistema de Indización y Segmentación Automática	<ul style="list-style-type: none"> -Proceso previo del fichero -Separa todas las palabras -Construye vocabulario -Formar colección de raíces -Representación numérica -Modelo de crecimiento del vocabulario -Modelo de distribución de frecuencias -Identifica cambios temáticos -Segmentación Automática: Genera árbol de temas -Indización Automática: Palabras de cada tema -Obtiene los Términos a un informe visual.
---------------	---	---

Cada uno de los ficheros de texto utilizados es a priori un fichero secuencial formado sólo por caracteres. La forma de tratamiento que hemos seguido es convertirlo en primer lugar en un fichero de acceso aleatorio para posibilitar la actuación sobre ellos de los programas informáticos (en concreto los mencionados en la tabla anterior que implementan el método vectorial entre otros). El proceso de convertir el fichero de secuencial a aleatorio es esencial para la posterior indización y segmentación automática y el análisis de las características que presenten los documentos exportados y llevados a estudio.

En líneas generales, el proceso que se ha aplicado a los ficheros de texto ha consistido en: convertirlos en ficheros de acceso aleatorio con formato fijo, para compatibilidad entre los distintos programas utilizados, dividirlos en una serie de documentos o fragmentos pequeños según distintos criterios, formar la colección ordenada de palabras que los componen prescindiendo de palabras vacías, obtener el vocabulario o colección de palabras distintas, sustituir el vocabulario por la colección de raíces, identificar saltos con posible cambio semántico para dividir el texto en partes, identificar palabras importantes para la caracterización e indización de cada parte, buscar pares de palabras asociadas para reforzar el valor semántico.

Las técnicas documentales y de procesado de la información textual junto con las diversas técnicas de cálculo y aplicación de métodos clásicos de recuperación de información han sido desarrolladas con aplicaciones creadas en Microsoft Access. Somos conscientes que nuestras aplicaciones se podrían perfeccionar utilizando otros lenguajes de programación como el C, pero esto no ha sido nuestro objetivo, el desarrollar nuestras aplicaciones con dicho lenguaje de programación suponía salir de la senda marcada y realizar una investigación más informática que basada en nuestro objetivo final: las técnicas documentales.

En la construcción de todas estas etapas, el plan de trabajo ha sido detenerse en cada una: explorar distintas posibilidades, obtener resultados cuantitativos, obtener conclusiones cualitativas, etc. Todo ello con la justificación estadística cuando proceda o con la explicación de su relación con resultados ya publicados, clásicos o recientes.

Finalmente, con vistas a mostrar la verosimilitud de los métodos estudiados, se ha recopilado en el sistema MALLOV el resto de bases de datos para obtener el Sistema de Indización y Segmentación Automática del Texto.

3. Antecedentes y estado del conocimiento

Un problema real al que nos enfrentamos en esta era de la información es el incremento exponencial que está sufriendo la información, sobre todo desde la aparición de la World Wide Web en 1965 y el incremento del número de documentos electrónicos. La llegada de la Web ha incrementado la importancia de la recuperación de información (RI). Internet puede ser considerada como una gran base de datos a la que se puede acceder desde cualquier lugar del mundo. Por estas razones el desarrollo de mecanismos de búsqueda eficientes tiene gran relevancia en el campo de las tecnologías de la información.

Calvin N. Mooers (1950) es conocido por acuñar el término de recuperación de información en 1950, una definición de RI puede ser la operación en la que se interpreta una necesidad de información de un usuario y se seleccionan los documentos más relevantes capaces de dar una solución a la necesidad de información planteada.

El proceso de RI acarrea la realización de varias tareas, una de ellas es la obtención de una representación adecuada de los documentos que permita un almacenamiento y una comparación con las consultas del usuario (proceso de indexado). Otra tarea consiste en el procesamiento de la consulta de entrada para la obtención de su representación en un lenguaje de consulta más o menos complejo. La última tarea consiste en la comparación de la consulta con los documentos para la obtención de la lista de documentos que satisfacen dicha consulta.

Según Sangkon Lee, [et al.] (2002) es común para un sistema de información recuperar pasajes, secciones, trozos de documentos en vez de examinar documentos enteros cuando los documentos almacenados son gran cantidad de textos su recuperación completa no suele ser del interés del usuario. Como Salton, Allan y Buckley (1993) indicaron *“en tales casos, la eficiencia y la efectividad de los resultados de la recuperación pueden ser obtenidos usando estrategias de recuperación por trozos, secciones, pasajes diseñadas para recuperar extractos de texto de diversos tamaños en respuesta a las declaraciones de los intereses de los usuarios”*.

Uno de los beneficios para los usuarios es el recibir extractos cortos de texto, que están relacionados con el tema de la consulta en vez de recibir artículos a texto completo.

Uno de los problemas que se da al realizar búsquedas en bases de datos que contienen texto completo es como mencionó Korfhage (1997) *“muchos documentalistas, buscadores de información están probablemente ofreciendo grandes volúmenes de datos, en los cuales se encuentran incluidos algunos documentos no relevantes”*. Este problema sucede cuando se utilizan los documentos a texto completo en lugar de documentos cortos.

En el campo de la RI existe multitud de autores que han revolucionado este campo con sus teorías, considerados como los pioneros de una ciencia que hoy en día debido a la explosión de la información y los medios de comunicación y almacenado de grandes cantidades de información esta adquiriendo una gran importancia.

En los años 40 se comenzó a plantear el problema del almacenamiento y recuperación de documentos. A finales de los 50 y principios de los 60 con el incremento de la producción de información científica, los métodos tradicionales de almacenamiento y recuperación fueron disminuyendo su efectividad. Los ordenadores se convirtieron en herramientas imprescindibles para el almacenamiento, tratamiento, recuperación y difusión de la información contenida en los distintos soportes.

Se consideran los años 50 como el comienzo de la RI, en estos años Luhn (1957) sugirió que los sistemas de recuperación de textos se debían diseñar basándose en la comparación entre los identificadores de contenido del texto y las peticiones de las preguntas, además Luhn propuso por primera vez las bases de la indexación automática, los métodos automáticos de asignación de pesos en los términos, los métodos de resumen automático basados en asignar factores de significado en las frases⁴, también propuso que las operaciones intelectuales del análisis de los textos se basaban en el análisis de la frecuencia de las palabras existentes en el texto, concretamente la importancia de una palabra estaba relacionada a su frecuencia de aparición en los textos. Así vemos que Luhn fue el precursor en la descripción de la recuperación estadística. Los primeros sistemas de recuperación, además de la comparación introdujeron el álgebra de Boole para expandir y limitar las búsquedas. Hoy día esto sigue presente en muchos sistemas de RI.

Es en los años 60 cuando empiezan a aparecer los primeros experimentos con procesamiento de lenguaje natural y con métodos estadísticos. Luhn como se ha mencionado introdujo los postulados de Zipf sobre el uso de la frecuencia de aparición de las palabras en los documentos, para determinar si son significativas como para representar el contenido del mismo. Es en esta época cuando se empieza a estudiar la frecuencia de aparición conjunta de los términos, lo que se conoce con el nombre de coocurrencia: cuando varios términos aparecen con cierta frecuencia juntos o muy próximos en un texto es porque hay cierto grado de relación entre ellos. Autores que investigaron esto en la década de los 60 y 70 son Maron y Kuhns, Suites, Spark Jones y Robertson, entre otros.

Poco después en 1960 Calvin N. Mooers (1950) conocido por acuñar el término de recuperación de información en 1950 publicó un artículo en el "American Documentation" (1950). Mooers introdujo ideas novedosas y audaces respecto al futuro de la recuperación de información, anticipando en su artículo ideas como entornos de recuperación on-line, interfaces expertos y adentrándose en la interacción flexible entre el usuario y el sistema, incluso predijo la automatización en la asignación de descriptores de contenido en los documentos. Pero Calvin N. Mooers además de esto señaló que los ordenadores podían ser utilizados para perfeccionar las consultas de los usuarios e incluso construir frases de consulta formales empezando con las consultas en lenguaje natural. A partir de este momento el autor centró sus estudios básicamente en la "máquina de inferencia inductiva", la cual creyó revolucionaría el campo del procesamiento de texto y la recuperación creyendo que dicha máquina reemplazaría o al menos supliría la habilidad del razonamiento humano mediante la ejecución de tareas de traducción del lenguaje, generando textos escritos y detectando vacíos en cadenas de

⁴ Basado en el número de palabras significativas en el texto y en la distancia entre estas palabras en las frases del documento.

argumentos, es decir Mooers supuso que la máquina de inferencia inductiva se utilizaría como una máquina de traducción y para la producción de nuevos ensayos y artículos de áreas temáticas particulares.

Es en los años 60 y 70 cuando Gerald Salton revoluciona los métodos de RI con el sistema vectorial y el clustering.

Es en la década de los 70 cuando tienen su apogeo los modelos de RI probabilísticos y el modelo de espacio vectorial. La indización probabilística fue propuesta en esencia por Maron & Kuhns (1960) en los años 60, estos autores definieron un modelo probabilístico de RI, posteriormente en 1981 la autora Spark Jones (1981) continuó con el estudio de este modelo, y ya en las décadas de los 80 y 90 otros autores como Robertson (1976) se han dedicado a investigar en el modelo probabilístico y en los métodos de evaluación.

A finales de los 80 se comienzan a usar técnicas basadas en el conocimiento como en la creación de los sistemas expertos. En esta misma época también surge otra línea de investigación que es la del procesamiento del lenguaje natural (PLN). Las últimas tendencias en RI combinan el PLN con métodos de análisis sintáctico, lematización, n-gramas, inteligencia artificial, sistemas expertos, redes neuronales y algoritmos genéticos.

En definitiva, la RI es una técnica en continuo crecimiento y de gran interés para los investigadores.

3.1. Modelos de Recuperación de Información

A continuación se describen los modelos de recuperación de información (RI) que se han desarrollado a lo largo de la historia y que conforman muchas de las técnicas que se utilizan hoy día en modelos de recuperación más avanzados.

3.1.1. Modelo booleano

El método booleano es un método simple de RI basado en la teoría del álgebra de Boole. Este método de recuperación utiliza expresiones booleanas para realizar las sentencias de búsqueda, las cuales tienen una semántica muy específica. Este método tuvo una gran importancia hace años y fue adoptado por los primeros sistemas comerciales bibliográficos.

La estrategia de recuperación de este modelo se basa en un sistema binario, es decir un documento puede encontrarse en el criterio de la sentencia de búsqueda o no puede hallarse en la expresión de búsqueda $\{1,0\}$. Si este documento coincide con la expresión de búsqueda tendrá un valor 1 y si por el contrario no se corresponde con dicho criterio, tendrá el valor 0. Es decir el método booleano considera que los términos indizados están presentes o ausentes en un documento.

Una ecuación de búsqueda q se compone de términos que se enlazan con tres conectores Y, NO, O (AND, NOT, OR) la figura muestra un ejemplo de los tres componentes conjuntivos para la ecuación de búsqueda q

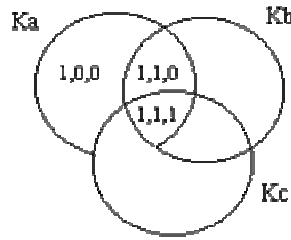


Figura 1. Método Booleano

Los tres componentes de la conjunción para la consulta $[q = k_a \wedge (k_b \vee \neg k_c)]$

Así las expresiones booleanas se forman por las ecuaciones de búsqueda que realiza un usuario. En algunos sistemas de recuperación los usuarios introducen las expresiones booleanas directamente, pero en algunos sistemas más sofisticados los usuarios introducen la expresión de búsqueda en lenguaje natural, el cual el sistema convierte en expresión booleana.

El siguiente ejemplo es una muestra de las expresiones booleanas (Frakes y Baeza-Yates, 1992, p. 268):

Considerando un conjunto de cinco documentos y suponiendo que estos contienen los términos que muestra la tabla 1:

La expresión {conductores y cobre} es:

$$\{d1, d3\} \wedge \{d1, d2, d4\} = \{d1\}$$

La expresión {conductores o cobre} es:

$$\{d1, d3\} \vee \{d1, d2, d4\} = \{d1, d2, d3, d4\}$$

y un ejemplo más complejo sería el siguiente: {conductores y cobre} o no{conductores y estaño}

$$(\{d1, d3\} \wedge \{d1, d2, d4\}) \vee \{d1, d2, d3, d4, d5\} - (\{d1, d3\} \wedge \{d2, d3, d4, d5\}) = \{d1\} \vee \{d1, d2, d3, d4, d5\} - \{d3\} = \{d1, d2, d4, d5\}$$

<i>Documentos</i>	<i>Términos</i>
d1	dispositivo, conductores, cobre
d2	cobre, estaño
d3	dispositivo, conductores, estaño
d4	presión, cobre, estaño
d5	estaño, dispositivo

Tabla 1. Ejemplo Método Booleano

En definitiva, el método booleano sitúa un 0 cuando no encuentra la raíz y un 1 cuando sí se encuentra, asignando un valor que corresponderá con el número de veces que aparece esa palabra en el documento, este método resulta ser demasiado drástico y no

suficientemente flexible aunque sigue utilizándose hoy día en gran cantidad de sistemas de RI.

Para dar mayor flexibilidad al método booleano se utilizan también otra serie de operadores de comparación como: mayor que, menor que, igual que, distinto que ($>$; $<$; $=$; \neq) y operadores posicionales que localizan la posición de un término en relación a los demás, la cercanía entre los términos puede reflejar una relación entre los conceptos.

Una propuesta de mejora de este método que combina el sistema tradicional de recuperación booleano con el modelo de espacio vectorial es el conocido como modelo booleano extendido⁵.

3.1.2. Lógica difusa

En 1965 Lotfi A. Zadeh, un matemático estadounidense de origen iraní, profesor de ciencia de computadoras en la Universidad de California en Berkley publica un documento en donde por primera vez se menciona el término *Fuzzy Logic*, conocida como Lógica difusa o Lógica borrosa.

La lógica difusa reconoce no sólo los valores de verdadero o falso, cero o uno de la lógica clásica, la lógica difusa admite un tercer valor posible o grado de pertenencia, de esta manera los valores cero y uno se subdividen en valores infinitos entre ellos existiendo entre el cero y el uno los siguientes valores posibles (**0**, 0,1,0,2...**1**)

El principal objetivo de la lógica difusa es tratar con los modelos de razonamiento que son más aproximados que exactos y en esto difiere en esencia de la lógica clásica.

Los sistemas tradicionales de RI han recurrido al empleo de herramientas como los operadores booleanos, operadores de proximidad, caracteres de truncamiento, etc. A fin de obtener resultados más precisos. Sin embargo el correcto uso y aplicación de dichas herramientas requiere de un conocimiento en la materia por parte del usuario que realiza la consulta de búsqueda al sistema o base de datos. Los motores de búsqueda que recurren a la lógica difusa permiten que los usuarios con poca experiencia en sistemas de RI logren buenos resultados en poco tiempo mediante el uso de lenguaje natural en sus búsquedas.

La lógica difusa aplicada a los sistemas de RI permite la posibilidad de realizar una búsqueda expresada en un lenguaje natural, siendo el motor de búsqueda el responsable de traducir esa búsqueda a un lenguaje estructurado. A estos sistemas se les conoce como sistemas expertos o de búsqueda difusa siendo capaz de generar información no explícita, posibilitando recuperar información similar a la solicitada, aunque los términos utilizados no tengan exacta coincidencia, así se consigue recuperar documentos que traten la misma temática aunque dichos términos no aparezcan en los documentos recuperados.

⁵ Vid. PEÑA, R.; BAEZA-YATES, R. Y RODRÍGUEZ MUÑOZ, J.V (col.). (2002). Gestión digital de la información: de bits a bibliotecas digitales y la Web. Madrid: Ra-ma

Si buscamos una aplicación de la lógica difusa o teoría de los conjuntos borrosos a la RI podemos afirmar que ésta consiste en representar para una combinación de dos términos el número de veces que éstos coinciden en un documento (c_{ij}).

Cada término puede aparecer en un documento o en más de uno, de este modo se puede representar este valor en una matriz de correlación sin normalizar en la cual se representará el número de veces que dos términos coinciden en el documento.

Así se expone un ejemplo explicativo con sólo cuatro documentos de tipo literario, poético.

<p><i>Documento 1</i> La luna en el mar ríela Y en la lona gime el viento</p>	<p><i>Documento 3</i> Esta la mar en calma La luna estaba crecida Solo se oían los remos</p>
<p><i>Documento 2</i> Sobre las aguas del mar Vio venir una galera Que a tierra quiere llegar</p>	<p><i>Documento 4</i> Amarrado al duro banco De una galera turquesca Ambas manos en el remo Y ambos ojos en la tierra</p>

Después de declarar artículos, verbos, etc. como vacías, quedarían los siguientes términos o palabras.

	<i>Términos</i>	<i>Documentos</i>
1	luna	1,3
2	mar	1,2,3
3	lona	1
4	viento	1
5	agua	2
6	galera	2,4
7	tierra	2,4
8	calma	3
9	remo	3,4
10	banco	4
11	mano	4
12	ojo	4

Tabla 2. Ejemplo Lógica Difusa

En la matriz de correlación sin normalizar se representa el número de documentos en los que aparecen conjuntamente los dos términos.

Hasta el momento no se ha resuelto un punto de vista muy importante, qué ocurre con los términos que son escasos en la totalidad de los documentos pero aparecen coincidentes en un documento, este caso es muy significativo. Al contrario que dos términos que son abundantes y por tanto no es extraño que coincidan.

La respuesta a esta cuestión se solventa con la necesidad de normalización. Para formar la matriz de correlación normalizada es necesario la utilización de una fórmula que

determinará que valor tiene una palabra o término en relación a un documento, esto se obtiene mediante un valor numérico comprendido entre 0 y 1, así se obtendrá el valor de pertinencia que tiene cada término en relación a los demás términos en cada documento. Es decir, se obtiene de este modo valores comprendidos entre 0 y 1, que representarán mejor la asociación entre términos.

Así pues para formar la matriz de correlación normalizada se divide cada término c_{ij} por (número de documentos en que aparece el término i + el número de documentos en que aparece el término j) – el número de documentos en que aparecen a la vez.

$$\frac{c_{ij}}{c_i + c_j - c_{ij}}$$

Esta fórmula es válida pero existe otra de su autor llamado *Dice* en la cual se puede obtener valores aproximados y resulta igual de eficaz, dicha fórmula es la siguiente:

$$\frac{2 \cdot c_{ij}}{c_i + c_j}$$

En este ejemplo concreto que se describe a continuación se ha utilizado la primera de las fórmulas expuesta anteriormente. De este modo se obtienen los datos siguientes:

<i>Antes de normalizar</i>	<i>después de normalizar</i>
C(2,9) 1	0,2
C(10,11) 1	1

En definitiva, con los valores de la matriz se obtienen las funciones para cada documento en función de cada término. Esta función proporciona un valor entre 0 y 1 que representa la relación entre los distintos términos que aparecen en el documento según la fórmula que sigue a continuación

$$M_{i,doc} = 1 - (1 - c_{ij}) \cdot (1 - c_{ij})$$

Con esta fórmula se obtendrá el valor de pertinencia de un documento sobre un término. A continuación se presenta un ejemplo con los datos expuestos en los documentos ejemplo.

CB banco

Mbanco,4=1
Mbanco,1=1-(1-0) (1-0) (1-0) (1-0)=0
 luna mar lona viento

Mbanco,2= 1-(1-0) (1-0) (1-0,5) (1-0,5)=0,75
Mbanco,3= 1-(1-0) (1-0) (1-0,5)= 0,5
 luna mar remo

Estos datos se pueden interpretar de manera que ante la petición de obtener los documentos que contengan el tema representado por el término *banco*, desde luego

aparece el documento cuatro que contiene esa palabra, pero aparece también como documento bastante interesante el dos, aunque no tenga la palabra en cuestión. En definitiva esta función dará valores altos cuando no apareciendo la palabra en el documento, hay otras palabras que coinciden con la palabra buscada en otros documentos. Este método requiere que el sistema tenga capacidad de memoria.

3.1.3. Clustering

Las técnicas de análisis de clusters y los sistemas de información tienen un mismo objetivo: organizar por temas la información almacenada. Basados en el cálculo de la similitud entre pares de términos es la operación de agrupar juntos documentos similares o afines en clases o grupos, con el propósito del almacenamiento de la información.

El análisis del cluster permite la identificación de grupos y clases similares en un espacio multidimensional. (Salton, 1971) en el SMART establece una jerarquía que se forma dividiendo el documento en unos cuantos clusters que a su vez se dividen en otros más pequeños y así sucesivamente, utilizando la medida de similitud para las búsquedas de los usuarios.

Los algoritmos de clustering estudian la forma en que se agrupan los términos de indexación asignados a los documentos o los propios documentos para revelar la relación que existe entre documentos de materias similares y crear de este modo grupos con características comunes.

En el campo de la RI el análisis de cluster ha sido utilizado para crear grupos de documentos con la meta de mejorar la eficiencia y la efectividad de la recuperación o determinar la estructura de la literatura de un campo. Los términos en una colección de documentos pueden ser agrupados para mostrar sus relaciones.

Para medir la intensidad relativa de las apariciones conjuntas de los términos o palabras en los documentos se tiene en cuenta las frecuencias de las dos palabras consideradas y se utiliza el denominado índice de equivalencia o de asociación, el cual mide la intensidad de la asociación entre dos palabras i y j realizada sobre el conjunto de documentos del fichero. Se obtiene el valor 1 cuando la presencia de i acarrea automáticamente la presencia de j , y viceversa, es decir, cuando las dos palabras están siempre juntas. Por el contrario es igual a 0 cuando la mera presencia de una de las dos palabras excluye la de la otra. Así llamaremos índice de equivalencia (E_{ij}) al coeficiente cuyo valor viene dado por la fórmula siguiente:

$$E_{ij} = \frac{C_{ij}^2}{c_i + c_j}$$

Donde, en un documento grande o fichero F , dividido en documentos pequeños n .

C_i = número de apariciones de la palabra clave i en la totalidad de documentos n

C_j = número de apariciones de la palabra clave j en la totalidad de documentos n .

C_{ij} = número de apariciones conjuntas de las palabras i y j en la totalidad de documentos n .

El cálculo de todos los coeficientes entre todos los pares de palabras posibles genera un número de relaciones importantes pero sería vano pretender visualizarlo. Por eso se utilizan algoritmos para identificar clusters que reúnan las palabras que están frecuentemente asociadas a otras, es decir, entre las cuales los índices de equivalencia son altos. Un método para construir estos clusters puede consistir en no tomar en consideración más que las relaciones existentes y en ordenar los clusters en función de la fuerza de las asociaciones entre las palabras que los constituyen (Callon, Courtial, y Herve, 1995). Existen algoritmos que seleccionan entre todas las palabras del cluster aquella que es la más central y que se retiene para asignar un nombre al cluster.

Los clusters tienen dos parámetros cuantitativos: la densidad y la centralidad, Según Callon, Courtial, y Herve (1995) nos ofrecen nociones de centralidad y densidad en los clusters: *la centralidad* se ocupa en un cluster de la intensidad de sus relaciones con otros clusters. La medición de la centralidad permite ordenar los diferentes clusters procedentes de un fichero por orden de centralidad creciente.

La densidad pretende caracterizar la intensidad de las relaciones que unen las palabras que componen un cluster determinado. Los clusters pueden ser colocados por orden de densidad creciente.

Según Ruiz-Baños y Contreras-Cortés (1998) la centralidad o índice de cohesión externa es la suma de los índices de equivalencia de todos los enlaces externos que posee el tema con otros. Y el concepto de densidad o índice de cohesión interna lo define como la intensidad de las asociaciones internas de un tema y representa el grado de desarrollo que posee.

Aparte de los dos parámetros cuantitativos: la densidad y la centralidad existe dos métodos principales de análisis de cluster, son los no jerárquicos, que dividen un conjunto de datos de N temas en M clusters, y los jerárquicos, que producen un anidamiento en los conjuntos de datos en que pares, parejas de temas o clusters son vinculados sucesivamente.

Los métodos no jerárquicos dividen un conjunto de N documentos en M clusters donde no se permite trasladarse

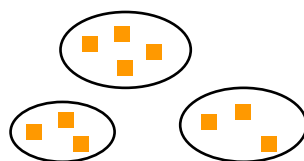


Figura 2. Clustering. Métodos no jerárquicos

Los métodos jerárquicos producen el anidamiento de un conjunto de documentos en los cuales pares de documentos o clusters son vinculados hasta que están relacionados todos los documentos de la colección.

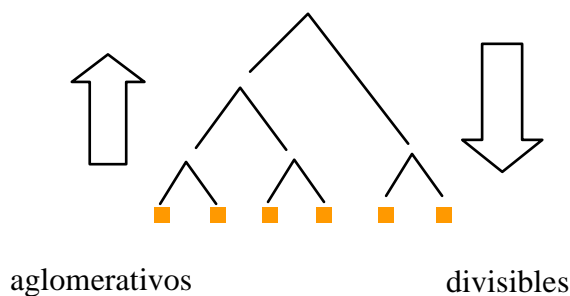


Figura 3. Clustering. Métodos jerárquicos

- ✓ Aglomerativos: documentos individuales son unidos en clusters, y los clusters pequeños en clusters más grandes.
- ✓ Divisibles: todos los documentos están en un cluster gigante, los cuales se van dividiendo sucesivamente.

El método no jerárquico y los métodos de realojamiento son heurísticos y naturales y requieren menos cálculos que el método jerárquico. Sin embargo los métodos jerárquicos son en general seleccionados para el clustering basado en la recuperación de documentos.

Entre las aplicaciones que tiene el clustering en la RI el análisis de cluster tiene la capacidad para clasificar temas por asignación que automáticamente crean grupos, los cuales dan una afinidad natural con el propósito del almacenamiento de la información. El análisis de cluster puede ser representado en los documentos por diferentes caminos (Frakes y Baeza-Yates, 1992, p.420):

1. Documentos que pueden ser agrupados por medio de los términos básicos que ellos contienen. El propósito de esta aproximación ha sido proporcionar una recuperación más eficiente y más efectiva, aunque haya sido usado después de la recuperación para proporcionar una estructura de grandes conjuntos de documentos recuperados. En los sistemas distribuidos el clustering puede ser usado para asignar documentos para el almacenamiento
2. Los documentos pueden ser agrupados basados en las coocurrencias de las citas para proporcionar la percepción general que existe en el campo de la literatura.
3. Los términos pueden ser agrupados en la base de los documentos en que estos tienen la coocurrencia para ayudar en la construcción de un tesoro o en la mejora de las sentencias de búsqueda

Si bien el análisis de cluster puede ser fácilmente implementado con paquetes de software asequibles, esto incluye:

1. Seleccionar los atributos en que temas serán agrupados y su representación.
2. Seleccionar un apropiado método de cluster y medida de similitud de entre los que se pueda disponer
3. Creación de los clusters o clusters jerárquicos, los cuales pueden ser caros en términos de recursos informáticos
4. Valorar la validez del resultado obtenido

5. En la colección que es agrupada de forma dinámica, los requisitos para ponerse al día deben ser considerados
6. Si el propósito es usar la colección agrupada como la base de la recuperación de la información, debe ser seleccionado un método para la búsqueda del cluster o del cluster jerárquico.

Por tanto el clustering de documentos es la operación de agrupar juntos documentos similares o afines en clases. A este respecto, el clustering de documentos no es una operación en el texto sino una operación en la colección de documentos.

La operación del clustering de documentos es en general de dos tipos según Baeza-Yates y Ribeiro-Neto (1999, p. 173): global y local. En una estrategia de clustering global, los documentos son agrupados con sus ocurrencias en toda la colección. En una estrategia de clustering local, el agrupamiento de los documentos se ve afectado por el contexto definido por la consulta en cuestión y su conjunto *local* de documentos recuperados.

Los métodos de clustering se utilizan en la RI para transformar las consultas originales, en el intento de mejorar la representación de las necesidades de información del usuario. Desde esta perspectiva, el clustering es una operación que es más afín a la transformación de las consultas que realiza el usuario que a la transformación del texto de los documentos.

Las metodologías empleadas en la automatización de la indización desde finales de los años 50 hasta la actualidad han variado. En los primeros momentos se utilizaba casi exclusivamente la estadística para obtener los términos de indización representativos de los documentos, pero a partir de los años 80 se incorporaron en las propuestas para la automatización de la indización técnicas de procesamiento del lenguaje natural como herramientas para conseguir las raíces de las palabras, etiquetadores morfológicos, así como analizadores sintácticos, entre otras. Pero lo habitual es que las propuestas o prototipos presentados por los investigadores incluyan una combinación de ambas aproximaciones, es decir, cálculo de la frecuencia y herramientas más o menos complejas para el procesamiento del lenguaje natural.

3.1.4. Modelo de espacio vectorial

Este método fue desarrollado a finales de los años 60 y comienzos de los años 70 por Gerald Salton, consiste básicamente en la representación de los documentos a través de vectores. De este modo en una colección o base de datos de N documentos, en el que existe un total de t términos, representamos cada documento como un vector de t -dimensiones cuyas coordenadas son los pesos de cada término.

Este método podemos definir su funcionamiento en tres pasos:

1. Representación de cada documento de una Base de datos como un vector de t -dimensiones cuyas coordenadas son los pesos de cada término.
2. Normalización de vectores
3. Obtención del producto escalar entre el vector pregunta y cada uno de los documentos de la Base de datos

1. Representación de cada documento de una Base de datos como un vector de t-dimensiones cuyas coordenadas son los pesos de cada término.

El modelo de espacio vectorial permite distinguir los términos de dos formas, de modo binario indicando la presencia o no del término ($t_i = 1$ ó $t_i = 0$) o de modo ponderado, calculando pesos en función de la importancia que tenga el término en el documento.

Este método considera las consultas de los usuarios realizadas a la base de datos y los documentos como vectores de t-dimensiones.

Un ejemplo sencillo sería el siguiente:

Definimos en una base de datos el siguiente vocabulario de indización compuesto por $t = 3$ términos:

$$V = \{t_1 = \text{ciencia}, t_2 = \text{Newton}, t_3 = \text{gravitación}\}$$

En esta base de datos existen 4 documentos, cuyos pesos específicos para cada término definen su correspondiente vector:

<i>Documentos</i>	<i>Pesos t_1, t_2, t_3</i>	<i>Temática de los documentos</i>
D ₁ =	(1,10,2)	Biografía de Newton
D ₂ =	(10,3,1)	Documento sobre ciencia
D ₃ =	(8,10,10)	Documento sobre las teorías de gravitación
D ₄ =	(4,5,0)	Documento sobre el cálculo diferencial e integral

Tabla 3. Ejemplo Método Espacio Vectorial

2. Normalización de vectores

El siguiente paso que realiza el modelo de espacio vectorial es normalizar los vectores de modo que todos tengan el mismo módulo con el fin de cuantificar la relevancia de cada término respecto al resto de los términos del documento

La normalización de los vectores se consigue dividiendo cada componente del vector (que como hemos dicho antes representa el peso del término en ese documento) entre su módulo. De esta manera, el módulo de todos los vectores es igual a 1.

Los vectores normalizados del ejemplo quedarían de la siguiente manera:

- D₁=(0'10, 0'98, 0'20)- Biografía de Newton
- D₂=(0'95, 0'29, 0'10)- Documento sobre ciencia
- D₃=(0'49, 0'62, 0'62)- Documento sobre las teorías de gravitación
- D₄=(0'62, 0'78, 0'00)- Documento sobre el Cálculo diferencial e integral

3. Obtención del producto escalar entre el vector pregunta y cada uno de los documentos de la base de datos

La mayoría de las funciones para medir la semejanza entre dos vectores está relacionada con el cálculo del coseno del ángulo que forman. Por esa razón, frecuentemente se emplea el producto escalar de los mismos.

El producto escalar de dos vectores (A y B) se define como el producto de sus módulos por el coseno del ángulo θ que forman. El resultado será siempre una magnitud escalar:

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}| \cos \theta = AB \cos \theta$$

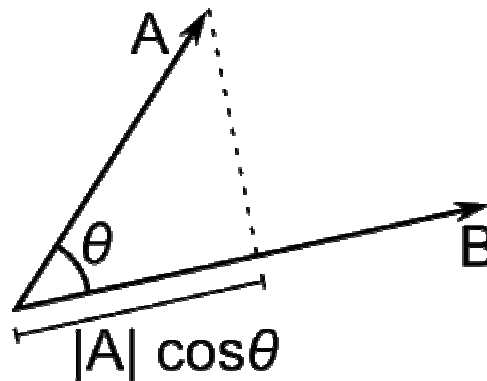


Figura 4. Método de Espacio Vectorial. Producto escalar de dos vectores

De este modo, dos vectores que coinciden formarán un ángulo $\theta = 0^\circ$, con lo que su coseno es igual a 1 y el valor de su producto escalar sería igual al producto de sus módulos. Del mismo modo, dos vectores ortogonales (es decir, que difieren completamente) formarán un ángulo $\theta = 90^\circ$, con lo que su coseno es igual a 0 y el valor de su producto escalar sería igual a 0.

Como todos los vectores están normalizados, es decir, su módulo es igual a 1, el producto escalar indica exclusivamente el valor del coseno (θ) que forman.

Por otro lado, como nosotros no sabemos de entrada el ángulo que forman los vectores, se puede emplear para el cálculo del producto escalar su expresión analítica.

Es decir, si los vectores \mathbf{A} y \mathbf{B} se expresan en función de sus componentes cartesianas rectangulares, o sea,

$$\mathbf{A} = A_x \mathbf{i} + A_y \mathbf{j} + A_z \mathbf{k}$$

$$\mathbf{B} = B_x \mathbf{i} + B_y \mathbf{j} + B_z \mathbf{k}$$

Entonces, teniendo en cuenta las propiedades anteriores:

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y + A_z B_z$$

De modo que el producto escalar de dos vectores es igual a la suma de los productos de las componentes cartesianas rectangulares correspondientes.

Volviendo al ejemplo que nos ocupa:

En el caso de que un usuario esté interesado en documentos relativos a la vida de Newton, la consulta a la base de datos podría tener los siguientes pesos:

$$P = (0,10,0)$$

Y normalizando este vector queda de la siguiente manera:

$$P_{norm} = (0,1,0)$$

Los vectores normalizados que representan a los documentos de nuestro ejemplo son los siguientes:

$D_1=(0'10, 0'98, 0'20)$ - Biografía de Newton

$D_2=(0'95, 0'29, 0'10)$ - Documento sobre ciencia

$D_3=(0'49, 0'62, 0'62)$ - Documento sobre las teorías de gravitación

$D_4=(0'62, 0'78, 0'00)$ - Documento sobre el Cálculo diferencial e integral

Por tanto, para la similitud entre estos documentos y la pregunta se calculará como el producto escalar de cada uno de ellos:

$$P \cdot D_1 = (0 \cdot 0'10 + 1 \cdot 0'98 + 0 \cdot 0'20) = 0'98$$

$$P \cdot D_2 = (0 \cdot 0'95 + 1 \cdot 0'29 + 0 \cdot 0'10) = 0'29$$

$$P \cdot D_3 = (0 \cdot 0'49 + 1 \cdot 0'62 + 0 \cdot 0'62) = 0'62$$

$$P \cdot D_4 = (0 \cdot 0'62 + 1 \cdot 0'78 + 0 \cdot 0'00) = 0'78$$

Como conclusión, se observa que el mayor valor del producto escalar corresponde al Documento 1 que consiste en la Biografía de Newton.

A pesar de que el Documento 3 tiene el mismo peso para el 2º término (*Newton*), el proceso de normalización relativiza la importancia del término frente al resto de términos del mismo documento. De este modo, el producto escalar del vector D_3 con P es menor que con el del D_1 .

3.1.4.1. Inverse Document Frequency. IDF

Parte del éxito del método vectorial se debe a que la ponderación de los términos se realiza de forma más compleja utilizando para ello la frecuencia inversa del documento (Inverse Document Frequency. IDF) introducido por Spark Jones (1972)

Delimitar el peso de un término en un documento por la frecuencia de aparición es inexacto, es por ello que existen varios métodos que consideran todos los términos de la colección y que aplican diversos procedimientos para la ponderación del peso, los cuales se detallan a continuación:

Ponderación por el tamaño del documento: en la que el peso de cada término se calcula como el cociente entre la frecuencia absoluta del término y el tamaño del documento.

Un término con igual frecuencia en dos documentos se considera más importante en el documento más corto.

$$P_{ij} = \frac{F_{ij}}{t_j}$$

Ponderación inversa de la frecuencia de documentos: (Inverse Document Frequency), este método se expone posteriormente con más detalle, *Idf* del término t_i viene determinada por la fórmula:

$$Idf_i = \log_2 \left(\frac{n}{d_i} \right) + 1$$

El *idf* de un término decrece de manera logarítmica con el aumento del número de documentos en que aparezca el término en la colección. Cuando d_i disminuye *idf_i* aumenta. Como d_i es un número comprendido entre n y 1 ($n > d_i > 1$), *idf_i* toma el valor mínimo 1 aumentando hasta un valor constante ($1 < idf < \log_2(n) + 1$), y el peso del término en el documento se estima como

$$P_{ij} = F_{ij} \cdot idf_i$$

De modo que entre dos términos con igual frecuencia absoluta en un documento se asigna mayor peso a aquel que aparece menos veces en la colección.

Discriminación: esta ponderación pretende otorgar más importancia (más peso) a los términos del vocabulario seleccionado que clasifiquen mejor a los documentos dentro de la colección concreta en que se están considerando.

La autora Spark Jones (1972, p.11-20) fue la descubridora del tipo de fórmulas capaces de expresar la importancia informativa de unas palabras en un documento, lo que se denominó como indización ponderada o por pesos. En estas fórmulas es esencial la presencia de la función logaritmo, como función matemática que atempera el crecimiento de una variable. Este logaritmo llamado *ITF* (*inverse term frequency*) o frecuencia inversa del término obtiene la importancia de un término dependiendo del número de veces que aparezca en el documento en relación con el resto de términos que tenga, desechando las palabras que aparezcan tanto en exceso como por defecto, al considerar que no aportan valor informativo suficiente como para describir el contenido del documento. De igual modo, desarrolló el *IDF* (*Inverse Document Frequency*) o frecuencia inversa del documento, cuyo procedimiento es similar al anterior pero se emplea en todos los documentos de una base de datos que contienen la colección. Con este logaritmo obtendríamos los documentos ordenados de mayor a menor relevancia ante una consulta en un corpus dado.

Para expresar el poder discriminatorio de una palabra existen varios tipos de formulaciones que ofrecen un valor numérico para cada palabra de un documento según sea su poder discriminatorio respecto a las demás.

Es evidente que algunas palabras resultan más significativas que otras, de manera intuitiva podemos pensar que no debe representar igual el contenido de un documento una palabra que aparece en casi todos los documentos y por tanto es muy frecuente, por tanto ésta palabra sería poco importante, que otra que aparece en pocos documentos pero muchas veces, por tanto esta palabra caracterizaría dicho documento. Como siempre se debe considerar la frecuencia relativa y no el número de veces que aparece.

El poder discriminatorio de una palabra o término es inversamente proporcional a su frecuencia de aparición en un documento y es directamente proporcional a su frecuencia de aparición en un documento. El peso de un término depende del inverso del número de veces que aparece el término en toda la colección y del número de veces que aparece el término en ese documento.

Para calcular este comportamiento de las palabras en los documentos utilizaremos algún tipo de coeficiente o peso que intente expresar el valor o importancia de cada palabra en cada uno de los documentos como el *IDF* (*Inverse Document Frequency*) o inverso de la frecuencia de aparición en documentos. Es un valor numérico asociado a cada palabra de una colección de documentos.

En una colección de N documentos la palabra p aparece en n documentos, siendo la frecuencia de aparición de la palabra p :

$$Fr_p = \frac{n}{N} \quad [1]$$

Como vamos a valorar mejor a las palabras que aparezcan en pocos documentos necesitamos un número que sea más grande cuanto menos aparezca la palabra, por eso tomamos el inverso de lo anterior [1]:

$$Fr_p = \frac{N}{n} \quad [2]$$

Con esta fórmula [2] básica obtendríamos valores muy exagerados, por ejemplo, si la colección es de 10.000 documentos, una palabra que aparezca en 1.000 obtiene un valor de $\frac{N}{n} = 10$, mientras que otra que aparezca sólo en 10 documentos $\frac{N}{n} = 1.000$. La relación 10 a 1.000 resulta muy exagerada, por ello se matiza con una función de crecimiento lento como el logaritmo y se define:

$$IDF_p = \log \frac{N}{n} \quad [3]$$

De este modo volviendo al ejemplo anterior obtendríamos utilizando la fórmula [3] la relación 2,3 a 6,9 que resulta más atenuada.

En el caso de palabras vacías que aparecen con mucha frecuencia en todos los documentos se obtendría $IDF = \log \frac{10.000}{10.000} = \log(1) = 0$

En el caso de palabras que aparecen en un solo documento se obtendría un valor máximo para esta determinada colección, $IDF = \log \frac{10.000}{1} = \log(10.000) = 9,2$

Como decíamos existen variedad de formulaciones para obtener el peso de los términos o palabras en la colección de documentos, todas ellas similares entre sí, la fórmula [3] que planteamos del *IDF* es la propuesta por Salton (1989) y en la que calcula el *IDF*:

$$idf = \log \frac{n}{df_i} \quad [4]$$

Donde df_i es el número de documentos en una colección de n documentos en la que el término t aparece.

Por otro lado la autora Spark Jones, K. (1972, p. 11-20) para el cálculo del *IDF* propone dos fórmulas:

$$idf = \log_2 \frac{N}{n_i} + 1 \quad [5]$$

Donde:

N =es el número total de documentos en la colección

n_i = el número total de aparición del término i en la colección

Otra de las fórmulas propuesta por la autora Spark Jones, K. (1972, p.133-144) para el cálculo del *IDF* es la siguiente:

$$idf = \log_2 \frac{\max}{n_i} + 1 \quad [6]$$

Donde \max_n es la frecuencia máxima de un término en una colección.

En definitiva, el valor del *IDF* puede calcularse para cada palabra y almacenarse junto a los índices para ser utilizado después en la indización automática y en las búsquedas. De las dos ideas a capturar (aparecer en pocos documentos y aparecer muchas veces en ellos) sólo refleja la primera, por lo que en el momento de responder a una búsqueda, aún debemos incorporar de algún modo la segunda idea: frecuencia de la palabra dentro del documento.

3.1.5. Indización semántica latente

Según Venegas (2003) la indexación semántica latente (LSI: Latent Semantic Indexing) o también llamada análisis semántico latente (LSA: Latent Semantic Analysis) es un método de análisis estadístico que estima la estructura latente. Esta técnica utiliza un valor singular de descomposición que segmenta una gran matriz de datos de asociación de término-documento y permite construir un "espacio semántico" en el que se asocian entre sí términos y documentos. El sustento estadístico en definitiva está determinado por la coocurrencia de palabras en la diversidad de documentos.

Según las definiciones aportadas por investigadores desde los inicios de la década de los 90 hasta nuestros días, la LSA según (Deerwester et al.,1990; Foltz, 1990), Foltz (1996) la definen como un modelo estadístico de uso de palabras que permite comparaciones de similitud semántica entre piezas de información textual.

Según (Landauer, Foltz y Laham, 1998, p. 263). La LSA *“es una técnica matemático-estadística totalmente automática para extraer e inferir relaciones de uso contextual esperado de palabras en pasajes de discurso. No es un procesamiento de idioma natural tradicional o programa de inteligencia artificial; no usa ningún diccionario construido humanamente, bases de conocimiento, redes semánticas, gramáticas, segmentadores sintácticos, o morfologías y toma como input sólo la segmentación del texto en palabras, pasajes, frases o párrafos”*

En trabajos más actuales, Kintsch (2001) define al LSA como un procedimiento totalmente automático de técnicas matemáticas estándar que sirve para analizar un gran corpus de texto digitalizado.

Para concluir, se puede establecer que el Análisis Semántico Latente o también llamada Indexación Semántica Latente se caracteriza por ser una técnica matemático-estadística que permite la creación de vectores multidimensionales para el análisis semántico de las relaciones existentes entre palabras, palabras y párrafos, y entre párrafos. En tanto que su valor como teoría de la representación del conocimiento humano, para algunos sólo explicaría parte del conocimiento del establecimiento de relaciones semánticas.

3.1.6. Redes neuronales

Según Gómez Díaz (2005) las redes neuronales consisten en una simulación de las propiedades observadas en los sistemas neuronales biológicos, como el cerebro humano (procesos de percepción, recuerdo, clasificación y decisión del conocimiento) a través de modelos matemáticos recreados mediante modelos informáticos. El objetivo es conseguir que las máquinas den respuestas similares a las que es capaz de dar el cerebro que se caracterizan por su generalización y su robustez. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales.

Una red neuronal⁶ se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones:

1. Una función de propagación, que por lo general consiste en el sumatorio de cada entrada multiplicada por el peso de su interconexión (valor neto).
2. Una función de activación, que modifica a la anterior. Puede no existir, siendo en este caso la salida la misma función de propagación.
3. Una función de transferencia, que se aplica al valor devuelto por la función de activación. Se utiliza para acotar la salida de la neurona y en general viene dada

⁶ Wikipedia: Red neuronal artificial
http://es.wikipedia.org/w/index.php?title=Red_neuronal_artificial&oldid=58620798. [ref. agosto 2012]

por la interpretación que queramos darle a dichas salidas. Algunas de las más utilizadas son la función sigmoidea (para obtener valores en el intervalo $[0,1]$) y la tangente hiperbólica (para obtener valores en el intervalo $[-1,1]$).

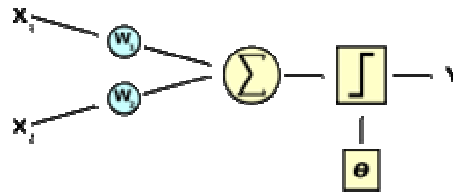


Figura 5. Red Neuronal

Las redes neuronales artificiales (RNA) tienen muchas ventajas debido a que están basadas en la estructura del sistema nervioso, como el cerebro.

- **Aprendizaje:** Las RNA tienen la habilidad de aprender mediante una etapa que se llama *etapa de aprendizaje*. Esta consiste en proporcionar a la RNA datos como entrada a su vez que se le indica cuál es la salida (respuesta) esperada.
- **Auto organización:** Una RNA crea su propia representación de la información en su interior, descargando al usuario de esto.
- **Tolerancia a fallos:** Debido a que una RNA almacena la información de forma redundante, ésta puede seguir respondiendo de manera aceptable aun si se daña parcialmente.
- **Flexibilidad:** Una RNA puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada (ej. si la información de entrada es la imagen de un objeto, la respuesta correspondiente no sufre cambios si la imagen cambia un poco su brillo o el objeto cambia ligeramente)
- **Tiempo real:** La estructura de una RNA es paralela, por lo cuál si esto es implementado con ordenadores o en dispositivos electrónicos especiales se pueden obtener respuestas en tiempo real.

Un ejemplo de red neuronal aplicada a los sistemas de RI, son los mapas autoorganizados o *SOM* (Self-Organizing Map), también llamados redes de Kohonen son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de normalmente dos dimensiones, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red los vectores de datos son introducidos en cada neurona y se comparan con el *vector de peso* característico de cada neurona. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora (o *BMU*) y ella y sus vecinas verán modificados sus vectores de pesos.

3.1.7. Modelo probabilístico

En los años 60 tienen su apogeo los modelos de recuperación de información probabilísticos y el modelo de espacio vectorial. La indización probabilística fue propuesta en esencia por Maron & Kuhns (1960), estos autores definieron un modelo probabilístico de recuperación de información que fue desarrollado por Robertson y Spark Jones (1976) en la década siguiente y estudiado por diversos autores en las décadas de los ochenta y noventa, entre los que destaca Salton (1988). Estos autores se han dedicado a investigar en el modelo probabilístico y en los métodos de evaluación.

Los métodos probabilísticos se basan en una estimación probabilística de la relevancia del documento obtenido con respecto a una pregunta de un usuario.

Los antecedentes de este modelo están en los estudios que Luhn (1957) y Zipf (1949) realizaron en torno a los años 50. El primero usaba la frecuencia de aparición de palabras en un documento para determinar si eran suficientemente significativas para representar el contenido o las características del mismo, por lo tanto, la frecuencia de aparición de esa palabra en el cuerpo del documento podía ser también utilizada para indicar el grado de significación. Esto proporcionó un simple esquema de pesado de palabras clave en cada lista y hacía posible la representación de un documento en la forma de “pesado de palabras clave-descripción”. En este modelo a diferencia del modelo de espacio vectorial, se estima la relevancia de un término en un documento en función de su frecuencia de aparición.

Cada combinación documento-término tiene un peso entre 0 y 1 que nos aporta la información del valor que el término tiene en el documento. El valor de este peso se calcula multiplicando la frecuencia del término en el documento (tf) por el número de veces que ese término aparece en la base de datos. La ecuación que propone Belkin (1987) es:

$$tf = \sum d_i q_j$$

Donde q_i es:

$$q_i = \log pr_i \frac{(1 - pnr_i)}{(1 - pr_i)}$$

La principal atracción de este modelo es la información que podemos extraer de los términos y las características de aparición conjunta de varios de ellos (coocurrencia de términos).

El criterio para recuperar un determinado documento o no es el cálculo de la probabilidad de que éste sea relevante para una pregunta dada. Para ello es preciso determinar las propiedades que definen el conjunto de documentos relevantes.

El modelo probabilístico parte de la presencia o ausencia de los términos de la consulta en los documentos de la colección. Utiliza índices de los términos descriptores con pesos definidos previamente. De esta manera se consigue que el sistema efectúe la

recuperación incidiendo sobre todo en los mejores descriptores de entre los empleados por el usuario en la consulta, minimizando la importancia de los peores.

Respecto a los pesos, el modelo probabilístico es capaz de calcular el grado de relevancia entre cada documento para una consulta dada. De esta manera permite ordenar los documentos de la colección en orden descendente de probabilidad de relevancia en relación a la consulta, superando así la gran deficiencia del modelo booleano.

Ahora bien, el modelo probabilístico necesita una hipótesis inicial para establecer los documentos relevantes y el peso de sus descriptores. Además, no tiene en cuenta la frecuencia de los términos índice y supone que estos son independientes entre sí. Por ello la estimación de las probabilidades iniciales sigue siendo una de las áreas más activas entre sus especialistas.

En definitiva, este es un modelo más teórico basado en la estadística para deducir fórmulas con mayor significado y no totalmente experimentales. Más que un método independiente consiste en utilizar ecuaciones y fórmulas de probabilidades para obtener valores de ponderación que luego se aplican a alguno de los métodos mencionados, por tanto el método probabilístico puede ser booleano pero también se puede aplicar al método vectorial.

3.1.8. Redes bayesianas

Una red bayesiana⁷ es un modelo probabilístico multivariado que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente influencia causal. Gracias a su motor de actualización de probabilidades el Teorema de Bayes, las redes bayesianas son una herramienta muy útil en la estimación de probabilidades ante nuevas evidencias.

Las redes Bayesianas son gráficos acíclicos dirigidos cuyos nodos representan variables y los arcos que los unen codifican dependencias condicionales entre las variables. Los nodos pueden representar cualquier tipo de variable ya sea un parámetro medible (o medido), una variable latente o una hipótesis. Existen algoritmos que realizan inferencias y aprendizaje basados en redes bayesianas.

3.1.9. Algoritmos genéticos

En los años 1970 surgió una de las líneas más prometedoras de la inteligencia artificial, la de los algoritmos genéticos⁸. Son llamados así porque se inspiran en la evolución biológica y su base genético-molecular. Estos algoritmos hacen evolucionar una población de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica (mutaciones y recombinaciones genéticas), así como también

⁷ Wikipedia: Red bayesiana http://es.wikipedia.org/w/index.php?title=Red_bayesiana&oldid=58723397. [ref. agosto 2012]

⁸ Wikipedia: Algoritmo genético http://es.wikipedia.org/w/index.php?title=Algoritmo_gen%C3%A9tico&oldid=57607054. [ref. agosto 2012]

a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados, que sobreviven y cuáles los menos aptos, que son descartados. También son conocidos como algoritmos evolutivos e incluye las estrategias evolutivas, la programación evolutiva y la programación genética.

Un algoritmo genético es un método de búsqueda dirigida basada en probabilidad. Bajo una condición muy débil (que el algoritmo mantenga elitismo, es decir, guarde siempre al mejor elemento de la población sin hacerle ningún cambio) se puede demostrar que el algoritmo converge en probabilidad al óptimo. En otras palabras, al aumentar el número de iteraciones, la probabilidad de tener el óptimo en la población tiende a uno (1).

Según Gómez Díaz (2005) los algoritmos genéticos surgieron para buscar soluciones a problemas que no podían ser solucionados por métodos matemáticos o analíticos y cuya única forma de resolverlos era a través del método ensayo-error dirigido. Tienen diversas aplicaciones a la RI, por ejemplo, en la construcción de preguntas para la realimentación por relevancia, indización de documentos, compresión de datos, recuperación de documentos, filtrado de documentos, etc.

El objetivo de este tipo de algoritmos es encontrar un conjunto óptimo de documentos en el cual la mejor coincidencia sea la que el investigador necesita.

3.1.10. Procesamiento del lenguaje natural

Según la autora Gómez Díaz (2005) el Procesamiento de Lenguaje Natural, PLN también denominado como NLP (*Natural Language Processing*) es la parte de la inteligencia artificial que se encarga del estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos, investigando mecanismos efectivos para facilitar la comunicación hombre-máquina, de manera que esta sea más fluida y menos rígida que con los lenguajes controlados.

La autora indica que en cualquier sistema de PLN se trata de simular el comportamiento lingüístico humano. Para lograr esto es preciso conocer las estructuras del lenguaje, cómo se forman y combinan las palabras, qué significan, tanto aisladas como en un contexto determinado, etc.

El PLN⁹ se estructura normalmente en cuatro etapas o niveles fundamentales: análisis morfológico, análisis sintáctico, análisis semántico y análisis pragmático:

- **Análisis morfológico.** El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos. Lo habitual es que el léxico sólo contiene la raíz de las palabras con formas regulares, siendo el analizador morfológico el que se encarga de determinar si el género, número o flexión que componen el resto de la palabra son adecuados.
- **Análisis sintáctico.** El análisis de la estructura sintáctica de la frase mediante una gramática de la lengua en cuestión. Tiene como función etiquetar cada uno

⁹ Wikipedia: Procesamiento de lenguajes naturales
http://es.wikipedia.org/w/index.php?title=Procesamiento_de_lenguajes_naturales&oldid=57346773. [ref. agosto 2012]

de los componentes sintácticos que aparecen en la oración y analizar cómo las palabras se combinan para formar construcciones gramaticalmente correctas.

- **Análisis semántico.** La extracción del significado de la frase y la resolución de ambigüedades léxicas y estructurales.
- **Análisis pragmático.** El análisis del texto más allá de los límites de la frase, por ejemplo, para determinar los antecedentes referenciales de los pronombres. El análisis pragmático añade información adicional al análisis del significado de la frase en función del contexto donde aparece. Se trata de uno de los niveles de análisis más complejos cuya finalidad es incorporar al análisis semántico la aportación significativa que pueden hacer los participantes, la evolución del discurso o la información presupuesta. Incorpora así mismo información sobre las relaciones que se dan entre los hechos que forman el contexto y entre diferentes entidades.

Como apunta Gómez Díaz (2005) los primeros experimentos con PLN se dieron a partir de los años 50 en el ámbito de la traducción automática, en los años 60 el PLN consistió en métodos de análisis de palabra clave, es también en estos años cuando se comienzan a dar las primeras aplicaciones del PLN a la RI. En este sentido, los primeros trabajos buscaban conseguir índices como los elaborados de manera manual como son los trabajos de Salton (1968) y Bely (1970), ambos estudios usaban tesauros para mostrar las relaciones entre los términos. Salton comparó su sistema con los de análisis estadístico, concluyendo que estos eran mejores. A finales de los 70 con el proyecto Cranfield se muestra cómo las descripciones lenguajes controlados no eran mejores que las que usaban términos simples o incluso raíces. En este sentido se estaba empezando a mostrar el gran potencial del PLN

El PLN es una técnica deductiva que necesita información previa (léxica, gramática, semántica, pragmática) para su correcto funcionamiento, así como un corpus de conocimiento (una colección de textos).

El PLN hoy día encuentra sus aplicaciones en varios campos como: la RI, los interfaces de lenguaje natural, la traducción automática, el reconocimiento del habla, la síntesis de voz y la generación de lenguajes naturales.

3.1.11. Sistemas expertos

Según Gómez Díaz (2005) los sistemas expertos (SE) también llamados sistemas basados en el conocimiento o sistemas inteligentes son programas informáticos que emulan el proceso de razonamiento humano

Un Sistema Experto¹⁰ está formado por:

- **Base de conocimientos (BC):** contiene conocimiento modelado extraído del diálogo con un experto.

¹⁰ Wikipedia: Sistema experto

http://es.wikipedia.org/w/index.php?title=Sistema_experto&oldid=58970288. [ref. agosto 2012]

- **Base de hechos** (Memoria de trabajo): contiene los hechos sobre un problema que se ha descubierto durante el análisis.
- **Motor de inferencia:** modela el proceso de razonamiento humano.
- **Módulos de justificación:** explica el razonamiento utilizado por el sistema para llegar a una determinada conclusión.
- **Interfaz de usuario:** es la interacción entre el SE y el usuario, y se realiza mediante el lenguaje natural.

Los Sistemas Expertos con su capacidad para combinar información y reglas de actuación son hoy día como una de las posibles soluciones al tratamiento y recuperación de información. La década de 1980 fue importante en investigación y publicaciones sobre experimentos de este orden, interés que continua en la actualidad.

Lo que diferencia a estos sistemas de un sistema tradicional de recuperación de información es que éstos últimos sólo son capaces de recuperar lo que existe explícitamente, mientras que un Sistema Experto debe ser capaz de generar información no explícita, razonando con los elementos que se le dan. Pero la capacidad de los SE en el ámbito de la recuperación de la información no se limita a la recuperación. Pueden utilizarse para ayudar al usuario en selección de recursos de información, en filtrado de respuestas, etc. Un SE puede actuar como un intermediario inteligente que guía y apoya el trabajo del usuario final.

3.2. Análisis Documental

3.2.1. El análisis documental y sus niveles

El proceso documental consta de tres fases claramente diferenciadas el de *entrada* en el que se realiza la selección y adquisición de los documentos y en el que el objeto principal es el de constituir un fondo, la fase de *tratamiento* en el que se realiza el análisis documental propiamente dicho y la búsqueda de los documentos por parte de los usuarios, dónde se procesa la información y se lleva a cabo la búsqueda documental por parte del usuario y la fase final de *salida* o *difusión* en el que los documentos son recuperados por los usuarios y la difusión de la información se completa.

La fase que ocupa toda nuestra investigación se centra en la fase de *tratamiento*, se entiende por tratamiento documental según García Gutiérrez (1984) la operación intelectual de aplicar técnicas específicas normalizadas (*análisis*) a un colectivo documentario con el fin de hacerlo controlable y utilizable (*recuperación*).

En la fase de tratamiento documental existen dos niveles de análisis documental que según la definición globalizadora de García Gutiérrez (1984) es aquella técnica que permite mediante una técnica intelectual objetiva la identificación y transformación de los documentos en productos que facilitan la consulta de los originales, en aras del control documental y con el objetivo último de servir a la comunidad científica. El análisis documental se subdivide en dos tipos: el *análisis formal* y el *análisis de contenido*, el análisis formal se ocupa de la descripción bibliográfica y de la catalogación y el análisis de contenido se ocupa de la clasificación, el resumen y la indización, este último proceso documental es el objetivo de nuestras investigaciones, el cual analizaremos en profundidad para desarrollar finalmente una aplicación conducente

a realizar la indización de documentos de modo automático, en definitiva la indización automática.

Básicamente ya se han comentado cuáles son los niveles del análisis documental pero se explican a continuación con mayor detalle, entre los niveles del análisis documental están el análisis formal y el análisis de contenido. En el análisis formal la finalidad principal es la de identificar los documentos para su localización y entre las técnicas utilizadas para ello se encuentran la Descripción Bibliográfica, la cual identifica los documentos para que no se confundan entre ellos y la Catalogación en la cual se traslada a un soporte informático los datos de la descripción bibliográfica y se ordena para su localización. El análisis formal básicamente se utiliza en las bibliotecas. En el análisis de contenido la finalidad principal es la utilización de técnicas documentales para hacer accesible al usuario el documento y entre las técnicas utilizadas para ello se encuentra la clasificación, mediante la cual se organiza el conocimiento con clasificaciones jerárquicas, facetadas o especializadas, la indización, la cual es la descripción del contenido de un documento por medio de descriptores que representan los conceptos del documento, y el resumen, el cual es una descripción del contenido sustancial en el cual se reduce el contenido del documento para facilitar su conocimiento. La indización y el resumen se utilizan en centros de documentación y bibliotecas especializadas

Por tanto podríamos enumerar el resultado del análisis documental, según sus operaciones

OPERACIONES	RESULTADOS
Descripción Bibliográfica	Referencias Bibliográficas
Catalogación	Catálogos
Indización	Índices
Resumen	Resúmenes

Tabla 4. Operaciones y resultados del análisis documental

Las referencias bibliográficas son el conjunto de datos mínimos que identifica a un documento, los catálogos son el conjunto de asientos bibliográficos ordenados, los índices son listas de palabras que representan conceptos de un documento y los resúmenes son la exposición abreviada del contenido de un documento.

De entre las operaciones que se describen, la que va a centrar esta investigación como hemos mencionado anteriormente es la indización. La indización es la operación que consiste en enumerar los conceptos sobre los que trata un documento y representarlo por medio de un lenguaje combinatorio.

Según García Gutiérrez (1984) la indización es la aplicación de un nivel de análisis documental cuyo producto final puede ser utilizado por los usuarios para conseguir o conocer la información que precisan o para elaborar lenguajes documentales.

Es primordial concretar claramente cual es el objeto y el sujeto del análisis documental, el objeto del análisis documental es el documento científico, éste debe cumplir varios requisitos como tener información científica y que se pueda transmitir, entre sus cualidades debe ser original y no una copia, la fuente debe ser fiable y debe ser accesible, igualmente el documento científico debe cumplir dos aspectos básicos uno

formal en el cual dicho documento se presenta en un soporte y otro aspecto de contenido en el cual estaría la información a transmitir. Respecto al sujeto del análisis documental se encuadraría el hombre ayudado por las tecnologías informáticas, ya que hoy en día el ordenador es una herramienta imprescindible para el análisis documental ya sea el análisis formal que utiliza operaciones más mecánicas que intelectuales y ya sea el análisis de contenido con técnicas intelectuales inicialmente desarrolladas por el documentalista y que actualmente se están desarrollando con aplicaciones informáticas, como pueda ser la indización automática.

3.2.2. Lingüística documental

En la disciplina de la Documentación afloró la necesidad de controlar los documentos a través de la Lingüística, pero la Lingüística documental si bien parte del cruce de dos términos que forman su denominación, es una disciplina impregnada de otros campos científicos como son la lógica, la estadística, la informática, la lexicografía, la biblioteconomía y la telemática. Por tanto la Lingüística Documental es una disciplina ligada a los procesos informativos-documentales (científico, informativos y profesionales) que tiene por objeto el establecimiento de un efectivo control documental mediante la utilización de mecanismos léxicos.

La Lingüística Documental resuelve problemas que tienen que ver con el *contenido*, lo importante es la información y por ello su tarea principal es la organización del conocimiento, prueba de esta organización del conocimiento son las Clasificaciones documentales, que estructuran jerárquicamente el conocimiento humano.

En definitiva, la Lingüística Documental es una disciplina que se ocupa del problema que plantea el almacenamiento y ulterior recuperación del contenido analítico de cualesquiera documentos. Su propósito es resolver dicho problema mediante agentes cualificados y especializados que se sirven sistemáticamente, corporativa e institucionalmente de unos medios semióticos llamados “Lenguajes documentales”.

La documentación tiene dos dimensiones: *la dimensión genética* en la que se encontraría la producción de documentos la cual no está controlada y una *dimensión instrumental* en la que se encontraría el uso de los documentos, tanto a nivel comunicativo limitándose a informar sobre la localización de los documentos como a nivel jurídico-histórico en la que se utiliza el documento como prueba o para justificar un hecho.

Entre las relaciones que se dan en la Comunicación-Información-Documentación podemos afirmar que: No hay Información sin Comunicación y No hay Información y Comunicación sin Documentación.

3.2.3. La terminología científica y técnica

El progreso científico y la utilización creciente de las lenguas nacionales en las ciencias exactas y humanísticas llevaron a los científicos especialistas a esforzarse en crear términos para los diferentes conceptos de sus especialidades, basándose en el latín y el griego, para ser comprendidos universalmente.

En 1969 la ISO elaboró la norma ISO/R 1087 Vocabulario de la Terminología, según esta norma la terminología es el conjunto de términos que representan el sistema de los conceptos ligados a un campo especial del saber. También se entiende la terminología como el estudio de los términos.

Según Marquet i Ferigle (1995), la terminología es el conjunto de términos de una especialidad (una ciencia, un arte, un oficio, etc.). Si hablamos de terminología es porque hay términos, y el conjunto de estos términos constituye la terminología.

Por tanto la terminología es el campo del saber consagrado a la formación y a la denominación de las nociones, ya sea dentro de un campo especial o en todos los campos. En resumen, se puede decir que la terminología es: el conjunto de términos y el estudio científico de estos términos.

Se entiende por término según la norma ISO/R 1087 como símbolo convencional de una noción que consiste en sonidos articulados o en su representación gráfica. Un término es una palabra o grupo de palabras

Así por tanto, la terminología se basa en la idea de conceptos y términos y sus relaciones.

En la relación existente entre concepto y término la comunicación sólo es posible si un término está permanentemente asignado a un concepto y viceversa, hay diferentes posibilidades de asignaciones: monosemia, sinonimia, cuasi-sinonimia, homonimia, polisemia. Si por el contrario, queremos relacionar dos lenguas o más, nos encontramos con otro tipo de relación: la equivalencia, que es la relación existente entre términos de diferentes lenguas que corresponden a un mismo concepto.

Hay normas internacionales que entran de pleno en el campo de la terminología, concretamente la ISO y su Comité 37 Terminología, ISO/TC 37. La ISO entre otras tiene establecidas la norma ISO/R 860 Unificación Internacional de conceptos y términos.

3.2.4. Orígenes de los lenguajes documentales

Los lenguajes documentales son el producto del análisis documental de contenido que caracterizan o describen el contenido de los documentos para posteriormente poder recuperar la información que éstos contienen. El fin de los lenguajes documental es recuperar los documentos.

Una definición globalizadora del lenguaje documental sería un sistema artificial de representación formalizada de documentos y demandas orientado a la recuperación de la información

Según García Gutiérrez (1984) el lenguaje documental es aquel conjunto normalizado y normativo de términos relacionados por principios comunes, declarados portavoces preferenciales de los mensajes encerrados en un colectivo documental con el fin de provocar una recuperación pertinente de información por aproximación temática.

El papel del lenguaje documental es precisamente mejorar la calidad del análisis, es la herramienta esencial para que se realimente el proceso documental, sin éste el proceso

documental sólo funciona por aproximación, sin rigor y sin eficacia y por tanto la calidad de los lenguajes condiciona la calidad de los resultados documentales.

Los lenguajes documentales son generados principalmente por el análisis de los documentos ya que el objetivo que persiguen es el control y manipulación de sus contenidos para poder satisfacer las potenciales demandas. Los lenguajes documentales influyen en el análisis y recuperación de documentos, son por tanto, lenguajes normalizadores y a su vez normalizados. Éstos a su vez se componen de una lista de términos que lo constituyen como vocabularios y de unas relaciones entre ellos que dinamizan esos vocabularios y le otorgan la categoría de lenguaje.

Entre las características de los lenguajes documentales éstos tienen una función principal y una función específica, la función principal es la de ordenar para recuperar y la función específica de los lenguajes documentales es básicamente la de ordenar por materias, la originalidad de los lenguajes documentales estriba en que el repertorio de términos que lo constituyen está destinado a la organización de los documentos en función de sus contenido. Son pues, instrumentos de características originales, cuya dificultad es esencialmente semántica.

La organización y ordenación de los documentos en función de la materia o del tema, es a la vez, lo más importante y lo más difícil. En general se distinguen dos tipos de análisis del contenido:

- 1.-La condensación o también llamado análisis, definida por las normas ISO 5122 (1979), AFNOR NF Z 44-004 (1963), UNE 50-112
- 2.-La caracterización analítica (indización) y sintética (clasificación), definida por la norma ISO 5963 (1985), AFNOR NP Z 47-102 (1978), UNE 50-121.

Por tanto los lenguajes documentales son léxicos artificiales construidos con un fin preciso, el objetivo para el que se elaboran condiciona su estructura y posibilidades. La estructura compleja de las materias hace necesario recurrir a un lenguaje documental preconstruido (precoordinado).

Debido a que la función específica de los lenguajes documentales es ayudar a seleccionar los documentos en función de su contenido temático, esto lleva consigo otra consecuencia no menos importante: la importancia de su estructura semántica.

Un campo semántico es un conjunto de términos unidos entre sí por relaciones de parentesco de significado, entre los campos semánticos se distinguen: los campos léxicos que agrupan el conjunto de las palabras de una lengua relativa a un dominio, y los campos conceptuales que agrupan los conceptos de un campo y sus relaciones semánticas.

En la selección de los conceptos de un documento hay que tener en cuenta no solamente los conceptos claramente expresados sino también las nociones implícitas respecto a las cuales el documento tiene un interés potencial.

Las relaciones de sentido entre conceptos se obtienen con la similitud: cuanto mayor sea la semejanza entre dos conceptos, más próximos y emparentados estarán los términos que los designan.

Las lenguas presentan comúnmente accidentes lingüísticos que pueden provocar ruido documental debido a los términos plurívocos, la sinonimia, la paráfrasis, la ambigüedad, la homonimia y la polisemia, entre otros. Y fundamentalmente la razón de ser de los lenguajes documentales es: recuperar fácil y rápidamente los documentos, por esto el lenguaje natural no es el instrumento ideal para asegurar el buen funcionamiento de la comunicación documental, ya que en el lenguaje natural podemos encontrar términos plurívocos (con varios sentidos), que provocan ruido documental en la recuperación. Los lenguajes documentales necesitan univocidad porque cuanto más cercanos del lenguaje natural mayores son los riesgos de silencio y ruido en la búsqueda documental.

En el lenguaje ocurren accidentes lingüísticos como la sinonimia, cuando dos palabras de la misma lengua son sinónimas y tienen la misma denotación y la misma connotación, es decir que dos sinónimos son sustituibles el uno por el otro en cualquier contexto y fuera de él, ya que comparten un mismo significado

La paráfrasis, al igual que la sinonimia la paráfrasis es una equivalencia de significado entre dos enunciados. La sinonimia y la paráfrasis tienen importancia en la riqueza de la lengua ya que permiten variar y matizar la expresión y evitar las repeticiones.

La ambigüedad en el lenguaje son producidas por la homonimia y la polisemia, la homonimia es la similitud formal de palabras diferentes. Cuando dos palabras con la misma forma tienen distinto significado. Si la similitud aparece en la forma oral de la palabra se habla de “homofonía”, si aparece en la forma escrita se habla de “homografía”. Es por ello que se consideran homógrafos a las palabras de distinto significado que se escriben igual

La polisemia se produce cuando los términos son similares y el sentido diferente, es decir varios significados. Se trata de una misma palabra que a partir de un determinado momento se ha enriquecido con un nuevo significado. Las lenguas generan continuamente nuevos casos de polisemia a causa de las posibilidades creativas del lenguaje, las dos causas más frecuentes son la metáfora y la metonimia.

Los lenguajes documentales se dividen según su funcionalidad en: *Lenguajes coordinados*, que se dividen en lenguajes precoordinados que son los que combinan los conceptos antes del almacenamiento (Clasificaciones, Listas de encabezamientos de materias) y los lenguajes postcoordinados que permiten yuxtaponer los conceptos en el momento de la recuperación (Tesauros, Listas de descriptores). También según su funcionalidad se encuentra los *Lenguajes de control*, son lenguajes documentales con el vocabulario controlado o libre. El vocabulario controlado se establece de antemano con una lista de términos exclusivos que pueden utilizarse unívoca y limitadamente. En los lenguajes libres no se define una lista de términos autorizados, sino que se extraen todos los conceptos tal como aparecen en los documentos. La ventaja de los lenguajes controlados es que la búsqueda es rápida y mucho más eficaz que con los lenguajes libres (como pueden ser los índices KWIC, KWOC, etc.). Y finalmente encontramos

según su funcionalidad los *Lenguajes de precisión*, son lenguajes documentales altamente elaborados con sintaxis y que necesitan de una herramienta informática.

Los aspectos de coordinación y control pueden mezclarse dando lugar a: Lenguajes precoordinados con vocabulario libre (Periodex, KWIC, KWOC.), Lenguajes postcoordinados con vocabulario libre (Uniters, listas de palabras clave.), Lenguajes precoordinados con vocabulario controlado (Clasificaciones jerárquicas, Listas de encabezamiento de materias, etc.) y los Lenguajes postcoordinados con vocabulario controlado (Tesauros).

La tipología de los lenguajes documentales se divide en:

- 1.- Lenguajes documentales de estructura jerárquica
- 2.- Lenguajes documentales de estructura combinatoria
- 3.- Lenguajes documentales de estructura sintáctica

La característica general de los *Lenguajes de estructura jerárquica* o arborescente es que todos los conceptos dependen de uno superior y así sucesivamente. Ejemplos de lenguajes con este tipo de estructura se encuentra las *clasificaciones enciclopédicas*, las *clasificaciones de facetas* y las *especializadas*.

Las *clasificaciones enciclopédicas* pretenden cubrir parte o la totalidad del conocimiento dividiendo en áreas y subáreas con una estricta jerarquía desde los conceptos más generales a los más específicos, ejemplos de este tipo de clasificación enciclopédica es la Clasificación Decimal Universal (CDU) utilizada en Bibliotecas. La CDU fue creada por Paul Otlet y Henri La Fontaine, dos jóvenes juristas que transformaron la Clasificación de Dewey con el propósito de organizar un Repertorio Bibliográfico Universal, así Otlet y La Fontaine crearon y propulsaron la magna Clasificación Decimal Universal. Ésta es una clasificación con una notación numérica ordenada según el principio que rige los números decimales, esta estructura numérica supone que un número pueda ser dividido y subdividido indefinidamente, también tiene carácter universal al abarcar todo el conjunto del saber, pensar y hacer humano La estructura jerárquica de la CDU sigue un orden sistemático que parte de lo general a lo particular, del todo a la parte, del género a la especie, además emplea signos de puntuación para poder relacionar de diversas formas los números asignados con diversos conceptos.

Las *clasificaciones de facetas* se basan en el sistema clasificatorio del Norteamericano Henry Evelyn Bliss, nacido en 1870 y formado en el Collage de Nueva Cork donde fue bibliotecario en el 1891. Propuso un nuevo esquema clasificatorio que fue usado en la Biblioteca del Collage. El sistema clasificatorio basa su ordenación en clases que aúnan la totalidad del conocimiento siendo las divisiones de las distintas disciplinas lógicas y conceptuales, es decir dividió las materias por aspectos o enfoques. Bliss divide cada disciplina desde cuatro puntos de vista: el filosófico, el teórico, el histórico y el práctico. Establece una división lineal de 28 clases principales representadas con una notación alfabética empleando como símbolos las letras mayúsculas del alfabeto latino, igualmente la clasificación de Bliss añade siete esquemas auxiliares para subdividir las clases principales según: Forma, Lugar, Geografía, Lengua, Periodos Históricos y Filología. Bliss logró que su sistema fuera el gran competidor del Sistema de Dewey y

del Sistema de la Biblioteca del Congreso de Washington, incluso tuvo gran incidencia en la Clasificación de Ranganathan.

Otra de las clasificaciones facetadas es la clasificación colonada de Ranganathan (Colon Classification), Shiyam Ramarita Ranganathan (1892-1972) bibliotecario y matemático de origen indio fue el creador de una de las más destacadas clasificaciones, la clasificación colonada muy desconocida en España. Su clasificación fue publicada en 1933, de la cual se publicaron siete ediciones más. Se trata del primer sistema de clasificación basado en el principio analítico sintético.

Ranganathan estableció cinco categorías principales, que son conceptos algo difíciles de concretar para los occidentales: la personalidad, la materia, la energía, el espacio y el tiempo. Y además desarrolla varias clases principales que son: matemáticas, física, agricultura, medicina, psicología, educación, historia y economía, éstas son representadas por las mayúsculas del alfabeto latino y otras clases se representan por las letras del alfabeto griego, además de ser la notación de las clases principales alfanumérica. La notación incluye numerosos signos de conexión para indicar una doble característica como la coma, el punto y los dos puntos o “colon” vocablo inglés que designa los dos puntos. Este vocablo ha dado la denominación de Clasificación Colonada al esquema de Ranganathan.

Y por último las *clasificaciones especializadas* forman parte de los lenguajes documentales de estructura jerárquica, éstas se estructuran de forma semejante a las enciclopédicas, aunque tienden a la especialización de un campo determinado, constan de un sistema jerárquico numérico, alfa-numérico o alfabético, este tipo de clasificaciones son muy utilizadas en Bibliotecas especializadas, empresas, instituciones de investigación, etc.

Una vez hemos revisado detalladamente los lenguajes documentales de estructura jerárquica entre los tipos de lenguajes documentales, ahora veremos igualmente otra tipología de los lenguajes documentales que son los *Lenguajes documentales de estructura combinatoria*. Los lenguajes combinatorios se componen de una lista de términos propios y representativos del ámbito científico y técnico del que se trate, y es dinamizado por una serie de relaciones lingüísticas sistematizadas entre esos términos, lo cual dota de grandes posibilidades al lenguaje documental. Entre las ventajas de los lenguajes combinatorios encontramos que al ser listas normalizadas de términos permiten ceñirse al conjunto de significados de un campo, también al incluir relaciones entre los términos facilitan el mejor sentido de las frases documentales y al ser normalizados evitan las duplicaciones. Ejemplos de lenguajes documentales de estructura combinatoria son las listas de materias, los léxicos Uniterms y los tesauros, entraremos más en detalle más adelante.

Otra tipología de los lenguajes documentales son los *Lenguajes documentales de estructura sintáctica*. Estos lenguajes de estructura sintáctica constan no sólo de un conjunto de descriptores (en lenguaje natural o artificial), sino también de una gramática que excluye la ambigüedad al relacionar conceptos. Existen los lenguajes de estructura sintáctica simple y los lenguajes elaborados, respecto a los primeros la falta de relación sintáctica entre los descriptores puede producir una cierta distorsión en la interpretación y es por ello que se establece un orden fijo en la aparición de descriptores, respecto a los

lenguajes elaborados, estos lenguajes constan además de un conjunto de términos convencionales, de unas depuradas reglas gramaticales que eliminan cualquier peligro de error o incomprensión en el diálogo documental.

3.2.5. Los nuevos lenguajes de representación del conocimiento

3.2.5.1. Ontologías

Según el Diccionario de la Real Academia Española¹¹ Ontología (del gr. *ὄν*, *ὄντος*, el ser, y *-logía*) es la parte de la metafísica que trata del ser en general y de sus propiedades trascendentales

Es a partir de los 90 cuando el término ontología empieza a utilizarse en el contexto de la Inteligencia Artificial como modelo de representación del conocimiento.

Las ontologías son una organización cognitiva que conforma un sistema de organización del conocimiento (Sánchez Cuadrado et. al., 2007). El fundamento de las ontologías es representar el conocimiento, ahora se conoce como ontologías a lo que previamente eran tipos de sistemas de organización del conocimiento o KOS (*Knowledge Organization Systems*). Las ontologías según los autores Sánchez Cuadrado [et. al.] (2007) constituyen una pieza clave para el modelado del conocimiento, dentro de la web semántica ya que éstas son un recurso fundamental para la web 2.0, de ahí su importancia creciente. En definitiva, las ontologías describen la semántica de una materia o ciencia en el entorno web de modo que pueda ser interpretado por las aplicaciones informáticas de forma inteligible para los usuarios.

El primer paso para el procesamiento informático del conocimiento es la representación formal de dicho conocimiento. Uno de los recursos disponibles para realizar esta labor es la ontología.

Según Arano (2003)¹², “Una ontología es una representación formal del conocimiento donde los conceptos, las relaciones y las restricciones conceptuales son explicitadas mediante formalismos en un determinado dominio. [...] la ontología es una representación formal y explícita de la estructura conceptual del campo sobre el que se trabaja. Este recurso lingüístico incluye como mecanismo de inferencia a la herencia, que implica una economía en la codificación de la información: los conceptos superiores transmiten sus características a los conceptos inferiores”.

Una ontología contiene definiciones que nos proporcionan un vocabulario para referirse a un determinado área de conocimiento, a un conjunto de conceptos (como cosas, propiedades, eventos y relaciones), que se especifican, por ejemplo, en lenguaje natural con el objetivo de crear un idioma común para intercambiar información. Ese vocabulario se define mediante un conjunto de términos básicos y relaciones entre

¹¹ 22 ed., 2001

¹² ARANO, Silvia. “La ontología: una zona de interacción entre la Lingüística y la Documentación” [en línea]. *Hipertext.net*, núm. 2, 2003. <<http://www.upf.edu/hipertextnet/numero-2/ontologia.html>> [ref. de 08 de noviembre 2012].

dichos términos, así como las reglas que combinan los términos y las relaciones que permiten ampliar las definiciones dadas en el vocabulario. Por tanto, una ontología es una forma de ver el mundo, ya que determina los términos a utilizar para describir y representar un determinado área de conocimiento, haciendo énfasis en compartir y reutilizar el conocimiento y el consenso en la representación de éste.

Como apunta García-Marco (2007) las ontologías son un campo de investigación de la inteligencia artificial y más específicamente de la rama relacionada con la representación del conocimiento, la ingeniería del conocimiento, que se ocupa de la construcción de sistemas expertos. Según el autor, el objetivo de la ingeniería del conocimiento es construir grandes bases de conocimientos sobre un tema en forma de declaraciones, reglas de inferencia y mecanismos de razonamiento (motor de inferencia) para resolver automáticamente problemas del dominio en cuestión. Las ontologías son herramientas para construir sistemas conceptuales o, por utilizar una terminología común, vocabularios estructurados, en los que se explicitan todas las relaciones entre los términos que se utilizan y otras restricciones de significado.

Las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesoro tradicional están elaboradas con una sintaxis comprensible para los ordenadores. Una ontología permite mayor riqueza en la definición de sus conceptos y sus relaciones que un tesoro. Las ontologías expresan conceptos con un lenguaje basado en lógica simbólica y susceptible de ser eventualmente interpretado por un ordenador (Pedraza-Jiménez, Codina y Rovira, 2007)

Según los autores Pedraza-Jiménez, Codina y Rovira (2007), las ontologías incluyen las definiciones de los conceptos, denominadas “clases”, de un dominio y las relaciones entre ellos. El lenguaje OWL (Web Ontology Language), que es una extensión de RDF (Resource Description Framework), es el lenguaje estándar de la web semántica para expresar y codificar ontologías. Por tanto, puede ser utilizado para representar explícitamente el significado de términos en vocabularios y las relaciones (semánticas) entre ellos. OWL consigue formalizar las relaciones entre las clases aún más que RDF, indicando aspectos básicos para el razonamiento como la existencia de conceptos o clases disjuntas en un dominio, también es posible expresar la cardinalidad, es decir, el número de elementos que pueden componer un concepto o clase, por ejemplo: un libro puede tener varios autores. Y también puede expresar igualdad o equivalencia entre clases, características y restricciones de las mismas. Por tanto OWL es una extensión de RDF que añade los elementos mencionados anteriormente para describir características y clases.

RDF es el sistema que permite utilizar metadatos para describir recursos en la web semántica, este lenguaje consiste en habilitar la extracción del significado de la estructura de un documento, descrita en XML, con el fin de garantizar la interoperabilidad entre aplicaciones sin necesidad de intervención humana (Senso, 2003). El sistema RDF parte de tres entidades lógicas: recursos, propiedades y valores. Los recursos pueden ser sitios o páginas web, pero también cosas que no están en la web, como personas u objetos. Las propiedades son las características relevantes de los recursos y los valores son los datos en los que se concreta un atributo determinado de un recurso determinado. Para que un ordenador pueda entender este tipo de estructuras, denominadas triples, será necesario representar dicha información mediante RDF/XML

y Dublin Core. Una de las importantes utilidades de RDF consiste en la descripción de recursos digitales utilizando Dublin Core, norma que, como es sabido, consiste precisamente en aplicar la filosofía documental a la descripción de recursos. (Pedraza-Jiménez, Codina y Rovira, 2007)

Los autores Pedraza-Jiménez, Codina y Rovira (2007), indican que XML es un estándar que junto con su norma asociada XML SCHEMA, permite definir tipos de documentos y los conjuntos de etiquetas necesarias para codificarlos. La idea es que una vez que están marcados o codificados con una colección de etiquetas XML, es posible procesarlos y explotarlos de forma automática con diversos propósitos. XML es un meta-lenguaje para la definición de estructuras textuales.

En la Web semántica, las ontologías capturan un conocimiento consensuado de un modo genérico, de forma que pueda ser compartido y reutilizado por distintos grupos de personas y aplicaciones de software. Una de las condiciones para que funcione la Web semántica es que el contenido de los documentos se presente por medio de la utilización de ontologías que sean públicas y accesibles, de uso común y, a ser posible, normalizadas. Así, estos documentos con contenido semántico podrán ser utilizados por robots software.

Para garantizar la interoperabilidad de las ontologías habría que definir unas ontologías genéricas comunes, útiles para el intercambio de información entre distintas partes, que sean compatibles con las ontologías particulares de cada área de interés como las ontologías de alto nivel o *top ontologies* (Moreiro González, Sánchez Cuadrado y Morato Lara, 2012) La idea es que la Web semántica está formada, al menos en parte, por una red de nodos tipificados e interconectados mediante clases y relaciones definidas por una ontología compartida por sus distintos autores.

3.2.5.1.1. Diferencias entre ontologías y lenguajes documentales

Según un artículo de Sánchez-Jiménez y Gil-Urdiciain (2007) el aspecto externo de una ontología es bastante similar al de un sistema de clasificación. En realidad, los lenguajes documentales permiten establecer relaciones entre conceptos, que pueden ser en principio similares a las que ofrece una ontología, aunque cubriendo eso sí un espectro menos amplio. Sin embargo, un análisis más detallado de las características de los lenguajes documentales y las ontologías muestra que existen diferencias importantes entre ambos. Las diferencias entre lenguajes documentales y ontologías complican el traslado del conocimiento existente en un tesoro o un sistema de clasificación a una ontología.

Según artículo de Moreiro González, [et. al.] (2008) Los lenguajes documentales intentan evitar las ambigüedades propias del lenguaje natural mediante un subconjunto denominado lenguaje controlado. Para facilitar la recuperación de información es necesario el análisis de la información de los documentos. Este proceso pasa por dos fases inseparables, la de análisis propiamente y la de síntesis de los resultados obtenidos en el análisis. Con la nueva web 2.0, las aplicaciones informáticas necesitan de unos lenguajes documentales basados en XML más potentes y formalizados como las ontologías. Uno de los rasgo de la web 2.0 es su carácter participativo, un espacio libre para la colaboración y la comunicación que exige una respuesta diferente de los

lenguajes documentales, con mayor riqueza de asociaciones y más adaptados al cambio mediante una mayor proximidad a las necesidades de los usuarios. Siguiendo con la aportación de los autores, estos consideran que los lenguajes documentales representan una pieza clave de los sistemas de gestión de la información para reducir la ambigüedad del lenguaje natural y en la web, los lenguajes controlados deben convivir junto con la indización y la búsqueda libre.

3.2.5.2. Folksonomías

Las folksonomías (Moreiro González, 2009) son conjuntos de palabras clave incorporadas y asignadas por cualquier internauta para colaborar en la indización de todo tipo de contenidos en un espacio compartido y abierto. La asignación de estas etiquetas públicas se realiza sin ánimo de lucro y sin la supervisión de un organismo centralizador.

Según Iglesia Aparicio y Monje Jiménez (2012) una folksonomía es la clasificación de un objeto de forma colectiva, mediante etiquetas simples y sin jerarquías ni relaciones determinadas. A este modo de clasificación también se le denomina etiquetado colaborativo, clasificación social, indexación social, etc. su uso se ha popularizado desde el año 2004 y actualmente forma parte de una gran mayoría de aplicaciones webs.

Esta forma “social” de clasificar las cosas supone una serie de problemas derivados de la propia naturaleza del lenguaje:

- Las máquinas no logran interpretar que una palabra significa un concepto independientemente de su género y de su número, tampoco logran relacionar sinonimias y no pueden diferenciar homonimias.
- Las distintas lenguas usan distintas palabras para un mismo concepto
- Expresar conceptos complejos, de más de una palabra
- Las personas pueden utilizar etiquetas muy críticas o excesivamente personales (nombres propios)

En definitiva, se pueden clasificar los problemas derivados del uso de folksonomías en problemas lingüísticos: distintos idiomas, un mismo concepto se escribe de diferentes formas. Género, número y derivaciones de las palabras. Tildes y otros símbolos ortográficos. Sinónimos. Homonimias. Y por otro lado los problemas derivados de los usuarios: etiquetas muy personales o equivocadas.

Las folksonomías no tienen que sustituir a las taxonomías, a esas clasificaciones que, con el paso del tiempo y el avance de las investigaciones, han sido consensuadas dentro de cada ámbito de conocimiento. Pero sí tienen algunos usos donde son realmente útiles: organizarnos nosotros mismos o un grupo de personas que colaboran en un mismo proyecto. Tener conocimiento de los términos más relevantes en cierto momento y SEO (Search Engine Optimization) que es el conjunto de técnicas cuyo objetivo final es lograr que un determinado sitio web tenga más visitas gracias a la correcta selección de los términos de búsqueda asociados a dicho sitio web.

Es de utilidad mantener buenas prácticas si trabajamos con etiquetas que facilitarán la clasificación de nuestros recursos como puede ser: antes de etiquetar hay que elaborar

una mínima planificación. Utilizar un número reducido de etiquetas. Usar un lenguaje natural. Crear etiquetas compuestas usando el guión bajo (_). Crear siempre una etiqueta comodín o neutra. Etiquetar con criterio. Revisar el etiquetado de forma periódica.

Un claro ejemplo de utilización de folksonomías son los marcadores sociales (social bookmarking) que permiten organizar, almacenar, gestionar y buscar enlaces de recursos en línea. Cada registro suele ir acompañado de información adicional aportada por el usuario como descripciones, anotaciones y folksonomías. Estos registros se pueden hacer públicos o privados y también se pueden compartir con un grupo de usuarios, muy recomendable si se trabaja en colaboración.

Muchas de las herramientas de marcado social proporcionan fuentes RSS¹³ (*Really Simple Syndication*) es el estándar de sindicación de contenidos más extendido. Su estructura es muy sencilla y se compone de pocos elementos. De estas listas de marcadores (los más recientes, por etiquetas, etc.) de forma que también se pueden difundir y compartir por medio de sindicación de contenidos. Una de las principales características de estas herramientas es que son los propios usuarios quienes clasifican los enlaces y cuántos más usuarios hayan guardado un enlace con una determinada etiqueta es más probable que sea un sitio realmente relevante y relacionado con el término de búsqueda.

Actualmente existen multitud de alternativas para practicar el marcado social diferenciándose dos tipos de herramientas: las de carácter general y los enfocados al ámbito universitario o de investigación.

✓ De carácter general

Delicious (<http://delicious.com>)

Google Bookmarks (<http://www.google.com/bookmarks>)

Faves (<http://faves.com/>)

Folkd (<http://www.folkd.com/>)

Mister Wong (www.mister-wong.es)

StumbleUpon (<http://www.stumbleupon.com>)

Bundlr (<http://www.bundlr.com>)

✓ Para investigadores

Connotea (www.connotea.com)

CiteULike (www.citeulike.org)

Bibsonomy (www.bibsonomy.org)

En definitiva, las folksonomías (Moreiro González, 2009) han venido a renovar las formas de indizar, pues han distribuido su responsabilidad entre los usuarios y han impuesto métodos descentralizados, alejados de cualquier jerarquía sistemática. Si bien actualmente se hacen necesarias técnicas que aproximen las folksonomías, propias de la web semántica, a unos lenguajes controlados, eliminándose problemas propios de los

¹³ Su especificación se puede consultar en <http://cyber.law.harvard.edu/rss/rss.html>.

lenguajes libres, como la sinonimia, la homonimia y la ausencia de niveles de estructuración de términos entre sí.

3.2.5.3. Taxonomías

Las taxonomías (Moreiro González, 2009) se aplican en el mundo empresarial e institucional para organizar y gestionar los recursos de información digitales que alojan en sus servidores web, buscando categorizarlos, hojearlos y navegar por ellos. Una taxonomía organiza no sólo los contenidos propios de una organización, sino también los servicios que ofrece, sus productos y cuanto se deriva de la experiencia y datos sobre los recursos humanos. Las taxonomías están presentes en todos los esquemas, tesauros, modelos conceptuales y ontologías.

3.2.6. La indización

La indización es uno de los campos relacionados con las técnicas documentales más tratados por los investigadores de la documentación, de la información científica y de otros sectores como la estadística.

La indización según la definición de la UNESCO consiste en describir y caracterizar un documento con la ayuda de representaciones de los conceptos contenidos en dicho documento, destinado a permitir una búsqueda eficaz de las informaciones contenidas en un fondo documental.

Según Gil Urdiciain (1996) la indización consiste en el análisis e identificación de los conceptos del documento, la selección de aquellas nociones que representen con mayor fidelidad la información que contiene y su traducción a un lenguaje documental.

Según García Gutiérrez (1984) que denomina a la indización como *caracterización* define la indización como la aplicación de un nivel de análisis documental cuyo producto final puede ser utilizado por los usuarios para conseguir o conocer la información que precisan o para elaborar lenguajes documentales.

El producto de la indización es el índice, el cual sirve para la consulta directa y concreta. Etimológicamente la palabra índice viene de la voz latina "*index-icis*" relacionada con el verbo "*indicere*": señalar, notificar. Por tanto se relaciona el moderno término indizar con el proveniente del verbo latino "*indicare*": dar u ocasionar indicios de alguna cosa.

Según el Diccionario de la Real Academia (2001) el término índice es una lista ordenada de capítulos, artículos, materias, voces (en un libro u otra publicación) en él contenidos, con indicación del lugar dónde aparecen. Y el DRA proporciona como definición de indizar, hacer índices, registrar ordenadamente datos e informaciones, para elaborar su índice.

El índice es un documento típicamente secundario mientras que el indizado o producto proveniente de la indización conforma la base de un documento terciario: El tesauro.

La indización es una técnica consistente en aislar el contenido de un documento y el producto de este aislamiento no son solamente los índices (documentos secundarios) sino también la lista de términos (documento terciario)

Todo parece indicar que fue en Inglaterra donde se utilizó por primera vez el término indización como la técnica de hacer índices, y además donde se institucionalizó también por primera vez esta actividad, con la fundación en 1957 de la Society of Indexers de Londres, prueba de ello es que los anglosajones emplean los latinismos “*indexing*” e “*indexer*” para designar la tarea de indizar o el sujeto indizador.

La indización es una técnica más investigada por estadísticos que por documentalistas, afirma García Gutiérrez (1984), según el autor define la indización como: una técnica del tratamiento documental utilizada para la descripción del contenido de documentos o demandas documentales que posibilita la elaboración de estrategias de recuperación mediante conceptos o materias.

La indización es la mayor fuerza motriz y el instrumento auxiliar más eficaz de la información científica, afirma el autor.

Según Van Dijk y Van Slype (1991), la indización es una operación que se cumple en cuatro estadios.

1. el conocimiento del contenido conceptual del documento
2. la extracción de los conceptos en lenguaje natural
3. la traducción de esos conceptos al lenguaje documental
4. la búsqueda de otros conceptos pertinentes, no expresados por el autor

Y los lenguajes de indización se dividen en la siguiente tipología:

1. lenguaje natural: es el lenguaje en el que se presenta el documento original
2. lenguaje documental: los lenguajes coordinados (precoordinados y postcoordinados)

Igualmente la indización se dividen en indización simple e indización controlada, la indización simple se toman los términos del documento tal como aparecen expresados, es decir en lenguaje natural, las palabras clave. Entre los sistemas de indización simple se encuentran los Unitérminos, Sistema de Taube (1955), en el que aparecen vocablos simples de un documento que se combinan formando palabras que expresan un gran número de ideas. También existen los sistemas de indización simple de concordancia, que son listas de palabras o índice alfabético de las palabras de un documento y los índice KWIC y KWOC, respecto a los índices KWIC (Key Word in context), son listas alfabéticas de las palabras significativas del título del documento de tal modo que se efectúa la permutación del título para que la palabra quede destacada frente a las demás. Los índice KWOC (Key Word out context) se diferencia del KWIC, en que la palabra significativa no va destacada dentro del título, sino que aparece encabezando todos los títulos que la contienen.

Respecto a la indización controlada, ésta extrae las palabras que están expresadas en el documento pero se traducen a un lenguaje documental, los lenguajes documentales y los tesauros son el resultado de la indización controlada.

Los modos de indización existentes según García Gutiérrez (1984) son la caracterización por materias y unitérminos: el Método Taube (1955) y la indización por descriptores, en cambio para los autores Van Dijk y Van Slype (1991), éstos recurren a una triple tipología de sistemas de indización que en definitiva coincide con lo mostrado por el autor García Gutiérrez

Tipología de sistemas de indización:

1. indización por materias
2. indización por unitérminos
3. indización por descriptores

Hoy en día, debido a la diversidad tipológica de los documentos y sobre todo a su profusión espectacular de los últimos cincuenta años, las tradicionales técnicas de descripción fueron superadas por otras nuevas, que a su vez eran rápidamente desplazadas, en una época en la que aún no se aplicaba ampliamente el ordenador en las tareas documentales. Hoy sin embargo no es posible concebir las técnicas de indización sin considerar la influencia informática.

La única forma de establecer la eficacia de los diferentes métodos de indización, es a través de sistemas de evaluación que juzguen la actividad de un método frente a un fondo documental comparativamente con otros.

El investigador estadístico Bloomfield (1970) acepta los métodos de evaluación como eficaces para averiguar su pertinencia, definiéndolos como: “aquella capacidad de desarrollar reglas o generalizaciones que puedan juzgar un término indizado como útil o inútil.

Uno de los estudios más citados y criticados que forma parte de los clásicos de la indización fue el denominado *Experimento Cranfield* (Inglaterra) dirigido por Cleverdon (1967), el cual analizó cuatro modos distintos de indización dictados por las necesidades de Clasificación Decimal Universal (CDU), de la clasificación facetada, y de la indización por materias y unitérminos. El estudio se aproximó a la evaluación de la indización a través de la recuperación documental basándose en métodos estadísticos y tasas de precisión y centrando la experiencia en las formulaciones de la demanda (indicando éstas) sobre un fondo documental concreto.

3.2.6.1. La indización por materias y unitérminos. El Método Taube

La indización existe desde que el primer documento fue denominado con un título representativo de su contenido, de forma que pudiera ser recuperado por éste entre otros documentos. Sin embargo las técnicas de indización son posteriores, aunque puedan localizarse ya bien definidas en el siglo XIX con las primeras clasificaciones enciclopédicas y facetadas, basadas en la sistematización de temas. Por esto son estas clasificaciones las que primero aportaron el concepto de indización por materias

(aunque no el término). La poca flexibilidad de las materias para realizar combinaciones dificulta la recuperación de información cuando éstas se presentan dispersas o en soportes no tradicionales

Van Dijk y Van Slype (1969) afirman que la indización por materias fue la primera que apareció, con el fin de hacer más eficaz la búsqueda, pero conforme la recuperación se hacía más predominante y regular surgió la necesidad de profundizar en la indización y según Van Dijk y Van Slype (1969) "...de esta forma se crearon los unitérminos, los cuales se lanzaron como reacción contra la indización por materias. Fue una etapa intelectual en el camino hacia la indización basada en los conceptos, pues los unitérminos carecían a veces de significado por sí solos. Los descriptores pueden presentar inconvenientes, pero no hay nada perfecto. Pueden existir falsas combinaciones que se evitan con la ayuda de otros procedimientos técnicos..."

Los unitérminos son en su mayoría sustantivos que se extraen del propio documento y para cada uno se abría una ficha cuadrículada. El conjunto de fichas se ordenaba alfabéticamente por uniterm.

Mientras que los unitérminos se extraían del propio documento convirtiéndose así en unitérminos o palabras en lenguaje natural, surge otro concepto que va más allá en el proceso documental y que designa una palabra elegida de un léxico preestablecido como palabra clave, éstas también se extraen del lenguaje natural tal y como aparece en el documento original o primario, y se definen como: términos o expresiones elegidas para representar una noción contenida en un documento.

Fondin (1977) llama a las palabras clave "términos de indización y los define como "vocablos o expresiones elegidas para representar una noción contenida en un documento". El autor distingue además tres tipos de palabras clave:

1. palabra clave o derivada: término o expresión de la lengua natural escogido sin ningún instrumento léxico. Se extrae directamente del texto analizado=UNITERMINOS (la aportación terminológica de Fondin identifica los distintos sistemas de indización, el primer tipo corresponde a los unitérminos)
2. palabra clave controlada o asignada: palabra clave elegida de un léxico preestablecido=DESCRIPTORES
3. término temático: palabra extraída de listas preestablecidas. Representa una noción en una clasificación=MATERIAS

García Gutiérrez cree que el término "palabra clave" está estrechamente relacionado con el término descriptor pues ambos frente a los demás tienen un factor común: la representación de conceptos.

3.2.6.2. La indización por descriptores

Si nos remontamos al significado y definición de "DESCRIPTOR", fue Mooers (1951) a finales de los cincuenta el primero en utilizar la denominación de descriptor para referirse al primer producto de la indización basada en conceptos.

Según el Diccionario de la Real Academia (2001) se denomina descriptor al término o símbolo válido y formalizado que se emplea para representar inequívocamente los conceptos de un documento o de una búsqueda.

Por tanto, se considera un descriptor al término o conjunto de términos normalizados o controlados, que expresan el contenido significativo de un documento. Son conceptos traducidos a un lenguaje documental (controlado).

Pero no es posible hablar de descriptores sin hacer referencia al tesoro que según el Diccionario de la Real Academia (2001) lo define como: diccionario, catálogo.

Los descriptores no se basan en el número de palabras, como los unitérminos, o en los lemas, como las listas de materias, sino en los conceptos. El descriptor se caracteriza por la flexibilidad tanto interna (composición) como externa (relaciones y combinaciones con otros conceptos)

Según García Gutiérrez (1984), si un concepto es representado por una sola palabra, el descriptor se denomina Unitérmino o Descriptor simple y estará en la misma situación operativa que los unitérminos de Taube (1955).

Si el concepto necesita varias palabras para ser expresado, el descriptor se denomina Descriptor sintagmático o compuesto (por ejemplo psicología experimental) cada vez se tiende más a la utilización de descriptores sintagmáticos y esto ocurre por la mayor necesidad de profundidad en los conceptos, dada la extensión bibliográfica en temas especializados y por la eliminación de la ambigüedad que conllevan los descriptores simples

La elaboración de los descriptores sintagmáticos se produce mediante dos métodos:

1.- Unión morfológica: (sintáctica para algunos autores) cuando dos o más términos se unen apoyados por preposiciones, artículos o por la adjetivación de uno de ellos (ejemplo: /tratamiento/ + /documentación/= /tratamiento documental/

2.- Unión lexicológica: (semántica para ciertos autores) dos o más términos se funden en un número menor de ellos por ósmosis recíproca de contenidos (ejemplo: /información a distancia/+ /ordenador/= /teledocumentación/

En los dos tipos de descriptores sintagmáticos la unión puede producirse “a priori” o precoordinadamente y “a posteriori” o por postcoordinación

Por otra parte, es conveniente que la lista de descriptores presente una forma normalizada, por lo cual la mayoría de los autores acuerdan las siguientes:

- Forma sustantiva
- Género masculino (si hay opción)
- Forma desarrollada (no abreviaturas)
- Secuencia lineal normal
- Término o sinónimo más comúnmente utilizado.

Respecto a las estructuras sintagmáticas de los descriptores cabe citar a Gil Leiva, Rodríguez Muñoz (1997) que realizan un interesantísimo estudio sobre éstas en

diferentes áreas del conocimiento (biblioteconomía y documentación, medicina, química, biología, psicología y física) de las bases de datos del Consejo Superior de Investigaciones Científicas (ISOC, IME e ICYT) analizando un total de 2077 descriptores, asignados a 450 referencias de artículos científicos.

Sus conclusiones más interesantes son que la categoría gramatical que está más presente en los descriptores es el SUSTANTIVO, SUST+ADJETIVO, SUST+DE+SUST por tanto los autores opinan que respecto a la hipótesis planteada por varios investigadores de diseñar un algoritmo que localice en los documentos a indizar estructuras preestablecidas como (SUST, SUST+ADJ, SUST+DE+SUST) supondría efectuar previamente sobre el texto un análisis morfológico y sintáctico para extraer las categorías gramaticales de cada una de las palabras y proceder a su desambiguación en el caso que fuera necesario. Obviamente esto implicaría extraer todos los posibles términos de indización de un documento partiendo de sus estructuras sintagmáticas y para ello habría que contemplarlas todas y en este artículo los autores denotan tras su investigación la gran variedad y cantidad de estructuras sintagmáticas que pueden coexistir en un documento.

Por tanto resulta un proceso complejo tanto de ejecución como de tiempo empleado recurrir a una herramienta que autorice o rechace un término o conjunto de ellos y que realice las complejas operaciones de análisis morfológico, sintáctico, localización y extracción de estructuras sintagmáticas.

Otra conclusión interesante que los autores apuntan es que de los 2077 descriptores el 38,1% se encuentran presentes en el título o en el resumen y el 61,9% no viene ni en el título ni en el resumen. Es por ello que en nuestra investigación a la hora de aplicar la indización automática a un documento hemos aplicado el algoritmo de análisis de los documentos a todo el texto y no nos hemos centrado en partes del texto, aparentemente, más significativas como puedan ser el título o el resumen, etc.

Pero volviendo al hilo de la indización por descriptores, además de las dos formas de presentación de descriptores: los unitérminos o descriptores simples y los descriptores sintagmáticos, existen unos elementos llamados INFRACONCEPTOS (infra, super, mini, etc.) que están exentos de significado pero que unidos a descriptores constituyen nuevos significados.

En ciertos lenguajes documentales, encontramos términos de significación general, que acompañando a los descriptores principales concretizan su sentido (ejemplo: /información científica/, /concepto/, /método/, etc.) donde /información científica/ sería el descriptor principal y el resto los descriptores auxiliares o secundarios.

Es necesario, a efectos prácticos, eliminar esos descriptores secundarios de los lenguajes documentales con el fin de dar cabida a los más representativos. Los descriptores auxiliares o secundarios son perjudiciales aisladamente ya que pueden inducir a ambigüedad o ruido documental. Sin embargo, en contexto focalizan el sentido de los descriptores principales y por tanto son útiles para la Recuperación. El problema estriba en elegir un lugar apropiado para situarlos en las listas de descriptores ya que pueden ir combinados con los más relevantes.

Podemos distinguir dos tipos de descriptores auxiliares o secundarios:

- 1.- los causados por la ambigüedad
- 2.- los causados por la homonimia

Los descriptores auxiliares deben distinguirse de los principales ya sea por su grafía (negrita, cursiva, etc.) o posición.

Según García Gutiérrez (1984) éstos descriptores secundarios servirán para clasificar las referencias bibliográficas, a este tipo de descriptores les denomina: Descriptor informativo, y aquellos descriptores que además de aparecer en las referencias, integran el lenguaje documental se llamarán Descriptores referenciales o direccionales, es decir, que orientan hacia un colectivo de documentos, mientras que los primeros, los descriptores informativos informan sobre documentos individuales.

3.2.6.3. La indización por palabras clave. Folksonomías

La indización por palabras clave ha evolucionado enormemente con la aparición de la web semántica o web 2.0, recordemos que ésta nueva web semántica es una extensión de la actual web 1.0, una web en la que los internautas sólo tienen acceso a la consulta y descarga de la información allí contenida y en la que personas cualificadas se encargaban de editar los contenidos, esto actualmente está cambiando y ahora cualquier internauta sin conocimientos cualificados tiene permisos y accesos para crear y editar un sinnúmero de información on-line.

Esta nueva socialización de la nueva web semántica se ha orientado hacia la codificación semántica de los documentos y a la aplicación de nuevas tecnologías y procedimientos de representación del conocimiento con el fin de mejorar el acceso a los recursos web.

La mayor usabilidad de la web y la socialización de ésta ha traído consigo que los nuevos navegantes de internet puedan reelaborar información en la web, y para ello a cobrado una gran importancia la utilización de la indización por palabras clave o las conocidas folksonomías

Como se ha mencionado anteriormente las folksonomías (Moreiro González, 2009) son conjuntos de palabras clave incorporadas y asignadas por cualquier internauta para colaborar en la indización de todo tipo de contenidos en un espacio compartido y abierto. La asignación de estas etiquetas públicas se realiza sin ánimo de lucro y sin la supervisión de un organismo centralizador.

3.2.7. La indización automática

La indización es una operación que persigue captar y representar el contenido de un documento en dos etapas: identificación de conceptos en lenguaje natural que representan el contenido de un documento y posteriormente, traslación de estos conceptos a su expresión por medio de un lenguaje controlado (normalizado). La indización es por tanto una de las operaciones más complejas del proceso técnico documental, esta dificultad se torna doble cuando se intenta obtener de forma

automática ya que la indización automática es una técnica interdisciplinar donde intervienen la lingüística, la estadística, la informática y la documentación.

Entre los años 60 y 90 los investigadores han debatido pródigamente una cuestión interesante sobre la conveniencia de efectuar la indización mediante la tecnología que ofrece una máquina o mediante un profesional indizador humano, alegando que una máquina no puede captar todos los matices conceptuales que puede descubrir un profesional de la documentación que realiza constantemente un análisis documental profundo y meditado de los documentos objeto de análisis y alegando por otra parte la cuestión de que una máquina puede ofrecer mayor objetividad aplicando siempre los mismos parámetros y por el contrario el humano puede variar su criterio a la hora de indizar documentos iguales y de este modo provocar errores, e incluso dos indizadores pueden realizar la indización de un documento de forma diferente.

La diferencia fundamental entre las dos es básicamente que en la indización humana el contenido de los textos es sometido a un análisis intelectual para su representación en lenguaje documental y en la indización automática, un algoritmo toma el lugar del indizador y se aplica sobre cada documento.

García Gutiérrez en 1984 creía necesario el método automático en la indización, ya que debido a la explosión y saturación de la información, la cantidad de documentos a indizar sobrepasaría la capacidad humana.

Pero además la indización automática es el método más factible, porque actualmente la mayoría de los documentos están disponibles en formato electrónico, incluso documentos que están en soporte papel han sido concebidos con un procesador de textos y elaborados por tanto con tecnología electrónica, este cambio que se ha producido en la actualidad provoca que al estar el documento completo en un soporte legible por ordenador, puede ser procesado por programas informáticos y es por ello que es posible plantearse una indización totalmente automática.

Para el autor García Gutiérrez (1984) el indizador, sea hombre o máquina deberá ceñirse a ciertas partes del documento para elegir los conceptos característicos de éste, mediante los cuales podrá ser buscado. Son varios investigadores los que apoyan esta reflexión, pero en nuestro caso no la compartimos ya que según la investigación realizada por los autores Gil Leiva, Rodríguez Muñoz (1997), comentada anteriormente sobre las estructuras sintagmáticas de los descriptores concluyen que el mayor porcentaje de descriptores se encuentran presentes en todo el texto y no en el título o resumen.

Por otra parte el autor Jones (1976), apunta respecto a los descriptores, ideas interesantes como que no es posible seleccionar mediante la indización aquellas palabras que aparecen raramente o demasiadas veces, pues esto significa que nadie las conoce o que son demasiado comunes. Según el autor ambos casos son igualmente perjudiciales para la indización y para la posterior recuperación documental.

Por tanto, al igual que la indización manual, el principio de la indización automatizada es identificar un documento por un conjunto de palabras claves representativas de su contenido, que pertenezcan a un conjunto de términos, (indización libre) o que pertenezcan a una lista de autoridad o un tesauro (indización controlada).

Las primeras investigaciones en indización automática se desarrollaban a finales de los 50 y durante los 60 y se fundamentaban principalmente en técnicas estadísticas y probabilísticas, posteriormente estas técnicas evolucionaron con técnicas más complejas originadas por las investigaciones de G. Salton que aportó técnicas como el valor de discriminación y la relevancia de términos, como el valor de *Idf*, en el que el peso de un término en un documento dado es inversamente proporcional a su frecuencia en la colección de documentos y directamente proporcional a su frecuencia en el documento en cuestión. Se asigna el peso de los términos calculando el grado de similitud entre cada documento almacenado en el sistema y la consulta formulada por el usuario.

$$Idf(p) = \lg\left(\frac{N}{n}\right)$$

Donde *N* son los documentos que hay en una colección

Donde *p*, es una palabra o término que aparece en *n* documentos

Otro de los modelos teóricos formulado inicialmente por Salton, Wong, Yang (1975), es el Modelo de Espacio Vectorial, en el cual un documento puede ser representado mediante un vector o lista de números en correspondencia con una lista de palabras, donde cada una de las palabras tiene un peso, un coeficiente que intenta expresar en qué medida esa palabra es representativa del contenido de ese documento, además de estimar la similitud entre el vector consulta y los vectores de los documentos.

Ya en los años 80 comienzan a proliferar otras técnicas más lingüísticas que estadísticas para el análisis de los textos, incluso se han desarrollado sistemas híbridos que conjugan los dos métodos.

A finales de los años 50 y 60 se produjo el crecimiento exponencial de la información científica, es por ello que proliferaron los sistemas de información aumentando el número de investigaciones sobre el tratamiento de la información con la finalidad de facilitar las necesidades de información de los científicos. Al mismo tiempo se daba el crecimiento tecnológico suficiente para que fuera considerada una herramienta indispensable y necesaria para las tareas documentales y entre ellas la indización automática.

En los años 70 ya se vislumbraba dos vertientes distintas en la indización automática, por un lado la vertiente y los estudios en métodos no lingüísticos en los que la estadística, la probabilidad, la asignación de pesos, el clustering, etc. se convertían en la base conceptual principal para desarrollar los sistemas de indización automática, estos métodos no lingüísticos fueron liderados por Salton (1989) y Luhn (1957), Rosenberg (1971), entre otros, y es la vertiente que nosotros apoyamos en esta investigación. Y por otro lado, la vertiente y las investigaciones basadas en análisis lingüísticos y su gran puntal de estudio como es el procesamiento del lenguaje natural (PLN).

Comencemos desgranando los métodos no lingüísticos o como preferimos denominar en esta investigación, los *métodos numéricos* ya que se basan en cálculos basados en las ciencias exactas.

Los métodos numéricos han desarrollado a lo largo de la historia interesantes avances en materia de indización automática y recuperación de información, desarrollando significativas teorías que aún hoy son la base de las investigaciones surgidas al respecto, los precursores de estos métodos numéricos para la indización automática y en consecuencia para la recuperación de información desarrollaron teorías sobre la frecuencia de aparición de los términos, las probabilidades de co-aparición de éstos en los documentos, el clustering y la agrupación de los términos según su temática mediante el modelo vectorial, el valor de discriminación de los términos, la relevancia de éstos en los documentos y la asignación de pesos, etc.

La complejidad de los métodos numéricos ha ido creciendo con el paso de los años hasta desarrollarse modelos más avanzados que utilizan multitud de cálculos y algoritmos en aras de conseguir sistemas de indización automática y recuperación de información más eficaces. Los avances en materia de indización automática hoy en día intentan unificar las dos vertientes en una sola, es decir métodos que aglutinan tanto métodos numéricos como métodos lingüísticos con técnicas de procesamiento del lenguaje natural (PLN)

Según Moreiro González, Méndez Rodríguez (1999) distinguen el proceso evolutivo de la indización automática en tres generaciones: una primera generación, donde las palabras se entendían como objetos (análisis estadísticos o probabilísticos), una segunda generación, donde lo que prima es el análisis lingüístico para la desambiguación de conceptos (PLN), y una tercera generación, a la que denominan indización “inteligente” en tanto que trata de abstraer no sólo conceptos sino modelos conceptuales fundamentados en bases de conocimiento (análisis estadístico, ponderación estadística, acceso directo a los documentos a través del procesamiento lingüístico automático y la utilización del lenguaje natural).

Respecto a los métodos lingüísticos es a comienzos de los años 50 cuando se inician los trabajos teóricos sobre formalización del lenguaje dirigidos principalmente por Noam Chomsky, en años posteriores estas técnicas avanzaron en el procesamiento del lenguaje natural (PLN) y la comprensión del lenguaje natural y el tratamiento de la sintaxis coincidiendo así la teoría lingüística y la práctica computacional.

Según Gil Leiva, Rodríguez Muñoz, (1996). El procesamiento del lenguaje natural (PLN) consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos. El PLN surge en la década de los 50, las aplicaciones generales y más básicas del PLN se han dado en la traducción automática de textos, la comunicación hombre-máquina mediante software que permiten mediante un micrófono dar una serie de órdenes a un robot o máquina, en la enseñanza asistida por ordenador y en los procesadores de textos que llevan incorporados correctores ortográficos, diccionarios de sinónimos, así como los verificadores automáticos de palabras que se teclean en el procesador de texto. En cuanto a las aplicaciones específicas del PLN en el campo de la Documentación destaca, la interrogación de bases de datos en lenguaje natural, simplificándose así las consultas de los usuarios que debían utilizar operadores booleanos, de truncamiento, de proximidad, etc. también destaca la generación automática de tesauros, que posibilita la identificación de relaciones sintácticas y semánticas entre palabras y frases y la categorización y difusión de la información. La elaboración automática de resúmenes y la indización automática de documentos.

Ejemplos de sistemas de indización automática elaborados con métodos numéricos son *INDEXD*, *SAPHIRE* y *Centro de Información Aeroespacial NASA*, sistemas de indización automática con métodos lingüísticos son el sistema *CLARIT* y el *Proyecto SIMPR* y ejemplos de sistemas de indización automática comercializados se encuentran *SPIRIT*, *DARWIN*, *ALETH*, *GOLEM*, *SINTEX* Y *ALEXDOC*, e *INDEXICON*, información detallada de estos sistemas se describe ampliamente en Gil Leiva, Rodríguez Muñoz, (1996).

Siguiendo con los métodos lingüísticos los autores Figuerola, et al. (2006), describen la indización morfológica como: el Stemming que realiza una reducción de las palabras a su raíz unificando palabras con similar raíz pero con distinto significado como por ejemplo (“*organ*”, “*organization*”, “*organism*”), estos problemas según los autores pueden evitarse contando con un lexicón computacional y un adecuado procesador morfológico. La morfología computacional ha experimentado un fuerte desarrollo en la pasada década y ha adoptado técnicas muy eficientes desde el punto de vista computacional como es la morfología de estados finitos. Este procesador lingüístico para llevar a cabo la normalización de las palabras en el momento de la indización obliga a incorporar un desambiguador categorial (Part of Speech Tagger), esto es, una herramienta capaz de asignar, para cada palabra de un texto, una única categoría gramatical, dado el contexto de ésta. Como ejemplo el autor propone la palabra “*casa*” como sustantivo y la palabra “*casar*” como verbo, por tanto para indizar dicho documento ha determinarse previamente cuál es la correcta: *casa/S* o *casar/V*, otro ejemplo de una palabra que puede tener diferentes categorías gramaticales (adjetivo, sustantivo y preposición) sería la palabra “*bajo*” en la que se ha de buscar una forma de representación diferenciada, en el momento de la indización como: *bajo/A*, *bajo/P*, *bajo/S*. Estas técnicas de Post-Tagging entre los efectos positivos en la indización son: con la desambiguación categorial se discriminan diferentes acepciones y realizan la eliminación coherente de palabras vacías (eje. *bajo/P*). Entre las desventajas de utilizar técnicas de Post-Tagging para acometer la indización morfológica como el Stemming es que aumenta considerablemente los recursos computacionales frente al uso de las técnicas no lingüísticas como el Stemming.

Figuerola, et. al. (2006), concluyen según diversos experimentos realizados con diferentes lenguas, que por ejemplo con el idioma inglés la indización con técnicas lingüísticas no aporta mejoras respecto a los métodos no lingüísticos, con lo que no resulta aconsejable el uso de las primeras dado la diferencia en el coste computacional, según otros estudios la indización con Post-Tagging producía mejoras inapreciables frente a la indización de palabras ortográficas. Respecto al Español, los resultados con técnicas de stemming y con herramientas lingüísticas para tratar la morfología flexiva, obtenían mejores resultados utilizando éstas últimas.

En resumen, la evolución histórica de la indización automática nos demuestra que no existe una única corriente en cuanto a estos sistemas, y por un lado hay investigadores que propugnan sistemas basados en métodos no lingüísticos o numéricos como reflejamos en esta investigación e investigadores que propugnan sistemas basados en métodos lingüísticos y por otro lado los hay que defienden sistema híbridos.

Por lo que respecta a esta Tesis Doctoral presentamos un nuevo sistema de indización automática basado en métodos numéricos o no lingüísticos como comúnmente se

denomina en la literatura, a texto completo con una indización exclusivamente full-text, nuestro sistema de indización automática es conocido como MALLOV.

3.2.8. Metadatos. Topics Maps o Mapas Conceptuales

Los metadatos (Méndez Rodríguez, 2002) están destinados a ordenar y describir la información contenida en un documento entendido como objeto, de tal forma que se erigen como reveladores tanto de la descripción formal, como del análisis de contenido, en aras a mejorar el acceso a esos objetos de información en la red. No son más que estructuras de organización de la información, legibles por máquina, cuya finalidad es hacer útiles los datos, de distintas formas, según las necesidades concretas de cada servicio de información digital y según la aplicación que se les otorgue.

Los metadatos (Iglesia Aparicio y Monje Jiménez, 2012) son simplemente datos que proporcionan información sobre los datos que se encuentran almacenados en una colección digital. Es decir, no describen completamente un ítem pero, sí proporcionan una mínima información del mismo. Y además, y esto es lo realmente importante, permite uniformizar la información soslayando el método de catalogación que se haya utilizado originalmente para incorporar el ítem a la colección digital.

Los autores anteriormente citados Iglesia Aparicio y Monje Jiménez (2012) declaran que actualmente existen varias especificaciones de metadatos pero la más extendida es Dublin Core¹⁴. Dentro de ella existen dos modelos de especificación. La primera, se denomina comúnmente *Simple Dublin Core* y está compuesto por 15 elementos descriptivos. La segunda, se llama *Qualified Dublin Core* y propone nuevos elementos descriptivos. Actualmente, la versión simple de Dublin Core es la más usada.

Se detalla brevemente cuáles son estos 15 elementos:

1. *Title*. El nombre o título de un recurso.
2. *Creator*. La persona u organización que ha creado el contenido intelectual del recurso, es decir, el escritor de un libro, el fotógrafo de una imagen, el ilustrador de un dibujo, etc.
3. *Subject*. La descripción de la temática del recurso mediante frases y palabras clave.
4. *Description*. Descripción textual del recurso: un resumen o descripción del contenido.
5. *Publisher*. La organización responsable de que un contenido se encuentre disponible en la red en el formato actual, es decir, el editor.
6. *Contributor*. Otras personas distintas del autor que hayan colaborado en la elaboración del contenido intelectual del recurso.
7. *Date*. La fecha en la que el recurso, en su formato actual, se puso a disposición del usuario.
8. *Type*. El tipo de recurso.
9. *Format*. El formato de digitalización del recurso.

¹⁴ Su especificación completa se encuentra en <http://dublincore.org/>. Una traducción al castellano se puede encontrar en <http://www.rediris.es/search/dces/>.

10. *Identifier*. Identificador único del recurso. Habitualmente es la dirección URL donde se localiza aunque, también puede ser el ISBN u otro identificador único reconocido.
11. *Source*. Identificador de la fuente o fuentes utilizadas para elaborar el contenido actual.
12. *Language*. Lengua o lenguas del contenido.
13. *Relation*. Identificación de otros recursos digitales relacionados con el presente recurso digital.
14. *Coverage*. Cobertura espacial y/o temporal del recurso. Por ejemplo, si se refiere a un país o al siglo XII.
15. *Rights*. Descripción de los derechos de autor del recurso digital.

Estos elementos son opcionales y, en general, basta con que sean rellenados con texto aunque, para algunos, se recomiendan usar vocabularios controlados (*Type*) y estándares (*Format, Date*).

Un ejemplo de cómo sería la descripción Dublin Core de la reproducción en PDF de la obra *Rinconete y Cortadillo* de Miguel de Cervantes publicada en la Biblioteca Virtual Miguel de Cervantes¹⁵

```
<dc:title>Rinconete y Cortadillo</dc:title>
<dc:creator>Miguel de Cervantes y Saavedra</dc:creator>
<dc:subject>Novela, literatura española, siglo XVII</dc:subject>
<dc:description>Novela del siglo XVII...</dc:description>
<dc:publisher>Biblioteca Miguel de Cervantes</dc:publisher>
<dc:contributor>Florencio Sevilla Arroyo</dc:contributor>
<dc:date>2001</dc:date>
<dc:type>Documento</dc:type>
<dc:format>text/html</dc:format>
<dc:identifier>http://www.cervantesvirtual.com/obra-visor/rinconete-y-cortadillo--0/html/</dc:identifier>
<dc:source>http://www.cervantesvirtual.com/</dc:source>
<dc:language>es</dc:language>
<dc:relation>Otras obras de Miguel de Cervantes</dc:relation>
<dc:coverage>España, siglo XVII</dc:coverage>
<dc:rights></dc:rights>
```

Los metadatos son actualmente la implementación más cercana a lo que será una futura web semántica. Ofrecen información adicional sobre los datos que contiene una página web. Y precisamente, esto es lo que la web semántica pide para poder dar como resultado búsquedas más precisas.

Elaborar un estándar de metadatos exige el consenso de un amplio grupo de personas. Es necesario ponerse de acuerdo en qué elementos contendrá, cuáles serán obligatorios u opcionales, en qué formato se escribirán (texto, número, etc.). Este es un esfuerzo que generalmente se realiza entre personas dedicadas a un mismo ámbito de estudio o de negocio.

¹⁵ <http://www.cervantesvirtual.com>

A continuación se muestran algunos ejemplos de metadatos aparte del ya conocido Dublin Core:

- ID3, actualmente en su versión 2.0, sirve para etiquetar la información que contiene un archivo audiovisual con datos como título, artista, álbum, año, etc. Aquí se puede consultar la información oficial: <http://www.id3.org/>.
- LOM (*Learning Object Metadata*), especificación dirigida a la descripción de objetos digitales de educación. Además de los datos descriptivos de un objeto permiten señalar sus características pedagógicas y didácticas. Esta es su especificación oficial: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf.
- SCORM (*Shareable Content Object Reference Model*), para definir contenidos pedagógicos estructurados.
- EXIF (*Exchangeable Image File Format*), para etiquetar información de fotografías como la fecha y la hora y las características de la cámara.
- PBCore (*Public Broadcasting Metadata Dictionary*), basado en Dublin Core, permite describir emisiones audiovisuales proporcionando información sobre la propiedad intelectual y el contenido.
- MARC (*Machine Readable Cataloging*) y MODS (*Metadata Object Description Schema*). Para describir recursos bibliográficos.
- CDWA (*Categories for the Description of Works of Art*), para describir información sobre obras de arte y otros objetos de cultura material.

Todos son para ámbitos muy específicos, sólo Dublin Core parece querer abarcar un espectro más amplio de documentos. Por lo tanto, vemos que ha habido muchos esfuerzos grupales que han dado lugar a especificaciones de metadatos aceptadas pero que, a su vez, estas especificaciones no se entienden entre sí. Es decir, nada tienen que ver la descripción ID3 de un archivo de sonido con la LOM de un objeto digital de educación. Y, por supuesto, es difícil que los robots buscadores sean tan inteligentes como para entender todas y cada una de las especificaciones de metadatos existentes. Pero al menos, las máquinas que alojan contenidos pertenecientes a una misma área de conocimiento parece que ya pueden entenderse entre sí.

Respecto a los mapas conceptuales estos son herramientas para la organización y representación del conocimiento que incluyen relaciones entre conceptos. Según Moreiro González, García Martul, (2005), en los mapas, los conceptos son representados por nodos y las relaciones por arcos etiquetados o por palabras que configuran unas proposiciones, en forma de enlaces. Las proposiciones son afirmaciones acerca de algún objeto que constan de dos o más términos conceptuales, unidos por palabras para formar una unidad semántica.

Los autores Moreiro González, García Martul, (2005), indican igualmente que para su construcción siguen unas pautas, como son identificar los conceptos importantes de un dominio, clasificar de los más generales a los más específicos y poner en relación el conjunto. Igualmente los autores indican el procedimiento para elaborar un mapa conceptual, el cual consistiría en determinar, en primer lugar, la materia y ámbito de aplicación, el empleo de ontologías, tesauros, diccionarios y todas las fuentes de información al uso que puedan ayudar para la elaboración del mapa conceptual, seguidamente la elaboración de un índice de términos sobre los que deseamos realizar las relaciones y describir las propiedades que nos interesa observar de los mismos, la

definición de clases y la jerarquía de clases: de conceptos más genéricos a más especializados; de clases más específicas a superclases más genéricas o una combinación de las dos. Y la definición de las propiedades de las clases, indicando una serie de atributos para cada una de ellas, para por último la definición de las fuentes de información.

Los mapas conceptuales son una solución a la mera presentación estática de los registros obtenidos en la recuperación de información, ya que éstos se presentan interrelacionados por proximidad semántica, así de este modo la recuperación o navegación del usuario en el sistema es más pertinente. El usuario ve mejorada la recuperación de información ya que los resultados que éste demanda en su búsqueda son relacionados con otros resultados de interés. Los mapas conceptuales se convierten así en un instrumento para la representación de la información.

3.2.9. Los Tesauros

Un texto, en su origen se crea para comunicar un mensaje, en el cual podemos distinguir dos tipos de palabras:

- 1.- Los términos gramaticales: conjunciones, preposiciones, artículos, pronombres, etc. sirven para hacer funcionar el discurso, para construir frases y organizarlas en un texto coherente. Son palabras vacías o gramaticales.
- 2.- Los términos léxicos: nombres, verbos, sustantivos, adjetivos. Son los que dan sentido al texto, son palabras llenas o léxicas.

Pero ¿qué queda de un texto cuando se le despoja de sus términos gramaticales? Una serie de términos léxicos, desprovistos de enlaces sintácticos, que constituyen, un conjunto poco inteligible. Así pues, la extracción de palabras llenas o léxicas conforma una lista ordenada de términos que designan conceptos, y este es el principio de la indización y de la construcción de tesauros.

Según Currás (1998) en 1976 se publicaron los manuales de la UNESCO dentro del programa UNISIST, donde se definen los tesauros, según su función y su estructura.

Según su función son un instrumento de control terminológico usado para trasladar desde un lenguaje natural de los documentos los descriptores a un sistema lingüístico y según su estructura: son vocabularios controlados y dinámicos de términos relacionados semántica y genéricamente, que cubren un dominio específico del conocimiento.

La norma UNE 50-106, de 1989 traducción de la norma ISO 2788 en su 2ª ed. 1986, donde se define entre otras acepciones, un tesoro como: un vocabulario de un lenguaje de indización controlado, organizado formalmente, con objeto de hacer explícitas las relaciones a priori entre conceptos.

En julio de 2005 se publicó la cuarta edición de la norma ANSI/NISO Z39.19:2005, *Guidelines for the construction, format and management of monolingual controlled vocabularies* surgida del Workshop on Electronic Thesauri, promovido por NISO y celebrado los días 4 y 5 de noviembre de 1999, la finalidad principal es la de redactar una norma que abarcara la elaboración de tesauros electrónicos.

Según Currás (1998) un tesoro es un lenguaje especializado, normalizado, post-coordinado, usado con fines documentarios, donde los elementos lingüísticos que lo componen –términos simples o compuestos-, se hallan relacionados entre sí sintáctica y semánticamente.

y entre las condiciones que debe cumplir un tesoro se encuentran:

- ser un lenguaje especializado
- estar normalizado: post-controlado
- las unidades lingüísticas adquieren la categoría de términos
- los términos o palabras clave se relacionan entre sí con relaciones: jerárquicas, asociativas y de equivalencia
- estos procesos de relación se podrán realizar con métodos de : pre-coordinación y post-coordinación
- se trata de lenguajes terminológicos, usados con fines documentarios por tanto se convierten en : lenguajes documentarios para la indización o la clasificación o la recuperación de información
- deben permitir la introducción o supresión de términos para mantener su actualidad constante
- deben servir para convertir el lenguaje natural de los documentos, ambiguo y libre, en un lenguaje concreto, normalizado, apto para controlar la información contenida en el documento
- han de servir de nexo de unión entre el documento y el usuario

La autora Currás (1998) define un lenguaje controlado como un sistema lingüístico, en el que las unidades que lo componen, se estructuran de acuerdo con unas normas prefijadas.

En el momento en el que se utilizan unas normas, se debe hablar de lenguaje normalizado. Para la autora un lenguaje normalizado es un lenguaje controlado, al que se le aplican unas normas lógicamente prefijadas y cuyas unidades lingüísticas son términos.

Igualmente define lenguaje documentario como un lenguaje controlado (normalizado) usado con fines clasificatorios, en el sentido amplio de esta expresión. Entre las operaciones documentarias se encuentra:

- la indización
- la recuperación de información

Corresponde ahora considerar en qué forma y manera se ha de controlar los términos de un lenguaje documentario. El control supone establecer relaciones jerárquicas, asociativas y de equivalencia. Y los términos pueden ser simples o compuestos, y los compuestos se podrá formar coordinando sus elementos a priori= lenguajes precoordinados y a posteriori= lenguajes postcoordinados

Las relaciones pueden ser:

- Jerárquicas
- Asociativas
- Equivalencia

Los términos pueden ser:

- Simples
- Compuestos: lenguajes precoordinados y los lenguajes postcoordinados

Las relaciones jerárquicas

- la relación genérica: a causa de la antisimetría, ofrece un caso típico de implicación fuerte en el sentido específico/genérico y de implicación débil en el sentido inverso. De este modo “caballo” implica necesariamente “vertebrado” lo cual implica “mamífero”, pero en el sentido inverso esto no es cierto. De ahí lo interesante de tener en cuenta los términos específicos de un tema antes que los genéricos, si se quieren obtener todas las relaciones pertinentes posibles.
- la relación partitiva: esta relación tiene numerosos puntos en común con la relación genérica, de forma que en la distinción se confunden, tanto en las clasificaciones como en los tesauros, bajo el término de jerárquicas. La relación todo/parte.

Las relaciones asociativas: la relación asociativa facilita una buena transmisión entre la relación genérica y la relación asociativa, ya que si no existiese semejanza jerárquica entre las dos primeras, ocuparía lógicamente su lugar en las relaciones asociativas.

La autora define los lenguajes precoordinados como lenguajes documentarios en los que los términos que los componen se coordinan en un proceso previo a su utilización (eje: las clasificaciones). Y lenguajes postcoordinados como lenguajes documentarios, en los que los términos que los componen se coordinan en un proceso posterior a su fijación, por ejemplo en el momento de su establecimiento o de su uso (eje: tesauros)

Según Currás (1998) los lenguajes de indización y de recuperación de información podrán ser pre o post-coordinados, según las necesidades del sistema de información o del centro de documentación dónde se usen.

3.2.9.1. Evolución histórica de los tesauros

Con el aumento exponencial de la información en la que controlar la abundante literatura con los sistemas jerárquicos o facetados no era la respuesta adecuada para la gran demanda de información, las bibliotecas o los centros de documentación se quedaban obsoletos y eran infrautilizados. La solución pasaba por encontrar, idear nuevos sistemas de clasificación que permitiesen más flexibilidad en el tratamiento de los temas contenidos en los documentos.

Entre las ideas surgidas se recurrió a clasificaciones sistemáticas utilizadas en el pasado en 1561 en “Libro de los epítomes” y el “libro de materias o proposiciones” de Hernando Colón y el “*Dictionarium Historicum de Charles Estieme*”. De esta forma surgieron los primeros métodos de clasificación utilizando conceptos sacados de los propios documentos, sin conexión previa. Recibieron el nombre original de Thesurus

La palabra Thesaurus se utiliza por primera vez a mediados de los años cincuenta. Coincide este hecho con el desarrollo del sistema Uniterm elaborado por el

norteamericano Taube (1955) cuando escribe su obra en la que concibe los unitérminos como los vocablos más pequeños y simples seleccionados del propio documento

Según García Gutiérrez (1984) los unitérminos son en su mayoría sustantivos que se extraen del propio documento, para cada uniterm se abre una ficha cuadrículada. El conjunto de fichas se ordena alfabéticamente por uniterm. El problema de los uniterms:

- 1.- las palabras aisladas no tienen en muchas ocasiones un significado concreto por lo que comportan ambigüedad
- 2.- las relaciones entre ellas pueden dar lugar a falsas combinaciones y distintas de las necesidades del usuario (ej. /documentación/ /psicología/ (psicología de la documentación o documentación psicológica)

La primera objeción ha sido superada por los descriptores, pero la segunda objeción suele plantear graves problemas de recuperación a la mayoría de los métodos de indización.

Con los descriptores se construían índices, por consiguiente “indizar” se llamó al proceso de obtener aquellas palabras de los documentos, con los que se confeccionaban los índices

Los primeros tesauros, formalmente contruidos, aparecen a partir de 1960, debido al desarrollo de la informática, aparecieron en el mercado ordenadores más manejables y económicos y esto potenció la automatización de las ciencias de la documentación.

A partir de los setenta hasta nuestros días se sigue sintiendo la necesidad de disponer de tesauros eficaces tanto para la indización, como para la recuperación de información.

Actualmente, tanto los estándares británicos como los americanos incluyen novedades con respecto a la cobertura, los tipos de vocabularios incluidos, la precoordinación y la postcoordinación y la necesidad de interoperabilidad entre sistemas de la norma ANSI/NISO Z39.19:2005.

En cuanto a la cobertura, tanto la norma británica BS 8723 elaborada por el BSI (British Standard Institution) como la norma americana ANSI/NISO Z39.19:2005, ambas normas, amplían su radio de acción en lo referente a tesauros, sino que abarcan también otros vocabularios controlados, como son las taxonomías, los anillos de sinónimos, encabezamientos de materia y ontologías.

La norma en lugar de hablar de descriptor o término preferente utiliza la palabra “término”. Un término se define como una o más palabras utilizadas para representar un concepto. La norma establece que la finalidad de un vocabulario controlado es lograr la consistencia en la descripción de contenido de los documentos y facilitar su recuperación. La norma distingue cuatro tipos de vocabularios controlados, determinado por su estructura compleja creciente: listas, anillos de sinónimos, taxonomías y tesauros.

Una lista es un grupo simple de términos preferentes sin estructura y suelen presentarse en orden alfabético u otra secuencia lógica.

Los anillos de sinónimos son un tipo de vocabulario controlado que se utiliza para la recuperación de información proporcionando acceso al contenido que se representa en lenguaje natural, no controlado. El anillo de sinónimos asegura que un concepto pueda ser descrito por múltiples sinónimos o términos equivalentes.

Una taxonomía se define como una lista de términos preferentes con estructura jerárquica.

Un tesoro se compone de una lista alfabética de todos los términos preferentes y no preferentes con tres tipos de relaciones para todos los términos (de equivalencia, jerárquicas y asociativas)

Autores como Sánchez Cuadrado, Colmenero Ruiz y Moreiro (2012) indican que las recomendaciones de la norma NISO Z39.19:2005 incluyen criterios para el mantenimiento de los sistemas de organización del conocimiento mediante gestores de tesauros automatizados. La propuesta de la NISO entiende que los vocabularios controlados suelen ser utilizados para describir el contenido por asignación de términos para representar metadatos asociados al contenido de los objetos. El estándar NISO Z39.19:2005 está pensado para ser aplicado a tesauros monolingües. Es decir, a diferencia de los estándares ISO aportados hasta ese momento, esta iniciativa incorporaba ya su aplicación en entornos informatizados, como demuestra el tratamiento de esquemas de metadatos como Dublin Core.

Otra cuestión que trata la norma ANSI/NISO Z39.19:2005 y que está estrechamente relacionado con la web 2.0 o web semántica es la interoperabilidad que según Soler Monreal (2009) consiste en desarrollar métodos que permitan utilizar vocabularios controlados en múltiples bases de datos y sistemas y permitirles ser compartidos por indicadores y buscadores.

3.2.9.2. Clases de tesauros

-generales	-monodisciplinarios	-monolingües	-principales
-especiales	-multidisciplinarios	-plurilingües	-auxiliares
-macrotesauros	-alfabéticos	-jerárquicos	-públicos
-microtesauros	-sistemáticos	-facetados	-privados

3.2.9.3. Estructura interna de los tesauros

La estructura interna que han de tener los tesauros y cuales son sus componentes básicos, como se forman y como se pueden relacionar entre sí. Según la norma UNE 50-106, que es la traducción y adaptación de la norma ISO 2788, 2ª ed, 1986.

La diferencia existente entre listas de términos, listas de encabezamiento de materias y tesauros a pesar de que los tres son vocabularios, lenguajes especializados constituido por términos, estriba en que:

La lista de términos: no contiene otra relación entre ellos, salvo la propia ordenación alfabética de los términos.

La lista de encabezamiento de materias: los términos que la componen se relacionan entre sí, a priori en un proceso de pre-coordinación, que le confiere cierta rigidez

El tesoro: los términos simples o compuestos se hallan relacionados entre sí de forma que permiten su combinación en un proceso de post-coordinación, son por tanto más flexibles en sí mismos y su actualización resulta más dinámica y rápida

3.2.9.4. Los componentes fundamentales de los tesauros

Son las palabras clave, los unitérminos, que cuando dichas palabras clave o unitérminos representan un concepto de un área del conocimiento, se le denomina término o descriptor.

Entre las clases de términos:

Según Curras (1998)	Según García Gutiérrez (1984)
Simple	Unitérmino o descriptor simple
Compuestos	Descriptor sintagmático o compuesto
Identificadores: palabras clave aisladas (ayudan a identificar un documento)	Descriptores informativos: informan sobre documentos individuales
Indicadores de función: indicadores clasificatorios	Descriptores referenciales o direccionales: orientan hacia un colectivo de documentos

Tabla 5. Clases de términos que componen un tesoro

Los términos pueden ser Principales: son los que describen el tema (también llamados términos preferentes) o Secundarios: términos sinónimos o cuasisinónimos

3.2.9.5. Tipos de relación entre términos

Desde el punto de vista semántico:

Clases de relación:

- Jerárquica – Término
 - Cabecera (TC)
 - Genérico (TG)
 - Específico (TE)
- Asociativa – Término
 - Relacionado (TR)
- Equivalencia
 - Sinonimia (USE UP)
 - Cuasi sinonimia (S. UP)

Desde el punto de vista sintáctico:

- Términos simples:
 - sustantivos (SUST)
 - Adjetivos sustantivados (ADJ)
 - Verbos sustantivados (VERB)
- Términos compuestos:
 - sustantivos nominales (SUST+SUST)
 - sustantivos adjetivados (SUST+ADJ)
 - frases preposicionales (SUST+PRE+SUST)
 - expresiones adverbiales (SUST+ADV+SUST)

- frases mixtas (SUST+ETC+SUST)

Si aplicamos un breve ejemplo de punto de vista semántico:

BAHÍA

NA-Geográficamente considerado - Nota de aplicación= definición

UP-ENSENADA - Relación de equivalencia

TC-OCEANOGRAFÍA

TG-MAR

TGG-OCÉANO - Relación jerárquica

TGP-GOLFO

TE-PUERTO DE MAR

TR-PUERTO DEPORTIVO

- Relación asociativa

3.2.9.6. Diferencias entre Tesauros y Ontologías

Según Soler Monreal (2009) la diferencia entre ontologías y tesauros estriba en que las relaciones entre los términos son diferentes, los tesauros constan de relaciones jerárquicas, asociativas y de equivalencia, mientras que las ontologías expresan relaciones de mayor variedad entre clases y propiedades (por ejemplo “escrito por”, “aparece en”, etc.) además las ontologías tienen un mayor nivel de abstracción que los tesauros y mayor complejidad formal, además de estar más orientadas al conocimiento automático.

Según los autores Pedraza-Jiménez, Codina y Rovira (2007), no hay que confundir una ontología con un tesoro, ya que al igual que un tesoro, las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesoro tradicional, están elaboradas con una sintaxis comprensible para los ordenadores. Por tanto, una ontología permite mayor riqueza en la definición de sus conceptos y sus relaciones que un tesoro.

Los autores Codina y Pedraza-Jiménez (2011), indican que los tesauros son una tecnología bien asentada en los sistemas de información, con modelos bien establecidos y con una buena implantación en la industria y en la profesión. Las ontologías son un tecnología prometedora y de un enorme potencial, ya probada en otros campos (lingüística y en ámbitos específicos de la inteligencia artificial), pero en su aplicación a los sistemas de información aún es una solución inmadura.

3.2.9.7. Diferencias entre Tesauros y Folksonomías

Soler Monreal (2009) apunta que la diferencia entre folksonomías y tesauros estriba en que las folksonomías son sistemas de clasificación muy simples, que no precisan del aprendizaje de reglas para su utilización, además de ser para los usuarios muy sencillas, las folksonomías las crean los usuarios y reflejan la frescura y dinamicidad de la lengua, éstas se caracterizan por su bajo coste y su escaso mantenimiento, igualmente son apropiadas para indizar grandes volúmenes de información, las folksonomías utilizan conceptos organizados por los usuarios, mientras que las ontologías utilizan los

conceptos organizados por los diseñadores, pudiéndose construir ontologías basadas en folksonomías.

3.2.9.8. Diferencias entre Tesauros y Taxonomías

Según Soler Monreal (2009) la diferencia entre tesauros y taxonomías estriba en que las taxonomías están vinculadas a la organización de la información en sitios web para facilitar la navegación y la recuperación de la información, además están vinculadas a entornos corporativos, en los que gestionan todo tipo de información empresarial. Tanto tesauros como taxonomías son vocabularios controlados y ambos poseen una estructura jerárquica que admiten la polijerarquía y las relaciones asociativas entre sus términos.

Se diferencian fundamentalmente en que las taxonomías se centran en el usuario y entre sus aplicaciones se encuentra que mientras las taxonomías clasifican los recursos para facilitar su recuperación y como estructura de navegación visual en la navegación web, los tesauros se utilizan en la indización de documentos y en la búsqueda de información.

3.2.10. La Web Semántica o Web 2.0

Hoy día se publica en la literatura y medios digitales la gran evolución de la web 2.0, en parte viene provocada por el aumento ingente de personas que tienen actualmente acceso a la web y como no, del fácil acceso a contenidos donde los usuarios tienen un papel protagonista y activo, debido también a la idea de la comunicación entre personas que cada vez es más evidente.

Todo apunta a que dejamos atrás la web 1.0, una web más estática en la que sólo personas muy cualificadas colgaban información y en la que los usuarios sólo podían consultar, acceder a servicios y leer los contenidos.

Un caso concreto en el que se ha producido un gran cambio en la web y en el que se evidencia un salto cualitativo en cuanto a la presentación de contenidos son los periódicos en línea, donde se evidencia un cambio sustancial respecto al periódico tradicional, dentro de los cambios más visibles se encuentra que los periódicos en línea tienen aplicaciones muy prácticas como puede ser una hemeroteca digital, en la que puedes consultar cualquier número de tu interés, otros cambios en mi opinión que mejoran el acceso y las características de los periódicos en línea son los blogs, entre los que se encuentran opiniones actuales sobre cualquier noticia, entre ellos aparecen también los blogs más leídos, las noticias más leídas por los usuarios, otro de los cambios visibles es la rapidez en la actualización de la noticia, la posibilidad de consultar prensa internacional, la posibilidad de consultar la prensa de tu comunidad, además de ofrecer servicios añadidos como traductores en línea, aplicaciones, hemeroteca, calendario, diccionarios en línea o servicios patrocinados como páginas amarillas, etc.

Todas estas mejoras en la accesibilidad y en el amplio abanico de servicios suponen un avance tecnológico importante y un avance en el acceso a la información, aunque lleve consigo ciertas desventajas como publicidad excesiva, etc.

Por otro lado, respecto a los nuevos navegantes de Internet que reelaboran información en la web, supone un crecimiento aún mayor, casi exponencial en cuanto al volumen de

información, la cuestión es que esto supone que la fiabilidad de la información contenida en la web puede minimizarse y esto puede llevar a que en los resultados de las búsquedas aparezca una menor precisión debido a la gran cantidad de ruido documental.

Este problema en el que nos encontramos es el cometido principal la W3C, dirigido por Berners-Lee (2001), en el que indica la nueva web semántica como una extensión de la actual, en la cual se orienta hacia la codificación semántica de los documentos y a la aplicación de nuevas tecnologías y procedimientos de representación del conocimiento con el fin de mejorar el acceso a los recursos web. Es decir que lo que parece a priori un problema de fiabilidad, precisión en los resultados, etc. debe verse como una oportunidad en la que los propios usuarios etiquetan los documentos que reelaboran aportando así una lingüística actualizada y concreta, y el cometido de los profesionales de la información es conocer las herramientas para orientar toda esa masa de información hacia una estructuración semántica.

La cantidad de información existente en la web crece imparablemente, aunque somos conscientes de ello asusta realmente ver los datos que justifican este hecho. Pero nosotros los profesionales de la información se nos abre un camino nuevo por explorar y descubrir, el documentalista y profesional de la información tiene ante sí una oportunidad de formarse y en consecuencia reciclarse adecuadamente para colaborar en esta nueva web 2.0 o web semántica que se está gestando. Es de vital importancia que esta nueva figura de documentalista o profesional de la información vaya de la mano de otras disciplinas necesarias para abordar este cambio tan importante. Entre estas se encuentra la informática (Inteligencia Artificial relacionada con la Representación del Conocimiento que se ocupa de sistemas expertos), es decir el éxito para alcanzar una web semántica, en la que se cumpla el objetivo de la interoperabilidad de los datos que circulan en la web y de la recuperación de información radica en la conexión e integración de estas dos disciplinas: la documentación y las nuevas herramientas informáticas.

De ahí la importancia de que los profesionales de las ciencias de la información debamos actualizar nuestros conocimientos y dirigir nuestra formación hacia las aplicaciones informáticas, estándares, lenguajes, etc.

Como decía al comienzo el crecimiento de Internet supone un problema, éste es un problema semántico: ¿cómo toda esa información que circula por la web se puede relacionar y utilizar de un modo más preciso? ¿cómo podemos hacer para que las máquinas parezcan realmente inteligentes? ¿cómo un buscador en la web puede ofrecerme resultados que estén estrechamente relacionados con mi consulta?, las respuestas a todo esto está en las ontologías: vocabularios controlados que aportan semántica y denominan conceptos de forma no ambigua y que están elaboradas con una sintaxis comprensible para los ordenadores.

Según los autores Pedraza-Jiménez, Codina y Rovira (2007), el lenguaje HTML al no ser marcado semántico, las máquinas no pueden identificar la información, el estandar XML soluciona esto utilizando etiquetas que expresen el significado de los elementos y no su formato.

Berners-Lee (2001) publicó un artículo en el que anunciaba el proyecto de la web semántica como una extensión de la actual, dotada de una estructura que permitiera expresar el contenido de las páginas de una forma que los ordenadores pudieran entenderlas y posibilitase tanto la interacción entre ordenadores como entre éstos y los usuarios. Berners-Lee actual director del W3C presenta ahora una visión más prudente orientada hacia la codificación semántica de los documentos y a la aplicación de nuevas tecnologías y procedimientos de representación del conocimiento con el fin de mejorar el acceso a los recursos de la web. Tim Berners-Lee expone su visión de la web en los siguientes videos¹⁶.

En definitiva, la web semántica (Moreiro González, 2009) no sería una nueva web sino la extensión de la existente, mediante la adición de metadatos que describan la semántica de las páginas de forma procesable por máquinas.

3.2.10.1. La interoperabilidad en la web semántica. SKOS en el entorno Linked Open Data

La interoperabilidad (Moreiro González, Sánchez Cuadrado y Morato Lara, 2012) es la capacidad de compartir datos y posibilitar el intercambio de información de los sistemas de información y los procedimientos a los que estos dan soporte a nivel de interoperabilidad técnica, semántica y organizativa. Los autores no cuestionan la necesidad de interoperabilidad a la hora de elaborar o de utilizar vocabularios controlados, tanto para acceder a la información precisada como para representar el contenido específico de un campo.

Siguiendo con el artículo de los autores antes citados la web semántica o web 2.0 da la oportunidad a usuarios, empresas e instituciones de aprovechar e integrar servicios proporcionados por otros de manera que las entidades que participan en la web 2.0 usan métodos incluyentes a la hora de agregar datos. Un usuario puede añadir y compartir información con otros usuarios, de forma que genera valor añadido como efecto colateral del uso ordinario de la aplicación.

Así Iglesia Aparicio y Monje Jiménez (2012) indican que la descripción de los ámbitos de conocimiento mediante metadatos u ontologías es básica para conseguir hacer real la web semántica. Pero, es igual de importante que los datos disponibles en la red se puedan relacionar entre sí de una forma estructurada y que, además, dichos datos sean accesibles por el mayor número de servicios y aplicaciones posibles. Por ello, el objetivo de la web semántica es la existencia de datos con enlaces estructurados y abiertos, y para ello es necesario que los datos disponibles en la web ya estén estructurados mediante metadatos, ontologías u otros estándares de la web semántica como RDF (Resource Description Framework) o SKOS (Simple Knowledge Organization System, Sistema para la Organización del Conocimiento simple). Es lo que se llama datos enlazados o, en inglés, *linked data* (LD).

¹⁶Youtube: La web semántica <http://www.youtube.com/watch?v=MEjRFXbqjLc>
<http://www.youtube.com/watch?v=5IxMov7StOI> [ref. de 08 de noviembre 2012]

Estructurar datos es un esfuerzo que no es asumible por un reducido número de personas. Necesita de equipos amplios, en muchos casos interdisciplinarios y también del apoyo de organismos e instituciones de prestigio. Actualmente, existe un número creciente de sitios web dedicados a esta tarea aprovechando incluso datos que están siendo categorizados de forma desinteresada por usuarios de todo el mundo. La iniciativa más fructífera y que se ha convertido en el núcleo central de esta red de datos enlazados es *Dbpedia* (<http://dbpedia.org>).

Dbpedia es una iniciativa alemana que se ocupa de extraer información estructurada de la enciclopedia Wikipedia. Durante muchos años, los usuarios de Wikipedia han ido organizando los artículos en categorías, subcategorías; han creado fichas estructuradas de datos; han relacionado unos conceptos con otros. Todo este esfuerzo es el que Dbpedia trata de aprovechar para facilitar la estructuración de los datos contenidos en otros sitios web. El producto final es una base de datos descrita en RDF y que además es de uso libre. Cada concepto ofrecido por Dbpedia tiene una URI como esta: <http://dbpedia.org/resource/Spain>.

El esfuerzo de estructuración de datos no sirve de nada si no disponemos de datos actualizados, fiables y disponibles. Si bien, no se puede pedir que todos los datos sean accesibles de forma gratuita y universal, la gran mayoría de las administraciones o de las instituciones de carácter público poseen datos de interés que deben de estructurar y de poner al servicio del conocimiento en la Web sin restricciones de copyright u otros mecanismos de control. Es lo que se llama datos abiertos (*open data*).

Por lo tanto, la red de sitios web que hará posible la existencia de una web realmente semántica debe de proporcionar datos abiertos y con relaciones estructuradas entre sí, es decir, usando la terminología inglesa, tienen que ofrecer *Linked Open Data*. (LOD)

Para que otros sitios web puedan hacer referencia a estos datos es crucial disponer de un punto de información acerca de qué datos están disponibles y quién los proporciona. Esta es la labor de la comunidad *Linked Open Data* (<http://linkeddata.org/>). Cada vez son más los nodos de información que se van incorporando a la red. Mientras que en 2007 solo había 12 nodos, actualmente, la red ha crecido considerablemente. El último grafo de la red está disponible en esta dirección: <http://richard.cyganiak.de/2007/10/lod/imagemap.html>.

Los datos públicos y enlazados son cada vez más relevantes y serán la base de una nueva forma de buscar y de obtener la información. Por supuesto, el buscador líder no podía obviar este hecho y ya proporciona un interfaz de consulta de este tipo de datos. De momento la información que muestra es limitada pero seguro que irá creciendo de forma progresiva. Se trata de su servicio Google Public Data Explorer <http://www.google.com/publicdata/directory>.

En este sentido, Saorín (2012) manifiesta que abrir los datos aumenta las posibilidades de participar en la innovación. Nuestra responsabilidad no era hacerlo todo directamente, era hacerlo posible. Eso es también voluntad de servicio, crear las condiciones para la acción desde la iniciativa social o empresarial. Linked data nos permitirá ser parte de la web, y no sólo estar en ella.

Según un reciente artículo de las autoras Méndez y Greenberg (2012) hablar de la interoperabilidad y en consecuencia de linked open data (LOD) o linked data (LD), *datos enlazados* se ha convertido en un tema de creciente interés en el entorno de la biblioteconomía y documentación y de la informática. LD es clave para la evolución de la web semántica para la reutilización y compartición de datos y también para enlazar el conocimiento.

Siguiendo con el interesante artículo de las autoras antes citadas, Tim Berners-Lee (1999) ha estado promocionando y defendiendo la idea de una potente estructura de conocimiento interconectado que enlaza información, documentos y datos. A esta idea le llamó primero la web (1989), luego la web semántica (1989) y ahora linked data (2006). Tal como predijo en su libro *Weaving the Web (Tejiendo la Web)* (1999), html hizo posible la web de documentos hipertextuales, y RDF y las tecnologías de la web semántica (OWL¹⁷, SKOS¹⁸, Sparql¹⁹) harán posible la web de datos, a través de conjuntos de datos enlazados definidos en tripletes RDF.

Las autoras Méndez y Greenberg (2012) indican que el desarrollo de la web semántica basada en estándares abre la posibilidad de una interoperabilidad más global a través de la web de datos enlazados. SKOS (Simple Knowledge Organization System, Sistema para la Organización del Conocimiento simple), como estándar recomendado del Consorcio Web desde 2009, implica un paso más para hacer los tesauros interoperables, permitiendo que se puedan compartir y poniéndolos a disposición de tal forma que su contenido se puede integrar y vincular. Es decir, donde los vocabularios se crean explícitamente para la web, vocabularios abiertos y enlazados.

Según los autores Codina y Pedraza-Jiménez (2011), un tesoro representado en SKOS sigue siendo un tesoro, pero el hecho de disponer de un lenguaje de alto nivel formal y exigencia lógica como RDF puede ayudar no solamente a desarrollar mejores lenguajes documentales, sino que facilita la interconexión de entre los lenguajes desarrollados y especificados mediante SKOS. Además, consideran los autores que adoptar las tecnologías de la web semántica, como las ontologías, pondrán las bases para la interconexión de datos y la prestación de mejores servicios de información.

La interoperabilidad entre SOC (Sistemas de Organización del Conocimiento) busca armonizar las relaciones conceptuales y terminológicas que pudieran establecerse entre ellos. Es lograr que estos SOC puedan intercambiar información, independientemente del contexto en el que han sido creados y mantener al mismo tiempo la eficiencia, junto a otros SOC, en la recuperación de la información (Martínez Tamayo, et. al., 2011)

Según un reciente artículo de Pastor-Sánchez, Martínez-Méndez y Rodríguez-Muñoz (2012) SKOS se define formalmente como una ontología OWL-full (extensión de RDF) que permite representar cualquier tipo de sistema de organización del conocimiento mediante RDF. Su ámbito de aplicación se extiende a la práctica totalidad de

¹⁷ Lenguaje para definición de ontologías, es una extensión de RDF para la descripción de aspectos lógicos de las relaciones entre recursos

¹⁸ Estándar recomendado por W3C desde 2009 para la interoperabilidad de datos. Sistema para la Organización del Conocimiento Simple.

¹⁹ Es un lenguaje de consulta para interrogar y recuperar datos RDF

vocabularios controlados: clasificaciones, tesauros, encabezamientos de materia, taxonomías, tesauros, glosarios, etc. el estudio realizado tiene como objetivo reflejar la presencia de SKOS en la publicación de este tipo de vocabularios y sus relaciones con otros vocabularios o conjuntos de datos abiertos para determinar entre tanto el alcance de esta tecnología en el ámbito del linked open data, como el acceso abierto a los datos, la consulta a través de un Sparql y la existencia de licencias, además de determinar el grado de interoperabilidad entre los vocabularios mencionados.

La nueva concepción de sistema de organización del conocimiento tras un detallado análisis de las normas de tesauros y KOS, Sánchez Cuadrado, Colmenero Ruiz y Moreiro (2012) indican que esta nueva concepción deja a un lado las diferenciaciones entre mono y multilingüismo. Requieren la representación formal del contenido, lo que implica un lenguaje de formalización más lógico. Este cambio de representación condiciona la estructura relacional. La función de los KOS se ha ampliado en la última década con las TIC, convirtiéndose la interoperabilidad, la vinculación entre KOS y la reutilización en factores esenciales.

Un caso real de la implementación de las técnicas LD y LOD en instituciones culturales, es el caso de Europeana (Ríos-Hilario, Martín-Campo y Ferreras-Fernández, 2012) siendo abanderada en la implementación de esta tecnología. Europeana es un portal de Internet que actúa como una interfaz de millones de libros, pinturas, películas, objetos de museo y registros de archivos digitalizados de toda Europa, siendo también una plataforma para el intercambio de conocimiento que promueve la colaboración entre los agentes culturales.

4. Métodos cuantitativos y Leyes clásicas en Recuperación de la Información. Estudios destacados.

Los métodos cuantitativos y leyes clásicas en Recuperación de información son la base de estudio en que se fundamentan muchos de los algoritmos de recuperación de información utilizados hoy día por software documentales, por sistemas gestores de bases de datos, Internet, etc. La creación de nuevas herramientas de software más complejas que utilizan métodos cuantitativos aplicados a la gestión del conocimiento está conduciendo a que se desarrollen aplicaciones más inteligentes en el ámbito de la indización automática y la recuperación de información.

Las tecnologías de la información y la comunicación facilitan el acceso a Internet y a bases de datos documentales accesibles fácilmente desde cualquier terminal conectado a la red. Esto ha posibilitado que los usuarios estén fácil y rápidamente informados y puedan utilizar para ello técnicas y software expertos para obtener la información puntual que necesiten.

Los métodos cuantitativos y leyes clásicas en recuperación de información comenzaron a desarrollarse en los años 40 mediante estudios destacados de varios autores que sentaron las bases de esta ciencia, estos autores y sus modelos cuantitativos se analizan minuciosamente en este y en posteriores capítulos.

Para conseguir el objetivo final de esta investigación, que es desarrollar e implementar un Sistema de Indización y Segmentación Automática para textos largos en español, se estudian y perfeccionan los métodos cuantitativos y leyes clásicas en Recuperación de Información, como son los modelos relativos al proceso de repetición de palabras (Zipf, 1949), (Mandelbrot, 1953) y al proceso de creación de vocabulario (Heaps, 1978). Se realiza una crítica de las circunstancias de aplicación de los modelos y se estudia la estabilidad de los parámetros de manera experimental mediante recuentos en textos y sus fragmentos. Se establecen recomendaciones a priori para los valores de sus parámetros, dependiendo de las circunstancias de aplicación y del tipo de texto analizado. Se observa el comportamiento de los parámetros de las fórmulas para vislumbrar una relación directa con la tipología de texto analizado. Se propone un nuevo modelo (Log-%) para la visualización de la distribución de frecuencias de las palabras de un texto.

4.1. Estudios destacados de Heaps. La ley de Heaps y su modelo de crecimiento del vocabulario respecto del tamaño del texto

La ley de Heaps es conocida en la lingüística como una ley empírica que describe la cantidad de vocabulario en un documento o conjunto de documentos, según los estudios de Heaps se demuestra que el número de palabras distintas no crece a la misma velocidad que pueda crecer el tamaño absoluto del texto ya que el vocabulario total es finito y las palabras se repiten.

Respecto a esta afirmación existen modelos experimentales que averiguan la cantidad de palabras distintas que aparece en un texto en relación con otro de tamaños diferentes, aunque el modelo más conocido y preciso para realizar estas observaciones es la proporcionada por Heaps (1978) y su fórmula:

$$V = K \cdot n^\beta = O(n^\beta)$$

La fórmula de Heaps establece el modelo de crecimiento del vocabulario al crecer el tamaño absoluto del texto. A grandes rasgos, en la fórmula n representa el número total de palabras en el texto, K y β son parámetros que dependen del tipo de lenguaje empleado, K puede tener un valor comprendido entre 10-100 y β un valor positivo entre 0,4 y 0,6. Por ejemplo valores típicos pueden ser $K = 40$ y $\beta = 0,5$. Con estos valores es de esperar que un texto de 10.000 palabras tenga un vocabulario:

$$V = 40 \cdot 10.000^{0,5} = 6.000$$

En la siguiente figura, como ejemplo puramente visual se muestra como el tamaño del vocabulario varía con el tamaño del texto.

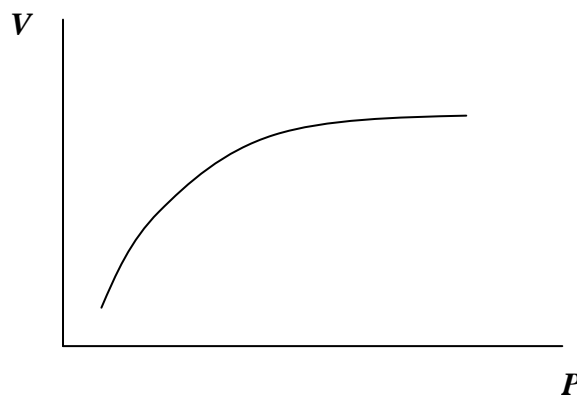


Figura 6. Ley de Heaps. Crecimiento del vocabulario respecto al tamaño de un texto

Se puede observar como el vocabulario acorde al aumento progresivo del tamaño del texto aumenta describiendo una curva potencial.

Si tomamos el modelo inicial mencionado anteriormente y aplicamos esta teoría de Heaps con datos provenientes de textos literarios-poéticos de diferentes tamaños; 3.000, 5.000, 20.000, 30.000, y 50.000 palabras, podemos observar como el gráfico se aproxima a la figura 6 anterior.

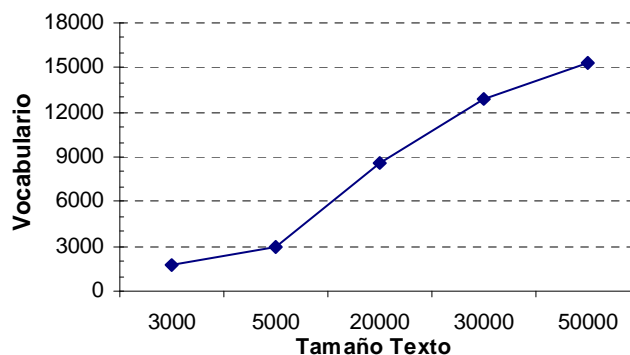


Gráfico 1. Ejemplo Ley de Heaps

Esencialmente esta ley intenta averiguar la cantidad de palabras distintas que hay en un documento, pero hay que tener en cuenta que este análisis no determina la útil y

provechosa información que un documento aporte, ya que aunque un documento contenga abundantes palabras distintas no significa que sea un documento más rico en su lenguaje, como decía Miguel de Unamuno en un escrito;

“Pues nos parece evidente que mayor riqueza de lenguaje son cuatro palabras precisas, de oro, que no cuarenta meramente sonoras o de calderilla. Y de aquí que no sean, verbigracia, los oradores de más abundoso vocabulario los que usan más palabras diferentes, los de lenguaje más rico. La riqueza no es el número precisamente”

Miguel de Unamuno
Publicado en Los Lunes de El Imperial
Madrid 21 de mayo 1917

“La abundancia de palabras diferentes no equivale a riqueza. Pues la riqueza de una lengua más se pesa que se cuenta. Una palabra de oro, precisa, concreta, sugestiva vale mas que una docena de palabras de plomo”

Miguel de Unamuno
Publicado en La Nación, Buenos Aires
14 de mayo 1917

En estos escritos de Miguel de Unamuno ya se vislumbra lo que en capítulos posteriores de esta investigación se verá como un método para caracterizar la importancia, la riqueza de las palabras en los documentos, este método consiste en aplicar un peso a las palabras de mayor relevancia en dichos textos, este peso designará la importancia del documento por la calidad de su vocabulario y no por el tamaño de este. Así de este modo, ya decía Miguel de Unamuno en 1917 que la riqueza de una lengua más se pesa que se cuenta.

Entre las publicaciones más destacadas de Heaps:

Heaps, H. S. Data Compression of Large Document Data Bases. Journal of Chemical Information and Computer Sciences 15 (1): 32-39, 1975

Heaps, H. S. A theory of Relevance for Automatic Document Classification. Information and Control 22 (3): 268-278, 1973

Heaps, H. S. Criteria for Optimum Effectiveness of Information Retrieval Systems. Information and Control 18 (2): 156-167, 1971

Heaps, H. S. Information Retrieval. Computational and Theoretical Aspects. Academic Press, 1978

4.2. Estudios destacados de Zipf. La ley de Zipf y su modelo de frecuencias de palabras en un texto

George K. Zipf lingüista norteamericano (1902-1950), de familia acomodada fue profesor de Filología de la Universidad de Harvard desde finales de los años veinte. Comenzó estudiando la economía humana del habla, esto es, las restricciones habituales

en la expresión oral cotidiana, para abordar más tarde el uso del vocabulario en la producción escrita.

Sus primeros trabajos, publicados en 1932 se basaron en análisis empíricos acerca de la regularidad con la que los términos aparecían en diversos textos. Comprobaciones que extendió a obras de diferentes autores y distintas lenguas, con resultados que venían a verificar sus hipótesis. El último trabajo de Zipf publicado en 1949, poco antes de su prematura muerte, abordó el *Ulises* de *James Joyce*.

Más preocupado por el comportamiento humano que por las matemáticas se definió como un estadístico de la ecología humana, argumentando su mencionada “ley del mínimo esfuerzo”, en la que constató que las palabras más cortas eran mucho más frecuentes en su uso escrito que las largas y del mismo modo los términos más conocidos adquirirían mayor protagonismo.

Verdaderamente se ha comprobado que en una colección de documentos las palabras más frecuentes en un texto tienden a ser las más comunes, o las que no contienen información, es decir las palabras vacías (*stopwords*), estas conviene que sean eliminadas usando una lista de palabras vacías. Por el contrario las palabras que sí contienen información, las palabras clave (*index term*) pueden aparecer aleatoriamente en cada texto pero con una menor frecuencia que las palabras vacías. De este modo resulta de gran interés tanto las palabras que se repiten con asiduidad como las palabras que aparecen una vez en el texto.

La ley de Zipf fue la primera descripción del modelo de frecuencias de aparición de palabras en un texto. Además, Zipf, se interesó desde los inicios de su carrera profesional por los cambios fonéticos que tenían lugar en las distintas lenguas y por las frecuencias del empleo de los distintos fonemas, que presentaban ciertas modificaciones cuando eran observados durante un tiempo suficientemente prolongado. Del estudio de las frecuencias relativas de los fonemas Zipf pasó a trabajar con las frecuencias relativas del empleo de las palabras en los textos.

La ley de Zipf está considerada como uno de los fenómenos más llamativos de la lingüística cuántica, según Zipf si ordenamos las palabras de mayor a menor frecuencia, el producto que resulta de multiplicar las frecuencias (f) de observación de las palabras de los textos por el valor numérico del rango (r) que ocupan estas palabras en una distribución de frecuencias de observación, permanece constante. Así la ley verifica que: $Rango * Frecuencia = Constante$. Igualmente se constata que dicha constante sufre una pequeña desviación, por ello la fórmula de Zipf incluye un exponente que tendrá un valor muy próximo a 1. Mediante la siguiente formulación aplicaremos la distribución de frecuencias de Zipf:

$$Fr = \frac{K}{r^e}$$

La siguiente figura es un ejemplo puramente visual en el que se muestra la relación entre la frecuencia de las palabras en un texto y el rango de cada una de ellas en dicho texto. Esta relación muestra claramente como cuanto mayor es la frecuencia mayor es el

rango, esto es que a la palabra más frecuente le corresponde el rango 1 y según va disminuyendo las frecuencias de las palabras el rango es menor (rango 76, 77, etc.).

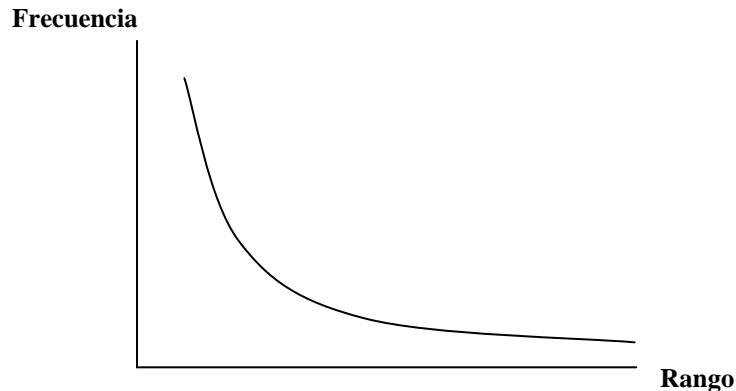


Figura 7. Ley de Zipf. Distribución de frecuencias de palabras

Es decir, conforme aumenta el rango disminuye la frecuencia, de modo que su producto se mantiene constante.

En el siguiente ejemplo gráfico se constata esta tendencia, siendo el texto escogido del autor Pérez Galdós de tipología literaria compuesto por un total de 4.125 palabras y un vocabulario de 2.590 palabras

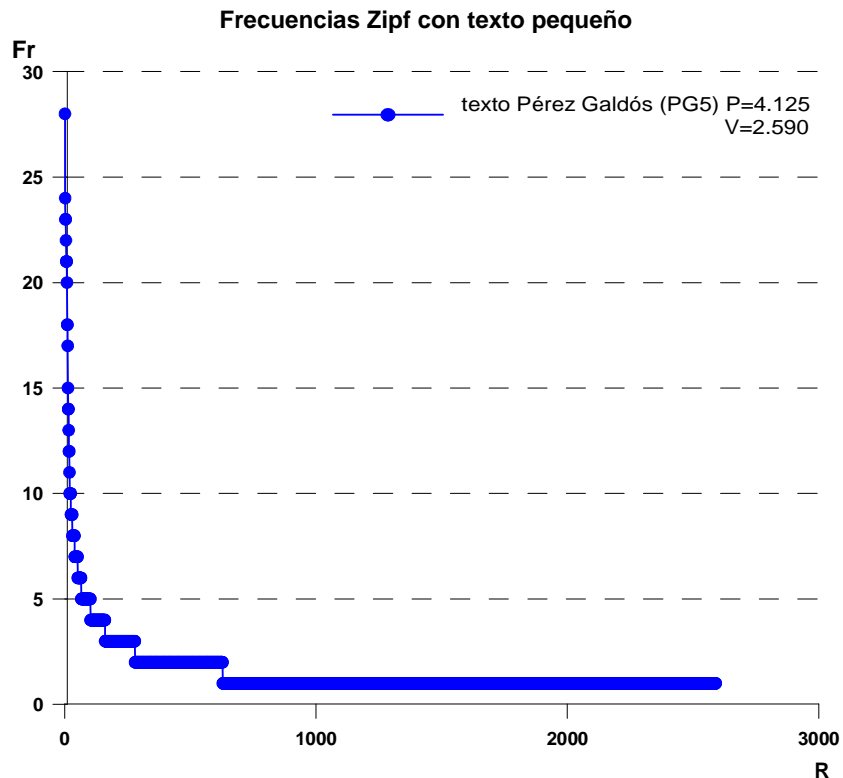


Gráfico 2. Ley de Zipf. Distribución de frecuencias de palabras

Al comprobar la validez de esta ley se puede observar en el ejemplo, que ésta se cumple sólo para un conjunto de palabras, situadas en los valores intermedios, es decir entre las palabras más frecuentes y las menos frecuentes, en cambio para las palabras más frecuentes, el producto se aparta sensiblemente del valor constante que parecen definir

las palabras intermedias. Investigaciones más desarrolladas de esta Ley de Zipf se abordarán en capítulos posteriores.

Entre las publicaciones más destacadas de Zipf:

Zipf, G.K. Selective Studies and the Principle of Relative Frequency in Language, 1932

Zipf, G.K. Psycho-Biology of Languages, 1935

Zipf, G.K. Human Behaviour and the Principle of Least-Effort, Addison-Wesley, Cambridge MA, 1949

4.3. Cantidad de palabras significativas en un documento

Uno de los objetivos de este estudio es investigar la asociación entre palabras significativas que pueden indicar conceptos manejados por el autor y es por ello que estamos procesando palabras extraídas de colecciones de documentos para fabricar instrumentos para la indización y segmentación automática de la información. La aparición o no de una palabra o término en un documento servirá para caracterizarlo con relación a la demanda de información de un usuario. Por ello es obvio comenzar planteándose desde el principio la eliminación de un gran número de palabras vacías.

Las palabras más frecuentes, si aparecen en todos los documentos, carecen de poder discriminatorio, pero cabe plantearse si son esas palabras muy frecuentes las que aparecen en más documentos. Puede una palabra ser globalmente muy frecuente por que en unos pocos documentos aparece muchas veces, mientras que en otros no aparece casi nunca. Este hecho nos alerta que puede ser significativa en cierto número de documentos.


Por otro lado hay palabras que siempre aparecen en la misma proporción. Por ejemplo las palabras vacías (*stopwords*), consideramos vacías no sólo los habituales artículos, preposiciones, adverbios, adjetivos indeterminados sino también muchas formas verbales que indican modo de la expresión o forma de hablar, y punto de vista, etc. como por ejemplo: *Así de este modo, no obstante, de este manera, observamos, indica*, etc. Las palabras vacías no sólo son muy frecuentes, sino además su frecuencia es aproximadamente la misma en cada fragmento de texto. No sirven, pues, para discriminar entre unos documentos y otros, además de no aportar un valor significativo para la recuperación de información.

En realidad, hay por lo menos, tres conceptos diferentes de palabra vacía, las palabras vacías gramaticales: artículos, preposiciones, etc. Las semánticamente vacías: expresiones de puntos de vista o forma de hablar como las mencionadas anteriormente y que no aportan información esencial. Por último las palabras vacías estadísticamente son las que tienen una distribución de frecuencias parecidas en cada documento, sea cual sea el motivo, gramatical, semántico u otro.

A la hora de construir un sistema de recuperación de información (SRI), o en nuestro caso un sistema de Indización y Segmentación Automática se debe prescindir de las palabras vacías, al menos las más frecuentes deben eliminarse, por el espacio que ocuparían en los índices.

Por tanto tras aplicar varios ejemplos, que un autor escriba más palabras significativas por página que otro autor depende más del criterio del analista sobre qué palabras son significativas, que de la forma de escribir del autor. El criterio del analista puede estar sesgado a favor de un autor.

Como estamos hablando de palabras bastante frecuentes, resulta que tienen influencia sobre el número total de palabras encontradas, pero muy poca sobre el número de palabras distintas o vocabulario, que son las obtenidas con la fórmula de Heaps. Así pues, la arbitrariedad introducida por el conjunto de palabras vacías tiene poca influencia en la sucesión de datos

Caracteres palabras vocabulario raíces palabras asociadas


Para no caer en esta alteración arbitraria basaremos nuestras fórmulas en el número de caracteres o bytes del texto, manteniendo alternativamente la fórmula en función del número de palabras para comparación con la expresión clásica de la fórmula de Heaps.

Para hallar la cantidad de palabras significativas de los documentos que van a formar el corpus del sistema de Indización y Segmentación Automática y una vez obtenido el vocabulario de estos documentos se realiza un proceso importante tanto para la capacidad de almacenamiento del sistema en cuestión como para la posterior recuperación de información en sí, como es el proceso de lematización o extracción de raíces.

Otro paso importante es determinar el poder discriminatorio de cada palabra del documento y medir la intensidad relativa de las apariciones conjuntas de las palabras en los documentos hallando así la cantidad de palabras significativas de entre las palabras asociadas.

Pero imaginemos que en un sistema de recuperación de información, el módulo de búsqueda además de dar una lista ordenada de documentos al usuario fuera capaz de realizar y mostrar la estructura temática del texto subdividido en partes y que en cada una de las partes incluyera la indización automática. Esto es nuestro objetivo final y lo que hemos desarrollado con el Sistema de Indización y Segmentación Automática MALLOV.

Es decir expresaremos el contenido de un documento, no por el resumen automático sino por la enumeración de sus partes principales que en definitiva el objetivo final será similar al que se obtiene con el resumen automático que es la categorización del texto o documento llevado a análisis. Igualmente en dicho sistema existe la posibilidad de mover las palabras entre los distintos bloques, según estén más asociadas con unos u otros, para realizar esto utilizaremos métodos de clustering y similitud entre palabras.

En definitiva, todas las experimentaciones realizadas hasta ahora con los textos se han llevado a cabo para formar una estructura coherente respecto al tamaño del documento y que se puedan seleccionar las palabras significativas, es decir, que definen mejor el contenido general del documento, pero no sólo eso, sino que también se pueda

seleccionar las palabras significativas que mejor definen el contenido de cada uno de los capítulos o subdocumentos en los que se divide el documento grande de partida u original, y así sucesivamente hasta seleccionar de igual modo las palabras más características de los párrafos, etc. Es decir indizaremos automáticamente la granularidad de los documentos objeto de análisis.

4.4. Poder discriminatorio de las palabras

El poder discriminatorio de las palabras o términos de los documentos es un proceso clave para desarrollar el Sistema de Indización y Segmentación Automática, ya que el peso de un término en un documento se caracteriza por varios principios, mayor frecuencia de aparición de un término en un documento implica mayor peso y si en dos documentos un término aparece el mismo número de veces, la ponderación del término será mayor en el documento más corto.

Para formar la matriz de asociación término-documento aproximada por frecuencias se calcularán las frecuencias de cada término en cada documento y en el global de la colección o corpus, así como el número de términos en cada documento y el número de documentos en que aparece cada término. Una vez obtenemos los cálculos en la matriz de asociación podremos discriminar los términos que se repitan demasiado en el corpus, es decir de muy alta frecuencia y por tanto son términos muy generales y por otro lado podremos igualmente discriminar los términos de frecuencia extraordinariamente baja, posiblemente muy específicos, igualmente podemos discriminar los términos tomando el criterio de eliminar los que tienen una frecuencia total en el corpus mayor que un umbral determinado ($>0,5$).

Por último la matriz de asociación se puede representar como una lista de registros tipo nodo, con tantos nodos como términos. Cada nodo se representa por un registro con tres campos: el nombre del término, la frecuencia total del término en el corpus y el número de documentos en que aparece. (*Ciencia*, 50, 4)

En definitiva, la discriminación pretende otorgar más importancia (más peso) a los términos del vocabulario seleccionado que clasifiquen mejor a los documentos dentro de la colección concreta en que se están considerando.

4.5. Fórmulas relativas a la similitud entre palabras

La similitud o asociación entre palabras es un aspecto trascendental para esta investigación, que se abordará con más detalle en el Capítulo 8. A continuación se detallan las fórmulas relativas a la similitud entre las palabras que se emplearán en las investigaciones realizadas en dicho capítulo.

4.5.1. Fórmula de Dice

Para medir la intensidad relativa de las apariciones conjuntas de las palabras en los documentos se tiene en cuenta las frecuencias de las dos palabras consideradas y se utiliza el denominado *índice de equivalencia o de asociación*, el cual mide la intensidad de la asociación entre dos palabras i y j realizada sobre el conjunto de documentos del fichero. Se obtiene el valor 1 cuando la presencia de i acarrea automáticamente la

presencia de j y viceversa, es decir, cuando las dos palabras están siempre juntas. Por el contrario es igual a 0 cuando la mera presencia de una de las dos palabras excluye la de la otra. Así llamaremos índice de equivalencia (E_{ij}) al coeficiente cuyo valor viene dado por la fórmula siguiente:

$$E_{ij} = \frac{2 \cdot C_{ij}}{C_i + C_j}$$

Donde, en un documento grande o fichero F dividido en documentos pequeños n :

C_i = número de apariciones de la palabra clave i en la totalidad de documentos n

C_j = número de apariciones de la palabra clave j en la totalidad de documentos n .

C_{ij} = número de apariciones conjuntas de las palabras i y j en la totalidad de documentos n .

4.5.2. Fórmula de Jaccard

Del mismo modo que la fórmula de Dice (1945) esta fórmula ha sido concebida tanto para evaluar la semejanza entre dos términos o conceptos, como para medir la semejanza entre documentos. La fórmula de Jaccard ha conseguido ser de las más utilizadas tanto en trabajos experimentales como en sistemas de funcionamiento.

Semejanza de Jaccard:

$$E_{ij} = \frac{C_{ij}}{C_i + C_j - C_{ij}}$$

4.5.3. Coeficiente de Semejanza del Coseno

El coeficiente de semejanza del coseno no se puede aplicar para medir la similitud entre dos términos sino que se utiliza sobre todo en el Sistema Vectorial para mediar la similitud entre documentos. La fórmula del Coseno es también de las más utilizadas, en la siguiente figura se muestra la medida de similitud del coseno

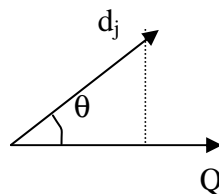


Figura 8. Medida de Similitud del Coseno

El coseno de θ es adoptado como la similitud de (d_j , q)

Semejanza del coseno:

$$E_{ij} = \frac{C_{ij}}{\sqrt{C_i^2 \cdot C_j^2}}$$

4.5.4. Medida de Información Mutua (Mutual Information. MI)

La información mutua mide la dependencia mutua de dos términos aleatorios en un corpus. La medida de la Información Mutua es una máxima probabilidad para la intensidad en la asociación estadística (logarítmica) entre dos componentes como en este caso parejas de palabras.

Fue introducida en el campo de la lexicografía computacional por Church & Hanks (1990) quien lo desarrolló a partir de la noción de información-teórica relativa a la cuestión de la Información Mutua (*point-wise*). Los valores positivos indican asociaciones positivas mientras que los valores negativos indican disociaciones (donde los componentes tienen una tendencia a no coocurrir o coincidir juntos).

La información mutua se define entre dos palabras w_1 y w_2 y mide estadísticamente la información que obtenemos sobre la posible aparición de un término a partir de la aparición de otro término

Evert (2005, p. 85) ofrece la siguiente formulación para la información mutua:

$$MI = \frac{O_{11}}{E_{11}}$$

En definitiva, la información mutua mide la dependencia mutua de dos términos aleatorios en un corpus.

4.5.5. La Media Geométrica

Del mismo modo que la fórmula de Dice y la fórmula de Jaccard, la media geométrica es otra medida que se puede utilizar para evaluar la semejanza entre dos términos o conceptos o en su caso para medir la semejanza entre documentos. Siendo la fórmula la siguiente:

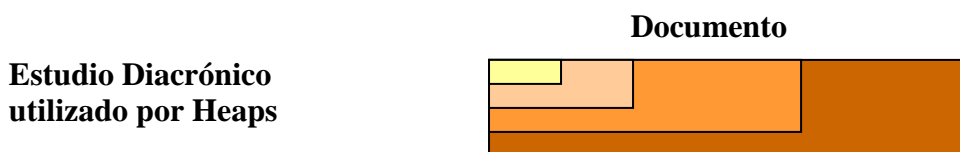
$$E_{ij} = \frac{C_{ij}}{\sqrt{C_i \cdot C_j}}$$

5. Modelo de crecimiento del vocabulario. Ley de Heaps

En este capítulo se exponen los estudios que contemplan el texto como una unidad y se obtienen conclusiones en función de su tamaño y su tipo de lenguaje o de contenido. Los textos utilizados en este capítulo que se aborda no requieren su división en documentos más pequeños, en cambio en el capítulo seis se incorpora la división del texto en documentos y se realizan los correspondientes análisis cuantitativos. Será precisamente en el capítulo ocho donde aparecen los estudios cuantitativos con las palabras asociadas.

Este capítulo como los siguientes son experimentales, es decir utilizamos una serie de textos reales para hacer recuentos o cálculos y utilizamos las aplicaciones informáticas desarrolladas para ello con los que obtenemos resultados numéricos o cuantitativos. Todos estos experimentos se plasman en gráficos para visualizar los resultados.

En este capítulo se realiza un enfoque diacrónico, este enfoque es el utilizado por Heaps ya que considera los fragmentos del texto como partes cada vez mayores del texto original. Se estudia el modelo de crecimiento del vocabulario, la estabilidad de sus parámetros de manera experimental aportando críticas y una nueva versión de la Ley de Heaps tanto para la interpolación, es decir para predecir el vocabulario de los fragmentos del texto como para extrapolación.



A esta investigación se añaden tres apéndices que ayudarán a tener un mejor conocimiento tanto de las aplicaciones utilizadas para llevar a cabo este estudio científico como de la tipología de los textos utilizados para tal fin.

En el Apéndice I se describe la base documental que se ha tomado, la colección de textos en los que se han basado los experimentos, y que es común a los capítulos cinco y seis.

En el Apéndice II se detalla la relación ordenada de los programas utilizados, éstos son mencionados por nombres simbólicos, normalmente constituidos por dos palabras separadas por un guión. Como se han realizado en Microsoft Access las dos palabras responden al nombre de la aplicación y el nombre del formulario.

En el Apéndice III se incluyen los formularios destacados de las bases de datos.

5.1. Estudios cuantitativos relacionados con palabras en un texto o colección de textos similares. Metodología estadística del estudio cuantitativo

5.1.1. El vocabulario o número de palabras distintas utilizadas en cada texto

Existen estudios que miden la distribución de las frecuencias de las palabras en un texto, (Zipf, 1949) la cantidad de palabras vacías y la influencia de estas en dicho texto, e igualmente existen estudios que miden la distribución de palabras en los documentos de una colección, también existen estudios sobre las palabras distintas en un documento o lo que podemos llamar el vocabulario en un documento. Para predecir la cantidad de vocabulario en un texto con un lenguaje natural de un tamaño cualquiera como puede ser P , aplicaremos la ley de Heaps. Según Baeza-Yates y Ribeiro-Neto (1999) esta Ley de Heaps se ha utilizado en estudios muy diversos para caracterizar a un autor o a una época u otra serie de características más propia de algún tipo de textos (Carpentier, 1982).

En este capítulo se va a aplicar la Ley de Heaps con máxima precisión para obtener detalles de cómo, cuanto y en qué circunstancias el modelo de Heaps se cumple y ver si podemos aprovechar el modelo de Heaps para obtener algo práctico y aplicarlo al Sistema de Indización y Segmentación Automática.

En toda la exposición se va a considerar el conjunto de palabras de un texto donde ya se ha eliminado una serie de palabras consideradas vacías de significado. Esto introduce una cierta arbitrariedad, ya que en distintas ocasiones o por distintas personas se pueden considerar ciertas palabras como vacías o no. Así pues, los valores de algunos parámetros pueden resultar distintos a los obtenidos en otra ocasión. Hasta cierto punto esto se convierte en una ventaja, ya que se trata de determinar los hechos con independencia del valor concreto que tome el parámetro.

No se va a considerar el hecho cronológico del autor escribiendo su texto. Los textos grandes que manejamos deben considerarse obra del autor o autores y del documentalista que los ha reunido. Cuando hablamos del crecimiento del vocabulario, no nos referimos a cómo el autor utiliza cada vez más palabras distintas, en el transcurso temporal de la creación de su obra. Aunque este es el punto de vista en muchas presentaciones elementales del tema (Parrondo, 2003) donde se representa el crecimiento del vocabulario al tomar sucesivos capítulos del Quijote, por ejemplo.

Hablaremos de la relación entre vocabulario y tamaño del texto en el siguiente sentido:

Reunidos un conjunto de textos del mismo autor o de distintos autores pero con características que se consideran parecidas, queda formado un texto grande, representativo del total de la literatura similar que podría haberse reunido. La colección, o población por decirlo en términos estadísticos, de los fragmentos de tamaño P , es decir de P palabras en total, después de haber quitado las palabras vacías, son todos los posibles fragmentos de ese tamaño que podrían encontrarse.

Cuando extraemos una muestra de N fragmentos de tamaño P , la consideramos una muestra aleatoria del total de la población de todos los fragmentos posibles de tamaño P , dentro de las características de autor o tipo de lenguaje que nos hemos marcado.

Utilizaremos estadísticos medidos sobre la muestra como estimadores de los valores reales en la población. Por ejemplo el valor para el vocabulario $V(P)$ lo estimaremos por la media muestral \bar{V} de los valores del vocabulario, contados efectivamente, sobre cada uno de los elementos de la muestra. Otro estadístico medido sobre la muestra como estimador de los valores reales de la población es la varianza muestral S^2

De esta forma, entenderemos por crecimiento del vocabulario la función $V = V(P)$ sobre el total de la población, que estimaremos a través de muestras con las garantías estadísticas que se indican mas abajo.

5.1.2. Estimación del valor del vocabulario V para un texto de tamaño P palabras

Naturalmente, los valores de V obtenidos por recuento en distintos ejemplos de fragmentos de texto de tamaño P , difieren entre sí: ya que las observaciones muestran que V es una variable aleatoria y como primera medida vamos a verificar la hipótesis de que su distribución es normal (Dunning, 1993).

En diversos estudios científicos se observan ciertas variables y aspectos de un tema cualquiera objeto de análisis dándose conclusiones muy a la ligera sin tener antes en cuenta si los valores objeto de estudio siguen una distribución normal o no. En el artículo de Dunning (1993) se advierte que en cualquier estudio estadístico en el que apliquemos fórmulas deducidas para una distribución normal que tenga un mínimo de rigor científico antes debe necesariamente comprobar que los datos que se van a evaluar siguen una distribución normal, ya que en el caso contrario estaríamos hablando de resultados sin un rigor ni calidad científica

Para verificar si sigue una distribución normal $V = V(P)$, cuando P es fijo, para el mismo tamaño de palabras, los valores siguen una distribución normal y siguiendo las recomendaciones de Dunning (1993) tomamos una muestra de 30 fragmentos de 4000 palabras cada uno del autor Joaquín Costa²⁰ y varias de sus obras sobre discursos políticos y contamos el vocabulario presente en cada uno (columna V). El promedio es 2478.167 y la varianza muestral 27843.32

Aplicamos el test de Kolmogorov-Smirnov en la variante de Lilliefors (Ugarte y Militino, 2002) que nos dará como resultado si la muestra observada sigue una distribución normal.

Si V sigue una distribución normal, entonces las diferencias entre la función de distribución empírica y la real para una muestra suficientemente grande no serán significativas.

La tabla siguiente contiene los resultados y su descripción somera es la siguiente:

En la columna V están los datos observados y en la siguiente su normalización según la media y la desviación típica calculadas. Es decir la columna V_{nor} representa el

²⁰ Véase Apéndice I

resultado de $\frac{(v - \bar{v})}{\sqrt{s}}$. La siguiente columna contiene la distribución normal acumulada

hasta esos valores y las dos siguientes la distribución empírica de la muestra, tal como exige el Test. En relación a la tabla de Ugarte y Militino (2002) se han omitido dos columnas, situando directamente en la última columna el valor correspondiente al máximo de las dos diferencias. El máximo de la última columna es 0.1088. Como valor crítico se toma el recomendado para muestras de más de 30 elementos, según Ugarte y Militino (2002), que en este caso es 0.1593

Así pues, los datos numéricos indican que el valor observado es menor que el crítico, por un amplio margen. Por lo tanto en lo sucesivo admitiremos que los valores del vocabulario en una población de documentos de tamaño fijo se ajustan a una distribución normal.

Si realizamos la misma operación con otra tipología distinta de texto como puede ser un texto de tipo científico el valor crítico es 0,1592 y el valor observado es 0,0968 con lo cual podemos afirmar igualmente que los datos observados en el texto científico sigue una distribución normal.

V	Vnor	FDN(0,1)	Ni/N	Ni-1/N	Dn
2150	-1,966682	2,327272E-02	3,333334E-02	0	2,327272E-02
2152	-1,954696	2,397334E-02	6,666667E-02	3,333334E-02	4,269333E-02
2232	-1,475261	6,869128E-02	0,1	6,666667E-02	3,130872E-02
2255	-1,337424	8,918193E-02	0,1333333	0,1	0,0441514
2257	-1,325438	9,115488E-02	0,1666667	0,1333333	7,551178E-02
2257	-1,325438	9,115488E-02	0,2	0,1666667	0,1088451
2366	-0,672208	0,2492846	0,2333333	0,2	4,928463E-02
2378	-0,6002927	0,2727383	0,2666667	0,2333333	0,039405
2417	-0,3665683	0,3556477	0,3	0,2666667	8,898105E-02
2439	-0,2347238	0,4059501	0,3333333	0,3	0,1059501
2448	-0,1807874	0,4270317	0,3666667	0,3333333	9,369836E-02
2454	-0,1448297	0,4412045	0,4	0,3666667	7,453785E-02
2468	-6,092866E-02	0,47453	0,4333333	0,4	7,453001E-02
2477	-6,992245E-03	0,4960582	0,4666667	0,4333333	6,272484E-02
2483	2,896536E-02	0,5100197	0,5	0,4666667	4,335305E-02
2514	0,2147463	0,5835724	0,5333334	0,5	8,357245E-02
2528	0,2986474	0,6159903	0,5666667	0,5333334	8,265701E-02
2529	0,3046404	0,6182776	0,6	0,5666667	5,161094E-02
2529	0,3046404	0,6182776	0,6333333	0,6	1,827761E-02
2530	0,3106333	0,6205607	0,6666667	0,6333333	4,610596E-02
2548	0,4185061	0,6608601	0,7	0,6666667	3,913994E-02
2548	0,4185061	0,6608601	0,7333333	0,7	7,247327E-02
2578	0,5982941	0,7238975	0,7666667	0,7333333	4,276921E-02
2594	0,6941811	0,7549663	0,8	0,7666667	4,503374E-02
2657	1,071736	0,8566785	0,8333333	0,8	5,667854E-02
2671	1,155637	0,8747086	0,8666667	0,8333333	4,137526E-02
2682	1,221559	0,8877031	0,9	0,8666667	2,103639E-02
2693	1,287482	0,8996915	0,9333333	0,9	3,364187E-02

V	Vnor	FDN(0,1)	Ni/N	Ni-1/N	Dn
2746	1,605107	0,9444588	0,9666666	0,9333333	2,220788E-02
2765	1,718973	0,955798	1	0,9666666	4,420203E-02

Tabla 6. Distribución normal del vocabulario en una población de documentos de tamaño fijo

Concretamente los cálculos presentados en esta tabla corresponden al texto del autor Joaquín Costa y varias de sus obras sobre discursos políticos.

5.1.3. Intervalo de confianza para el promedio de valores del vocabulario en la población de fragmentos de tamaño P

Obtenido el valor promedio del vocabulario sobre la muestra se considera que es una estimación del promedio sobre toda la población como se ha mencionado anteriormente. Se considera que es una estimación del promedio sobre toda la población no sólo de la población de fragmentos extraídos del texto del que disponemos, sino de toda la literatura con parecidas características, siempre que nuestra selección haya sido adecuada. Es decir, el proceso de extracción de una muestra se ha hecho en dos fases: primero hemos constituido nuestro texto y después hemos extraído N fragmentos de tamaño P.

Conocidos los valores del promedio \bar{V} y la varianzas muestrales S^2 se determina el intervalo de confianza al nivel 95% por ejemplo, donde se encontrará el verdadero promedio de la población (Ugarte y Militino, 2002, p. 213). Se obtienen siempre valores muy razonables.

Por ejemplo para una muestra de tan solo 5 fragmentos del texto larra3, de 4000 palabras cada uno, se obtuvo:

$$\begin{aligned} \text{Media de la muestra } \bar{V} &= 2499 \\ \text{Varianza muestral } S^2 &= 16556 \end{aligned}$$

Utilizando el caso más desfavorable con varianzas de la población desconocida se obtiene el intervalo de confianza 2499 ± 194 es decir (2305,2693)

5.1.4. Variabilidad en el vocabulario

Cuando un autor escribe varios tipos de párrafos provoca una variabilidad en el vocabulario. En unos casos utilizará gran cantidad de palabras distintas de modo que el valor del vocabulario se acercará al número total de palabras, mientras que en otros por conveniencia de la exposición de las ideas repetirá abundantemente algunas palabras o incluso frases. Cuando analizamos una muestra de párrafos esta variabilidad viene expresada por la varianzas muestrales, como estimador de la varianzas de la población.

En los ejemplos tratados se ha calculado la varianzas muestral y obtenido un intervalo de confianza para la varianzas de la población. Como ya se ha determinado que la población es normal, se ha seguido el esquema de Ugarte y Militino (2002, p. 228), los resultados muestran variaciones amplias.

Por ejemplo, para una muestra de 50 fragmentos de 2000 palabras extraída de un texto sobre discursos políticos del autor Joaquín Costa denominado para este análisis con el nombre de costa9 se ha obtenido el intervalo de confianza al 95% para la varianza de la población como (7112,16438). Expresado en desviación típica es (84,128). Hay que notar que la muestra es pequeña; de hecho, el cálculo de intervalos de confianza en media para tener un error menor que el 1% de P, es decir 20, recomienda tomar muestras de, al menos 96 elementos.

Cuando se toman fragmentos mucho más grandes la varianza se reduce relativamente. Es normal que el autor haya utilizado varios de sus recursos estilísticos y se haya compensado la variabilidad acercándose a una tendencia promedio, que en algún caso podrá encontrarse como característica del autor.

5.1.5. Relación entre los parámetros básicos del vocabulario y el tamaño del fragmento

Visto que en tamaños pequeños hay mas variabilidad que en tamaños grandes, para estudiar la variación de cualquier magnitud en relación a P, tomaremos mas datos en valores de P pequeños. Una forma de conseguirlo es tomarlos a intervalos regulares en $\log(P)$. Para unificar las tablas y gráficos que se van a elaborar a continuación, decidimos tomar los valores de P cuyos logaritmos valen aproximadamente 3, 3.5, 4,..., n. Seleccionamos el valor de P, dónde su logaritmo sigue un crecimiento pequeño y constante. De este modo, los gráficos construidos automáticamente con las herramientas habituales quedan en escala logarítmica. Los valores adecuados son:

P	Log P	Eje abscisas
20	2,9957	1
33	3,4965	2
55	4,0073	3
90	4,4998	4
148	4,9972	5
245	5,5012	6
403	5,9989	7
665	6,4997	8
1097	7,0003	9
1808	7,4999	10
2981	8,0000	11
4915	8,5000	12
8103	8,9999	13
13360	9,5000	14
22026	9,9999	15
36315	10,4999	16
59879	11,0000	17

Tabla 7. Muestra del tamaño de los textos en escala logarítmica

En todos los gráficos que se expondrán a continuación el eje de abscisas corresponde a los valores de la columna logaritmo del tamaño (Log P) reflejado en la tabla 7.

Según la ley de (Heaps, 1978) el vocabulario es una función potencial del número de palabras. Como estamos estudiando la relación entre el vocabulario y el tamaño del texto vamos a utilizar con carácter general y de manera aproximada la fórmula de la

Ley de Heaps donde manifiesta que el vocabulario va a ser una función potencial del número de palabras

$$V = a \cdot P^b$$

Lo que tomando logaritmos se convierte en una función lineal

$$\text{Log}(V) = \log(a) + b \cdot \log(P)$$

Y si construimos gráficos para $\log(V)$ en los valores de P indicados anteriormente, es decir si utilizamos la escala log-log se pondrá de manifiesto la función lineal porque los puntos dibujaran una línea recta.

Para el cálculo de la función lineal o potencial que mejor se ajusta, en el sentido de los mínimos cuadrados a la secuencia de los datos que se analizan, en este caso del texto empleado Costa9 se ha utilizado la aplicación²¹ desarrollada para obtener los datos que se representan en los gráficos que se muestran a continuación.

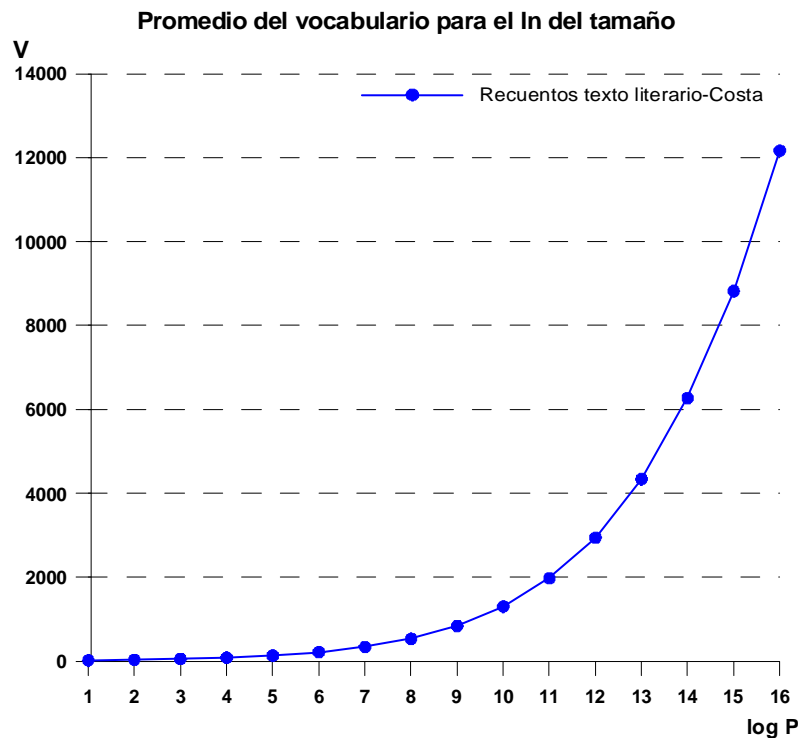


Gráfico 3. Promedio del vocabulario para el logaritmo del tamaño

Vamos a comprobar en lo sucesivo que es así en una primera aproximación, posteriormente se verá que no es totalmente cierto si se analiza con mas detalle. El siguiente gráfico corresponde a valores promedio del vocabulario en muestras de 50 fragmentos para cada uno de los valores de P mencionados. Por tanto es una representación de V frente a logaritmo de P . Esta distribución indica la similitud con la de Heaps sólo que al utilizar logaritmos muestra una apariencia distinta.

²¹ Aplicación RENOS. Véase apéndice II.

Ya que si tomamos logaritmos se obtiene una función lineal y no una función potencial como es el caso de arriba, este hecho se pondrá de manifiesto si se representa en la siguiente gráfica $\log(V)$ frente a $\log(P)$. La tendencia que sigue tiene que asemejarse a una función lineal y realmente se aprecia que casi es una recta, pero no completamente. En este ejemplo los números corresponden al texto del mismo autor Costa9.

5.2. Deducción del modelo de representación del vocabulario respecto al tamaño del documento

5.2.1. Relación entre los parámetros básicos del vocabulario y el tamaño del fragmento

Si representamos en la siguiente gráfica $\log(V)$ frente a $\log(P)$ se advierte como se asemeja a una función lineal, aunque si se observa con mas detalle se puede apreciar a primera vista el tramo final más curvado hacia la derecha y hacia abajo.

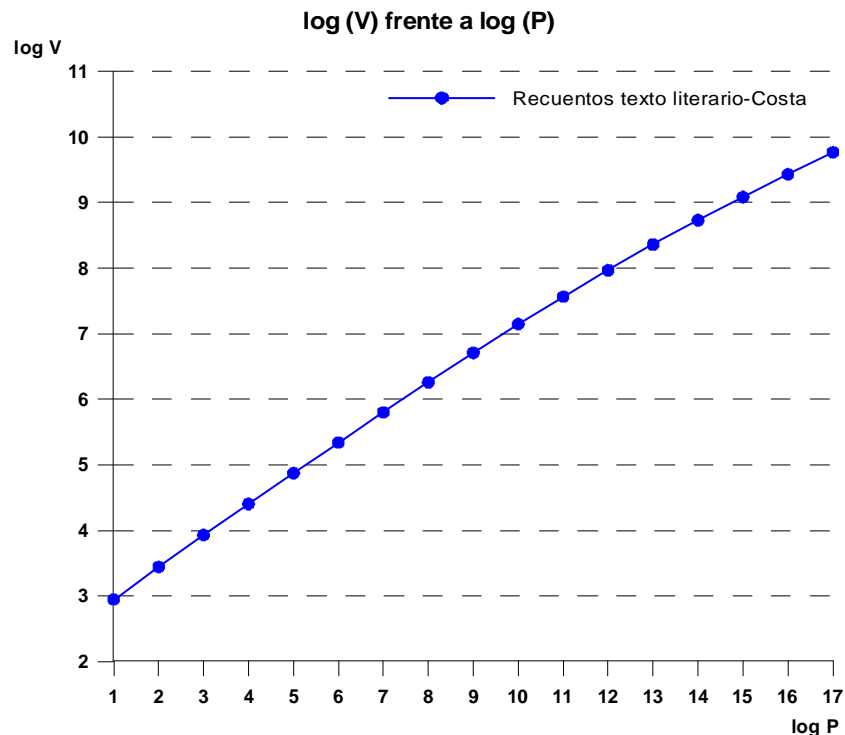


Gráfico 4. Crecimiento del vocabulario en escala logarítmica

En el siguiente gráfico se ha superpuesto la línea recta que mejor se ajusta para comprobar que existe una tenue diferencia y que la distribución queda algo curvada hacia la derecha y abajo. No tendría importancia si fuera de carácter aleatorio, pero en todos los experimentos revisados la tendencia es la misma: la función $\log(V)-\log(P)$ queda algo curvada hacia la derecha y abajo, es decir, para valores grandes de P crece mas lentamente el vocabulario.

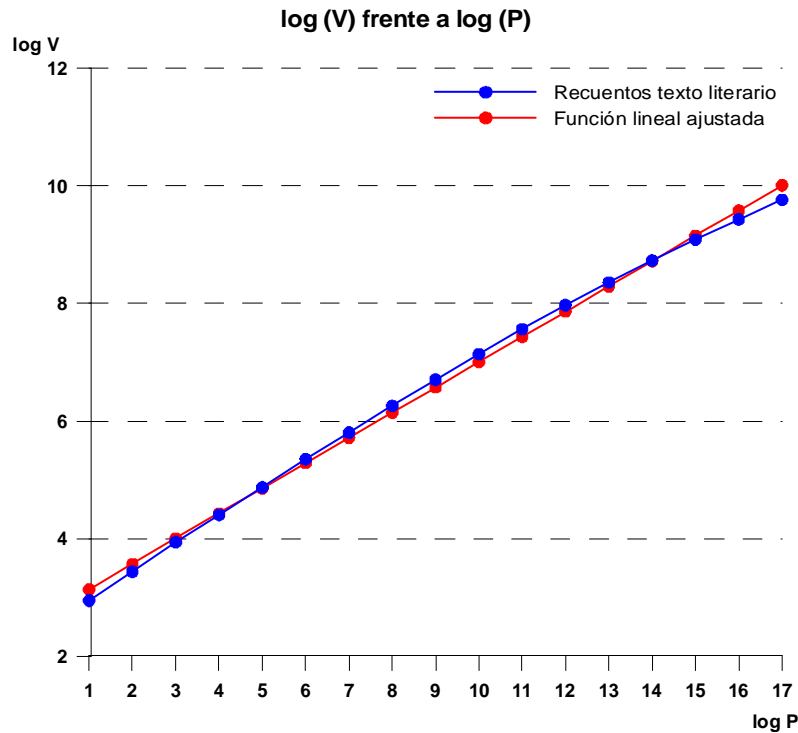


Gráfico 5. Crecimiento del vocabulario en escala logarítmica ajustada a una función lineal

Una revisión minuciosa de varias de estas gráficas correspondientes a distintos ejemplos parece revelar que la línea obtenida se compone de dos tramos, cada uno de ellos mucho más parecido a una recta que la línea total. Así pues, como una primera hipótesis de trabajo, consecuente con el comentario que se hizo mas arriba sobre la varianza, consideramos la curva $\log(V)$ - $\log(P)$ como la unión de dos rectas en distintos tramos. Para verificarlo ajustamos una recta a la mitad de los datos y otra recta distinta a la segunda mitad. Vemos que el error relativo cuadrático promedio en estos últimos casos es del orden de E-06, cuando la recta ajustada al total de los datos tiene un error del orden de E-04.

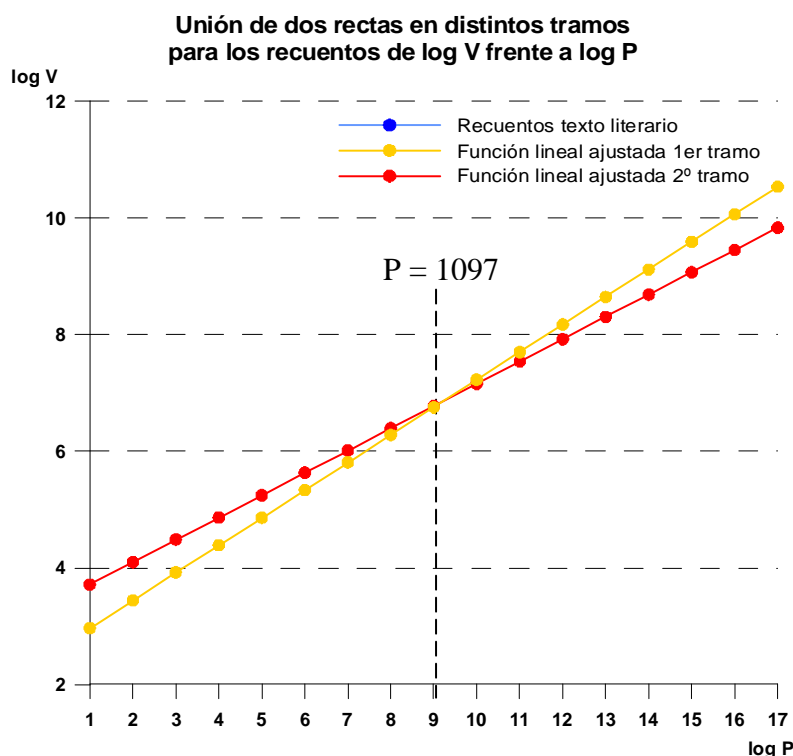


Gráfico 6. Crecimiento del vocabulario en escala logarítmica ajustada a dos funciones lineales

Con varias observaciones y cálculos hemos puesto de manifiesto que la fórmula de Heaps con una función potencial se ajusta bastante a la relación $V(P)$ pero tiene algunas imperfecciones que podrían subsanarse con el uso de dos, o incluso tres funciones potenciales en distintos tramos. En tamaños más grandes de P , se observa una disminución del vocabulario, esto es lógico debido a que el autor al escribir utiliza menos vocabulario nuevo y repite palabras conforme aumenta el tamaño del texto.

Para describir este fenómeno del incumplimiento de la Ley de Heaps, ajustamos dos líneas rectas, la línea coloreada en amarillo del gráfico vemos que queda más ajustada a los valores bajos y la línea coloreada en rojo queda más ajustada a los valores altos

Así pues, vamos a tomar la representación $V = V(P)$ como yuxtaposición de dos, o incluso tres funciones potenciales en distintos tramos, como puede apreciarse en la gráfica. Correspondiendo el punto de variación a un valor aproximado de $P = 1097$.

Utilizando la función potencial de Heaps como primera medida para observar la relación entre el vocabulario y el tamaño del fragmento podemos observar que en fragmentos pequeños la distribución que mejor se ajusta al crecimiento real que sigue el vocabulario sería mediante la utilización de dos, o incluso tres funciones potenciales. El crecimiento del vocabulario quedaría representado con más exactitud con dos funciones potenciales, pero aún así no es perfecto, podría mejorarse utilizando tres funciones potenciales, aunque la mejora es muy pequeña. Y para más precisión podría utilizarse cuatro funciones potenciales con lo cual mejoraremos el ajuste de manera casi imperceptible. Esta conclusión será expuesta más detalladamente en apartados posteriores de este mismo capítulo.

5.2.2. Otros problemas consecuentes a la representación $V=V(P)$ como función potencial

Vamos a realizar más comprobaciones para determinar el incumplimiento de la Ley de Heaps, es decir que el crecimiento del vocabulario no queda representado con total exactitud $V = V(P)$ por una función potencial y para ello vamos a utilizar la derivada, como expresión de la pendiente. Se pretende insistir en métodos alternativos para corroborar ese incumplimiento de la Ley de Heaps. Para ello, utilizamos el texto del autor Benito Pérez Galdós²² denominado *epis3* ajustando a los recuentos efectuados para el vocabulario una función potencial, obtenemos:

$$V = 1,7041 \cdot P^{0,8679}$$

Calculando la derivada de esta función veremos lo que crece el vocabulario en función del tamaño, por consiguiente se obtiene

$$V' = 1,478988 \cdot P^{-0,1321}$$

Por otra parte, utilizando la expresión elemental para derivación numérica, aunque sujetos a irregularidades si la muestra es pequeña, se obtienen como resultado de recuentos, es decir sobre el texto real y no sobre un modelo representado por la función potencial, unos valores parecidos pero no iguales.

La derivada numérica se calcula como sigue: Estimando un incremento de P , relativo a su tamaño, por ejemplo el 5%, al que llamamos I , se toman muestras de fragmentos de $P+I$ palabras y de $P-I$. Se cuentan sus vocabularios y se calculan los promedios:

$$\frac{V(P+I) - V(P-I)}{2 \cdot I}$$

Lo que corresponde a la diferencia central, así llamada en Cálculo Numérico. Así de este modo, en el gráfico se distinguen los datos reales obtenidos de un fragmento de un texto literario de Benito Pérez Galdós frente a la derivada de la función potencial, la cual se obtiene por derivación simbólica. En el gráfico siguiente se pueden comparar.

²² Véase Apéndice I

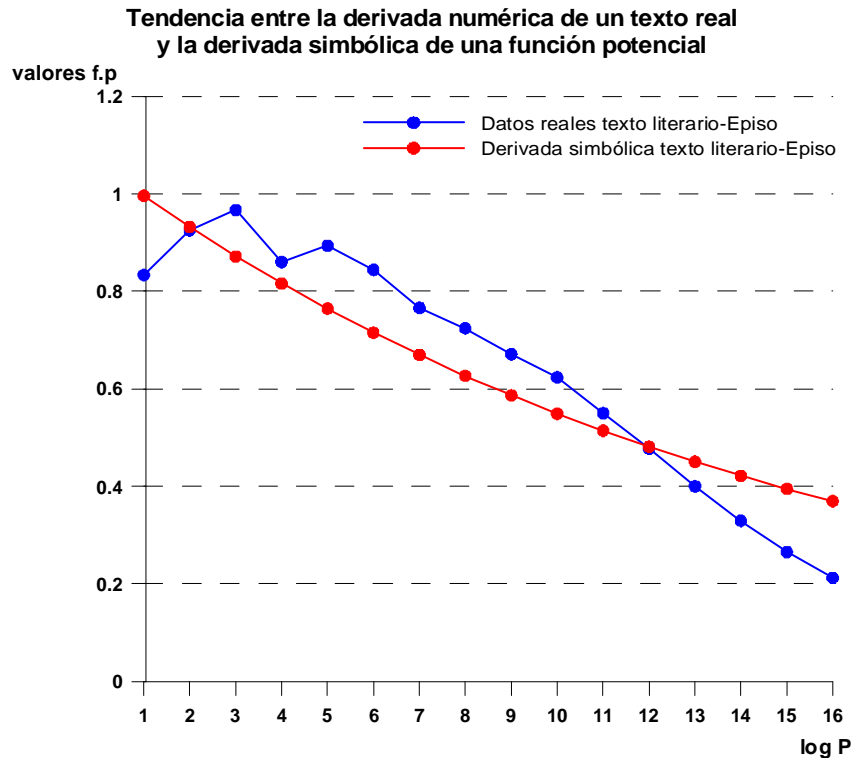


Gráfico 7. Derivada numérica y derivada simbólica de un texto

En definitiva de este gráfico se deben sacar dos conclusiones interesantes, en primer lugar la existencia en el texto real de dos tramos. A pesar de las irregularidades debidas al cálculo numérico de la derivada puede separarse un tramo inicial en el que los valores de la derivada son más o menos constantes, acotados entre 0,8 y 1; y después un segundo tramo de disminución constante (en escala logarítmica). En segundo lugar mirando solo el segundo tramo se concluye que la representación potencial es incompatible con el análisis del crecimiento. Es cierto que la fórmula potencial ofrece unos valores que son aproximadamente el valor del vocabulario para cada valor de P. Pero su derivada no representa adecuadamente la tasa de incremento real del vocabulario, en cada punto. Sobre todo en la parte final existe siempre una gran diferencia en la predicción entre la derivada simbólica y los datos reales, lo cual indica que no se acerca a la realidad. La derivada simbólica siempre queda por arriba de los datos reales, con lo cual predice valores mayores. El que la derivada simbólica siempre quede por encima de los datos reales significa que predice valores mayores.

5.2.3. Otras evidencias a favor de la existencia de dos tramos de características distintas

Visto que las diferencias numéricas son muy tenues y para corroborar la existencia de dos tramos de características distintas en la representación como función potencial del crecimiento del vocabulario frente al tamaño del fragmento se van a aportar otras pruebas procedentes de otros datos.

La causa de la existencia de dos o más tramos es la variabilidad del texto directamente relacionado con la varianza muestral de la variable aleatoria de V. Sus valores son crecientes en relación a P y por tanto también a V. Pero si relativizamos la varianza

muestral según V : exactamente tomando la desviación típica dividida por V , obtenemos unos valores como los siguientes correspondientes al texto *episo3*.

Dentro de la irregularidad de los valores puede observarse un primer tramo bastante estable, alrededor de algo menos de 0,05

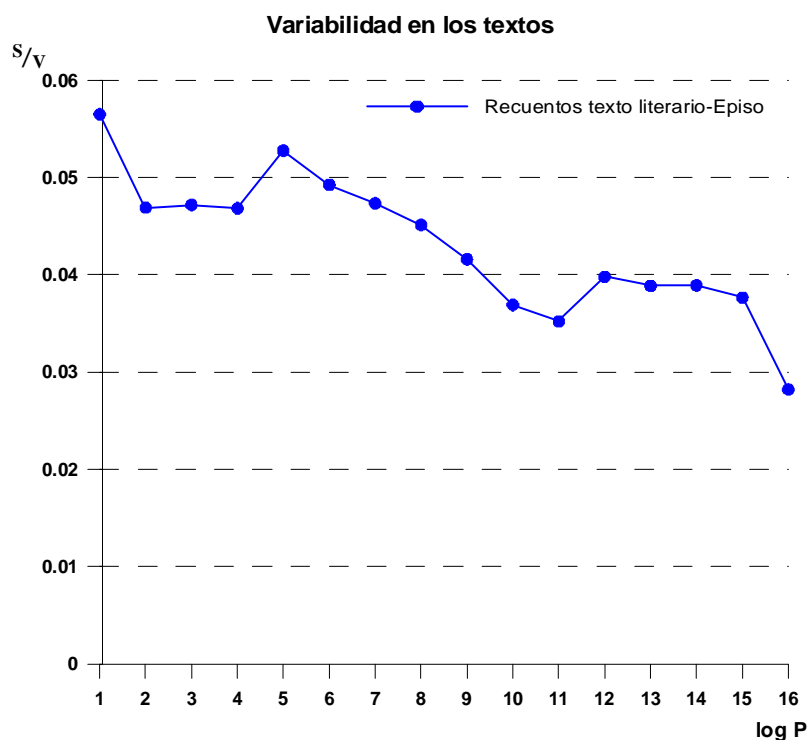


Gráfico 8. Variabilidad en los textos

Vistos los datos podemos preguntarnos ¿Porqué en tamaños pequeños de texto hay más variabilidad del vocabulario? Por que es lógico pensar que el autor al comienzo del texto utiliza o no alguno de sus recursos estilísticos y cuando el texto alcanza un tamaño mayor entonces se repite más el estilo y se estabiliza el crecimiento del vocabulario.

Vamos a estudiar también otro aspecto igualmente relacionado con la variabilidad del vocabulario. Para ello contamos la frecuencia de la palabra más frecuente, en función del vocabulario, cada una tomada con los tamaños de P indicados anteriormente para conseguir la representación logarítmica. Lo mismo con la siguiente palabra en frecuencia y con la tercera. Obviamente una vez ya se han quitado las palabras vacías.

La gráfica siguiente, muestra que este nuevo parámetro utilizado obtiene una gran regularidad, en principio adecuada para modelizarla por una función potencial.

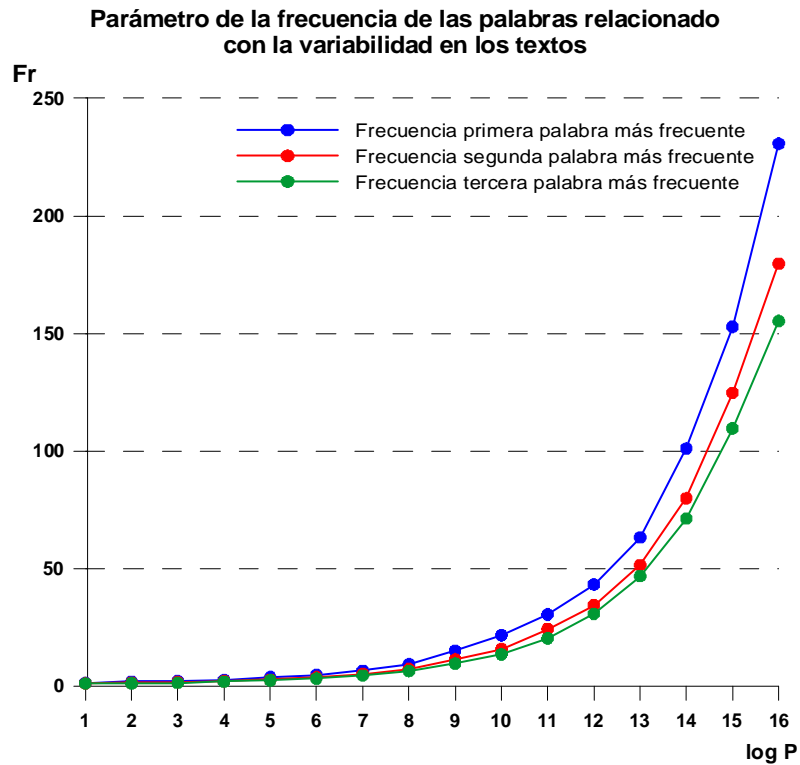


Gráfico 9. Frecuencia de las palabras y la variabilidad en los textos

Para confirmar que se puede modelizar a una función potencial, pasamos a la representación log-log y vemos su parecido con una línea recta; pero con más detalle, reconocemos dos tramos distintos cada uno de ellos mucho más parecido a una recta que el total

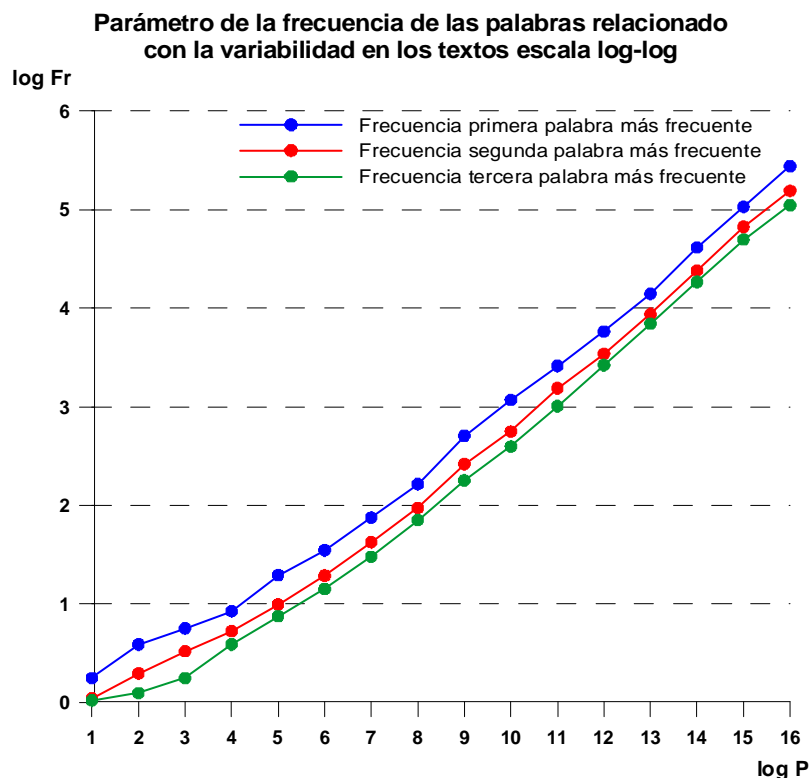


Gráfico 10. Frecuencia de las palabras y la variabilidad en los textos en escala logarítmica

La siguiente gráfica muestra los cocientes f_1/f_2 , f_1/f_3 , f_2/f_3 entre las anteriores magnitudes, y en ella se observa una cierta constancia. ¿Podemos afirmar que f_1/f_2 se mantiene alrededor de 1,3 independientemente del tamaño del fragmento?

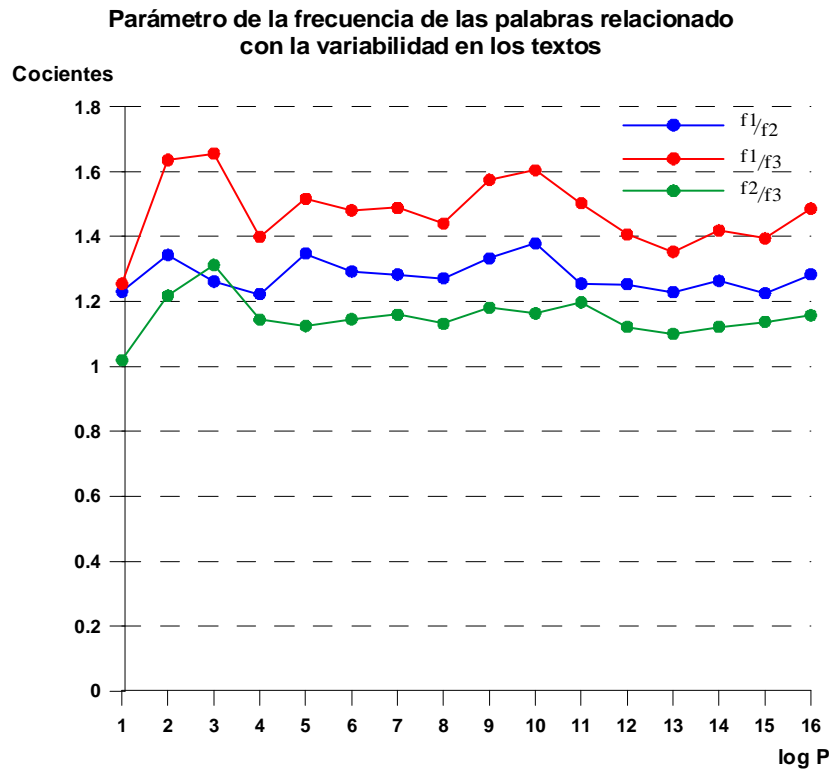


Gráfico 11. Cociente de las frecuencias de las palabras y la variabilidad en los textos

La siguiente gráfica tomada de valores reales es sorprendente. Visto que la frecuencia de la palabra más frecuente aumenta con el tamaño del fragmento, tomamos valores relativos dividiéndola por el vocabulario. Los recuentos muestran un mínimo en las gráficas. La dibujada aquí corresponde al ejemplo episo3, pero en otros ejemplos se observa una situación parecida. Esto es incompatible con la representación del crecimiento del vocabulario como una función potencial, ya que el cociente de dos funciones potenciales, que es otra función potencial, nunca tiene un mínimo.

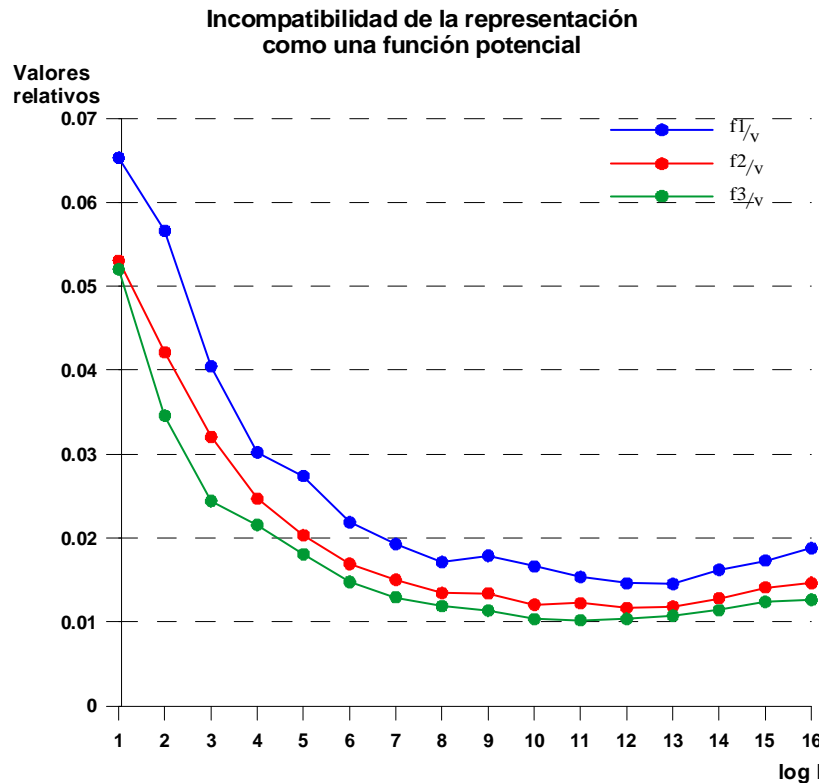


Gráfico 12. Incompatibilidad de la representación como una función potencial

5.2.4. Representación de $V=V(P)$ como dos funciones potenciales, una en cada tramo

Una vez obtenidos los resultados que indican claramente la distinción en dos tramos para la representación de $V=V(P)$ no factible por la representación de una función potencial, disponemos a representar la función $V=V(P)$ como dos funciones potenciales una en cada tramo, y para ello se emplea el texto *episo3*, utilizado hasta ahora, en el que determinamos el punto de subdivisión en aproximadamente $P=2500$, a la vista de algunas de las anteriores gráficas. Exponemos el modelo y ajustamos cada uno de los dos tramos, obteniendo

$$V = 1,2419 \cdot P^{0,9316} \quad \text{para } P < 2500$$

$$V = 5,9305 \cdot P^{0,7299} \quad \text{para } P > 2500$$

La función en la gráfica resulta inapreciable de la representación de los valores reales de V . De hecho, el error cuadrático relativo de cada uno de los dos ajustes es de 0,0004 y 0,0005. Dibujada en escala log-log resulta también indistinguible de los valores reales de $\log(V)$.

Obtenida por derivación simbólica la derivada se obtiene:

$$V' = 1,156954 \cdot P^{-0,0684} \quad \text{para } P < 2500$$

$$V' = 4,328672 \cdot P^{-0,2701} \quad \text{para } P > 2500$$

Cuya representación conjunta con los valores reales de V' se puede ver en la siguiente gráfica más ajustada que la similar vista anteriormente, es decir con la utilización de la doble función potencial la derivada simbólica con doble función potencial va más pareja a los datos reales con lo cual la predicción sería más acertada en este caso que en el anterior, aunque igualmente sigue prediciendo valores mayores la derivada en tamaños de textos más grandes.

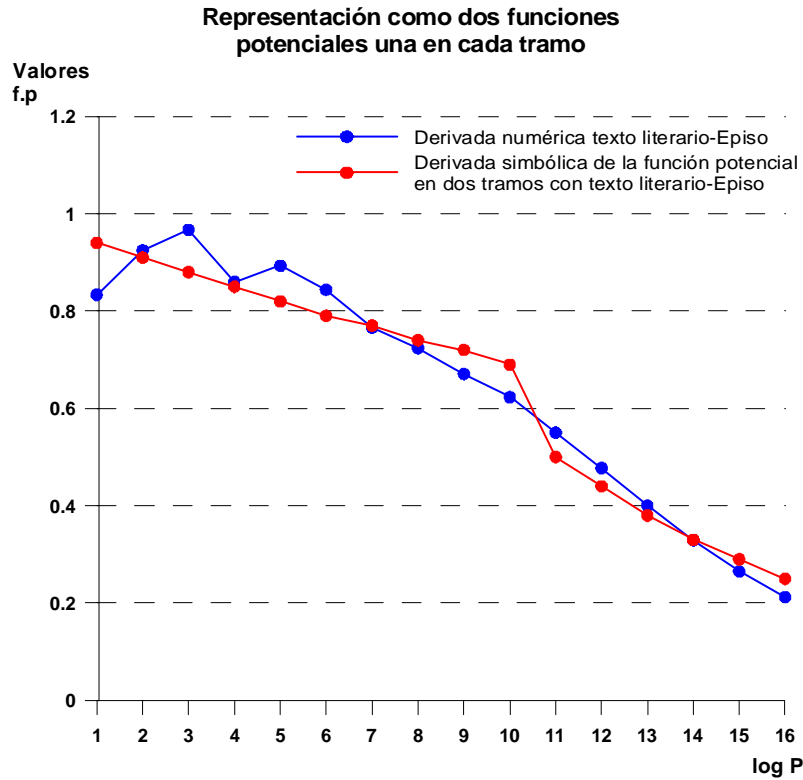


Gráfico 13. Representación como dos funciones potenciales

En este modelo de doble función potencial vamos a comprobar si el cociente de dos funciones potenciales, que es otra función potencial, no tiene mínimo, al igual que ocurría con una función potencial. Así, por último, dividimos las frecuencias reales, de la palabra mas frecuente, la segunda y la tercera, por el valor de V deducido del modelo con doble función potencial y obtenemos las gráficas con mínimo que serían imposibles en la representación con una sola función potencial.

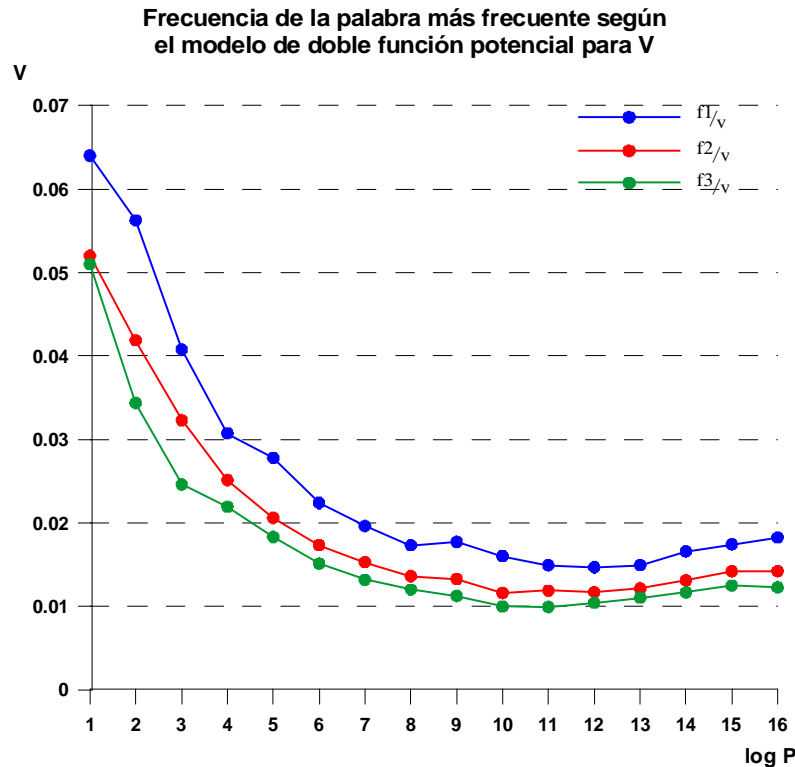


Gráfico 14. Representación como dos funciones potenciales de la palabra más frecuente

Esta gráfica es prácticamente igual a la anterior, esto confirma que el modelo con doble función potencial coincide con el crecimiento del vocabulario observado en textos reales²³. Así podemos concluir que el modelo con las dos funciones potenciales se ajusta bastante bien. Aunque podría mejorarse utilizando tres funciones potenciales, siendo esta mejora muy pequeña, incluso para más precisión podría utilizarse cuatro funciones potenciales con lo cual mejoraremos el ajuste de manera casi imperceptible.

5.2.5. Posibilidades de la representación como función potencial

Una vez establecido el modelo de la doble función potencial y considerando que estamos interesados en textos grandes vamos a centrar la discusión en la segunda de las funciones. Por ello en las siguientes páginas, las gráficas abarcan el tramo $P = 3.000$ en adelante, casi siempre hasta $P = 50.000$.

²³ Se ha buscado la posibilidad de mejorar aún más con la utilización de tres fórmulas distintas, pero no se aprecia mejora sustantiva, por esa razón se mantendrá las dos funciones potenciales, aunque en alguna ocasión, más adelante, se utilicen tres funciones potenciales.

Para las gráficas en escala logarítmica se tomarán estos valores que se detallan a continuación:

P	Log P	Eje abscisas
2981	8,00	1
3828	8,25	2
4915	8,50	3
6311	8,75	4
8103	8,9999	5
10405	9,25	6
13360	9,50	7
17154	9,7499	8
22026	9,9999	9
28286	10,2499	10
36315	10,4999	11
46630	10,7499	12
59874	10,9999	13

Tabla 8. Muestra del tamaño de los textos en escala logarítmica

En todos los ejemplos observados, con todo tipo de lenguajes y con varios tamaños, siempre a partir de 3.000 palabras (se han tratado ejemplos hasta 300.000 palabras aprox.), se puede conseguir un buen ajuste en el tramo considerado. Pero siempre la función potencial mejor ajustada es algo inferior a los datos reales en el centro del tramo y algo superior en los extremos. La gráfica siguiente es representativa y corresponde a fragmentos del texto sobre discursos políticos y otros tipos de escritos del autor Emilio Castelar denominado castela7 de tamaño 3.000 a 50.000 palabras

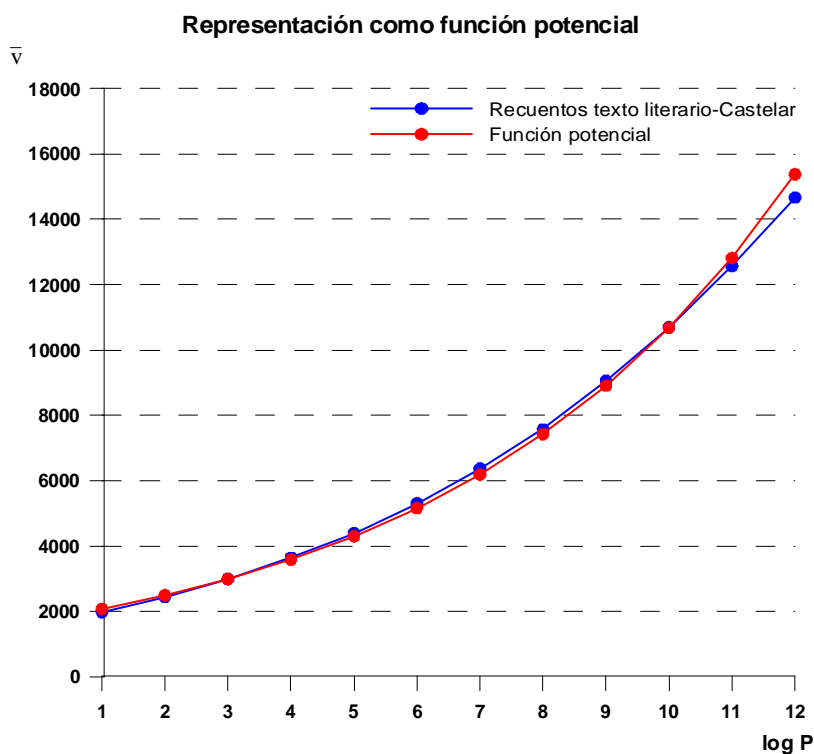


Gráfico 15. Representación como una función potencial

Este efecto nos indica que puede ser un buen modelo para interpolación, es decir para predecir el vocabulario de un fragmento de tamaño cualquiera, pero comprendido en el rango de tamaños sobre el que se ha llevado a cabo el ajuste. Sin embargo, resulta completamente inútil para extrapolación, es decir para predecir el vocabulario de un fragmento de mayor tamaño; ya que el comportamiento en el extremo de la derecha de la curva le lleva a predecir valores exageradamente grandes.

La misma conclusión se obtiene del estudio de la derivada. En todos los ejemplos revisados la derivada de la función potencial ajustada es superior para los valores mas altos del tramo, a la derivada real y por tanto, su prolongación hacia la derecha produciría valores excesivamente altos, es decir una extrapolación incorrecta.

Véase, a continuación, una muestra de las gráficas obtenidas de textos literarios y científicos, en las que la derivada simbólica extrapolaría siempre valores superiores al vocabulario real, ya que siempre queda por encima de la derivada de los recuentos en los datos reales o derivada numérica.

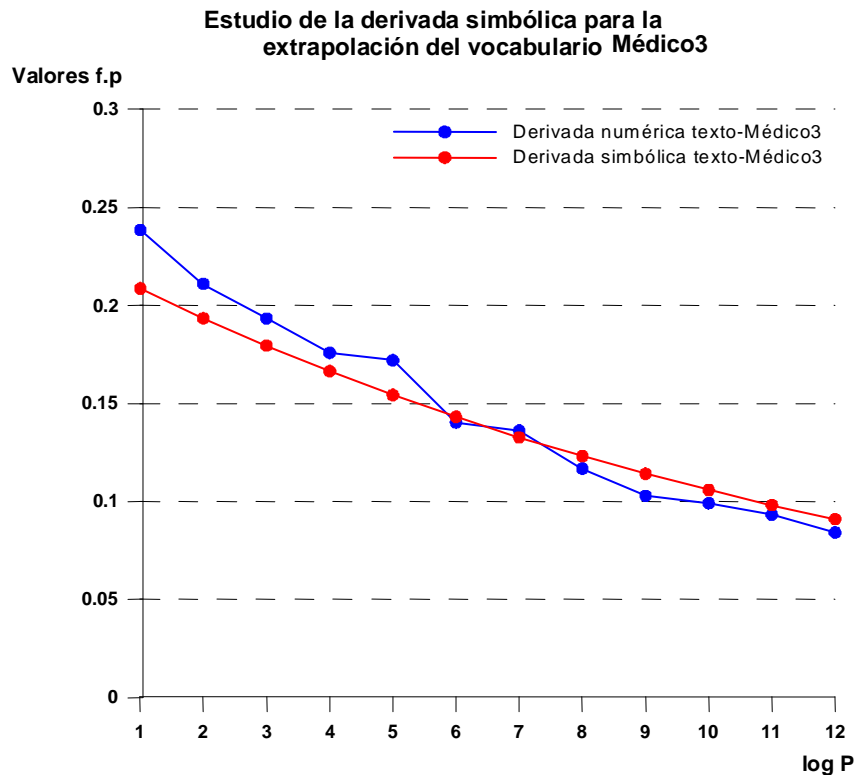


Gráfico 16. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico

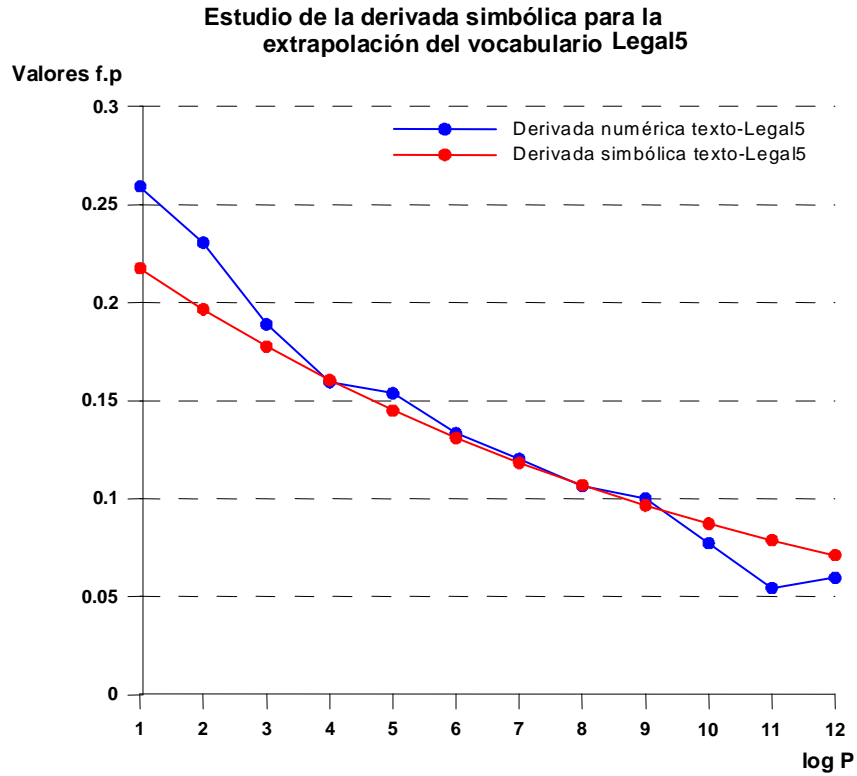


Gráfico 17. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico

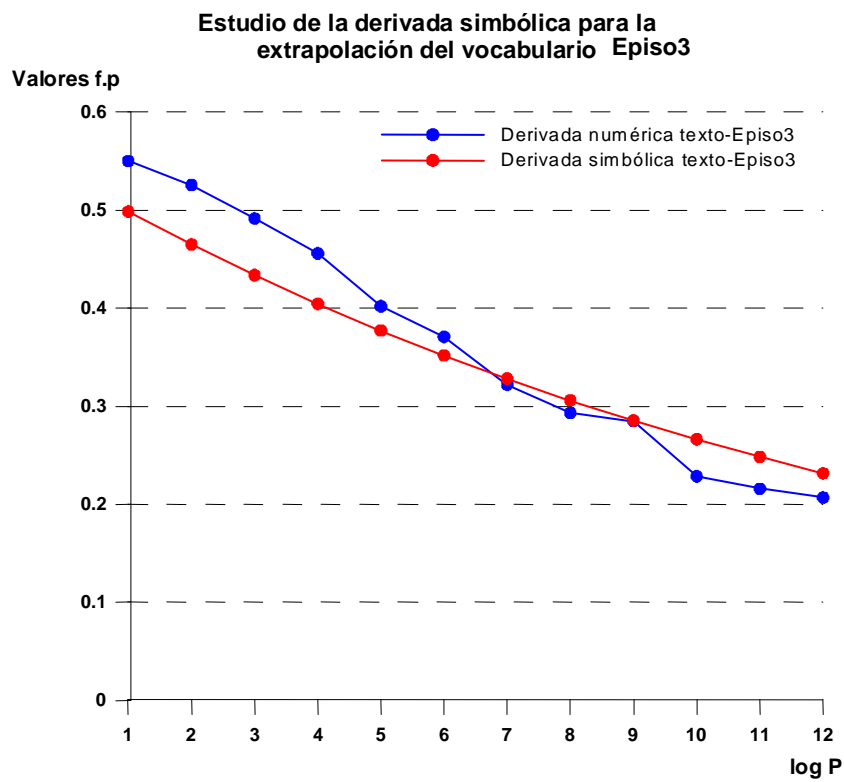


Gráfico 18. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario

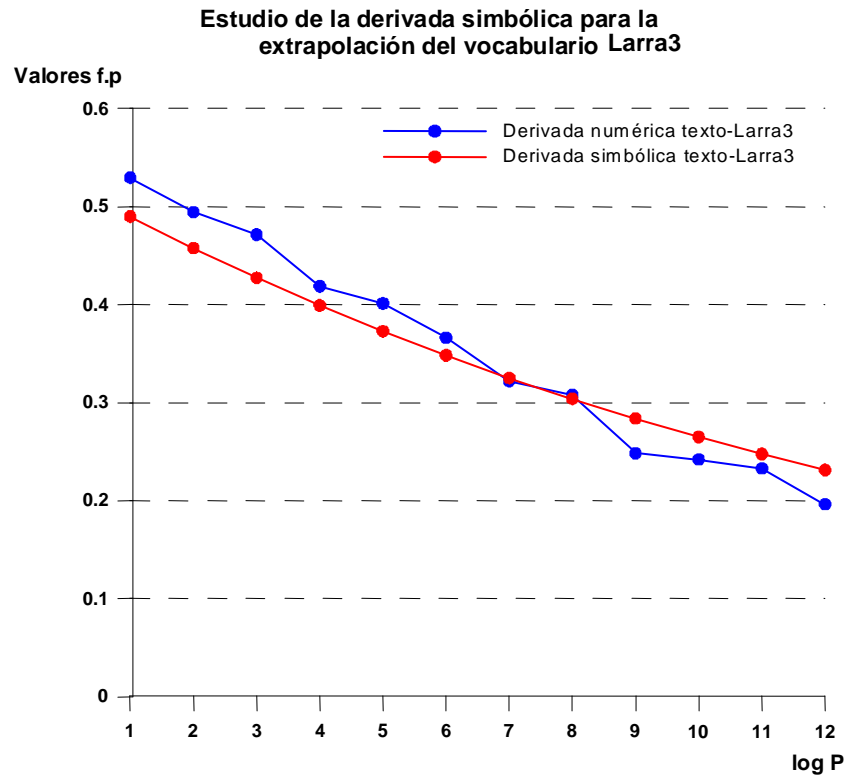


Gráfico 19. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario

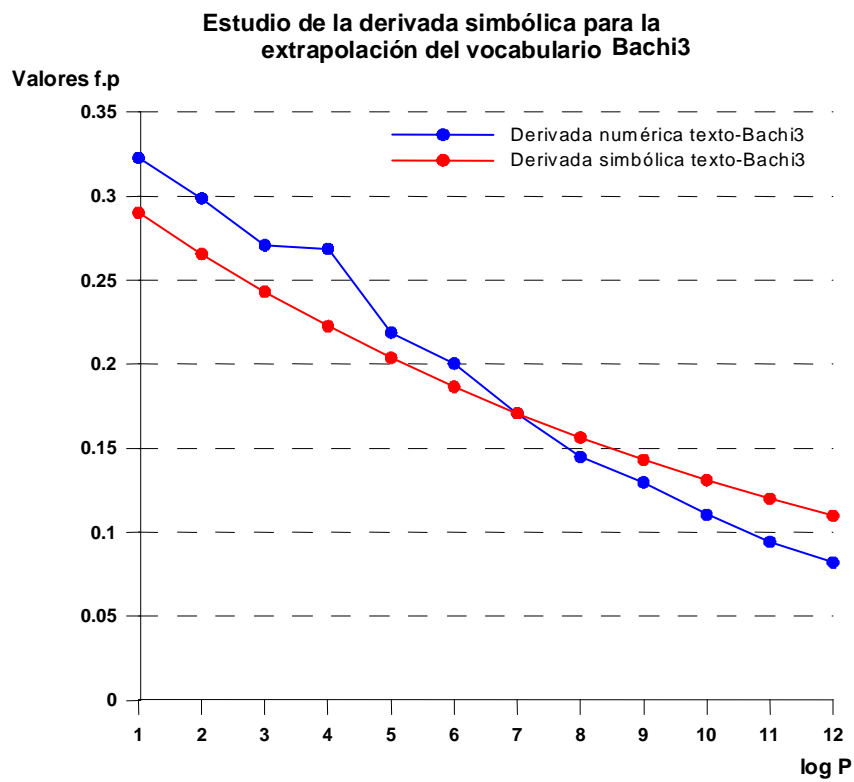


Gráfico 20. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico

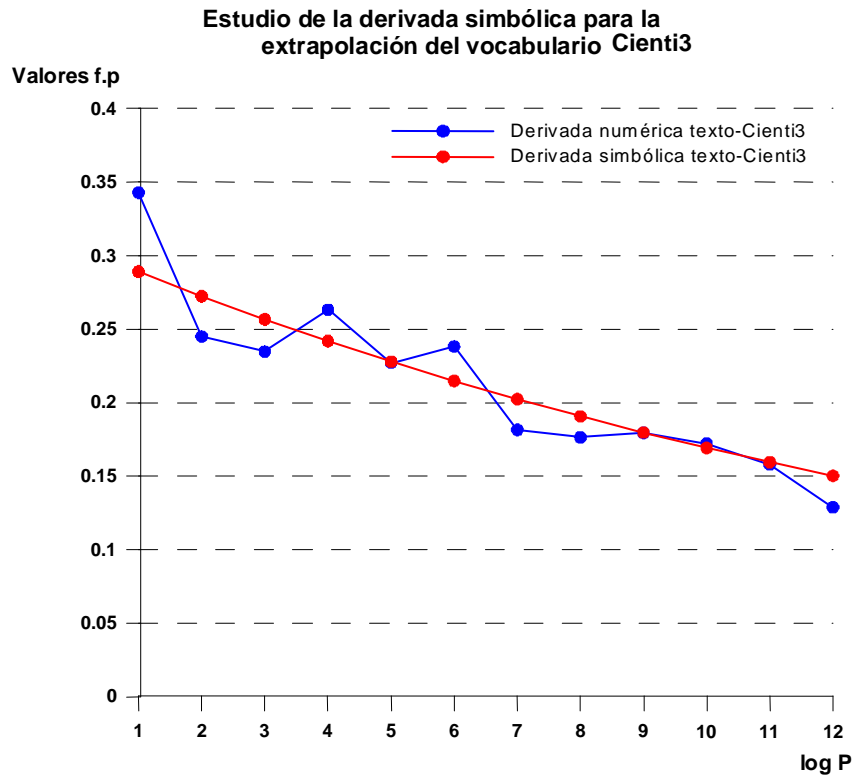


Gráfico 21. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico

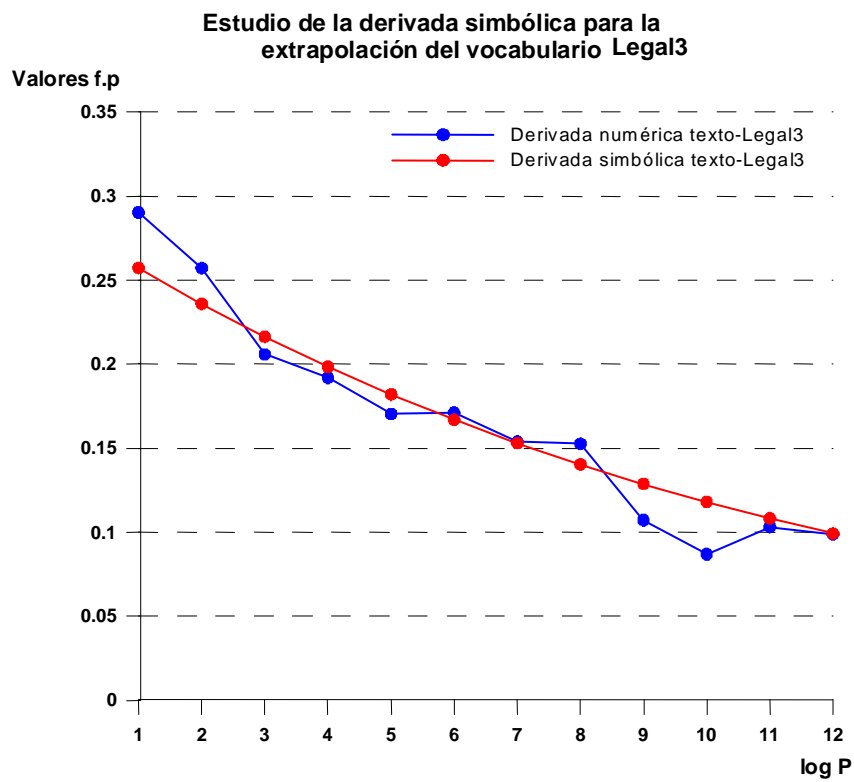


Gráfico 22. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto científico

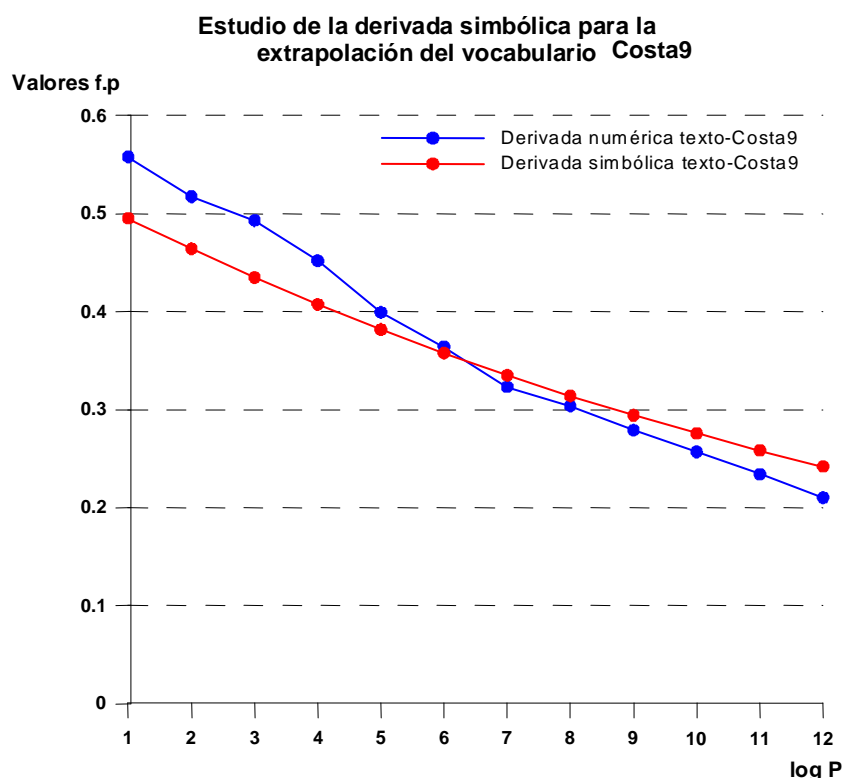


Gráfico 23. Derivada simbólica para la extrapolación del vocabulario ejemplo con texto literario

La utilización de las funciones potenciales ajustadas a los valores en un tramo predecirán valores excesivamente altos para un fragmento de tamaño mayor. Para ver la magnitud de esta desviación hemos ajustado una función potencial a los valores de fragmentos entre 3000 y 50.000 palabras de cada uno de los textos que se indican en la tabla. La penúltima columna es el valor real del vocabulario en el total del texto y la última es el valor predicho por la función potencial

Texto	Palabras	Vocabulario	Predicción
Medico3	58618	6926	7137
Legal5	55441	6006	6197
Episo3	59885	17003	17926
Larra3	56495	16375	17049
Bachi3	67801	8794	10087
Paten3	67263	11041	11626
Cienti3	85136	13198	14541
Costa5	79629	20978	22950
Legal3	88106	10631	10732
Costa9	92198	22751	25262
Costa7	179942	35849	42572
Castela7	149687	29307	35971
Costa3	224791	41621	54203
Castela5	275031	38211	55286

Tabla 9. Cantidad de palabras predichas por la función potencial en varios textos

Como conclusión provisional deducimos que la función potencial es una buena herramienta y sobre todo sencilla para interpolación, es decir, para predecir el vocabulario de un fragmento de tamaño comprendido en el rango de tamaños sobre el que se ha calculado el ajuste. Por lo que vamos a elaborar unas pautas para realizar ese ajuste con el mínimo esfuerzo.

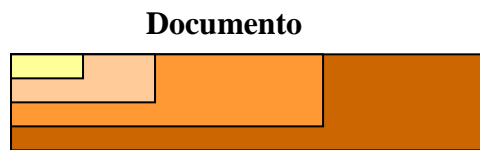
Por otro lado, visto que el modelo de la función potencial no es viable para extrapolación ya que se aleja mucho de los datos reales, vamos a trabajar con el segundo tramo, admitido el modelo de doble función potencial.

5.3. Estudio del Modelo y utilización práctica

En la fórmula de Heaps $V = A \cdot P^b$ cualquier intento de determinar experimentalmente cuáles serán los valores característicos de A y b para un determinado tipo de texto fracasa por la inestabilidad del valor obtenido para A , respecto a cuestiones muy accidentales. Es decir, se ha observado la gran disparidad del valor obtenido para A utilizando extractos de textos de distintos tamaños muy similares y obtenidos del mismo fichero de texto. Así de este modo, cabe preguntarse si realmente esta inestabilidad que presenta el valor de A es muestra del incumplimiento de la ley de Heaps o si por el contrario responde a características de tipo matemático en las formulaciones.

Para obtener la estabilidad de los parámetros tomaremos las medidas más objetivas y un mayor número de muestras de modo aleatorio y de distintas posiciones en el texto

*Enfoque Diacrónico
utilizado por Heaps*



El enfoque utilizado por Heaps y que muestra una inestabilidad mayor en los coeficientes de la fórmula se podría denominar enfoque Diacrónico ya que Heaps considera los fragmentos del texto como partes cada vez mayores del texto original. En cambio en capítulos posteriores de esta investigación se utiliza un enfoque Sincrónico para ajustar los coeficientes tomando un mayor número de muestras aleatorias y de distintas posiciones en el documento

*Enfoque Sincrónico
utilizado en esta investigación*



5.3.1. Estudio de las propiedades de la función potencial como modelo de $V=V(P)$

En este segundo apartado el objetivo que se propone llegados a este punto es deducir la doble función potencial con los menos cálculos posibles, así observamos que la función potencial como representación de $V = V(P) = a \cdot P^b$ tiene algunas limitaciones lógicas: en primer lugar, no es posible que V sea mayor que P . Dado un valor del coeficiente a , no es posible que el exponente sea tan grande que haga que el vocabulario sea mayor

que el número de palabras. En la siguiente gráfica se ha representado, para valores del coeficiente $a = 3, 4, \dots, n$ los valores del exponente que, para $P=50.000$ o para $P=3.000$ consiguen este efecto. Estaría representada en el gráfico por la franja azul y nos dice que los valores razonables del exponente, para un coeficiente dado, deben ser menores que este valor. (Basta considerar entonces el límite definido por el caso $P=3000$ que es inferior)

Como cota inferior, no tan drástica es difícil que un texto de 50.000 palabras tenga menos de 5.000 palabras distintas, de modo que los valores razonables del exponente estarán en la franja roja y amarilla dibujada en la gráfica en 3D.

Así pues, en un diagrama $a-b$, los valores de estos parámetros para la fórmula de Heaps, no pueden situarse en cualquier punto, sino sólo en la estrecha franja delimitada en la figura.

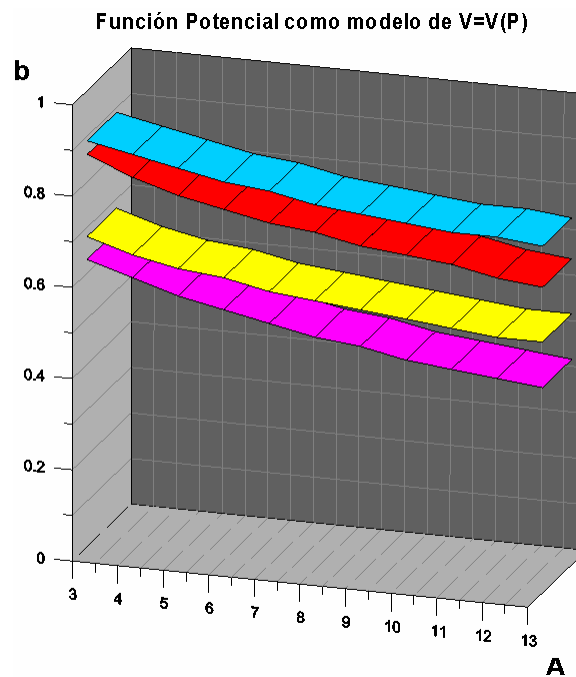


Gráfico 24. Función potencial como modelo de $V=V(P)$ en 3D

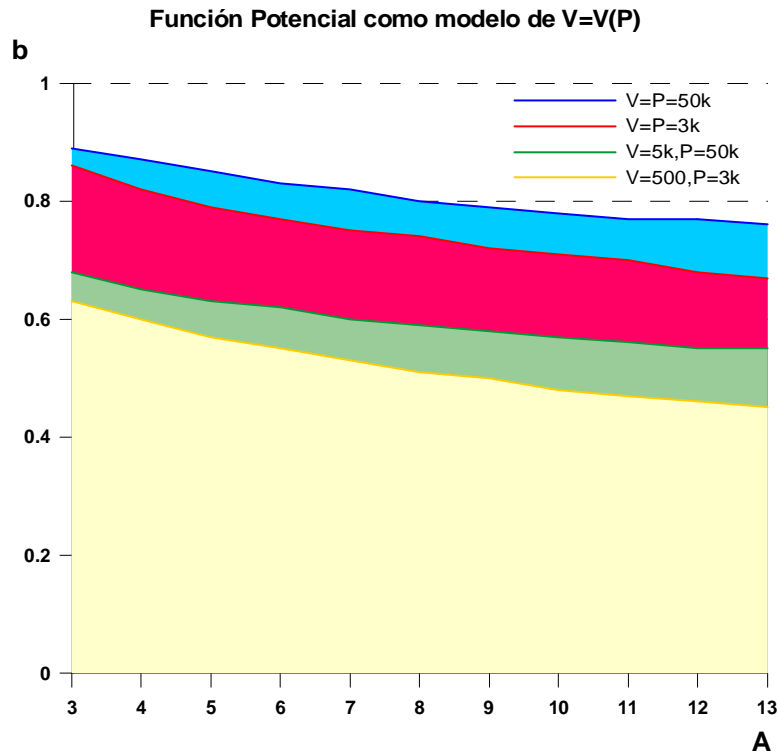


Gráfico 25. Función potencial como modelo de $V=V(P)$

Vista la forma estrecha y alargada de la zona de valores permitidos para a, b , tratamos de dar una interpretación de ello. Primero tomaremos dos parejas (a, b) y (a', b') situadas en una línea paralela a la dirección principal a la franja.

En la siguiente gráfica se presentan los valores de dos funciones potenciales, una con coeficiente y exponente iguales a $(7, 0.7)$, y la otra corresponde a $(5, 0.73)$, se ha obtenido ajustando a los valores de la primera una función potencial con coeficiente forzado a valer 5.

El error cuadrático relativo promedio entre ambas es 0.0009, lo que nos indica que tanto una como la otra valdría como modelo de la misma función $V = V(P)$, dado el orden de magnitud de la aproximación.

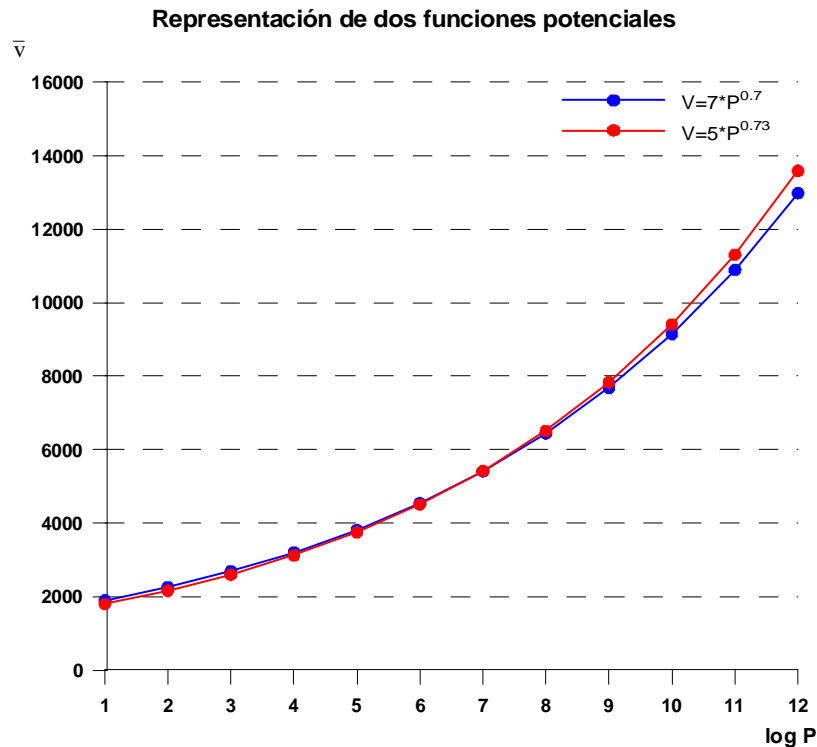


Gráfico 26. Representación de dos funciones potenciales

En la gráfica se observa claramente que los valores de estas dos funciones potenciales con diferente coeficiente, son prácticamente iguales.

De este modo, deducimos que en la función potencial sus parámetros, no son valores que caracterizan a un texto.

Prosiguiendo con este análisis se han obtenido todos los valores adecuados del exponente para función potencial de coeficiente 3, 4,..., n (observando que aún en los casos peores: 3 y 13 el error es de 0,002) obteniendo la línea central en el siguiente gráfico, donde la superior y la inferior son los límites lógicos ya obtenidos anteriormente.

Cada uno de los puntos de la línea azul es una función potencial distinta, pero todas pueden representar el crecimiento del vocabulario del mismo texto y se observa como cada uno de los valores de los parámetros (a,b) sigue una línea, representada entre unos límites en la que cada una de las funciones potenciales son muy parecidas entre sí, esta afirmación se podrá comprobar si representamos en un gráfico las derivadas de cada una de las funciones potenciales, como vemos en el gráfico 28.

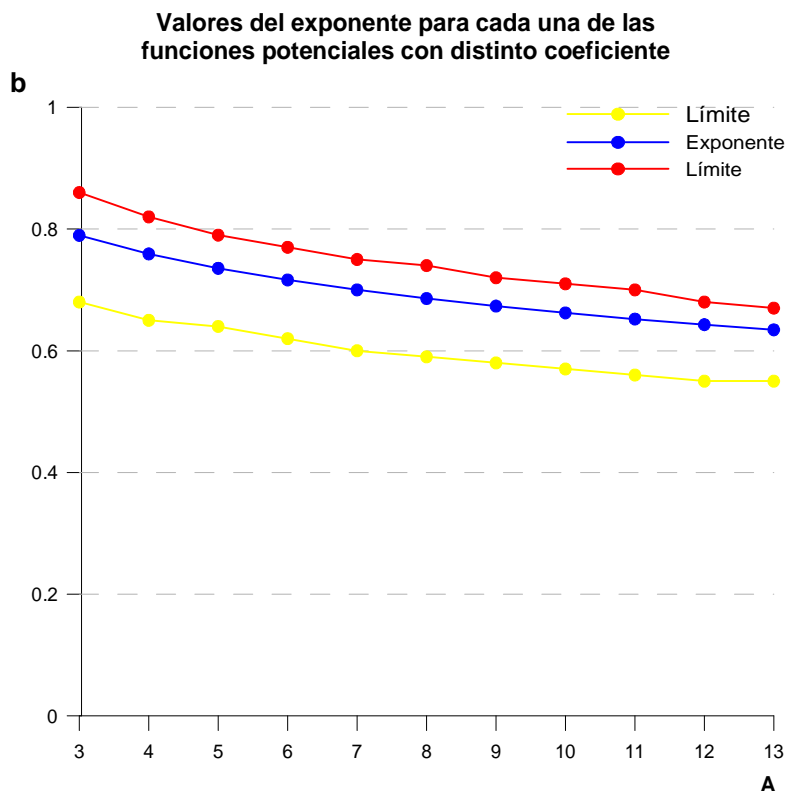


Gráfico 27. Representación de los valores del exponente para distintas funciones potenciales

Este gráfico podría reflejar claramente el argumento con el que hemos comenzado este apartado, en dicho gráfico se han obtenido todos los valores adecuados del exponente para una función potencial de coeficiente 3, 4,..., n, representado por la línea azul. Si la línea roja y amarilla reflejan los límites, esto significa que si decidimos coger un coeficiente 9 y un exponente 0.9 ya sabemos de antemano que esta posible función potencial sería imposible porque predice valores fuera del límite, es decir, estaríamos ante un texto con más vocabulario que palabras totales y esto es imposible.

Por otro lado, si decidimos coger un coeficiente 8 y un exponente 0.2, igualmente ya sabemos de antemano que esta posible función potencial sería también errónea ya que no es posible que un texto tenga tan poco vocabulario, a no ser que fuese un texto conteniendo una sola palabra repetida, caso inexistente.

Vista la similitud de todas las funciones potenciales cuyos parámetros quedan representados en el esquema $a-b$ por una línea paralela a los límites, nos planteamos si son idénticas a todos los efectos y entonces su determinación tiene un solo grado de libertad, o si hay cierta diferencia, aunque pequeña, entre ellas.

Para ello, de todas las funciones potenciales equivalentes a una dada calculamos su derivada simbólica y las representamos conjuntamente para ver en que detalle se diferencian.

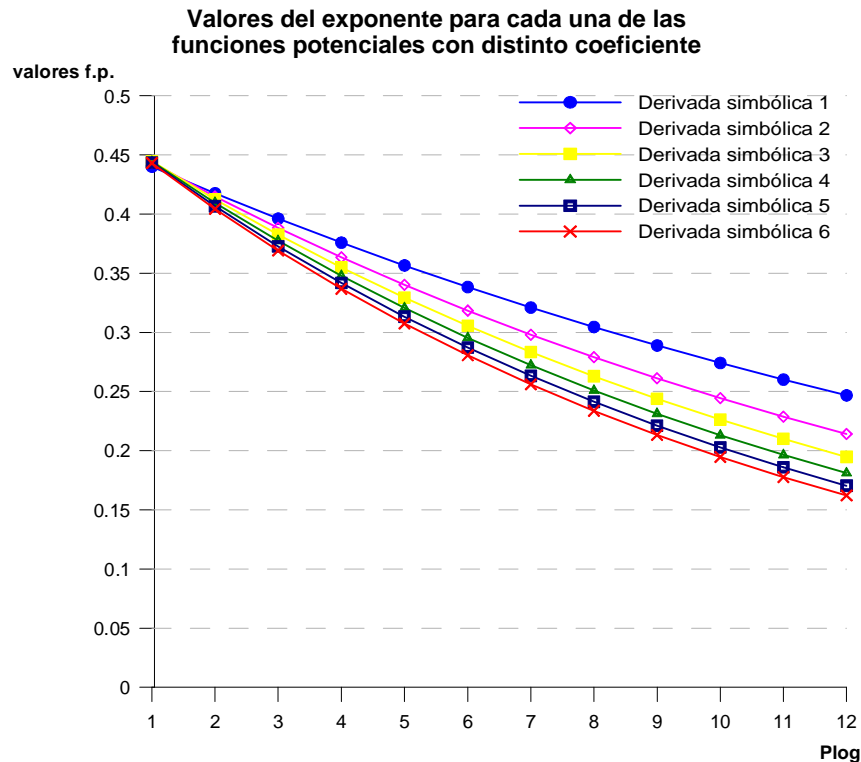


Gráfico 28. Representación de los valores del exponente para distintas funciones potenciales

Así, se observa que cada una tiene un ritmo distinto de disminución de la derivada. Quizá pueda aprovecharse esta libertad en la elección para conseguir algún otro efecto provechoso, por ejemplo corregir la diferencia entre la derivada de la función real $V=V(P)$ y el modelo potencial que se le ajusta.

Vemos claramente que al utilizar las derivadas simbólicas las funciones potenciales se diferencian muy poco. Concretamente el crecimiento que indica el modelo no difiere tanto cuando utilizamos valores pequeños de P , en cambio el crecimiento que indica el modelo de la función potencial utilizando las derivadas observamos que en valores $Plog$ más grandes difiere más.

Una de las aplicaciones²⁴ creadas para los cálculos incluye la determinación de una función potencial que tenga el mejor ajuste posible, tanto ella como su derivada, a los valores del vocabulario promedio en muestras de fragmentos de distinto tamaño del texto estudiado. Puede indicarse en dicha aplicación el tamaño deseado para las muestras. La derivada de la función potencial se obtiene por derivación simbólica y la de los valores muestrales por derivación numérica con diferencia central, para lo que se toman dos muestras suplementarias para cada valor de P , una para $P+I$ y otra para $P-I$. La obtención del mejor ajuste se realiza por minimización numérica.

²⁴ Véase Apéndice II

5.3.2. Interpretación de los coeficientes de la función potencial $V=V(P)$

Obtenidos por el procedimiento indicado anteriormente, (ajuste al crecimiento del vocabulario y a su derivada numérica observada), los valores de los parámetros a y b para el modelo $V=V(P)= a \cdot P^b$ para una serie de textos se comprueba que no siguen una distribución aleatoria, sino que hay ciertas pautas que convendría interpretar. Vemos que los distintos tipos de textos se agrupan con una tendencia diferente según su tipología.

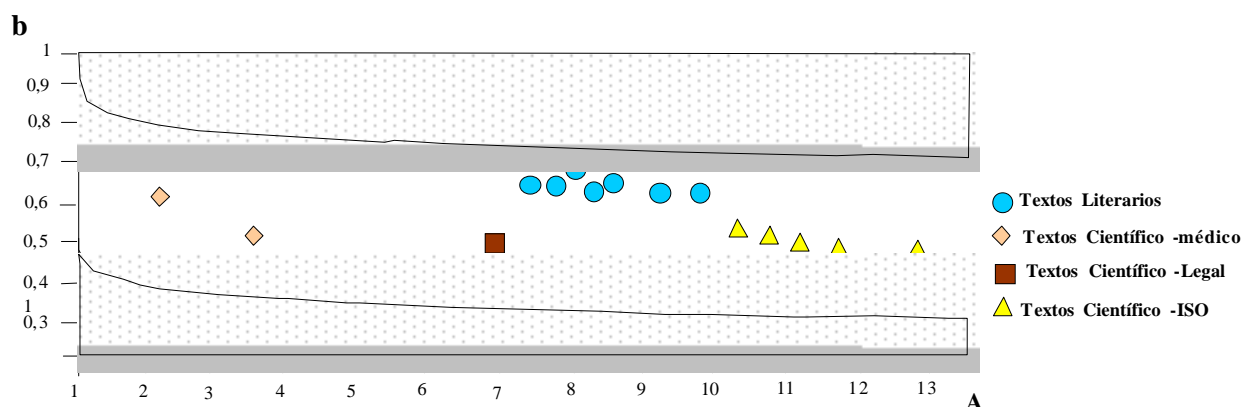


Figura 9. Interpretación de los valores de los parámetros a y b de la función potencial

A la vista del dibujo podemos distinguir varias categorías de textos:

- textos literarios, ensayos que se han redactado para componer unidades largas (Joaquín Costa, Emilio Castelar, Benito Pérez Galdós, Larra, etc.)
- ▲ textos compuestos por yuxtaposición de resúmenes cortos, extraídos de referencias bibliográficas tomando solo el campo resumen; tanto los seleccionados por algún elemento ajeno a la materia, como el apellido del autor, como los seleccionados por materias o palabras de significado amplio (turismo, administración, riesgo,...)
- ◆ textos formados por yuxtaposición de referencias bibliográficas incluyendo muchos campos, como título de la revista, lugar de trabajo, etc y de campos científicos especializados con su vocabulario propio.
- Textos de lenguaje muy formalizado, que por tanto, resultan pobres en cuanto a vocabulario. Se incluyen aquí los textos legales y los formados por resúmenes de patentes.

Representamos en la siguiente gráfica algunas de las funciones potenciales seleccionadas correspondientes a elementos característicos de las categorías:

La línea roja y azul corresponden a la categoría 1 y vemos que no hay diferencia apreciable entre ellos, a pesar de corresponder a valores de a, b (6,79 0,72) y (10,39 0,67). La línea amarilla es paten3 (9,15 0,64), la línea representada por el color marrón es legal3 (6,47 0,65) y la línea verde es legal5 (13,52 0,56)

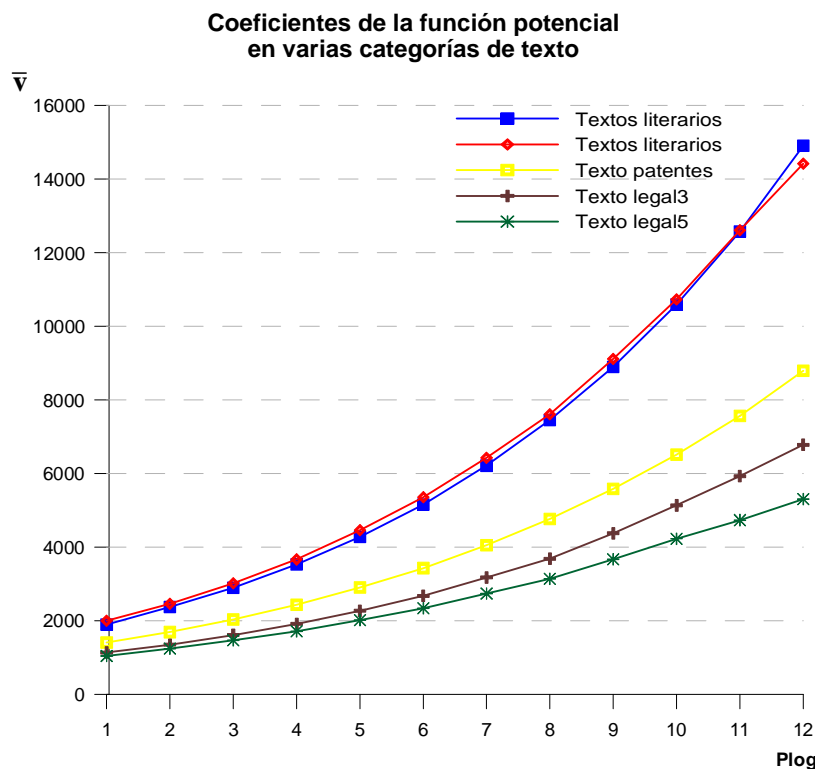


Gráfico 29. Coeficientes de la función potencial en varias categorías de texto

¿Puede establecerse una relación entre los valores de los parámetros y alguna característica visible de los textos?

De la comparación de estas cinco gráficas y del dibujo anterior se deduce que:

En la gráfica, las líneas roja y azul representan a los textos literarios, la línea amarilla representa a los textos de patentes y las líneas verde y marrón representan dos ejemplos de textos legales, aclarado esto se deduce que:

1. El exponente b es mayor cuanto mayor es el vocabulario para un tamaño fijo de los fragmentos. Por ejemplo en fragmentos de 3.000 palabras como indica la posición 1 en el eje de Plog en el gráfico. También se observa en el dibujo los círculos azules agrupados y en una posición muy alta.
2. Dentro de un mismo grupo de textos el coeficiente a , es mayor cuanto menor es el crecimiento del vocabulario, por ejemplo medido por la inclinación del último tramo de las gráficas, entre los puntos 11 y 12 (legal3, con un valor de 6,47 tiene mas inclinación que legal5 con un valor de 13,52). Esto puede observarse igualmente en el dibujo anterior, por ejemplo el grupo de textos legales (cuadros rojos) o en el gráfico (líneas marrón y verde), claramente se ve que cuanto mayor es el coeficiente el vocabulario crece más lentamente. Este efecto “inverso” en el significado del coeficiente a está ligado a la casi-equivalencia de varias funciones potenciales. De las varias funciones casi equivalentes, las que tienen valor de a más pequeño, tienen valor de b más grande y por ello contribuye a una mayor velocidad de crecimiento del vocabulario.

Estas dos características visibles de los textos pueden materializarse en las siguientes cantidades de VI , DI

VI = vocabulario de los fragmentos de 2981 palabras dividido por 1000 (por ejemplo 1,94 para larra3, o 1,048 para legal5)

DI = inversa de la derivada del crecimiento del vocabulario aproximadamente en $P = 50000$, que especificamos con precisión como 1 dividido por $(\sqrt{46630} - \sqrt{36315}) / (46630 - 36315)$

Ajustando, por separado para cada categoría de textos los valores conocidos de a , b a los de VI , DI se obtiene para textos de la primera categoría

$$A = 0,87 \cdot VI \cdot DI \qquad b = 0,357 \cdot VI$$

Para la segunda categoría

$$A = 0,98 \cdot VI \cdot DI \qquad b = 0,369 \cdot VI$$

En estas dos categorías la precisión de la fórmula dada es bastante alta, con un error cuadrático relativo promedio de alrededor de 0,001 en los ejemplos estudiados. Las otras dos categorías pueden contener textos de muy variables características y por ello la precisión alcanzada es menor, pero aún así, puede darse

Para la tercera categoría

$$A = 0,4 \cdot VI \cdot DI \qquad b = 0,6 \cdot VI$$

Y para la cuarta categoría

$$A = 0,67 \cdot VI \cdot DI \qquad b = 0,55 \cdot VI$$

El sentido de estas fórmulas aproximadas es:

Hasta ahora los valores de los parámetros a y b de la función potencial se han obtenido como hemos mencionado anteriormente consiguiendo un ajuste perfecto mediante la recopilación de varias muestras y obteniendo la derivación numérica con diferencia central, calculando el vocabulario, etc. Todo ello en cada caso. Este proceso realizado por las aplicaciones desarrolladas para tal fin llevaba horas de trabajo de ordenador. Por esta razón se plantea otro método menos laborioso e igual o similar en exactitud para tratar de adivinar el valor de los parámetros a y b prediciéndolo con una fórmula sencilla.

5.3.3. Determinación, a priori, de la distribución de los valores del vocabulario en fragmentos extraídos de un texto

En este apartado resumimos varios de los resultados ya obtenidos anteriormente, orientándolos a la situación práctica en presencia de un texto que abordamos por primera vez. Se intenta dar los pasos esenciales a seguir a la hora de abordar este estudio o similares.

Al hablar de fragmentos de tamaño P palabras nos referimos a la colección de todos los fragmentos de ese tamaño que se pueden extraer del texto y que, en muchas circunstancias, puede considerarse representativa de la posible colección de fragmentos de ese tamaño existentes en ese tipo de literatura.

El fragmento más grande que es posible considerar en este estudio es el propio texto en su totalidad. Para este caso o los cercanos a él, los resultados no son muy precisos y menos aún para textos más grandes, del mismo tipo de literatura. Ello requiere otro tipo de modelo, que se aborda mas adelante.

Distinguiremos varias situaciones según el conocimiento que se tenga o pueda tenerse, del texto

<i>S1</i>	Sólo conocemos el número de caracteres
<i>S2</i>	Conocemos el número de palabras total
<i>S3</i>	Cuando conocemos P y V del total y utilizamos la ley de Heaps para predecir el Vocabulario que va a tener un trozo más pequeño o fragmento

Tabla 10. Situaciones posibles para averiguar el vocabulario de un texto

Situación 1.

Sólo conocemos el tamaño del texto en bytes o caracteres y el tipo de contenido. Como primer paso hay que estimar el número de palabras. Podemos tomar para la mayoría de textos en español un valor de 14,5 bytes por palabra (esto no significa que las palabras tengan en promedio ese número de letras, sino que ya se ha tenido en cuenta el efecto de los espacios en blanco, signos de puntuación y palabras vacías más generales).

Si el texto tiene abundantes encabezamientos, líneas en blanco, etc. como en la composición de textos literarios, el valor puede aumentar hasta 17. Por el contrario, si el texto está formado por yuxtaposición apretada de notas escuetas, como resúmenes o referencias, puede disminuir hasta 12. Una vez estimado este valor nos encontramos en la situación 2

Situación 2.

Conocemos el número total de palabras del texto ya descontadas las palabras vacías más generales. Como primera medida hay que estimar el posible vocabulario correspondiente al texto total. Este es un problema de solución incierta: En efecto con

los valores habituales de la varianza²⁵ tenemos que considerar una muestra tan pequeña que tiene un solo elemento: el texto total. El intervalo de confianza se extiende a unos límites muy amplios.

En consecuencia sólo damos unas indicaciones globales que no son más que valores promediados entre los distintos textos que hemos tomado como ejemplo

Tamaño en palabras	Valor de coeficiente a	Valor de exponente b
Alrededor de 100.000 o menores	10	Literario 0,67
		Resúmenes 0,64
		Otros tipos 0,62
Alrededor de 300.000	20	Literario 0,60
		Resúmenes 0,59
		Otros 0,58
		Legal 0,56

Tabla 11. Indicaciones globales sobre el vocabulario de los textos

A la vista de la tabla, si tenemos un texto literario y su función potencial tiene un coeficiente 7 y un exponente 0,9 esa función potencial no es que sea incorrecta, sino que a la hora del ajuste será tan buena como la que se indica en la tabla con coeficiente 10 y exponente 0,67

La idea es tomar el valor de a mas adecuado al tamaño, no variarlo ya que poco se conseguiría con ello, y por otra parte escoger un valor adecuado de b , este sí modificándolo entre los límites propuestos, si tenemos algún criterio para ello. Por ejemplo, para un texto literario de 230.000 palabras, que estimamos un poco pobre en vocabulario (porque es monótono en su temática, por ejemplo) podemos tomar $a = 20$ y $b = 0,596$ con lo que estimamos su vocabulario en $20 \cdot 230000^{0,596} = 31377$ palabras.

Situación 3.

Ya hemos pasado el texto a analizar por una aplicación desarrollada para tal fin en el que se extrae el vocabulario y las palabras y hace recuentos de éstas, con lo que sabemos el número de palabras y el vocabulario exacto que tiene.

²⁵ Al tener una muestra tan exageradamente pequeña que sólo consta de un elemento, los resultados son estadísticamente inciertos.

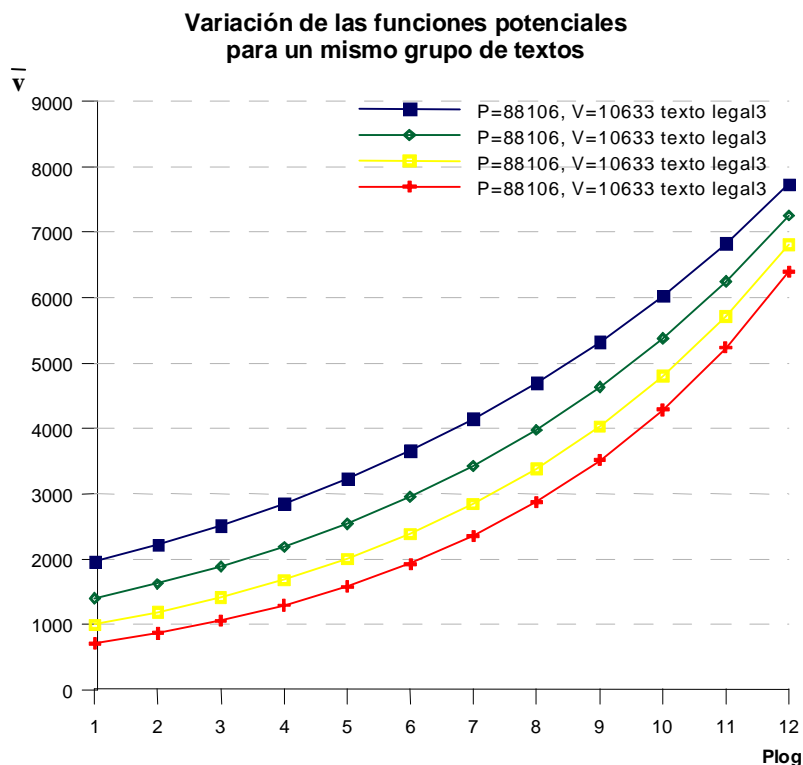


Gráfico 30. Variación de las funciones potenciales para un mismo texto

Lo que tratamos de predecir es el vocabulario de sus fragmentos de menor tamaño que el total. Como hay varias funciones potenciales distintas que predicen el mismo vocabulario total recurrimos aún a la tabla anterior de valores recomendados, pero ahora tomamos sólo la recomendación para el exponente b y calculamos el valor de a .

Por ejemplo, para el texto legal3 de 88.106 palabras y vocabulario = 10.633 tomaríamos el valor recomendado $b = 0,62$ y con el calculamos $10.633 = a \cdot 88.106^{0,62}$ de donde

$$a = \frac{10.633}{88.106^{0,62}} = 9,1359 \text{ y así tenemos la fórmula } V = 9,1359 \cdot P^{0,62}.$$

Compararemos con la fórmula obtenida con gran esfuerzo y precisión utilizando el programa RENOS que es $V = 6,2854 \cdot P^{0,6537}$. Para ver la magnitud de la diferencia entre ambas, calculamos unos cuantos valores:

Fragmento tamaño P	Vocabulario calculado por RENOS en muestras muy amplias	Vocabulario predicho por la fórmula recomendada
3500	1282	1439
4000	1409	1563
5000	1651	1795
5500	1746	1904
8000	2213	2402
10000	2608	2759
15000	3439	3547
25000	4723	4869
80.000	10.007	10.015

Tabla 12. Diferencia entre los resultados obtenidos por la fórmula recomendada y la aplicación

De aquí se concluye:

- a) Para tamaños del fragmento cercanos al texto total, todos los valores son parecidos, puesto que hemos forzado a que sea así, por medio del cálculo de a
- b) Para hacerse una idea del vocabulario de cada fragmento, los valores obtenidos con poco esfuerzo por el método recomendado son suficientes, teniendo en cuenta la variabilidad inherente al asunto.
- c) Quizá podríamos haber observado que el texto legal³ contiene un plan de estudios de bachiller con enumeración de los temas de cada una de las asignaturas. Esto le da una riqueza de vocabulario mayor que la usual en un texto legal. Teniendo esto en cuenta, podríamos haber decidido subir la recomendación a 0,64 y habríamos obtenido una predicción mucho más ajustada.
- d) Comprobamos que con cálculos sencillos obtenemos casi lo mismo que haciendo muchos recuentos con la aplicación desarrollada para tal fin.

Por tanto y a la vista de la tabla anterior, comprobamos que con cálculos sencillos obtenemos casi lo mismo que haciendo muchos recuentos con la aplicación desarrollada para tal fin.

5.3.4. Breve resumen de este capítulo.

En la 1ª parte: Vemos la existencia de dos tramos en la representación del crecimiento del vocabulario de un texto y utilizamos la doble función potencial, cogemos un segundo tramo y realizamos la interpolación, porque sabemos que falla la extrapolación.

En la 2ª parte: Sin hacer cálculos encontramos la función potencial que mejor sirve para la interpolación. Es decir para un fragmento del texto

En la 3ª parte (Estudios complementarios): ¿Qué hacemos para la extrapolación?, ¿Qué hacemos si nos encontramos un texto mucho mayor?

5.4. Estudios complementarios: Posibilidades de utilización del modelo para la extrapolación

Como hemos observado, la utilización del modelo para predecir el vocabulario de un fragmento de mayor tamaño de los comprendidos en el ajuste predice un crecimiento de la función potencial en el último tramo que lleva a predecir valores exageradamente grandes, por ello será necesario para la extrapolación modificar la función potencial para que no crezca tanto. Para ello utilizaremos la función potencial con logaritmo para intentar reducir dicho crecimiento.

Cogemos fragmentos hasta 60.000 palabras y se calcula la función potencial que mejor se ajusta $A \cdot P^b$ si utilizamos esta función potencial para ver el vocabulario comprobaremos como predice valores mayores que el vocabulario real y por ello cogemos el mismo ejemplo pero en este caso para menos palabras 36.315 (penúltima fila de la tabla).

Para conseguir una función potencial que consiga una buena extrapolación, es decir para conseguir acercarnos al tamaño del vocabulario con textos más grandes, entonces modificamos la función potencial para que esta no crezca tanto, cuando el valor de P es más alto y así la predicción del vocabulario sea más ajustada.

Para ello utilizaremos la función potencial con logaritmo para intentar reducir dicho crecimiento. Así pues, cogemos fragmentos de texto grandes hasta 60.000 palabras y calculamos la función potencial que mejor se ajusta $A \cdot P^b$, así comparamos el valor de esta función potencial para ver si la predicción ha sido acertada, buena o por el contrario sigue prediciendo valores más altos de los reales.

Tras probar el modelo para la extrapolación se observa que la función potencial que mejor se ajusta igualmente predice valores altos.

Entonces decidimos coger el mismo ejemplo pero con textos grandes y con menos palabras por ejemplo: 36315 (penúltima fila de la tabla) y se observa que este caso igualmente predice valores más altos que los valores reales del vocabulario pero incluso la predicción es mayor o es peor que la función potencial anteriormente citada

A continuación se detallan con precisión varias de las pruebas realizadas con distintas fórmulas de la función potencial que han sido modificadas por una función con logaritmo para intentar reducir el crecimiento, se detalla cual de estas fórmulas ofrece un resultado más positivo para la extrapolación y así por consiguiente una mejor aproximación a los datos reales con tamaños más grandes o cuando el valor de P es más alto.

P	Log P
20	2,9957
33	3,4965
55	4,0073
90	4,4998
148	4,9972
245	5,5012
403	5,9989
665	6,4997
1097	7,0003
1808	7,4999
2981	8,0000
4915	8,5000
8103	8,9999
13360	9,5000
22026	9,9999
36315	10,4999
59879	11,0000

$A \cdot P^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	5,495	0,7359				24.741
40	4,84	0,75			18.633	25.706
30	4,39	0,762		14.056	19.143	26.537
25	4,141	0,769	11.432	14.260	19.475	27.078

$A \cdot P^b \ln P$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	1,624	0,628				24.354
40	1,467	0,639			18.364	25.101
30	1.357	0.649		13.890	18.760	25.741
25	1,294	0,654	11.323	14.049	19.018	26.158

$A \cdot P^b (\ln P)^2$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	0,474	0,521				24.016
40	0,439	0,530			18.131	24.568
30	0,414	0,537		13.751	18.429	25.044
25	0,399	0,541	11.235	13.871	18.623	25.354

$A \cdot (\ln P)^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	0,00122	6,863				22.316
40	0,00135	6,815			16.969	22.021
30	0,00149	6,771		13.022	16.792	21.755
25	0,00161	6,735	10.735	12.922	16.640	21.527

$A \cdot \sqrt{P} (\ln P)^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	0,377	2,193				23.923
40	0,321	2,267			18.049	24.391
30	0,277	2,335		13.714	18.336	24.842
25	0,257	2,370	11.211	13.817	18.498	25.095

$A \cdot P^b (\ln P)^c$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60						24.006
40					18.116	24.533
30				13.770	18.472	25.135
25			11.225	13.851	18.580	25.270

$A \cdot (P \ln P)^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	2,452	0,664				24.475
40	2,160	0,676			18.452	25.297
30	1,966	0,684		13.942	18.879	25.989
25	1,864	0,689	11.351	14.107	19.145	26.417

$A \cdot (P(\ln P)^2)^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	1,249	0,606				24.291
40	1,111	0,615			18.315	24.985
30	1,017	0,621		13.855	18.678	25.568
25	0,964	0,626	11.299	14.002	18.913	25.945

$A \cdot (P(\ln P)^4)^b$	A	b	30	40	60	$P = 92.128$ $V = 22.751$
60	0,439	0,515				24.003
40	0,395	0,521			18.123	24.524
30	0,365	0,526		13.745	18.410	24.999
25	0,349	0,528	11.226	13.853	11.226	25.268

RESULTADOS:

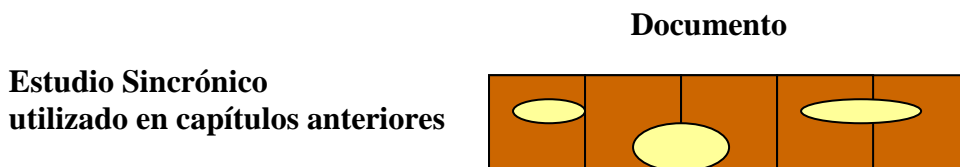
FORMULA	RESULTADOS
$A \cdot P^b$	La predicción sale alta
$A \cdot P^b \ln P$	Las predicciones igualmente son altas pero mejor que en el caso anterior
$A \cdot P^b (\ln P)^2$	el resultado es alto aunque mejora el anterior, se podría seguir mejorando esta expresión modificando el exponente, pero los resultados apuntan que no es significativo la variación de dicho exponente.
$A \cdot (\ln P)^b$	Predice valores del vocabulario más pequeños que el texto real
$A \cdot \sqrt{P} (\ln P)^b$	Podríamos considerar de todas ellas la que mejor resultado ofrece, aproximándose bastante al tamaño real del vocabulario, aunque el valor sigue siendo alto, es la mejor en todos los casos.
$A \cdot P^b (\ln P)^c$	Valores altos y mejorando a las anteriores salvo la fórmula 5.
$A \cdot P^b (\ln P)^c$	Valores más altos que la segunda fórmula pero no la mejora
$A \cdot (P(\ln P)^2)^b$	Valores más altos que la tercera fórmula pero no la mejora y predice valores más altos que la fórmula anterior
$A \cdot (P(\ln P)^4)^b$	Alto pero menos

Tabla 13. Fórmulas para la extrapolación

6. Modelo de distribución de frecuencias. Ley de Zipf

En este capítulo se aplica el modelo de distribución de frecuencias de Zipf cotejando sus resultados en distintas circunstancias y tipos de textos. Se obtiene los pormenores de la variación de los parámetros de la fórmula de Zipf. Igualmente se desarrolla y expone un método gráfico y sencillo denominado Modelo Log-% para visualizar las pequeñas desviaciones propias de la formulación de Zipf. Los textos utilizados en este capítulo requieren su división en documentos más pequeños, y se realizan los correspondientes análisis cuantitativos.

En este capítulo los cálculos realizados pueden considerarse sincrónicos en el sentido de que se han tomado y mezclado resultados en fragmentos de texto situados en cualquier posición, para ajustar los coeficientes tomando un mayor número de muestras aleatorias y de distintas posiciones en el documento.



Las experimentaciones realizadas llevan el siguiente proceso: se realizan cálculos, estos cálculos se trasladan a los gráficos para visualizar los resultados y con los resultados obtenidos poder aprovechar el modelo de Zipf o su incumplimiento para alcanzar nuestros objetivos.

6.1. La ley de Zipf en relación con el estudio de palabras en los documentos. Estudios destacados

Es evidente que en una colección cualquiera de documentos la cantidad de palabras distintas que éstos contengan es de gran interés para la fabricación de índices, de este modo el tamaño de un índice y por tanto su rapidez de utilización en un sistema de recuperación de información va a depender de la cantidad de palabras que incluyamos en él. El número de palabras distintas en un documento no crece a la misma velocidad que pueda crecer el tamaño absoluto del texto, ya que el vocabulario total es finito y las palabras se repiten. Esta afirmación y amplias investigaciones sobre esto se han desarrollado en el capítulo cinco en el que se tratan estudios cuantitativos relacionados con palabras en un texto o colección de textos similares utilizando las formulaciones de Heaps. Asimismo en una colección de documentos resulta de gran interés las veces que se repite una palabra, tanto las que se repiten con asiduidad como las palabras que aparecen una vez en el texto.

En este sentido, en los postulados de Zipf existe una relación directa con el lenguaje vislumbrándose características de la esencia del lenguaje humano como que la cantidad de palabras que forman el vocabulario de un texto es función del total de palabras que lo componen, según las fórmulas de Heaps de crecimiento amortiguado, que se han estudiado en capítulos anteriores. El hecho de que el vocabulario sea menor que el número de palabras totales es debido a que el autor repite palabras. Es consecuencia en

el sentido más directo de causa-efecto. Salton (1983) ya explicaba al respecto de la ley de Zipf como “*Principio de menor esfuerzo*”, este principio se basa en que la tendencia final del orador o escritor es repetir ciertas palabras en vez de crear palabras nuevas o diferentes a lo largo de su discurso o escrito. Por esta razón el autor no tiene como objetivo principal alcanzar cierta cantidad de vocabulario, sino que en un momento determinado decide utilizar cierta palabra que ya había utilizado con anterioridad, porque así conviene al sentido de lo que este autor está escribiendo.

Es decir, la repetición de palabras modelizada por la Ley de Zipf es el hecho primario y la cantidad de palabras distintas o cantidad de vocabulario modelizada por la Ley de Heaps su consecuencia.

Desde este punto de vista, como fundamento de la composición de los textos es como pueden verse los trabajos de Zipf y en este sentido los hechos son análogos a los que gobiernan la distribución de frecuencias en otros fenómenos naturales o artificiales.

Como se muestra en Denisov (1997) una secuencia de símbolos generada artificialmente, con cierta base estadística, produce a largo plazo frecuencias de aparición de grupos de símbolos (palabras) que responden a la ley de Zipf y esto puede interpretarse como aplicación del principio físico de máxima entropía.

Se observa además que existe una relación directa y numérica en las fórmulas con las que se medirá la similitud entre dos palabras con los postulados de Zipf ya que para ello se utilizará la frecuencia de cada palabra según indica la ley de Zipf. Así la distribución de los valores de las similitudes tendrá relación con la distribución de las frecuencias de las palabras.

El hecho de delimitar el contenido de los documentos mediante las palabras que estos contienen, por medio de la frecuencia de aparición de las palabras en dichos documentos y otras características afines es un modo en principio simple para conseguir el fin último de la recuperación automática de la información.

Por esta razón y para obtener resultados más fiables en la indización y posterior recuperación de información, las investigaciones van encaminadas a experimentar con conjuntos de palabras, es decir; con palabras asociadas que son las palabras que aparecen conjuntamente en un mismo documento. De este modo conseguiremos delimitar los clusters y agrupamientos de las palabras que en definitiva categorizan los documentos para obtener finalmente unos buenos resultados en la indización automática y en la recuperación.

Para determinar cuales son las palabras asociadas y cuales no, existe diversidad de mediciones, por ejemplo, para medir la intensidad relativa de las apariciones conjuntas de las palabras en los documentos se tiene en cuenta las frecuencias de las dos palabras consideradas y se utiliza el denominado índice de equivalencia o de asociación, el cual mide la intensidad de la asociación entre dos palabras i y j realizada sobre el conjunto de documentos del fichero. Se obtiene el valor 1 cuando la presencia de i acarrea automáticamente la presencia de j , y viceversa, es decir, cuando las dos palabras están siempre juntas. Por el contrario es igual a 0 cuando la mera presencia de una de las dos palabras excluye la otra.

Aunque el estudio con palabras asociadas se tratará en profundidad en el capítulo ocho. El tema que ahora nos ocupa no menos importante es la ley de Zipf en relación con el estudio de palabras únicas en los documentos que es el tema que se aborda en detalle a continuación. Para ello, debemos enmarcar el trabajo en el contexto del descubrimiento y aplicaciones de estas leyes, así hacemos un esbozo histórico de los científicos que han contribuido a ello.

La primera descripción del modelo de frecuencias de aparición de palabras en un texto fue la Ley de Zipf. George K. Zipf (1902-1950) profesor de Filología de la Universidad norteamericana de Harvard, se interesó desde los inicios de su carrera profesional por los cambios fonéticos que tenían lugar en las distintas lenguas y por las frecuencias del empleo de los distintos fonemas que presentaban ciertas modificaciones cuando eran observados durante un tiempo suficientemente prolongado. Del estudio de las frecuencias relativas de los fonemas Zipf pasó a trabajar con las frecuencias relativas del empleo de las palabras en los textos. Sus primeros trabajos publicados en 1932 se basaron en análisis empíricos acerca de la regularidad con la que los términos aparecían en diversos textos.

La ley de Zipf está considerada como uno de los fenómenos más llamativos de la lingüística cuántica, según esta ley, si ordenamos las palabras de mayor a menor frecuencia, el producto que resulta de multiplicar las frecuencias (f) de observación de las palabras de los textos por el valor numérico del rango que ocupan estas palabras en una distribución de frecuencias de observación permanece constante. Así se verifica que: *Rango * Frecuencia = Constante*. Siendo, la *frecuencia* igual al número de veces que se repite una palabra en el documento. Y el *Rango* igual al valor que corresponde a cada palabra ordenadas de mayor a menor frecuencia. Igualmente se investiga que dicha constante sufre una pequeña desviación, por ello la fórmula de Zipf incluye un exponente que tendrá un valor muy próximo a 1. La siguiente fórmula corresponde a la distribución de frecuencias de Zipf:

$$fr = \frac{K}{r^e}$$

Zipf observó que dentro de un texto se puede advertir que el uso de las palabras está claramente definido en términos estadísticos, por valores constantes. Esto significa que, la frecuencia con la que los vocablos aparecen en un texto están sujetas a unas relaciones matemáticas. Igualmente trató de extender la validez de su ley a otros campos en el orden de la naturaleza y de la vida, como la talla demográfica de las ciudades, la intensidad de los terremotos, etc. La explicación de Zipf se ha desacreditado, pero no la validez de los resultados empíricos a los que, posteriormente se han dado otras interpretaciones de otros científicos aportando mejoras a la fórmula inicial de Zipf.

Otro de los científicos que han contribuido al estudio de las frecuencias de las palabras ofreciendo una ley más completa que la dada por Zipf ha sido Benoit Mandelbrot (1924-2010), nacido el 20 de noviembre en Warsaw, Polonia. La fórmula de Mandelbrot (1953) principalmente trata de mejorar las deficiencias de la Ley de Zipf, incluyendo nuevos parámetros y una teoría más completa para la frecuencia de las palabras en los textos. Una de las aportaciones más importantes fue la de incluir un nuevo sumando a la fórmula inicial de Zipf.

$$fr = \frac{K}{(r + a)^b}$$

Donde K es una constante que se ajusta a la magnitud de las frecuencias. La constante K sería el tamaño del texto. Donde r es una constante que se suma al rango a y donde b es el exponente característico de la distribución de Mandelbrot.

Mandelbrot (1982) destaca también por su aportación a la Teoría de Fractales siendo uno de los precursores en esta materia. En su obra *Les objets fractals: forme, hasard et dimensions* publicada en el 1975 introdujo el término fractal para designar figuras geométricas especiales caracterizadas por su forma irregular repetitiva. Es asimismo autor de *The fractal Geometry of the Nature* publicado en 1982, obra en la que revisó y amplió su ensayo de 1975. Fue el principal responsable del auge de este dominio de las matemáticas desde el inicio de los años ochenta, y del interés creciente en esta materia. Utilizó la herramienta que se estaba democratizando en esa época, el ordenador, para trazar los más conocidos ejemplos de geometría fractal: el conjunto de Mandelbrot, así como los conjuntos de Julia descubiertos por *Gaston Julia* quien inventó las matemáticas de los fractales y que más tarde fueron desarrollados por Mandelbrot.

El matemático Benoit Mandelbrot acuñó la palabra fractal en la década de los 70, derivándola del adjetivo latín *fractus*. El correspondiente verbo latino: *frangere*, significa romper, crear fragmentos irregulares.

Los fractales fueron concebidos aproximadamente en 1890 por el francés *Henri Poincaré*. Sus ideas fueron extendidas más tarde fundamentalmente por dos matemáticos también franceses, *Gastón Julia* y *Pierre Fatou*, hacia 1918. Se trabajó mucho en este campo durante varios años pero el estudio quedó congelado en los años veinte.

El Fractal es matemáticamente una figura geométrica que es compleja y detallada en estructura a cualquier nivel de magnificación. A menudo los fractales son semejantes a sí mismos; esto es, poseen la propiedad de que cada pequeña porción del fractal puede ser visualizada como una réplica a escala reducida del todo. Existen muchas estructuras matemáticas que son fractales: el triángulo de Sierpinski, la curva de Koch, el conjunto Mandelbrot, los conjuntos Julia y muchas otras. Pueden darse de forma natural debido a que hay objetos en la naturaleza que tienen esta propiedad de ser parecidos a sí mismos cualquiera que sea la escala a la que sean observados, pero también pueden ser creados por ordenador. Como ejemplos de fractales naturales tenemos las nubes, los rayos, las líneas costeras, la estructura alveolar de los pulmones o incluso algunas superficies de ciertas proteínas (se las llama superficies fractales).

La característica que fue decisiva para llamarlos fractales es su dimensión fraccionaria. No tienen dimensión uno, dos o tres como la mayoría de los objetos a los cuales estamos acostumbrados. Los fractales tienen normalmente una dimensión que no es entera, por ejemplo 1,55.

Los objetos reales no tienen la infinita cantidad de detalles que los fractales ofrecen con un cierto grado de magnificación.

En un estudio desarrollado en 1974 en IBM e impulsado fuertemente por el desarrollo de la computadora digital. El Dr. Mandelbrot de la Universidad de Yale, con sus experimentos de computadora es considerado como el padre de la geometría fractal. En honor a él, uno de los conjuntos que él investigó fue propuesto con su nombre.

Como investigador en IBM se enfrentó a un problema que sorprendía a los ingenieros de la multinacional. Estos notaron que se producen siempre errores de transmisión de información entre ordenadores, sea cual sea la cantidad de información transferida. Además vieron que existía cierta regularidad en la distribución temporal de estos: tras un período de ausencia venía otro repleto de errores. Mandelbrot analizó el fenómeno, y descubrió que a medida que acortamos los intervalos de tiempo estudiados, sigue apareciendo más complejidad en la distribución. Así, si durante una hora no se producen errores, en la siguiente sí se producen. Si estudiamos únicamente esta hora, y la dividimos en intervalos de 20 minutos, veremos que habrá intervalos sin errores, y otros con muchos errores; estudiando aquellos con errores volvíamos a ver nuevos intervalos, etc. Mandelbrot asoció la imagen con un objeto estudiado por el matemático Georg Cantor, el conjunto de Cantor. Descubrió otros procesos con complejidad creciente a medida que nos aproximamos a ellos y los llamó fractales. Mostró cómo los fractales están presentes en muchos lugares tanto en las matemáticas como en la naturaleza.

Otros matemáticos como Douady, Hubbard y Sullivan trabajaron también en esta área explorando más las matemáticas que sus aplicaciones. Desde la década de los 70 este campo ha estado en la vanguardia de los matemáticos contemporáneos. Investigadores como el Dr. Robert L. Devaney, de la Universidad de Boston ha estado explorando esta rama de la matemática con la ayuda de los ordenadores modernos.

Respecto a la teoría de fractales, en esta investigación podemos observar la estructura fractal de las frecuencias de las palabras de un documento y en consecuencia ver si aplicando la ley de Zipf se puede afirmar que las palabras y sus frecuencias tienen una estructura fractal esto significaría que la tendencia de la frecuencia de las palabras seguiría una misma estructura compleja a cualquier nivel de magnificación.

Y lo que observamos es que a distintos niveles de magnificación, la distribución de las frecuencias de las palabras siguen una estructura diferente, incluso varía la distribución obtenida si los documentos analizados han sufrido distintos procesos de análisis lingüístico como puede ser la extracción en raíces, las eliminación de palabras vacías, etc.

Así pues, si tenemos un texto grande y aplicamos métodos más agresivos de agrupación de palabras, de manera superficial, observamos cómo el exponente de Zipf ofrece una dependencia directa con la intensidad del agrupamiento de palabras: a menos raíces resultantes, mayor valor del exponente, e igualmente con la extracción de distintas cantidades de palabras vacías: a más palabras vacías eliminadas del texto, menor valor del exponente, esto indicaría la negación como estructura fractal del texto respecto a la ley de Zipf.

Vemos igualmente que si la ley Zipf no sigue una estructura fractal la ley de Mandelbrot todavía se aleja más de esta hipótesis por tener su fórmula dos parámetros el exponente y el sumando, así pues en el caso de Mandelbrot con los dos parámetros (e) y (Σ) la

distribución de frecuencias si variamos los valores de (e) y (Σ) es si cabe todavía más dispar.

Con lo cual podemos afirmar tras los resultados obtenidos que se desarrollarán en profundidad a lo largo de este capítulo, que la frecuencia de las palabras en los textos no sigue una estructura fractal.

Básicamente a partir de este punto se expone en profundidad los modelos de Zipf y Mandelbrot, para observar entre otros, las fluctuaciones que sufrirá la distribución de Zipf según distintos valores del exponente y en qué modo la predicción de Zipf se ajusta a los valores reales de los textos.

6.2. Interés actual por los postulados de Zipf. Bibliografía básica y bibliografía reciente.

Entre la bibliografía básica de Zipf y Mandelbrot destacan los siguientes:

-Zipf, G.K. *Selective Studies and the Principle of Relative Frequency in Language*, 1932

-Zipf, G.K. *Psycho-Biology of Languages*, 1935

-Zipf, G.K. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley Publishing. Massachusetts, 1949

-Mandelbrot, B. "An Information Theory of the Statistical Structure of Language". *Communication Theory*, ed. By Willis Jackson,. New York: Academic Press, p 486-502, 1953

-Mandelbrot, B. "Simple Games of Strategy Occurring in Communication through Natural Languages". *Transactions of the IRE Professional Group on Information Theory*, 3, 124-137, 1954

-Mandelbrot, B. "A probabilistic Union Model for Partial and temporal Corruption of Speech". *Automatic Speech Recognition and Understanding Workshop*. Keystone, Colorado, December, 1957

-Mandelbrot, B. *Les objets fractals, forme, hasard et dimension*, París: Flammarion, 1975

-Mandelbrot, B. *The fractal geometry of nature*, W. H. Freeman & Company, New York, 1982

Aún hoy en la actualidad la ley de Zipf tiene una importancia creciente en multitud de estudios de distintas disciplinas. Desde que Zipf acuñó la ley sobre el modelo de frecuencias de aparición de palabras en los textos, ha sido expuesta a variedad de críticas, modificaciones y estudios científicos sobre la viabilidad de dicha ley, tal es así, que aún hoy en la actualidad la ley de Zipf sigue estando presente en investigaciones de diversa índole y lo más importante aún hoy resulta evidente que la ley de Zipf se cumple en otras materias llevadas a estudio, un ejemplo de esto lo tenemos en varios artículos que demuestran el interés que existe en la actualidad en el estudio y comprobación de la viabilidad de esta ley.

Adriano de Jesús Holanda et. al. (2003), en su investigación aplican dos tratamientos distintos para la clasificación de palabras, para obtener una herramienta de recuperación automática de información, así consideran las palabras que son los nodos y sus

relaciones son los enlaces, esto es la estructura básica que permite la conformación de un tesoro las estadísticas de dichos enlaces entrantes y ascendentes son obtenidas por funciones simples de ajuste, así se expresa que el número de enlaces de cada palabra en un tesoro, sigue la ley de Zipf e igualmente se prueba con una gran cantidad de artículos del periódico on-line *The New York Times*, que las palabras en lenguaje natural se distribuyen siguiendo la ley de Zipf.

Popescu (2003), muestra como el valor del factor de impacto de las revistas en una colección suficientemente amplia se distribuye según una variante de la ley de Zipf, conocida como fórmula de Lavalette (Popescu et. al., 1997).

El mismo autor, Popescu (2003) explica la modificación a la ley de Zipf conocida fórmula de Lavalette, a la que responde la distribución de factores de impacto de revistas.

Otro artículo científico (Bronlet y Ausloos, 2003) afirma como la fórmula de Zipf ha resultado de aplicación a ciertos problemas de física, sobre todo en relación a largas series de valores que muestran una regularidad a largo plazo en medio de un aparente caos. Por ejemplo el tiempo transcurrido entre dos terremotos o el régimen de caudal de los ríos, etc. Se ha encontrado relación con la aparición de palabras en un texto, que se considera como una larga serie de palabras.

Otro artículo publicado en 2002, (Malacarne, Mendes, Lenzi, 2002) estudia la distribución de la población en las ciudades de más de cien mil habitantes, estos autores indican que la distribución de ciudades según el número de habitantes sigue una distribución q-exponencial, como predice la ley de Zipf-Mandelbrot.

Siguiendo con el caso de las ciudades y su población Matteo Marsili y Yi-Cheng Zhang (1998) indican como actuaciones individuales que corresponden a decisiones tomadas por habitante de las ciudades, contribuyen a conformar los tamaños de las ciudades según la ley de Zipf.

Otro artículo científico (Lyra et. al. 2003) demuestra como también la ley de Zipf es aplicable a las elecciones, así indica que en países donde las elecciones son a listas abiertas (por lo tanto excluimos a España), el número de votos obtenidos por cada candidato responde a la ley de Zipf.

Según Tsallis (2002), La ley de Zipf también se cumple en la especialidad de la termodinámica.

Por último un artículo científico de Malacarne y Mendes (2000) explica como la ley de Zipf también es aplicable al número de goles marcados por cada jugador de fútbol, igualmente aunque no se mencione en este artículo, también el número de campeonatos de liga ganadas por cada equipo español, responde a la ley de Zipf.

La bibliografía reciente es cuantiosísima, sobre todo después de lo escrito con ocasión del centenario del nacimiento de Zipf en 2002. Una recopilación exhaustiva se mantiene en <http://www.nslj-genetics.org/wli/zipf/>.

6.3. Formas de presentación gráfica de la Ley de Zipf/Mandelbrot y significado de los valores de los parámetros

6.3.1. Representación clásica

Si tomamos un texto cualquiera para aplicar la Ley de Zipf y utilizamos para su representación una herramienta básica, en la cual representamos las frecuencias de las palabras en relación a su rango podemos obtener fácilmente la distribución de una curva hiperbólica, como puede apreciarse en el siguiente ejemplo correspondiente al texto PG5.txt (Texto: Episodios Nacionales; Autor: Pérez Galdós, Tamaño del texto escogido: 4.125 palabras; Vocabulario: 2.590 palabras)

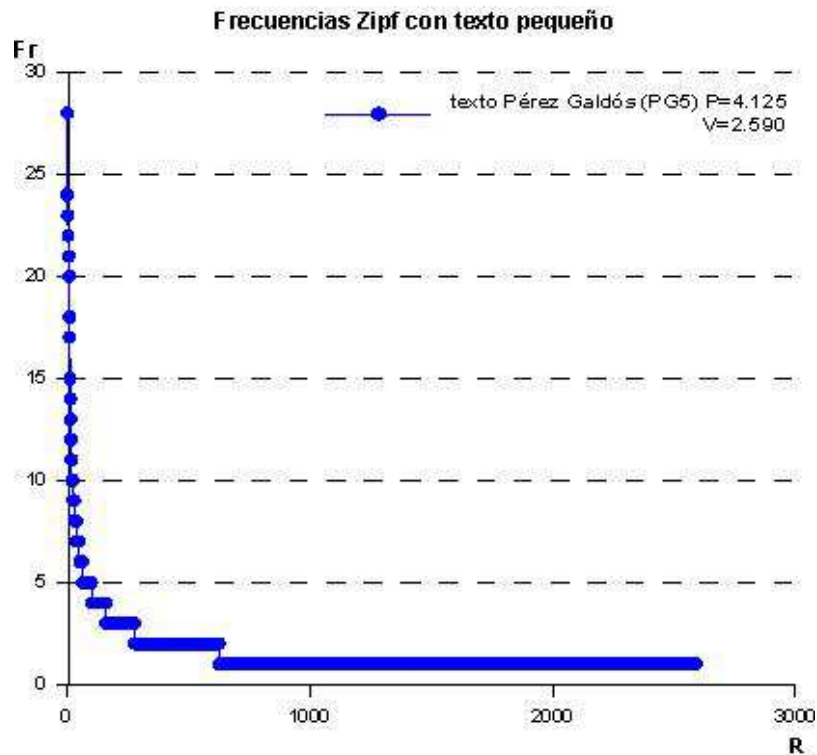


Gráfico 31. Distribución de frecuencias de Zipf con texto P=4.125

Este ejemplo de representación de curva hiperbólica se puede observar también con el siguiente ejemplo correspondiente a datos de población de las ciudades de más de un millón de habitantes

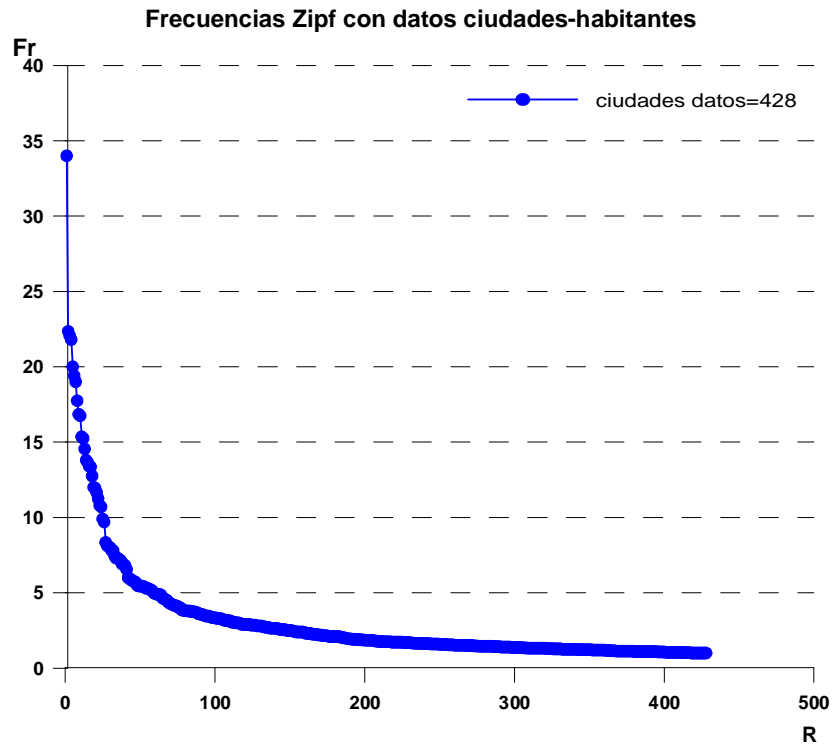


Gráfico 32. Distribución de frecuencias de Zipf con texto $P=428$

La representación clásica resulta útil a la hora de representar gráficamente una cantidad de datos pequeña, por el contrario resulta evidente que en el siguiente gráfico no alcanza para visualizar adecuadamente lo que ocurre en ejemplos voluminosos. Por ejemplo, para un texto de 379.945 palabras, la gráfica aparece completamente pegada a los ejes como puede observarse

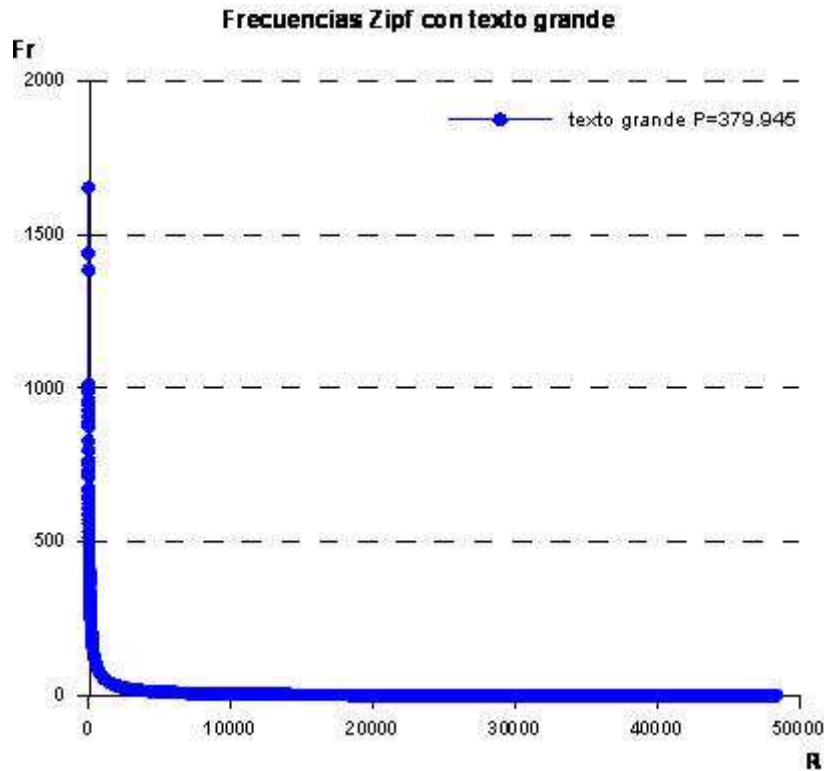


Gráfico 33. Distribución de frecuencias de Zipf con texto P=379.948

Dado que la utilización de estas distribuciones con textos grandes no pueden apreciarse debido a que quedan literalmente pegadas a los ejes, se han utilizado otras formas de representación más adecuadas, en las cuales no sea un impedimento el tamaño del texto que se analice. (véase toda la bibliografía citada anteriormente).

6.3.2. Representación logarítmica

La representación logarítmica muestra en el eje de abscisas el logaritmo del rango y en el eje de ordenadas el logaritmo de la frecuencia esta representación en los años en que trabajó Zipf era una técnica habitual en los años 40 y 50, y muchas series de datos de todo tipo se representaban en papel logarítmico, con unas escalas que corresponden exactamente a esta transformación.

Obviamente, al utilizar la representación logarítmica de la ley de Zipf el resultado visible no será una curva hiperbólica sino una línea recta con pendiente -1 . Esto es debido a que la fórmula de Zipf en su versión simple con exponente, $F = \frac{cte}{r^e}$ puede

transformarse tomando logaritmos en: $\log(f) = \log(cte) - e \cdot \log(r)$. Lo que representa una línea recta de pendiente $-e$ si tomamos el eje horizontal para representar $\log(r)$ y el eje vertical para representar $\log(f)$.

Una vez representados los datos reales de esta manera resulta muy visible si cumplen o no la ley de Zipf: basta observar si quedan en línea recta. Además, como el valor de e debe ser próximo a 1, esta recta tiene pendiente -1 aproximadamente. Al contrario que se aplicaba en el capítulo cinco, la representación logarítmica de la Ley de Heaps era una línea recta con pendiente 1, debido a que la fórmula de la Ley de Heaps es $V = a \cdot P^b$

De este modo se empleará la representación logarítmica de la ley de Zipf para poder apreciar una óptima visión de las distribuciones aunque se trabaje con textos grandes o con gran cantidad de datos.

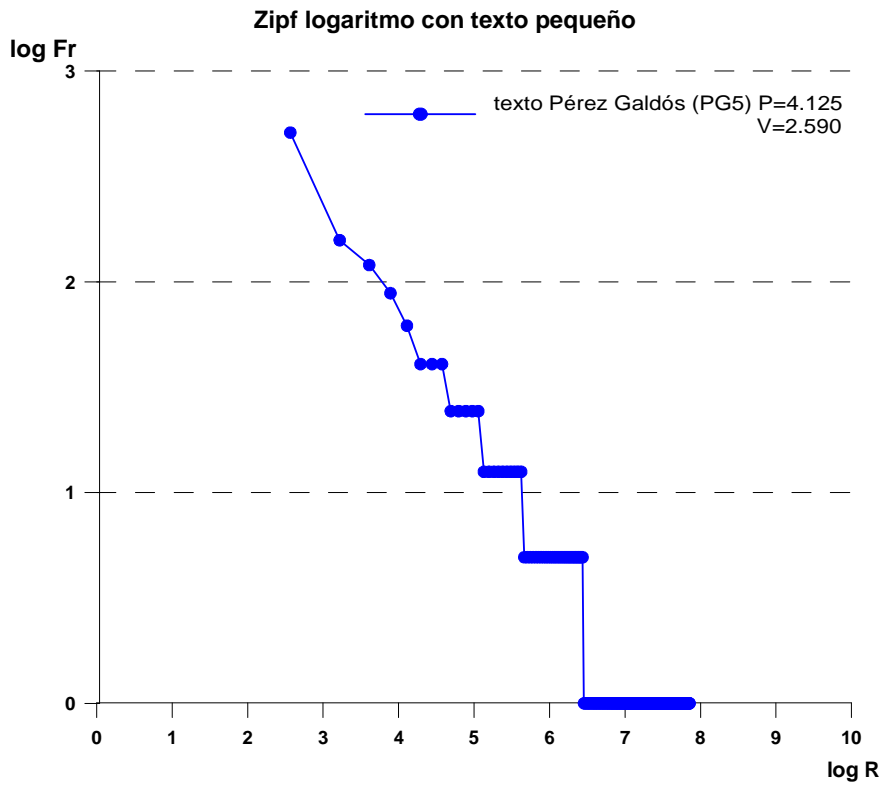


Gráfico 34. Distribución logarítmica de frecuencias de Zipf con texto P=4.125

Ejemplo de representación logarítmica sobre un conjunto de datos de población de las ciudades de más de un millón de habitantes.

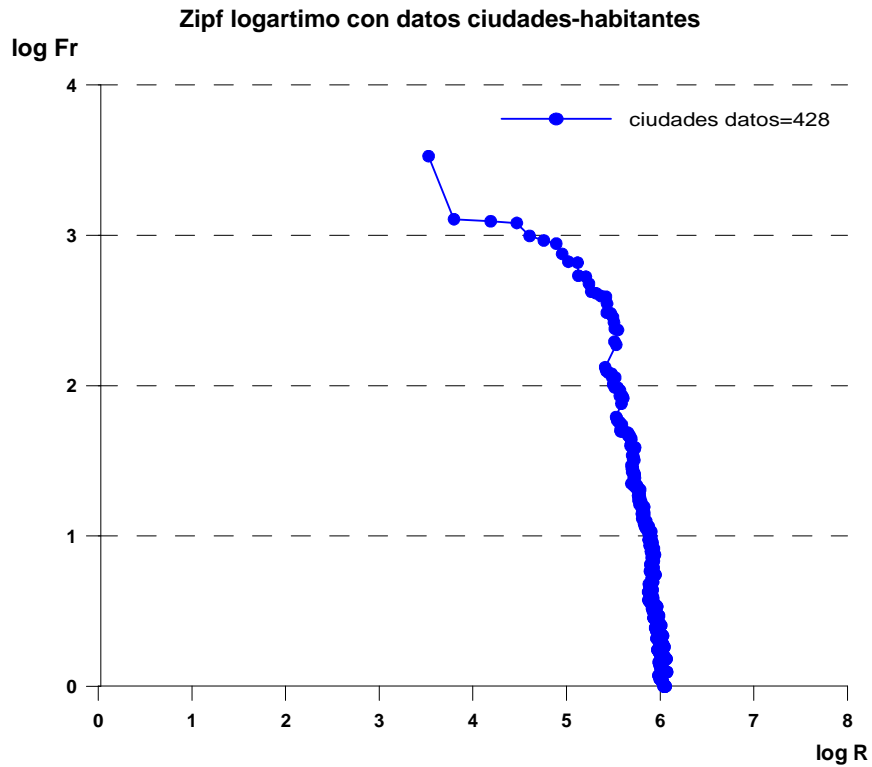


Gráfico 35. Distribución logarítmica de frecuencias de Zipf con texto $P=428$

Ejemplo de representación logarítmica sobre un texto grande de tamaño total 379.945 palabras.

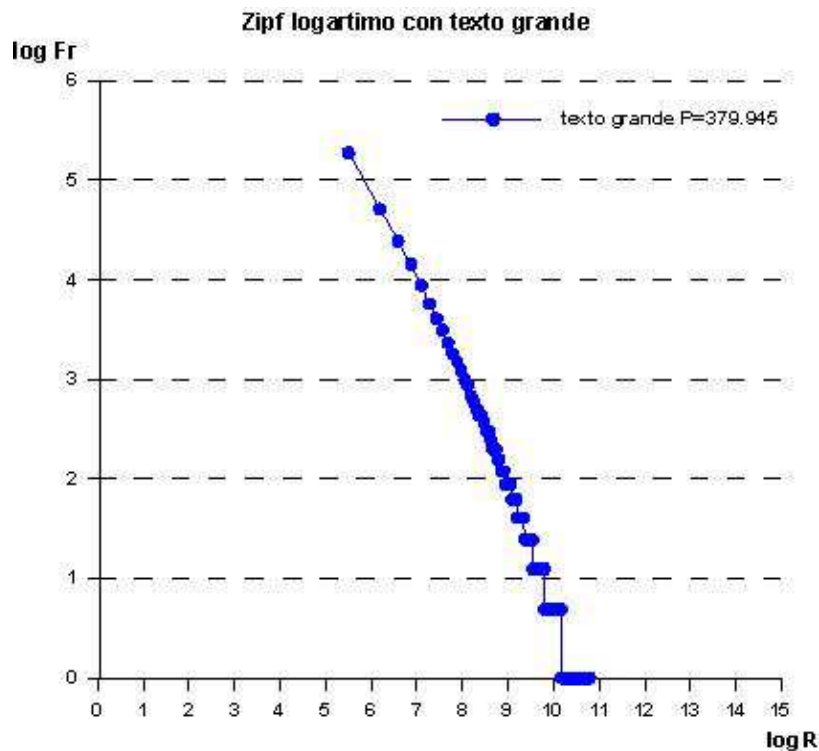


Gráfico 36. Distribución logarítmica de frecuencias de Zipf con texto $P=379.945$

Hasta ahora, las gráficas vistas representan datos de textos reales pero seguidamente en esta investigación se comparará la tendencia de los datos reales según la Ley de Zipf (como hemos visto en las tres gráficas anteriores), con la tendencia de los datos *teóricos* obtenidos con la fórmula de Zipf mejor ajustada²⁶.

6.3.3. Representación transformada o Espectro de frecuencias

Es interesante mencionar otras representaciones de la Ley de Zipf, como la *Representación transformada*, donde no se utiliza rango-frecuencia, sino frecuencia-cantidad de palabras que tienen esa frecuencia. En realidad esta representación se presta a agrupar por tramos de frecuencia presentando para cada tramo, por ejemplo frecuencias comprendidas entre 5 y 10, el número de palabras que tienen alguna de estas frecuencias.

A veces denominada segunda ley de Zipf a la que justifica esta representación. Demuestra Kornai (1999), que esta representación transformada describe un modelo matemático que justifica la ley de Zipf tanto en su versión “primera” como en la “segunda”, la llamada representación transformada es una consecuencia matemática de la “primera” ley de Zipf. Kornai descompone la curva de Zipf en tres tramos a los que aplica modelos matemáticos distintos. Así por tanto concluye que si un texto cumple la primera ley de Zipf necesariamente cumple también la segunda.

Distinguimos tres modelos distintos, en primer lugar, si el texto es tan perfecto que la distribución de las palabras sigue la distribución estándar armónica según Baayen (2001), esto es equivalente a la fórmula de Zipf con exponente 1 y con algunas simplificaciones más se demuestra que la *Distribución Transformada* sigue esta fórmula:

$$V(fr) = \frac{V}{fr * (fr + 1)}$$

Las palabras de frecuencia 1, son las palabras que sólo aparecen una vez en el texto y se les denomina hapaxes, éstos se caracterizan por el número de palabras en un texto de N palabras que se repiten m veces $V(m, N)$, siendo el número de hapaxes $V(1, N)$. En el caso de que la frecuencia sea 1, esto indica que el número de hapaxes es la mitad del vocabulario. Mediante los hapaxes también se estudian los modelos LNRE (Large Number of Rare Events), los eventos raros (LNRE) como los hapax, etc. conducen a otros modelos matemáticos más sofisticados.

Esta fórmula se utiliza para todas las frecuencias, en particular para los hapax:

$$V(m, N) = \frac{V(N)}{m * (m + 1)}$$

²⁶ Véase apartado 6.4

En segundo lugar, como una solución adaptada a los textos reales, donde la colección de valores $V(1, N)$, $V(2, N)$, ... para un texto determinado, lo que se denomina su espectro de frecuencias, queda representada por la expresión:

$$V(m, N) = \frac{K}{m^e}$$

Donde el exponente (e) toma valores cercano a 2 (Baayen, 2001). Esta fórmula es conocida como segunda Ley de Zipf.

En tercer lugar, como una solución mejor adaptada a los textos reales, los modelos LNRE (Large Number of Rare Events) son una sofisticación de la segunda fórmula de Zipf, lo que proporcionan es un modelo de cuantas palabras hay de cada frecuencia, es decir de $V(m, N)$. Como consecuencia, sumando todos $V(m, N)$ obtendremos el vocabulario.

Entre los modelos matemáticos más sofisticados como se ha mencionado existen los modelos de tipo LNRE (Large Number of Rare Events), (Baayen, 2001), más orientados a la forma de la distribución de frecuencias de las palabras, pero que también como ya hemos mencionado proporcionan como uno de sus resultados el tamaño del vocabulario. Ahora bien, tal como se reconoce en Evert y Baroni (2007)

“...LNRE models are rarely if ever employed in linguistic research and NLP applications. We believe that this has to be attributed at least in part, to the lack of easy-to-use but sophisticated LNRE modeling tools...”

“...los modelos LNRE son pocas veces empleados en la investigación lingüística y aplicaciones NLP. Nosotros creemos que esto se atribuye por lo menos en parte, a la ausencia de herramientas de Modelos LNRE de “fácil-de-usar” pero sofisticados...”

Las recomendaciones simplificadas que se han dado en el capítulo cinco de esta tesis doctoral interpolando o extrapolando cuando proceda son comparables en cuanto a precisión a las que hacen los autores mencionados Evert y Baroni (2007) con los LNRE, aunque con menos esfuerzo, menos o más ajustadas según el esfuerzo que se aplique.

A continuación se muestran ejemplos de la representación transformada de la Ley de Zipf con el texto bachi.txt (Texto: Científico-legal; temática: Real Decreto del currículo de Bachillerato, Tamaño del texto escogido: 73.173 palabras; Vocabulario: 9.161 palabras).

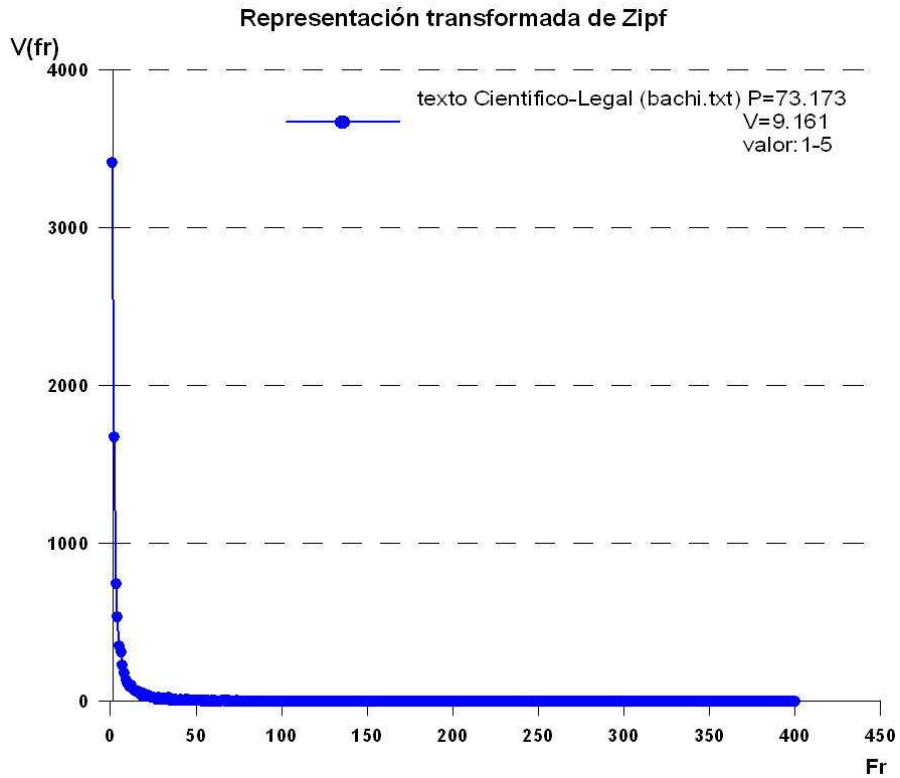


Gráfico 37. Distribución transformada de frecuencias de Zipf con texto $P=73.173$

Ejemplo de representación transformada de la Ley de Zipf con escala logarítmica en ambos ejes, concretamente en las frecuencias para las que hay 0 palabras, como no se puede calcular su logaritmo se han borrado de la gráfica.

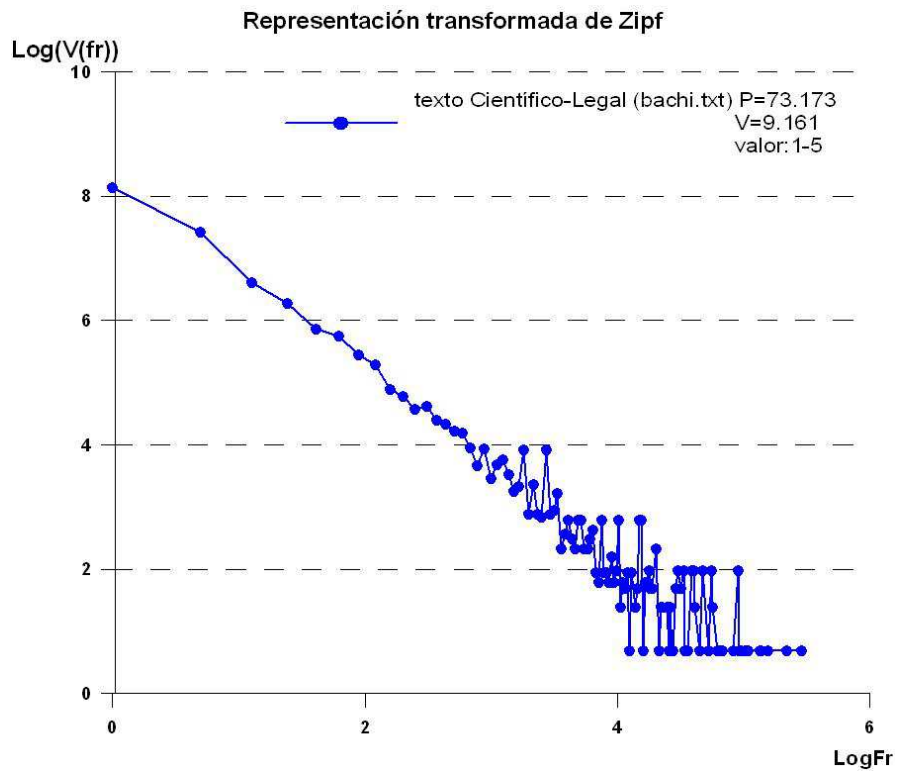


Gráfico 38. Distribución logarítmica transformada de frecuencias de Zipf con texto $P=73.173$

Señalar que en el presente trabajo no se empleará este tipo de representación, tan sólo en el punto 6.4.1 de este Capítulo visualizaremos algunos ejemplos con este tipo de Representación Transformada o Espectro de Frecuencias.

6.3.4. Gráficas sintéticas y significado de las fórmulas de Zipf y Zipf-Mandelbrot

Denominamos *Sintético* a que se obtienen valores a partir de cálculos sobre las fórmulas, sin efectuar ningún recuento sobre un texto real, estos cálculos se han realizado con la aplicación TOPOS y el formulario SignificadoZipf desarrollado para obtener entre otros estos resultados. Es decir hasta ahora hemos visto la distribución de frecuencias de textos reales aplicándoles la Ley de Zipf e incluso hemos podido observar distintos modos de representación: clásica, logarítmica y transformada o espectro de frecuencias, pero a partir de ahora vamos a observar otro tipo de gráficas denominadas sintéticas que muestran explícitamente la distribución de frecuencias que trazan las fórmulas a partir de cálculos sobre ellas mismas, y estas gráficas sintéticas nos van a servir para estudiar fielmente como afecta a la distribución; el exponente y sumando de las fórmulas de Zipf y Mandelbrot, es decir de qué modo un valor mayor o menor de exponente y/o sumando hace variar la curva en la distribución obteniéndose así conclusiones al respecto.

6.3.4.1. Zipf

Para dar una interpretación a los valores de los parámetros en las fórmulas de Zipf o Mandelbrot generamos la distribución de frecuencias que predicen las fórmulas con distintos valores y comparamos las gráficas obtenidas. Observaremos el comportamiento del exponente según distintos valores tanto en las fórmulas de Zipf y Mandelbrot (siempre teniendo en cuenta que no lo realizamos sobre datos de textos reales, sino con cálculos aplicados a las fórmulas de Zipf) y analizaremos cómo afecta a las frecuencias según el valor del exponente. Los datos se elaboran con el formulario SignificadoZipf de TOPOS²⁷. Las gráficas son de rango en el eje horizontal y frecuencia en el vertical utilizando para este caso, la representación clásica y la visualización de la curva hiperbólica.

Para valores fijos Vocabulario=50 y Palabras=400, se representan cuatro distribuciones de Zipf correspondientes a los valores del exponente 0,6 0,7 0,8 0,9.

²⁷ Ver Apéndice II

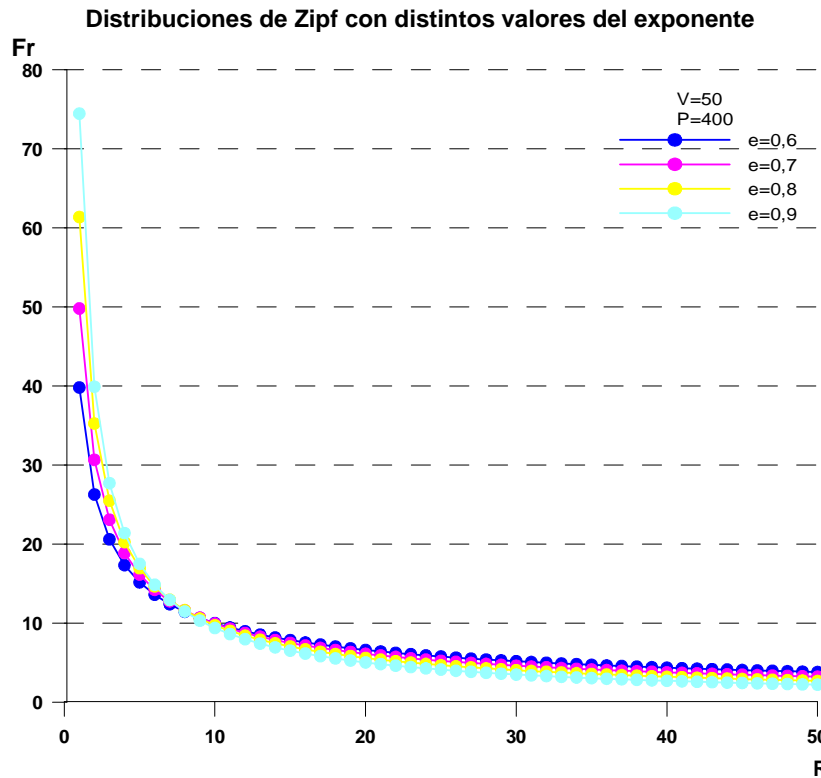


Gráfico 39. Distribución de frecuencias de Zipf con distintos valores del exponente (e)

Se observa que si el exponente es mayor las palabras más frecuentes del texto aparecen con valores altos de frecuencia, es decir destacan más las frecuencias máximas (palabras en los primeros rangos) sobre las demás. Por el contrario si el exponente es menor, las palabras más frecuentes del texto aparecen con un valor menor en su frecuencia. En resumen observamos a la derecha del gráfico una tendencia distinta que la observada a la izquierda del gráfico que es totalmente contraria.

A continuación se amplía el gráfico anterior a la zona de los rangos 3 a 20 y se puede observar con mayor detalle como se invierten las magnitudes, en este caso un menor exponente de Zipf significa más frecuencia.

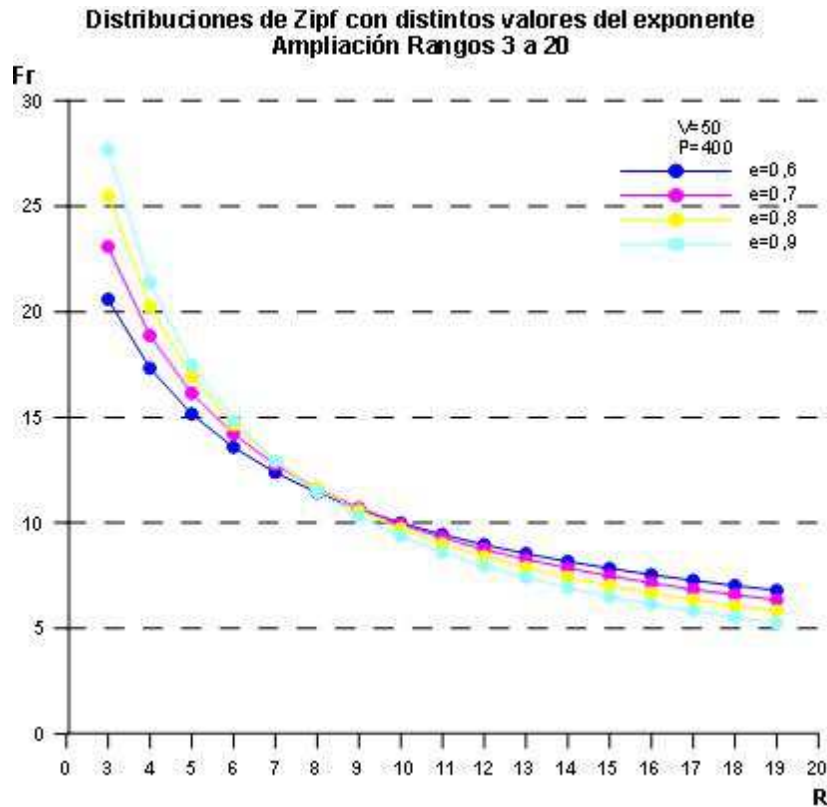


Gráfico 40. Distribución de frecuencias de Zipf con distintos valores del exponente (e). Rangos 3-20

Se advierte claramente como alrededor del rango 7, sobre el 14% del vocabulario se invierten las magnitudes. En las palabras de medias y bajas, un menor exponente de Zipf significa más frecuencia. También puede interpretarse como: menor exponente de Zipf significa menos palabras de frecuencia = 1

Se amplía el gráfico anterior a la zona de los rangos 40 a 51 se puede observar como la tendencia decreciente se mantiene

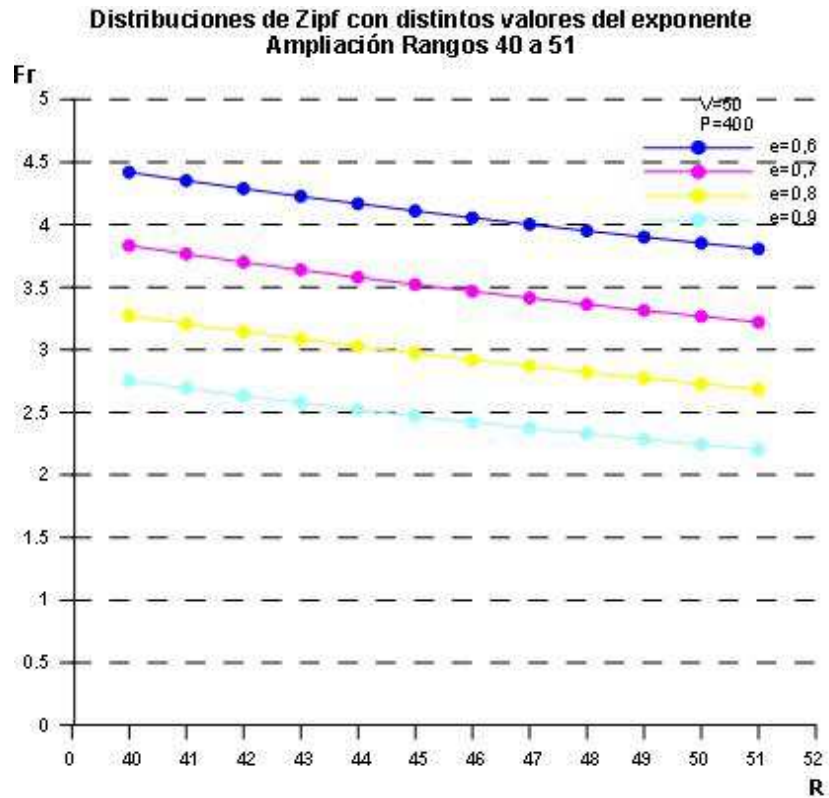


Gráfico 41. Distribución de frecuencias de Zipf con distintos valores del exponente (e). Rangos 40-51

Obviamente se mantiene la tendencia decreciente con menores valores cuanto mayor es el exponente. En este ejemplo de vocabulario tan pequeño no llega a verse la abundancia de palabras de frecuencia = 1

Con valores más distintos entre si, exponentes igual a 0,4 0,7 1 1,3 la tendencia es la misma. Un valor muy alto como 1,3 indica que la palabra más frecuente lo es mucho más que la segunda palabra.

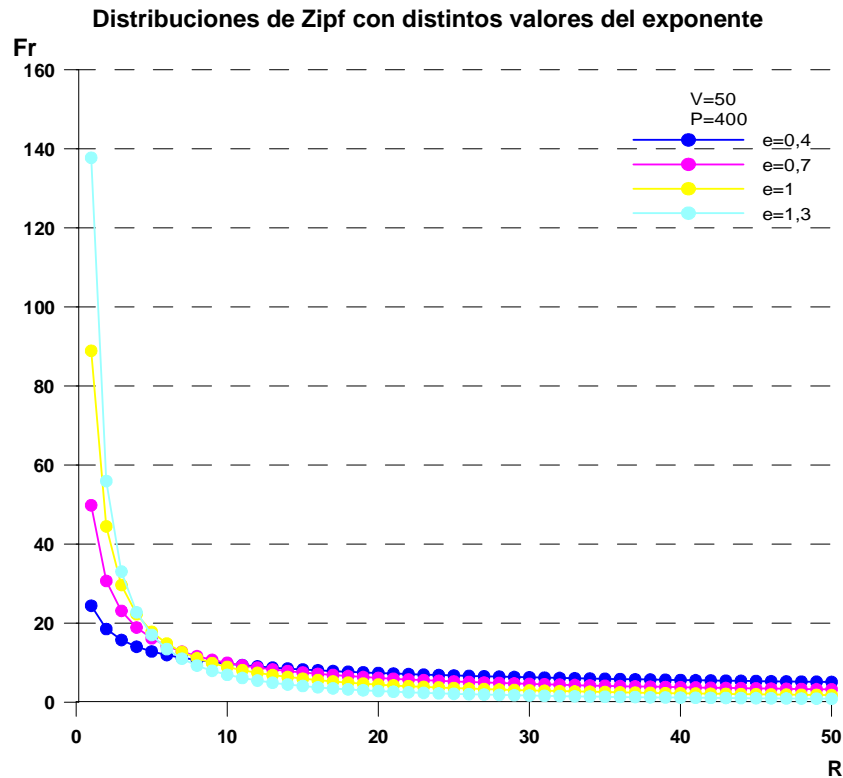


Gráfico 42. Distribución de frecuencias de Zipf con distintos valores del exponente (e)

Así se concluye que el valor del exponente (e) es importante ya que si utilizamos valores de (e) mayores destacarán las palabras más frecuentes respecto a las demás palabras. Por lo tanto la parte izquierda de la gráfica sirve para distinguir los primeros rangos y en la parte derecha de la gráfica la curva que observamos va dando valores $fr = 0,9$; $fr = 0,899$; $fr = 0,895$ y ofrece una tendencia distinta a la observada a la izquierda del gráfico.

En resumen lo que se está analizando es el comportamiento del exponente según distintos valores en las fórmulas de Zipf para examinar cómo afecta a las frecuencias el valor del exponente (e).

6.3.4.2. Mandelbrot

Una vez expuesto el resultado de los valores de los parámetros con la Ley de Zipf, seguidamente se evaluarán igualmente los resultados de los valores de los parámetros con la fórmula de Mandelbrot. Así en el caso de Mandelbrot para el mismo valor del

exponente, la presencia de un sumando $fr = \frac{K}{(r+a)^b}$ hace que disminuya el valor de

las palabras con frecuencias altas, tomando así valores más bajos.

Todas las gráficas que se exponen a continuación corresponden al exponente 0,7 con sumandos 0 3 10 30 y como puede apreciarse hacen disminuir los valores de las frecuencias mas altas. El sumando (Σ) de Mandelbrot se curva de otro modo distinto al (e) de Zipf, la curva se estira dando mayor o menor curvatura a la gráfica, es decir exagerando más o menos las frecuencias altas.

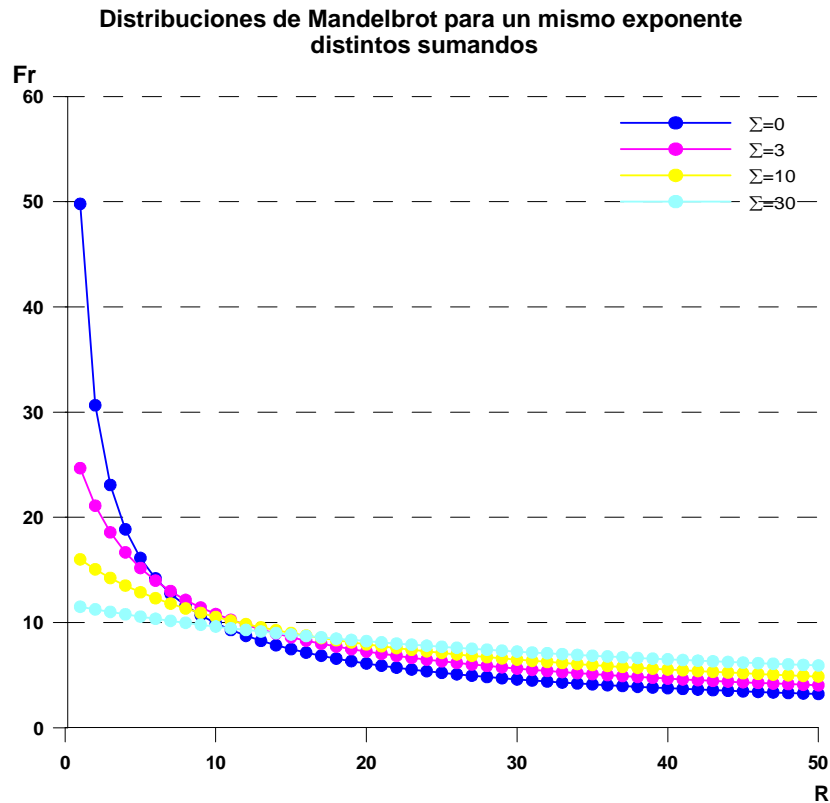


Gráfico 43. Distribución de Mandelbrot para un mismo exponente (e) distintos Σ

El exponente (e) de Mandelbrot al igual que Zipf logra aumentar los valores de las palabras más frecuentes respecto de las demás, pero el sumando (Σ) logra compensar esta tendencia porque cuanto mayor es el valor del (Σ) la distribución tendrá valores más bajos para las palabras más frecuentes.

En las siguientes gráficas se representa la fórmula de Mandelbrot con un $\Sigma=30$, y distintos valores para el (e). Se puede observar como se consigue un efecto parecido a lo que hacemos en Zipf; se balancea la curva manteniendo formas similares

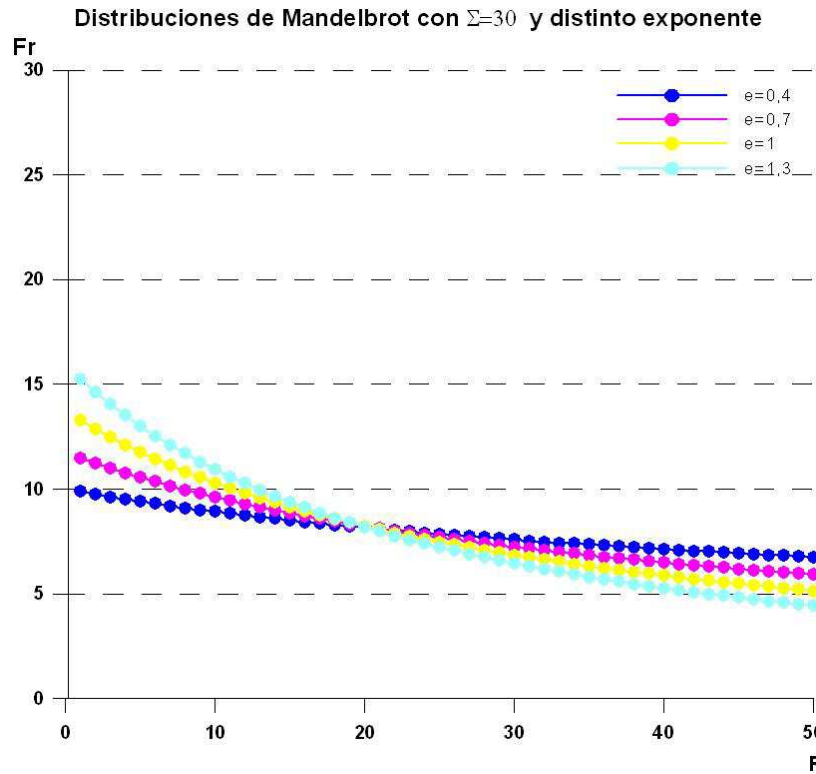


Gráfico 44. Distribución de Mandelbrot con $\Sigma=30$ y distinto exponente (e)

Como se trabaja con gráficas sintéticas, para comparar una distribución de Zipf y otra de Mandelbrot que puedan aplicarse al mismo texto se pueden escoger arbitrariamente los valores de todos los parámetros, de modo que las gráficas queden lo más superpuestas posible, con lo cual estamos haciendo verosímil que puedan corresponder al mismo texto. Al hacerlo vemos que a cambio de aumentar un parámetro hay que disminuir otro. Se amplía el gráfico anterior a la zona de los rangos 13 a 33

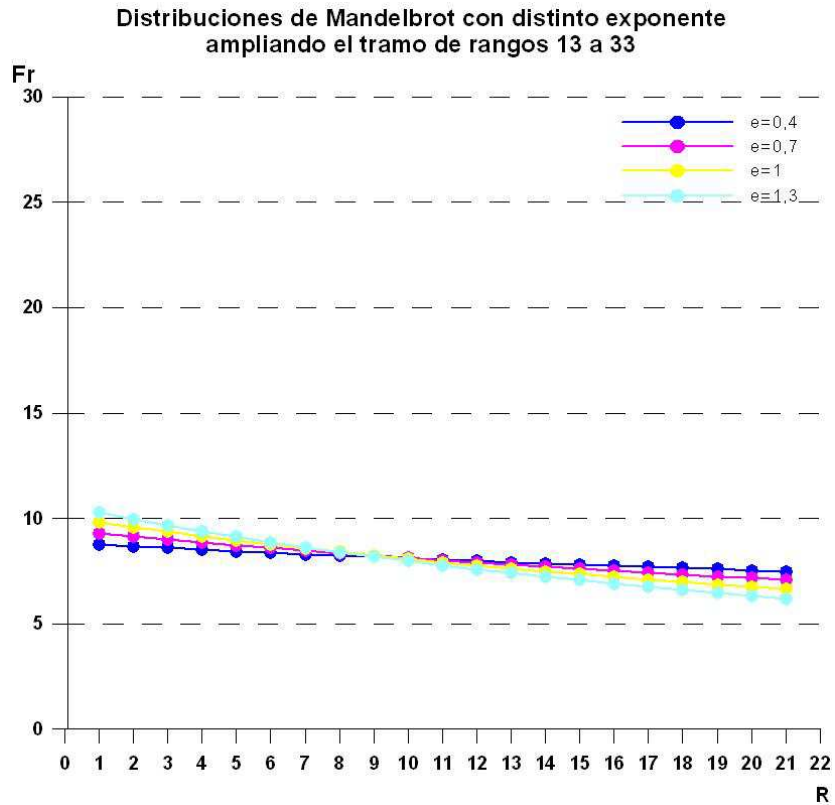


Gráfico 45. Distribución de Mandelbrot con distinto exponente (e) rangos 13-33

En estas frecuencias intermedias pero altas ha quedado por arriba el de exponente más alto compensado por el sumando más alto. En el gráfico siguiente se amplía el gráfico anterior a la zona de los rangos altos o frecuencias bajas

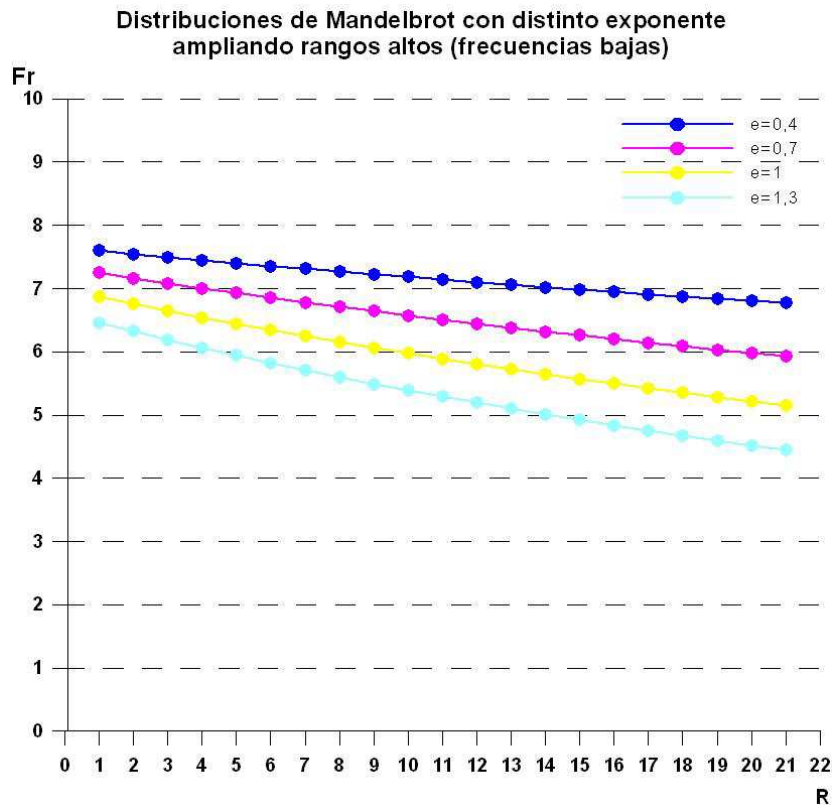


Gráfico 46. Distribución de Mandelbrot con distinto exponente (e) rangos mayores

Puede observarse como se ha invertido la relación quedando por debajo el de exponente más alto, al igual que ocurría con Zipf, pero en el caso de Mandelbrot la presencia de un (Σ) hace disminuir los valores de las palabras más frecuentes.

En resumen concluimos según lo que se ha explicado arriba que si aumentamos el valor del exponente (e), la distribución o curva que obtenemos en los gráficos tiende hacia arriba, tiende a dar valores muy altos para las palabras más frecuentes. Por tanto podemos decir que para un mismo (Σ) y distinto (e) la distribución queda para arriba.

Y si aumentamos el valor del (Σ) para un mismo (e) la distribución o curva que obtenemos en los gráficos tiende hacia abajo, tiende a dar valores muy bajos para las palabras más frecuentes, así pues podemos concluir que un sumando alto contrarresta a un exponente alto ya que hace el efecto contrario.

En las gráficas siguientes son simples recordatorios visuales de lo explicado hasta ahora, en ellas observaremos la tendencia de la curva respecto al vocabulario si variamos el exponente, el sumando e incluso observamos la constante que figura en el

numerador de las fórmulas, $fr = \frac{K}{(r+a)^b}$ este numerador sirve para ajustarse

globalmente a la magnitud de las frecuencias, lo que nos lleva en definitiva a la relación entre el número de palabras y el vocabulario. Se observa que a mayor amplitud de vocabulario, para el mismo tamaño de texto, las frecuencias son menores, la constante es menor y la curva queda como la rotulada en esta gráfica en color amarillo.

Esto es una consecuencia matemática de que la suma de las frecuencias debe ser igual al total de palabras. Vemos que variar el valor de la constante equivale a desplazar la gráfica arriba-abajo. Aunque en las experimentaciones realizadas decidamos no variar la K como sí hemos hecho con el exponente (e) y el sumando (Σ).

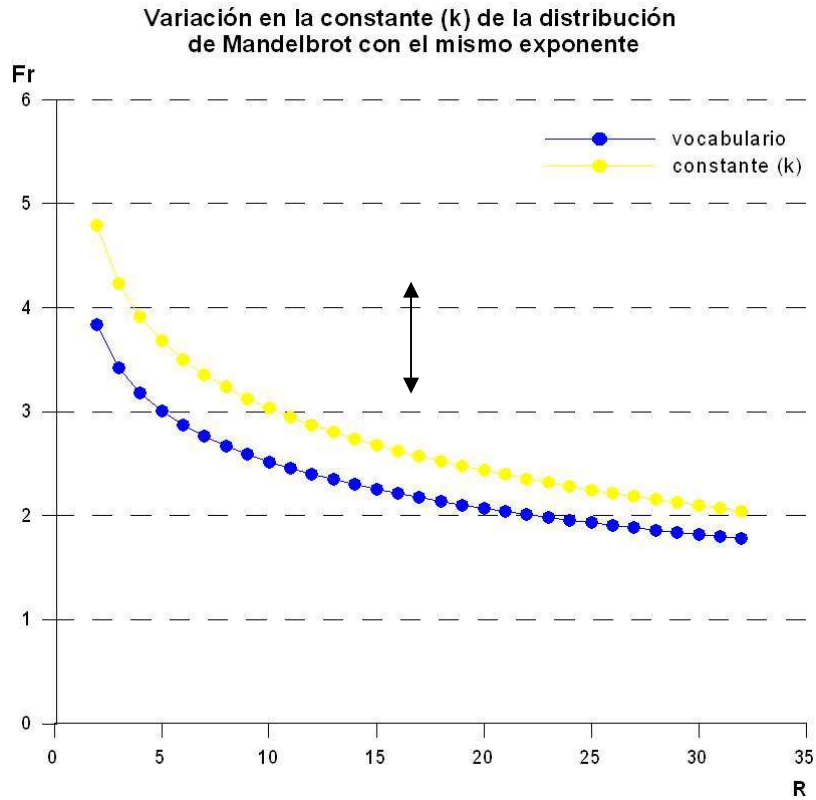


Gráfico 47. Tendencia de la constante (k) en la distribución de Mandelbrot

Suponiendo ajustada la constante los distintos valores del exponente nos dan gráficas más o menos agudas, indicando si la frecuencia de las palabras más frecuentes es exageradamente mayor o no que las siguientes frecuencias. En la siguiente gráfica esto se expresa por una inclinación u oscilación del dibujo balanceándose la curva manteniendo la misma forma o similares

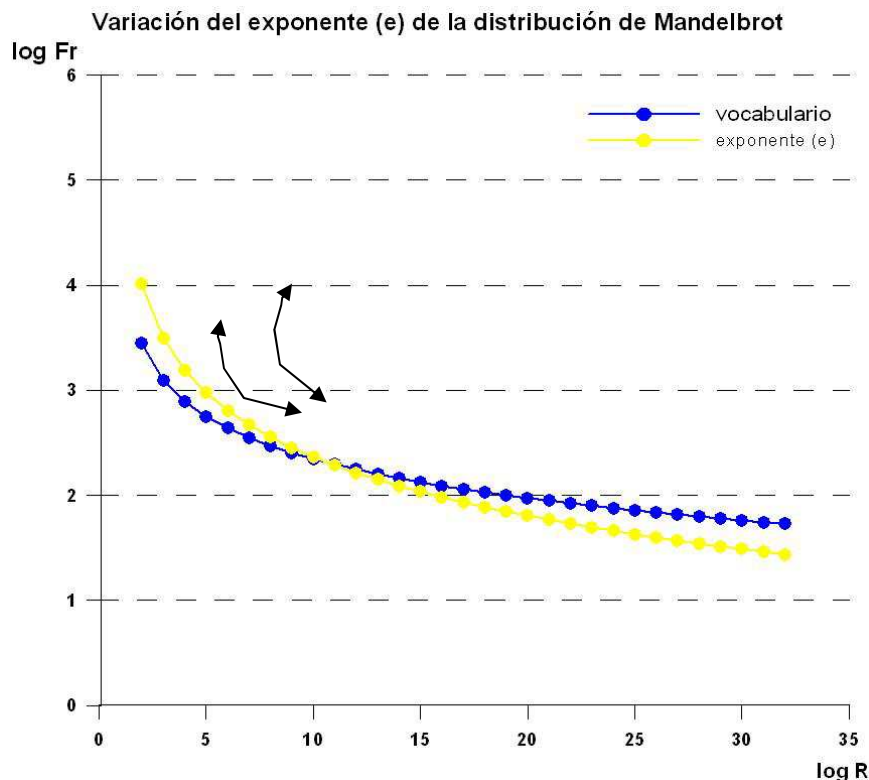


Gráfico 48. Tendencia de la exponente (e) en la distribución de Mandelbrot

Por último, una vez ajustados la constante y el exponente variando el sumando de la fórmula de Mandelbrot y que vale 0 en la de Zipf, obtenemos gráficas en la misma posición aproximadamente pero con mayor o menor curvatura exagerando más o menos las frecuencias altas.

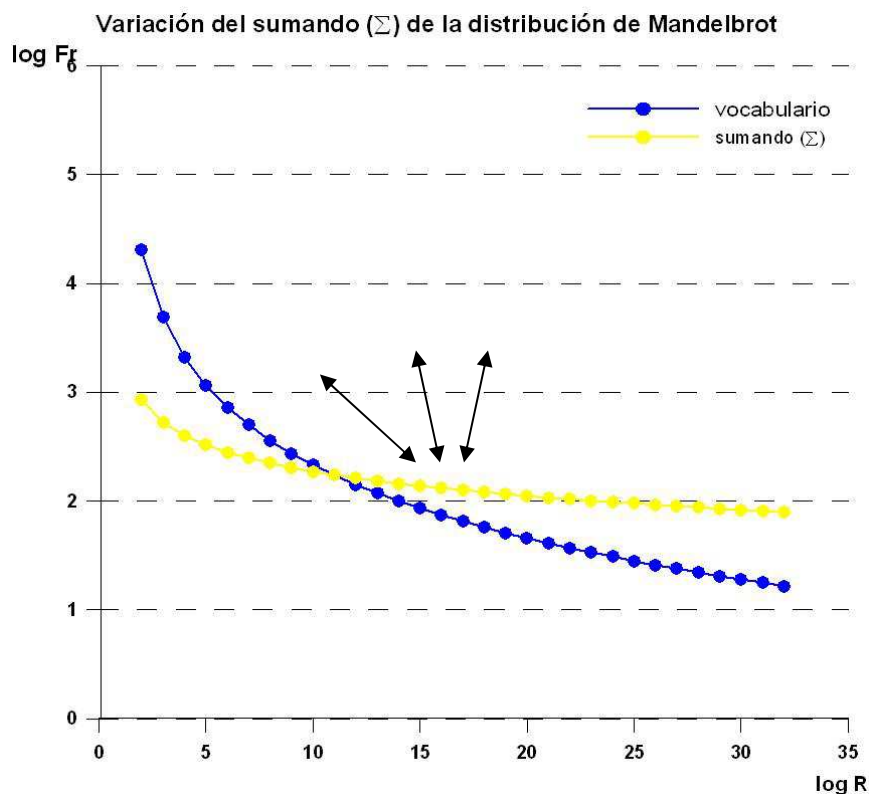


Gráfico 49. Tendencia del Σ en la distribución de Mandelbrot

6.3.4.3. Comparativa de gráficas sintéticas para Zipf-Mandelbrot con Zipf como medio de visualización de la tendencia para el exponente (e) y el sumando (Σ).

Una vez vista la tendencia tanto con la fórmula de Zipf como con la fórmula de Mandelbrot realizaremos las comprobaciones para distinto (e) y distinto (Σ), con la finalidad de ajustarlo lo máximo posible y que se asemeje a la fórmula de Zipf.

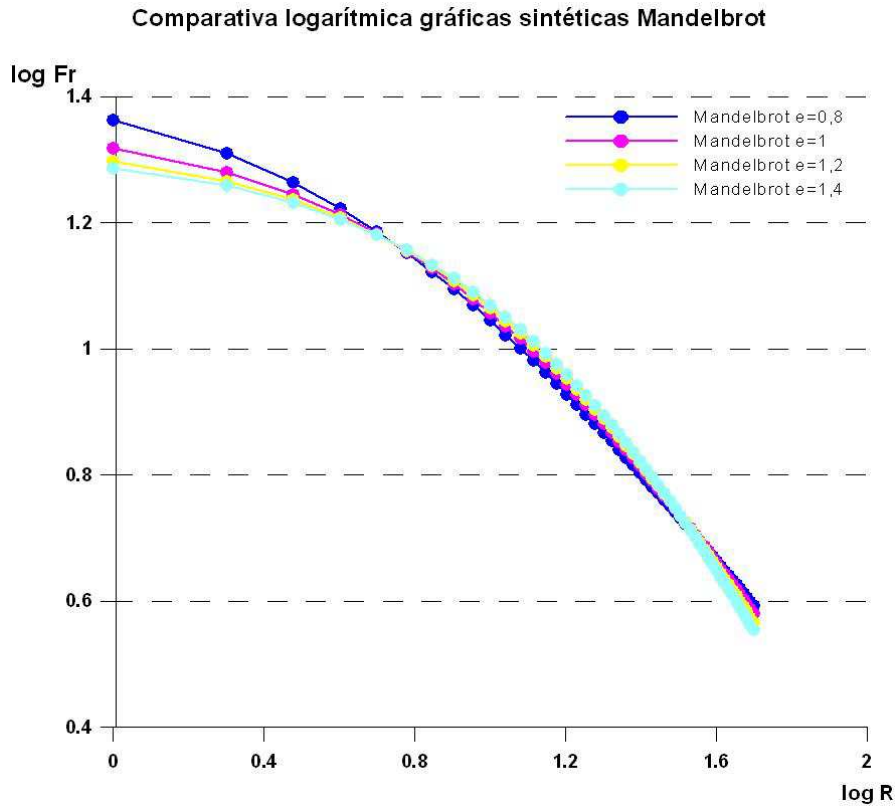
Para demostrar visualmente cuales son las mayores posibilidades que Zipf-Mandelbrot con dos parámetros: (e) y (Σ), ofrece sobre Zipf con un solo parámetro: (e), vamos a representar varias distribuciones de Zipf-Mandelbrot a las que se ajusta la misma distribución de Zipf. Para conseguirlo basta, con algunos tanteos, repetir la experiencia anterior hasta conseguir varios ejemplos.

Si tenemos una distribución de Zipf-Mandelbrot con sumando ¿Cuál es la de Zipf que más se le parece? Ya que de las gráficas anteriores vemos que al incrementar el sumando el efecto que hace es disminuir las frecuencias altas y aumentar un poco las bajas; Zipf no puede hacer algo exactamente igual, pero lo mas parecido es lo que ya hemos visto en las primeras gráficas, que es disminuir el exponente, así por ejemplo, para ZM $e=1$ $\Sigma=10$ seguimos el procedimiento y obtenemos 0,55 para el exponente de Zipf.

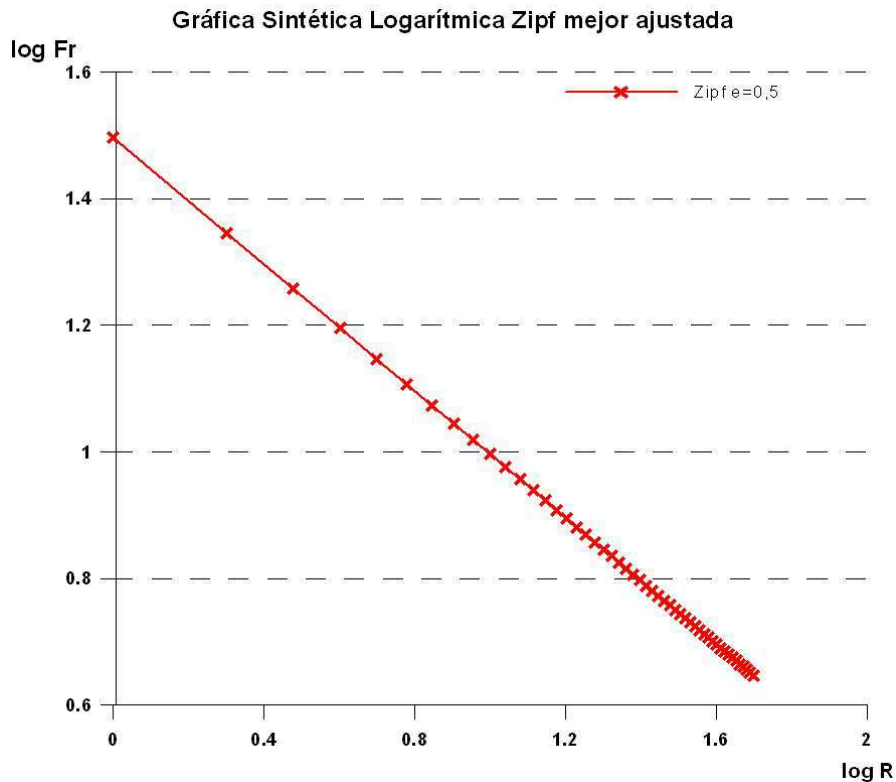
En concreto, tenemos las distribuciones de Zipf-Mandelbrot con los siguientes parámetros, estos se dibujan en la siguiente gráfica utilizando la representación logarítmica.

(e)	(Σ)	K
0,8	5	96,82194
1	10	228,4536
1,2	15	552,7856
1,4	20	1373,934

Tabla 14. Parámetros de la distribución de Zipf-Mandelbrot



A cada una de estas la de Zipf que mejor se ajusta es la de exponente 0,5 que se dibuja en la siguiente gráfica en este caso también utilizando la representación logarítmica



La representación conjunta de las distribuciones Zipf-Mandelbrot comparándola con la distribución de Zipf mejor ajustada se observa en el siguiente gráfico.

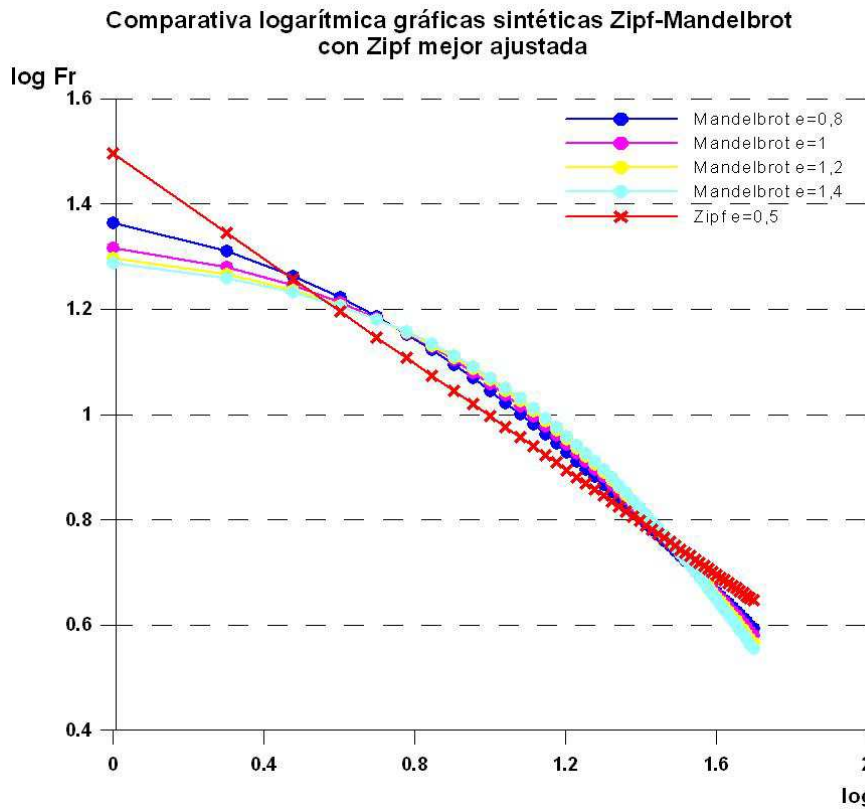


Gráfico 52. Gráficas sintéticas Zipf-Mandelbrot. Comparativa logarítmica Zipf ajustada

El gráfico anterior podemos visualizarlo igualmente utilizando la representación clásica. Vemos como el efecto es el de deformar la curva haciendo más altos los valores de las frecuencias intermedias y más bajos los de las extremas, en distinto grado según los valores de los parámetros.

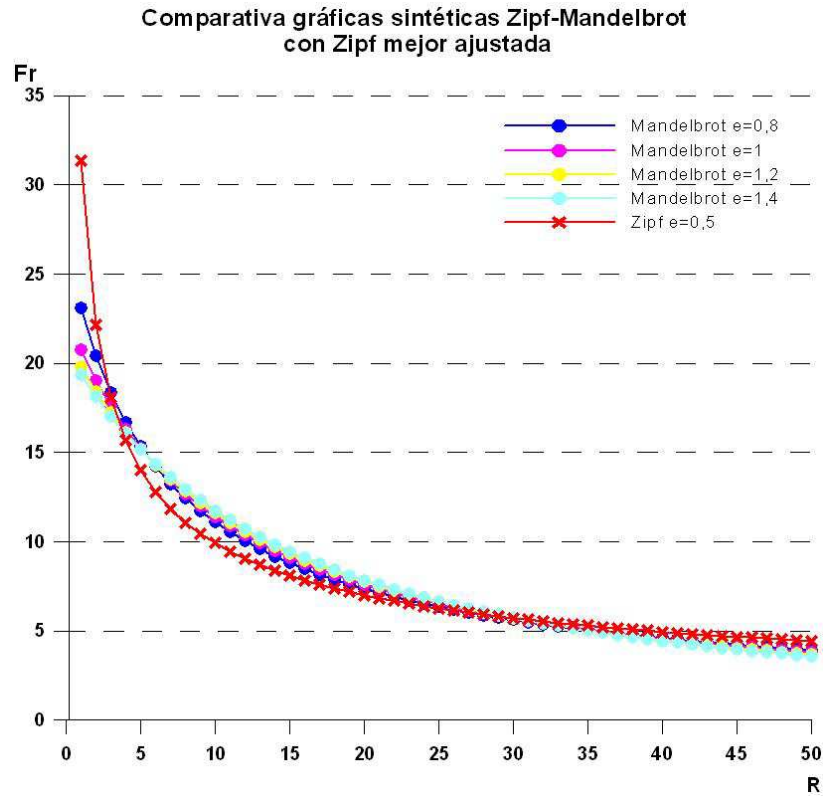


Gráfico 53. Gráficas sintéticas Zipf-Mandelbrot. Comparativa Zipf ajustada

En resumen, podemos afirmar que utilizar una fórmula de Mandelbrot con un sumando y a cambio aumentar el exponente para seguir ajustándose al mismo texto, lo que ocasiona son mayores valores de las frecuencias intermedias en detrimento de las más altas y de las más bajas.

6.4. Valores de los parámetros en las fórmulas de Zipf/Mandelbrot para el ajuste a las frecuencias de las palabras de un texto

La averiguación de los valores de los parámetros tanto del exponente como de los coeficientes en las fórmulas de Zipf y Mandelbrot se ha llevado a cabo mediante un proceso o técnica estándar que nos permitirá obtener un valor verosímil del exponente (e) y coeficientes, ajustado matemáticamente que permitiría ofrecer una predicción del valor de dichos parámetros y de este modo se podrán comparar los resultados obtenidos de las predicciones mejor ajustadas y de los datos obtenidos de un texto real.

Para determinar dichos valores de los parámetros en las fórmulas de Zipf y Mandelbrot y conseguir de este modo su ajuste a las frecuencias de las palabras en un texto, el método empleado es un proceso de cuatro pasos que se detalla pormenorizadamente en el punto 6.8 de este capítulo *Estudios complementarios: Cálculos fórmulas ajustadas de Zipf y Mandelbrot. Pasos 1-4*. Mediante la utilización de esta técnica estándar nos permitirá comparar la tendencia de los valores de los datos reales de los textos y los valores de los datos mejor ajustados a las fórmulas de Zipf y Mandelbrot, para estudiar finalmente las leyes de Zipf y Mandelbrot en qué aspecto se ajusta más o menos a la realidad.

A continuación se muestran los resultados obtenidos para que, a la vista de ellos, se juzgue hasta qué punto se ajustan las fórmulas a los valores reales en los textos. Vamos a realizar un análisis de resultados de predicciones de Zipf y Mandelbrot, ¿Hasta qué punto se ajusta un texto real a la fórmula de Zipf?

En un primer ejemplo utilizamos las frecuencias de las palabras de un texto grande texto: *vari1.txt* tipología: literario-varios autores, con un total de 379.945 palabras y con un vocabulario de 48.389 palabras. El error obtenido es el error relativo cuadrático promedio que se obtiene:

Error = valor predicho - valor real

Error relativo = error / valor predicho

Error relativo cuadrático = error relativo elevado al cuadrado

Error relativo cuadrático promedio = suma de los errores relativos cuadráticos para cada uno de los rangos, dividida por el total de ellos, es decir, por el vocabulario

Tras aplicar el formulario AnalizaZipf de TOPOS se obtienen los siguientes resultados:

	Parámetros fórmulas ajustadas
Zipf	$e = 0,9652$ $K = 27811$ $error = 0,076$
Mandelbrot	$e = 1,2618$ $K = 568646,9$ $\Sigma = 254,8$ $error = 0,031$

Tabla 15. Resultados de los parámetros fórmulas Zipf y Mandelbrot ajustadas

Debido al tamaño del texto analizado y por ello lo exagerado de la colección de valores, donde la mayoría de palabras tienen frecuencia igual a 1, 2, 3,..., no permitiría una apreciación visual, es por ello que para una comparativa entre Zipf, Mandelbrot y datos reales del texto, se utiliza la escala logarítmica en el siguiente gráfico.

Las siguientes gráficas muestran para un texto de $P = 379.945$ y $V = 48.389$, la distribución de los valores reales del texto, por otro lado la distribución según la fórmula de Zipf y por último la distribución según la fórmula de Mandelbrot.

Comparación logarítmica de las distribuciones con valores de textos reales y valores predichos por las fórmulas de Mandelbrot y Zipf mas ajustada

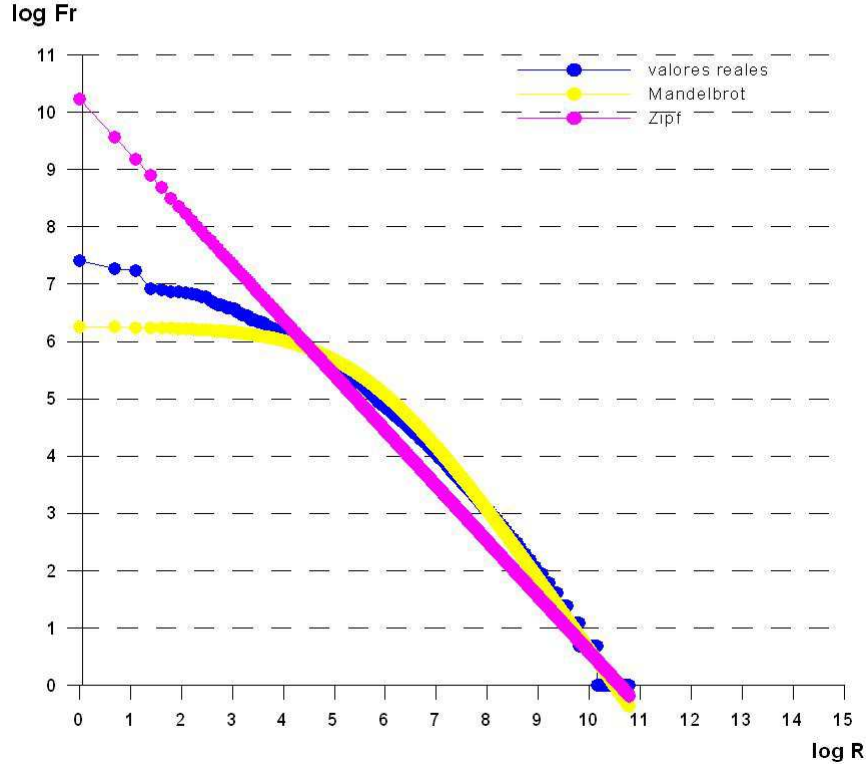


Gráfico 54. Comparación logarítmica distribuciones Zipf, Mandelbrot y Zipf ajustada

Seguidamente para tratar de comprender la tendencia real de estas tres distribuciones puestas al común se amplían los resultados de las gráficas anteriores. Se amplían valores desde los rangos 11 al 50, en este caso las gráficas siguientes utilizan la distribución clásica de Zipf y no la logarítmica.

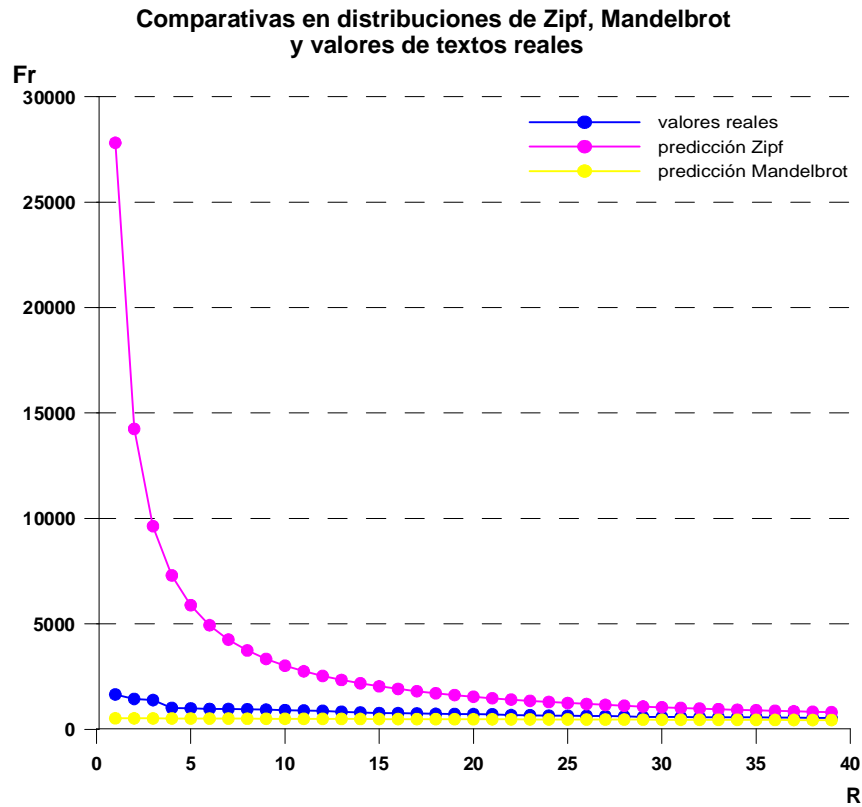


Gráfico 55. Comparación distribuciones Zipf, Mandelbrot y textos reales

En esta ampliación vemos que, en frecuencias muy altas la predicción de Zipf queda muy por encima de los valores reales y la de Mandelbrot por debajo. En rangos intermedios, no muy altos, las discrepancias ya se han invertido quedando la predicción de Mandelbrot por encima y la de Zipf por debajo de los valores reales. Seguidamente se amplían valores desde los rangos 1000 al 1200.

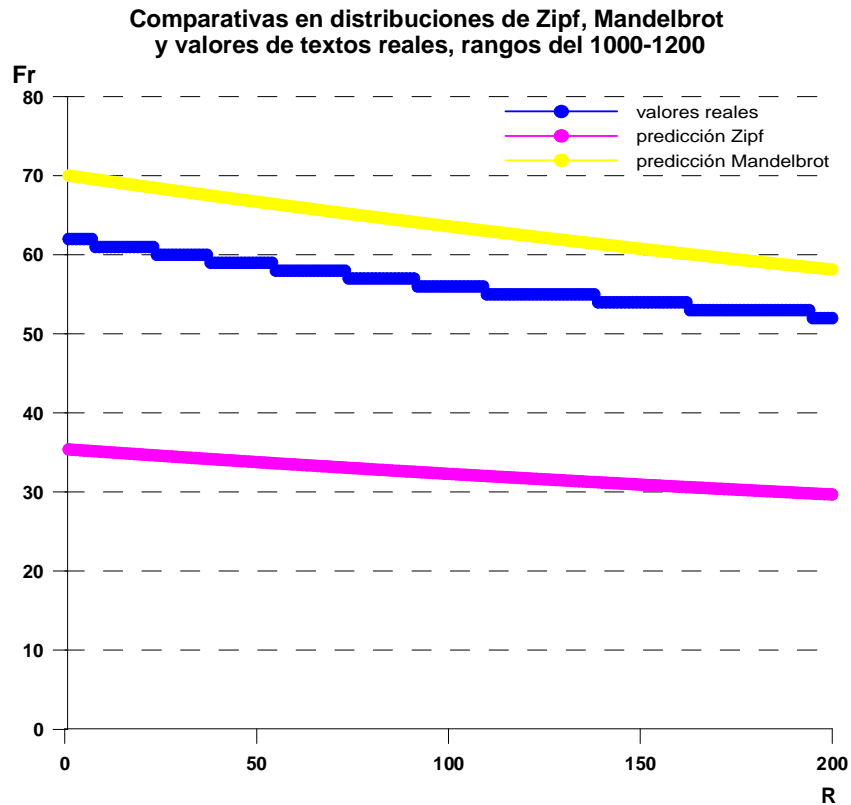


Gráfico 56. Comparación Zipf, Mandelbrot y textos reales rangos 1000-1200

Para rangos altos, acercándose al 25% del vocabulario (que en este ejemplo corresponden a palabras de frecuencia 5), ambas predicciones quedan por debajo de los valores reales. A continuación se amplían valores desde los rangos 10000 al 10200.

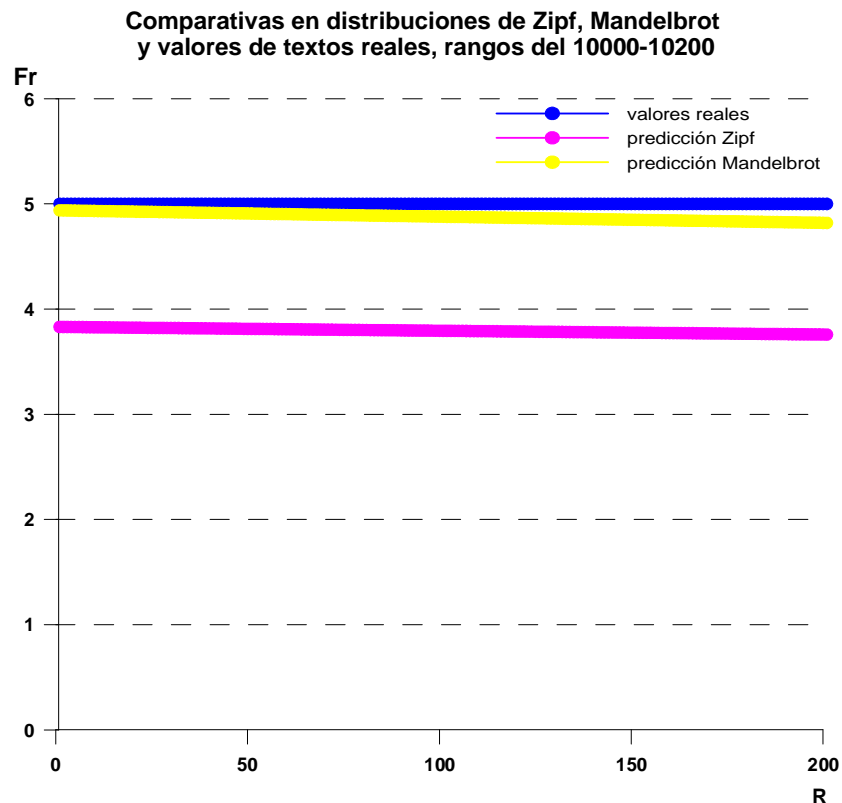


Gráfico 57. Comparación Zipf, Mandelbrot y textos reales rangos 10000-10200

En rangos muy elevados, correspondientes a palabras de frecuencia 1, las dos predicciones quedan por encima de los valores reales. Si ampliamos valores desde rangos 30000 al 30200.

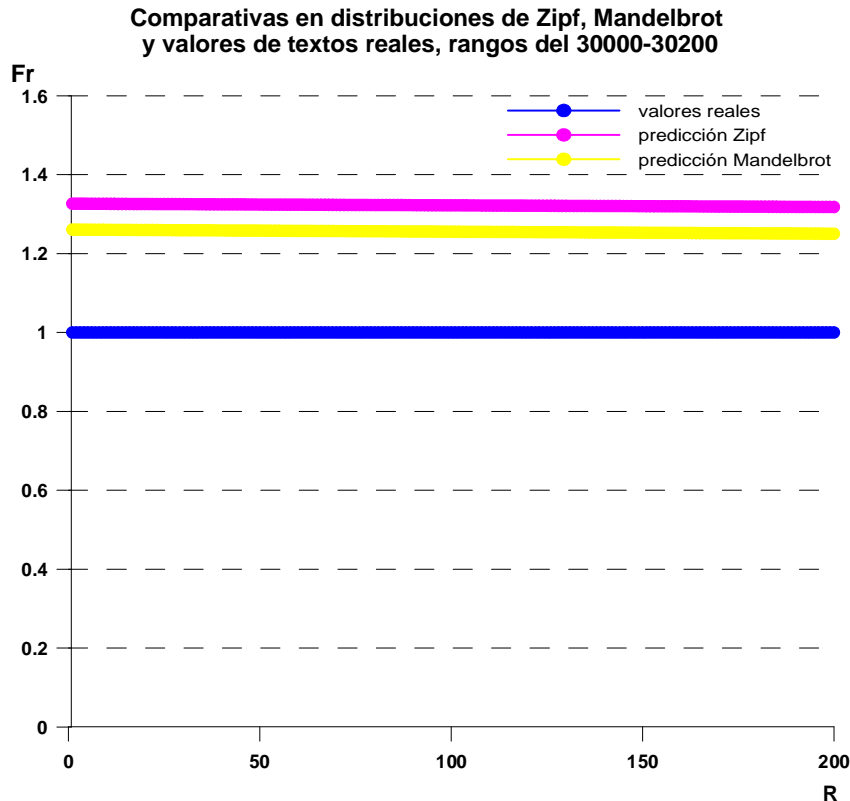


Gráfico 58. Comparación Zipf, Mandelbrot y textos reales rangos 30000-30200

Sorprendentemente, observamos que en rangos aún más elevados, correspondientes también a palabras de frecuencia 1 las dos predicciones quedan por debajo de los valores reales. Finalmente si ampliamos todavía más los rangos observados para averiguar si continúa la tendencia del gráfico anterior frente a las demás, por ello se amplía valores desde rangos 48100 al 48300

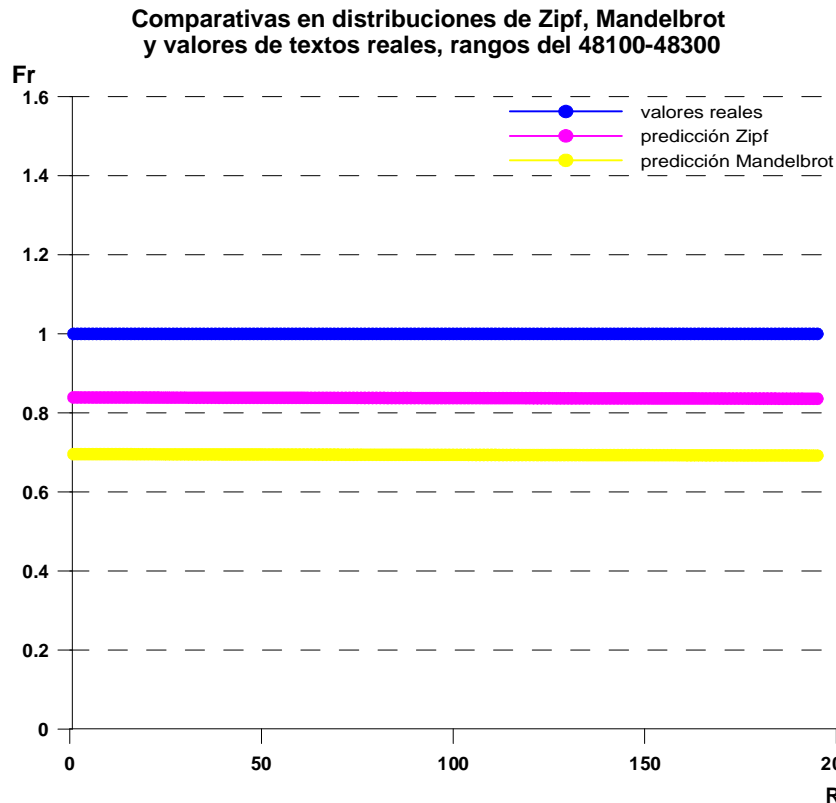


Gráfico 59. Comparación Zipf, Mandelbrot y textos reales rangos 48100-48300

Observamos que en rangos más elevados, correspondientes a palabras de frecuencia 1 las dos predicciones quedan igualmente por debajo de los valores reales.

Todo ello puede verse de manera muy comprimida si pasamos a escala logarítmica en los dos ejes, obteniendo así una visión de conjunto de lo explicado y utilizando la forma habitual de representación como es la escala logarítmica.

En la gráfica siguiente sólo mostramos la distribución logarítmica de los valores reales del texto analizado con un total de $P = 379.945$ y $V = 48.389$.

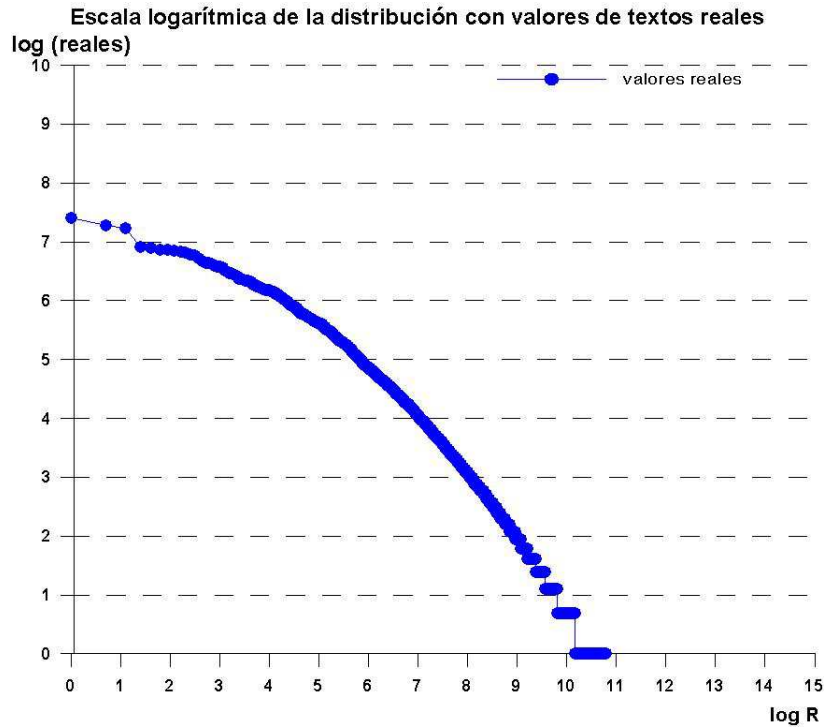


Gráfico 60. Distribución logarítmica de Zipf con textos reales

En la gráfica siguiente se realiza la comparación con los valores predichos por la fórmula de Zipf mas ajustada. Dado que la fórmula de Zipf, función potencial, debe dar una recta al dibujarla en ejes log-log y que los valores reales dibujan claramente una curva se comprenden las discrepancias en el ajuste.

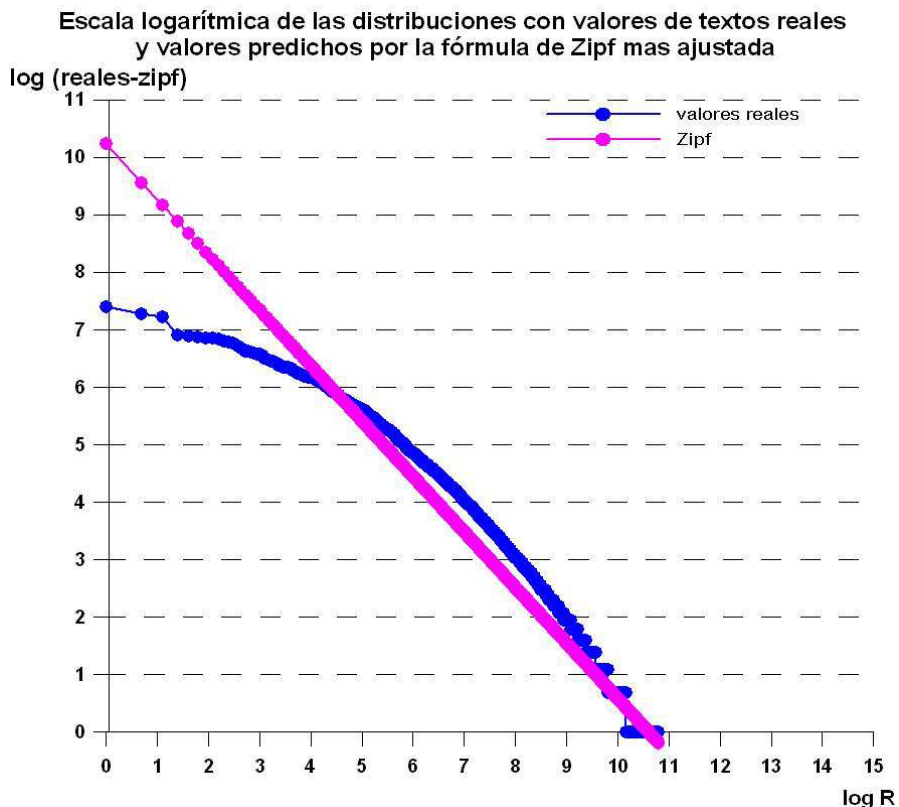


Gráfico 61. Distribución logarítmica de Zipf con textos reales y Zipf ajustada

En la gráfica siguiente se realiza igualmente la comparación con los valores predichos por la fórmula de Mandelbrot más ajustada. Esta fórmula permite una modificación limitada de la curvatura y así se ajusta más a los datos reales. Debido a la forma de definir el ajuste, minimizando la suma de los errores correspondientes a cada una de las palabras, en la deformación visual que provoca la escala logarítmica nos parece que se ha preferido ajustar a las palabras de baja frecuencia.

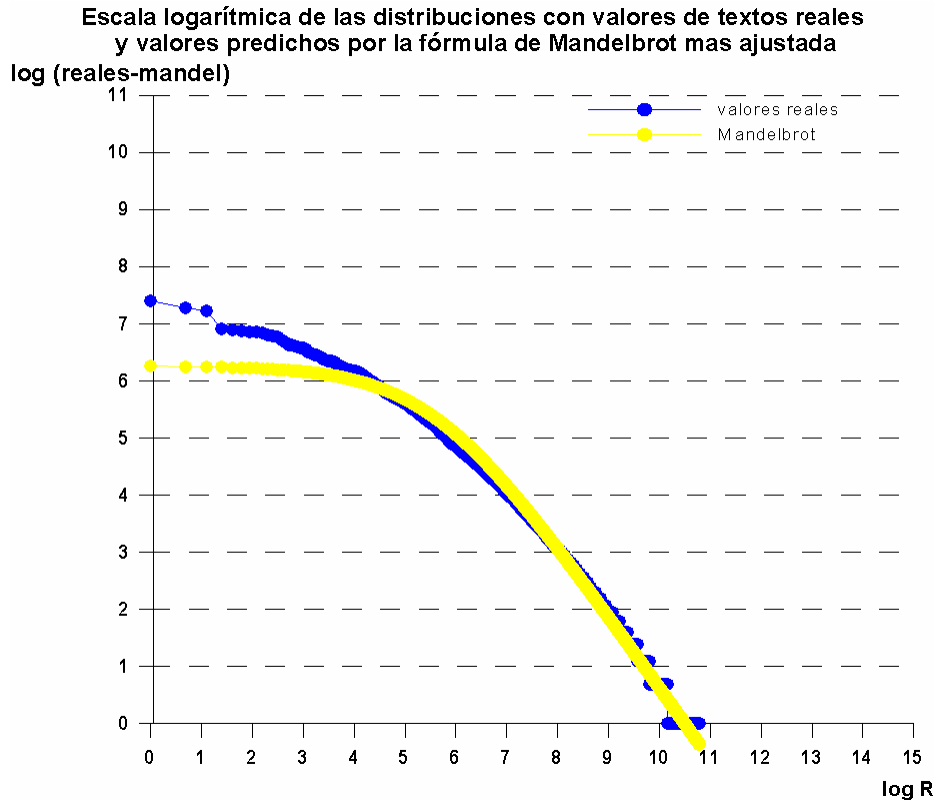


Gráfico 62. Distribución logarítmica de Zipf con textos reales y Mandelbrot ajustada

6.4.1. Representación Transformada. Visualización de ejemplos

Visualizaremos ejemplos para ver la comparación entre los valores reales de un texto y los predichos por la fórmula de Zipf mejor ajustada en la Representación Transformada o segunda Ley de Zipf.

Si recordamos la Representación Transformada de Zipf o también llamada segunda Ley de Zipf, otra forma clásica de presentación que reduce el volumen de datos es mostrar cuántas palabras hay de frecuencia 1, cuántas de frecuencia 2, etc (o también puede definirse un intervalo, por ejemplo 5, y mostrar cuántas palabras hay de frecuencias entre 1 y 5, cuántas de 6 a 10, etc). Estas series decrecientes de valores también pueden ser aproximadas por fórmulas de Zipf o Mandelbrot, con valores de sus parámetros distintos a los anteriores.

En el presente trabajo no se va a utilizar estas representaciones transformadas, por lo que nos limitamos a exponer su aspecto general sobre el mismo ejemplo. En la primera parte de la gráfica, en este caso sin utilizar la escala logarítmica para su visualización, correspondiente a frecuencias de 1 a 100, donde se confunden los valores reales y los predichos por la fórmula de Zipf más ajustada

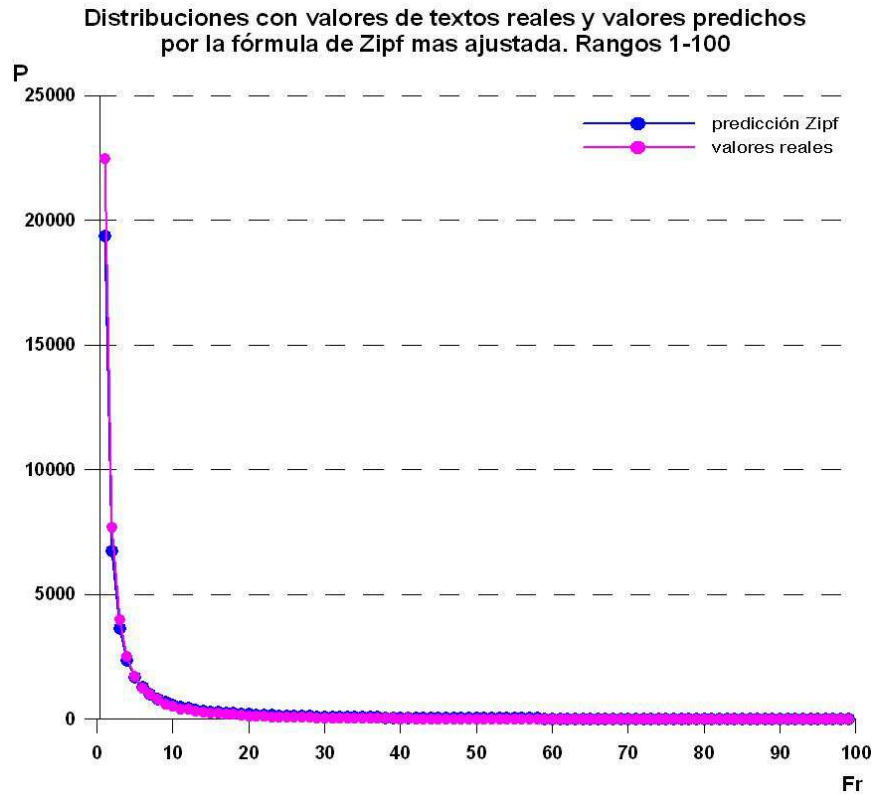


Gráfico 63. Representación Transformada. Distribución de Zipf con textos reales

Ampliación de una parte intermedia de la gráfica, frecuencias 500 a 600, donde los valores reales presentan discontinuidades cuando no hay ninguna palabra con un valor dado de frecuencia

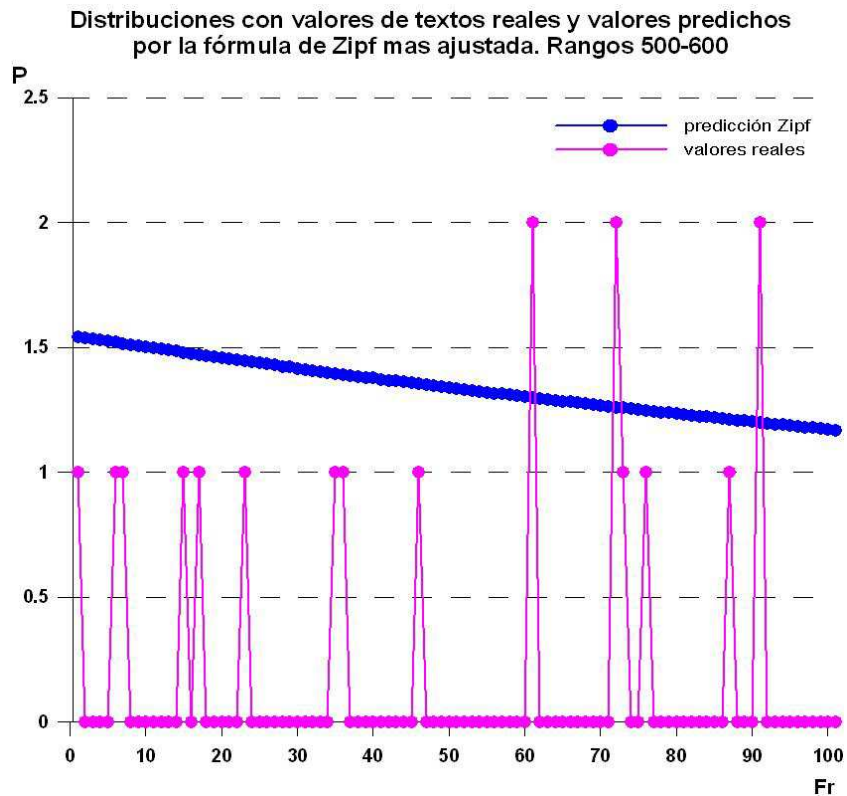


Gráfico 64. Representación Transformada. Distribución de Zipf con textos reales y Zipf ajustada

Finalmente, añadir que podrían hacerse agrupamientos, por ejemplo palabras de frecuencia entre 500 y 550, pero la escala de agrupamiento debería ser distinta según rangos de frecuencia; requeriría demasiada complicación para tener un método intuitivo de visualización, por eso se descarta a favor del método que se expone mas adelante, con el nombre Log-%.

6.5. Variación de los valores de los parámetros según el tipo de texto, según el tamaño del texto y según el tratamiento dado al texto

Hasta este punto en el que nos encontramos se ha interpretado con mucho detalle y profundidad la teoría de Zipf y Mandelbrot y sus distintos tipos de representación, hemos visualizado la tendencia para el exponente y el sumando en las fórmulas y hemos contemplado como conseguir su ajuste a las frecuencias de las palabras de un texto.

A partir de ahora observaremos con más detalle el comportamiento del exponente según la tipología del texto, el tamaño y el tratamiento dado a dicho texto. Realizaremos un ajuste parcial de Zipf por tramos, debido a la propia ambigüedad del estudio de las palabras; si damos más importancia a las palabras distintas, o a todas las palabras que aparecen, o si ajustamos una función que tiene decimales a unos valores que son enteros por definición (la frecuencia de una palabra es 2 ó 3, pero nunca 2,5), toda esta ambigüedad es lo que nos autoriza a explorar distintos criterios y lo expresaremos por medio de tramos que consideraremos en los análisis y tramos que no, como por ejemplo podemos decidir como una posibilidad, a complementar con otras, no tener en cuenta las palabras de baja frecuencia. En definitiva a partir de este momento ya no se analizan los textos en su conjunto sino que se ajustará por tramos dependiendo de lo que se quiera priorizar.

Finalmente se expondrá una nueva forma de representación, un modelo propio desarrollado en esta investigación y el cual se ha denominado Log-%.

6.5.1. El comportamiento del exponente (e) de Zipf en los documentos

El comportamiento del exponente (e) en la fórmula de Zipf es estudiado íntegramente para ver si el valor de (e) dependerá del tamaño del texto y de su tipología y lo que es más importante si el valor del exponente es significativo de una variedad concreta de documentos según su temática y tipología. Esta idea nos lleva a afirmar que la duda no es si se cumple la ley de Zipf o no, ya que estudios demuestran que esta se cumple igualmente con textos aleatorios, es decir utilizando un lenguaje inventado, seleccionado aleatoriamente de un conjunto de palabras dado, (Wentian, 1992).

En la literatura publicada sobre este tema específico se han elaborado distintos modelos que justifican o que tratan de discernir los motivos por los que debe cumplirse la ley de Zipf y precisamente con exponente cercano a 1, que aumenta ligeramente al aumentar el tamaño del texto. En general observamos que en la literatura publicada se asegura que el valor de (e) depende del tamaño, al igual que dicho valor por regla general está alrededor de 1.

Según Wentian (1992), mencionado anteriormente los textos generados aleatoriamente cumplen la Ley de Zipf con exponente ligeramente superior a 1, aún variando la forma de generar el texto, el exponente crece todo lo máximo hasta 1,5.

En este mismo sentido Debowski (2002), combina un modelo matemático con evidencia experimental para mostrar que hay un tamaño de texto para el que el exponente de Zipf vale 1 y a partir de aquí, al aumentar el tamaño, el valor del exponente aumenta ligeramente.

Así Ferrer i Cancho (2005), utilizando un modelo de teoría de la información justifica que debe cumplirse la fórmula de Zipf con exponente próximo a 1 y por qué en casos anómalos el exponente se aparta de ese valor²⁸.

Según Powers (1998), basándose en un modelo de coste de utilización de cada palabra (como el esfuerzo de recordar una palabra y el esfuerzo de escribirla), justifica que su distribución corresponde a la fórmula de Zipf con exponente aproximado = 1. Powers es uno de los primeros autores que observó que el exponente de Zipf aumenta con el tamaño de los documentos, es el primer estudio que se conoce al respecto aunque este no ofrece datos concretos y ejemplos muy detallados.

En realidad, se demuestra experimentalmente y se demuestra en los modelos teóricos, que aumentando el tamaño del texto se alcanza siempre un tamaño donde el exponente vale exactamente 1 y a partir de aquí los textos mayores tienen exponente >1 Según Debowski (2002)

Dado que diversas publicaciones inciden en que el valor del exponente en la fórmula de Zipf sobre todo en tamaños mayores es mayor a 1, vemos que efectivamente y según las experimentaciones realizadas el exponente tiene un valor alrededor de 1.

En las gráficas expuestas anteriormente, las inclinaciones en la distribución de frecuencias que aparecen al principio y al final de ésta se han explicado por el tratamiento del texto (Baayen, 1991), aunque este argumento es parcial y requiere aportar más elementos.

Pero la conclusión a la que se pretende llegar es descubrir si el valor que obtenemos del exponente de Zipf ¿está asociado a la tipología y/o tamaño del texto o dependerá de otros aspectos muy distintos?, esta hipótesis inicial es la que vamos a corroborar en breve con los experimentos realizados, así las conclusiones mostrarán que el valor obtenido del exponente de Zipf, dependerá no sólo de la tipología del documento (es decir, si es un documento de tipo literario, científico, etc.) sino del tratamiento que realicemos con él, como apunta Baayen (1991) haciendo referencia a las inclinaciones o curvaturas del principio y final que aparecen en las gráficas y que según el autor son la causa del tratamiento dado al documento, así el valor del exponente dependerá de la cantidad de palabras vacías y raíces que se utilicen en el proceso de análisis del texto.

²⁸ Este autor utiliza la versión transformada de Zipf por esa razón en el artículo se habla de $\beta=2$ que resulta equivalente a exponente $(e) =1$

Respecto al tratamiento dado al texto, cada analista puede utilizar una cantidad de palabras vacías y un método distinto de extracción de raíces o lematización, de modo que el tema que nos ocupa, la eliminación de las palabras vacías como parte esencial de la metodología antes de realizar recuentos y obtener las conclusiones oportunas respecto a qué influye en el valor del exponente, es un hecho que no se le da la importancia que merece ya que de ello depende en gran medida los valores de los coeficientes de las fórmulas. Es decir, en la mayoría de estudios se escoge una cantidad fija de eliminación de palabras vacías y un procedimiento fijo de obtención de raíces, etc. Y se aplica a las fórmulas obteniendo unos valores de los parámetros demostrativos de que los valores obtenidos dependen del tipo de documento analizado, es decir de si los documentos son de tipo científico, literario, técnicos: patentes, etc. Hemos descubierto que el valor de los coeficientes en las fórmulas objetivamente dependen poco de la tipología del documento analizado, sino del tratamiento dado al documento, es decir de si se eliminan más o menos palabras vacías, o si el proceso de extracción de raíces es más exagerado o menos. Entonces que un autor escriba más palabras significativas por página que otro autor, depende más del criterio del analista sobre qué palabras son significativas, que de la forma de escribir del autor, los experimentos con tablas de palabras vacías distintas, parecen dar cuenta de esta variación en los coeficientes.

Así pues, planteamos tres hipótesis iniciales de las que realizaremos una comprobación directa.

1. El valor del exponente es aproximadamente igual a 1.
2. El valor del exponente depende del tamaño del texto
3. El valor del exponente depende del tratamiento previo del texto

Incluso vamos a aportar una cuarta comprobación que respondería a una cuarta hipótesis

4. El valor del exponente depende de la tipología del texto

Respecto a la hipótesis planteada número 4: “El valor del exponente depende de la tipología del texto”

Se efectúan varios ensayos en diversos documentos para comprobar si el valor del exponente es característico del tipo de texto, es decir si los valores de (e) dependen de si el documento a tratar es de tipo literario, científico, etc. para ello a los documentos objeto de estudio se les ha quitado las palabras vacías y se ha extraído las raíces por el método de sufijos de Lovins con algunas adaptaciones al Español, el cual se aborda ampliamente en el capítulo siete de esta tesis doctoral.

PROCEDIMIENTO FIJO									
<i>Tabla palabras vacías</i>	<i>Tabla raíces</i>	<i>Texto</i>	<i>Tipología</i>	<i>Total palabras P</i>	<i>Total vocabulario V</i>	<i>Total raíces R</i>	<i>Valor (e)-raíces</i>	<i>Valor (K)-raíces</i>	<i>error-raíces</i>
386	358	aralpi.txt	Científico-legal	12.862	2.758	1.709	0,894- 0,676	993,007- 176,400	0,055- 0,191
386	358	bachi1.txt	Científico-legal	73.173	9.161	5.255	0,956-0,880	6203,283-2394,459	0,148-0,345
386	358	Cali1.txt	Científico- Resum	22.133	4.309	3.725	0,923-0,683	1797,144-263,6811	0,043-0,130
386	358	Cali2.txt	Científico- Resum	11.745	3.054	2.667	0,864- 0,578	780,6323- 88,00017	0,035- 0,086
386	358	Cali3.txt	Científico- Resum	7.621	2.152	1.906	0,838- 0,503	483,8007- 41,56239	0,031- 0,063
386	358	Cali4.txt	Científico- Resum	2.155	908	825	0,738- 0,272	110,9651- 5,671521	0,038- 0,027
386	358	cienti1.txt	Científico- Resum	101.164	13.668	9.915	1,001- 0,850	10094,95- 2070,099	0,033- 0,098
386	358	icyt3.txt	Científico- Resum	8.148	2.294	1.980	0,880- 0,526	612,5018- 46,33439	0,045- 0,048
386	358	icyt4.txt	Científico- Resum	66.260	10.515	7.933	0,976- 0,804	6055,22- 1138,598	0,034- 0,093
386	358	imaa2.txt	Científico- Resum	86.931	13.476	7.033	0,952- 0,926	6888,101- 4166,73	0,074- 0,461
386	358	imab2.txt	Científico- Resum	86.349	14.437	7.567	0,935- 0,917	6231,976- 3968,635	0,070- 0,398
386	358	imac2.txt	Científico- Resum	46.362	9.545	5.472	0,894-0,843	2892,845-1544,084	0,057-0,311
386	358	ime1.txt	Científico- Resum	12.345	2.280	1.986	0,998-0,615	1477,197-88,97209	0,040-0,067

PROCEDIMIENTO FIJO									
<i>Tabla palabras vacías</i>	<i>Tabla raíces</i>	<i>Texto</i>	<i>Tipología</i>	<i>Total palabras P</i>	<i>Total vocabulario V</i>	<i>Total raíces R</i>	<i>Valor (e)-raíces</i>	<i>Valor (K)-raíces</i>	<i>error-raíces</i>
386	358	ime2.txt	Científico- Resum	55.696	6.324	5.164	1,060-0,813	7561,628-930,3285	0,033-0,118
386	358	imea2.txt	Científico- Resum	95.374	16.448	8.886	0,924-0,920	6402,317-4508,252	0,068-0,353
386	358	imeb2.txt	Científico- Resum	81.463	14.721	8.058	0,918- 0,907	5386,514- 3546,305	0,069- 0,337
386	358	imec2.txt	Científico- Resum	79.744	14.765	8.020	0,909- 0,900	5022,677- 3433,016	0,066- 0,327
386	358	isoc2.txt	Científico- Resum	70.158	13.999	7.994	0,896- 0,873	4171,534- 2572,407	0,059- 0,259
386	358	legal1.txt	Científico- legal	97.670	11.080	5.636	0,998- 0,838	9830,914- 2464,791	0,108- 0,177
386	358	medico1.txt	Científico- Resum	67.818	7.170	5.857	1,068- 0,833	9373,55- 0,833	0,032- 0,115
386	358	natalia.txt	Científico- Patentes	104.536	16.994	12.893	0,973- 0,877	8964,082- 2825,911	0,053- 0,115
386	358	paten1.txt	Científico- Patentes	78.514	11.386	6.471	0,965- 0,914	6763,432- 2968,514	0,060- 0,293
386	358	sonia.txt	Científico-Monografía	32.852	6.348	3.753	0,913- 0,812	2405,956- 792,6857	0,063- 0,217
386	358	upala.txt	Científico- Resum	76.990	13.273	7.136	0,926- 0,905	5392,343- 3365,175	0,070- 0,384
386	358	upalb.txt	Científico- Resum	73.560	13.970	7.448	0,818- 0,898	2110,272- 3184,969	0,122- 0,367
386	358	upalc.txt	Científico- Resum	70.523	11.598	6.134	0,942- 0,910	5443,08- 3174,542	0,070- 0,441

PROCEDIMIENTO FIJO									
<i>Tabla palabras vacías</i>	<i>Tabla raíces</i>	<i>Texto</i>	<i>Tipología</i>	<i>Total palabras P</i>	<i>Total vocabulario V</i>	<i>Total raíces R</i>	<i>Valor (e)-raíces</i>	<i>Valor (K)-raíces</i>	<i>error-raíces</i>
386	358	calderon.txt	Literario-Calderón	172	142	130	0,268-0	3,398928-1	0,036-0
386	358	castea.txt	Literario- E. Castelar	74.939	17.979	8.600	0,857- 0,858	3416,215- 2699,391	0,050- 0,283
386	358	casteb.txt	Literario- E. Castelar	93.270	20.385	9.282	0,871- 0,878	4515,609- 3596,761	0,054- 0,303
386	358	castec.txt	Literario- E. Castelar	89.556	16.794	7.438	0,868- 0,868	4402,016- 3404,714	0,051- 0,390
386	358	casted.txt	Literario- E. Castelar	86.666	19.315	8.184	0,871- 0,878	4239,11- 3469,074	0,051- 0,385
386	358	castee.txt	Literario- E. Castelar	106.079	16.488	6.790	0,950- 0,928	8131,483- 4804,206	0,056- 0,523
386	358	castela.txt	Literario- E. Castelar	450.574	43.816	16.092	0,995- 1,053	39096,97- 36035,28	0,093- 0,617
386	358	cervantes.txt	Literario- Cervantes	385.213	28.317	11.703	1,086- 1,041	52094,31- 19687,18	0,062- 0,593
386	358	costa1.txt	Literario- J. Costa	279.141	42.169	17.790	0,938- 0,988	17872,55- 17188,66	0,070- 0,347
386	358	costaa.txt	Literario- J. Costa	93.140	20.798	9.756	0,869- 0,879	4451,41- 3562,564	0,058- 0,270
386	358	costab.txt	Literario- J. Costa	94.283	20.147	9.536	0,882- 0,889	4885,789- 3740,202	0,056- 0,290
386	358	costac.txt	Literario- J. Costa	91.714	24.533	11.925	0,830- 0,852	3353,891- 3093,15	0,044- 0,199
386	358	dos.txt	Literario- E. Castelar	73.790	17.171	9.089	0,863- 0,832	3522,069- 1972,973	0,047- 0,190

PROCEDIMIENTO FIJO									
Tabla palabras vacías	Tabla raíces	Texto	Tipología	Total palabras P	Total vocabulario V	Total raíces R	Valor (e)-raíces	Valor (K)-raíces	error-raíces
386	358	espiso1.txt	Literario-B.P. Galdós	81.398	17.217	7.561	0,889- 0,880	4490,578- 3140,067	0,050- 0,362
386	358	episogrande.txt	Literario-B.P. Galdós	204.757	29.554	12.031	0,953- 0,980	14879,08- 12078,55	0,063- 0,461
386	358	erudi1.txt	Literario-Menéndez P.	53.940	13.470	7.429	0,852- 0,818	2522,069- 1575,518	0,050- 0,228
386	358	fpoe.txt	Literario-poesía	609	412	360	0,489-0	14,8973-1	0,043-0
386	358	gg1.txt	Literario-Gabriel Galán	6.821	3.171	2.004	0,690- 0,534	187,81- 57,63411	0,044- 0,086
386	358	ksoti.txt	Literario-J.M ^a de Pereda	56.854	14.045	6.899	0,866- 0,832	2857,434- 1740,304	0,045- 0,255
386	358	ksoti1.txt	Literario-J.M ^a de Pereda	28.582	9.370	5.184	0,798- 0,741	1063,734- 585,9381	0,045- 0,169
386	358	ksoti2.txt	Literario-J.M ^a de Pereda	28.279	8.555	4.596	0,827- 0,752	1263,232- 608,7352	0,046- 0,200
386	358	larra1.txt	Literario-Larra	76.550	16.521	7.841	0,884- 0,869	4136,664- 2725,862	0,056- 0,301
386	358	seis.txt	Literario-San F. Borja	25.696	9.319	5.209	0,768- 0,706	798,6755- 407,3827	0,043- 0,114
386	358	vari.txt	Literario-varios	95.045	20.132	9.530	0,887- 0,873	5055,996- 3224,317	0,056- 0,241
386	358	vari1.txt	Literario-varios	380.054	48.636	19.426	0,963- 1,018	27601,41- 25912,22	0,075- 0,391

Tabla 16. Relación del exponente en la fórmula de Zipf con la tipología del texto

Las tablas muestran los resultados obtenidos con dos grandes grupos de textos, los que pertenecen a una tipología científica y los que pertenecen a la tipología literaria, pero mejor si vemos los resultados obtenidos mediante las siguientes gráficas

Aunque cabría pararse a pensar qué características tiene un texto científico respecto a un texto literario, siempre hablando en términos generales, porque siempre existe la excepción²⁹ y en este caso lo podremos comprobar.

Según hemos estudiado en cada uno de los textos llevados a análisis concluimos que los textos científicos, de resúmenes, de patentes, legales, etc. Tienen entre sus características más comunes que insisten mucho en el mismo tema repitiendo conceptos y por ello las palabras más importantes, las más frecuentes, se repiten mucho más que el resto de palabras, es decir existe una cantidad de palabras que se repiten mucho más que el resto de palabras. Esto es consecuencia también de que los textos científicos igualmente tienen por lo general menos vocabulario que los literarios. Y por su parte los textos literarios tienen entre sus características más comunes que tienen más variedad de vocabulario, tienen por tanto más variantes gramaticales que los científicos y además como característica los literarios suelen tener muchas palabras de frecuencias bajas ($fr=1$, $fr=2$).

Resumiendo tenemos dos grandes bloques en los que agrupamos la tipología de un documento:

<i>Tipo de texto</i>	<i>Características esenciales</i>
Científicos	<ul style="list-style-type: none"> - menos vocabulario respecto a P (P= tamaño) - las palabras más frecuentes tienen mayor frecuencia respecto a las demás (se debe a que los textos científicos insisten mucho en el mismo tema y por ello repiten conceptos) - menos variantes gramaticales
Literarios	<ul style="list-style-type: none"> - más vocabulario respecto a P (P= tamaño) - las palabras menos frecuentes, de frecuencias bajas $fr=1$, $fr=2$ son muy abundantes - más variantes gramaticales

Tabla 17. Tipología y características de los textos

²⁹ En el caso del texto literario de Cervantes y su conocida obra El Quijote, *cervantes.txt* es un texto que a pesar de ser una novela y sea considerado sin lugar a dudas como un documento de tipología literaria, su comportamiento según el estudio realizado es como un científico, por tanto estamos ante una excepción.

En esta gráfica³⁰ se comparan los textos científicos en color azul y los literarios coloreados en rosa, cada punto coloreado corresponde a un texto diferente y en el eje de abscisas se distribuye el tamaño de dichos documentos, así podemos comparar en las gráficas por un lado la diferencia entre los valores del exponente de Zipf obtenidos para los textos científicos y literarios y además podemos comparar las diferencias según el tamaño de los textos, así pues el valor 1 en el eje de abscisas indica un tamaño menor que va creciendo hasta el valor 23 que correspondería al texto más grande.

A simple vista después de lo mencionado podríamos observar que los textos científicos ofrecen valores del exponente mayores que los textos literarios y no parece afectarles el tamaño que tenga el texto para ofrecer valores más altos o menos, en cambio si podemos observar que en los documentos literarios parece que cuanto mayor es el documento el valor del exponente aumenta, es decir la variable tamaño sí que afectaría a los textos literarios.

Según hemos estudiado anteriormente³¹ [Gráfico 39: Distribución de frecuencias de Zipf con distintos valores del exponente (e)], un exponente mayor de Zipf indica mayor frecuencia en las palabras más importantes, así por tanto que el valor del exponente (e) para los textos científicos sea mayor que en los textos de tipo literario es porque los textos científicos insisten mucho en el mismo tema repitiendo los mismos conceptos y por ello las palabras más importantes, de mayor rango las repiten mucho. Y en cambio en los textos literarios existe más variedad en el vocabulario.

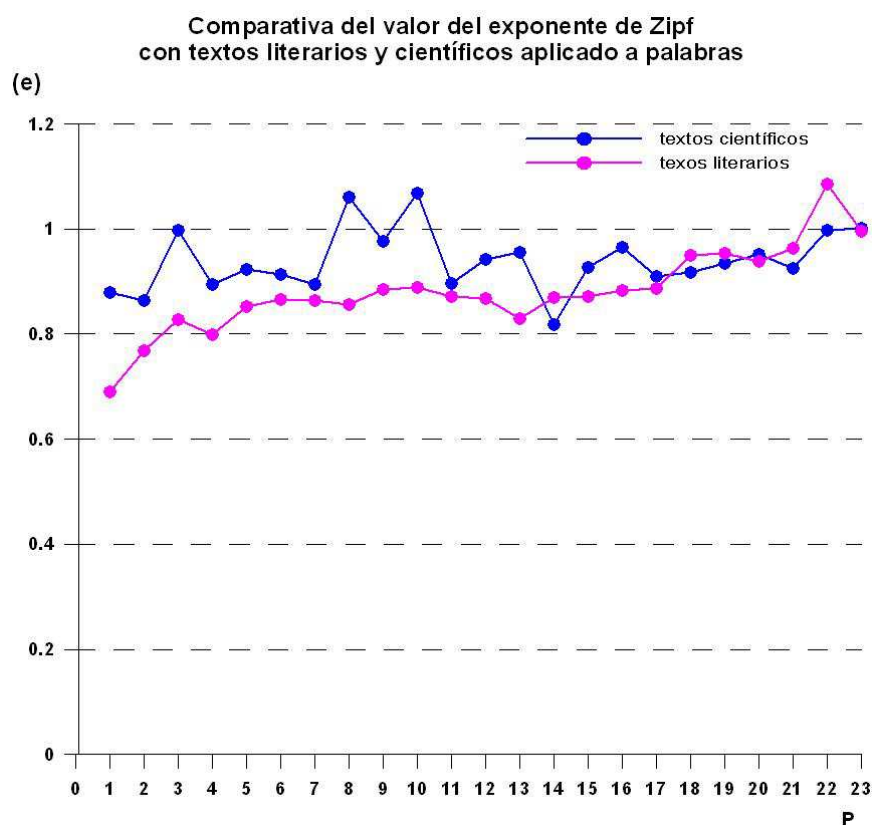


Gráfico 65. Valor del exponente (e) de Zipf con textos literarios y científicos

³⁰ En las gráficas no aparecen representados los textos más pequeños, por dar valores demasiados bajos. Así pues valoramos sólo 23 textos de tipo científico y 23 textos de tipo literario.

³¹ Véase gráfico pág. 145.

Leyenda gráfico:

Eje de abscisas (P)	Texto Científico	P Científicos	Texto Literario	P literarios
1	icyt3	8148	gg1	6821
2	cali2	11745	seis	25696
3	ime1	12345	ksoti2	28279
4	aralpi	12862	ksoti1	28582
5	cali1	22133	erudi1	53940
6	sonia	32852	ksoti	56854
7	imac2	46362	dos	73790
8	ime2	55696	castea	74939
9	icyt4	66260	larra1	76550
10	medico1	67818	episo1	81398
11	isoc2	70158	casted	86666
12	upalc	70523	castec	89556
13	bachi1	73173	costac	91714
14	upalb	73560	costaa	93140
15	upala	76990	casteb	93270
16	paten1	78514	costab	94283
17	imec2	79744	vari	95045
18	imeb2	81463	castee	106079
19	imab2	86349	episogrande	204757
20	imaa2	86931	costa1	279141
21	imea2	95374	vari1	380054
22	legal1	97670	cervantes	385213
23	cienti1	101164	castela	450574

Tabla 18. Leyenda gráfico núm. 65

Así por ejemplo, en los documentos científicos los que han obtenido un mayor valor del exponente son los situados en el eje de abscisas en la posición 3 que corresponde al texto: *ime1.txt* con $e = 0,998$, a la posición 8 que corresponde al texto: *ime2.txt* con $e = 1,060$, a la posición 10 que corresponde al texto: *médico1.txt* con $e = 1,068$ y por el contrario el texto que ha obtenido un menor valor del exponente es el que se sitúa en la posición 14 que corresponde al texto: *upalb.txt* con $e = 0,818$.

En los tres primeros casos son documentos científicos de resúmenes, eso implica que las palabras más frecuentes son las etiquetas de cada campo como las palabras: *Documento*; *Título*; *Revista*; *Idioma*; *Lugar-Trabajo*; *Datos-Fuente*; *Tipo*; *Autores*; *Idioma*; *ISSN*; *Descriptor*; por tanto las palabras más frecuentes obtenidas han sido:

ime1.txt		ime2.txt		medico1.txt	
pala	fr	pala	fr	pala	fr
DOCUMENTO	561	DOCUMENTO	2527	DOCUMENTO	3087
ESPAÑA	477	TÍTULO	2067	TITULO	2532
TÍTULO	463	REVISTA	1420	ESPAÑA	2033
BARCELONA	370	AGUDO	1393	REVISTA	1707
REVISTA	286	ESPAÑA	1364	AGUDO	1398
IDIOMA	236	IDIOMA	1046	IDIOMA	1281
TRABAJO	232	FUENTE	1035	FUENTE	1265
FUENTE	229	DATOS	1029	DATOS	1257
TIPO	229	TIPO	1029	TIPO	1257
AUTORES	228	AUTORES	1022	AUTORES	1251
AGUILAR	228	TRABAJO	1007	TRABAJO	1237
DATOS	228	LUGAR	1005	LUGAR	1230
LUGAR	226	ESPAÑOL	1005	ESPAÑOL	1223
ISSN	219	ISSN	975	ISSN	1194
ESPAÑOL	218	HOSP	927	HOSP	1109
SERV	215	INFARTO	814	DESCRIPTORES	994
DESCRIPTORES	185	MIOCARDIO	814	REF	974
HOSP	182	DESCRIPTORES	809	SERV	965
HOSPITAL	174	REF	802	INFARTO	816
REF	172	SERV	750	MIOCARDIO	813
MODO	105	MODO	481	BARCELONA	639
NEUROLOGIA	82	HOSPITAL	395	MODO	586
BELLVITGE	70	ESPAÑOLA	338	HOSPITAL	570
PRINCIPES	70	MADRID	332	ESPAÑOLA	446
HOSPITALET	70	MEDICINA	316	CARDIOLOGIA	391
PUJOL	66	CARDIOLOGIA	312	MEDICINA	373
BADALONA	66	RENAL	270	MADRID	351
LLOBREGAT	65	BARCELONA	269	GARCIA	298
CLINICO	64	GARCIA	257	RENAL	273
GERMANS	64	ESPA	240	SUPL	252
TRIAS	64	SUPL	222	CLINICO	251
NEUROL	61	ABDOMEN	219	AGUILAR	240
MEDICINA	57	TRATAMIENTO	203	TRATAMIENTO	238
TARRASA	55	UCI	203	ABDOMEN	219
ESPAÑOLA	49	UNIDAD	180	CLINICA	213

Tabla 19. Ejemplos palabras más frecuentes en textos tipología científico de resúmenes

En el caso contrario tenemos el texto *upalb.txt* que es un documento igualmente científico de resúmenes, pero en este caso la palabra más frecuentes es únicamente: *Resumen*; que corresponde a la única etiqueta de cada campo.

Esto viene a corroborar lo explicado anteriormente, en los tres casos anteriores los documentos *ime1.txt*; *ime2.txt*; *medico1.txt*, obtienen un valor mayor del exponente que en este ejemplo de texto *upalb.txt*, porque como bien hemos explicado al comienzo de este capítulo, un valor alto del exponente indica que las palabras más frecuentes

destacan más respecto a las demás palabras, y eso es realmente lo que ocurre, en los tres primeros casos las palabras: *Documento; Título; Revista; Idioma; Lugar-Trabajo; Datos-Fuente; Tipo; Autores; Idioma; ISSN; Descriptores*, tienen una frecuencia mucho mayor que el resto de palabras, este hecho es lo que tiende a aumentar el valor del exponente en la fórmula de Zipf, y por el contrario pese a que el texto *upalb.txt* es igualmente científico y de resúmenes como únicamente tiene una palabra que destaca más respecto a su frecuencia que el resto, como es la palabra *Resumen* (por la característica inherente del tipo de texto); existe más variedad en las palabras y sus frecuencias, por ello tiende a disminuir el valor del exponente en la fórmula de Zipf.

Upalb.txt	
pala	fr
EXITO	1146
RESUMEN	1032
TRABAJO	363
HAN	334
ARTICULO	314
RESULTADOS	251
PROCESO	243
EMPRESAS	233
DESARROLLO	227
EMPRESA	218
ESTUDIO	201
AÑOS	199
GESTION	192
ANALISIS	191
PARTE	174
TANTO	167
ESTAS	155
FORMA	152
FACTORES	151
MAYOR	147
SOCIAL	146
MODELO	139
FORMACION	138
INFORMACION	137
PRESENTE	135
INVESTIGACION	134
GRAN	126
RELACION	124
PROBLEMAS	123
OBJETIVO	122
ELLO	122
SISTEMA	121
PUEDEN	121
CAMBIO	120
RECURSOS	119

Tabla 20. Ejemplo palabras más frecuentes en texto tipología científico de resúmenes

En el caso de los textos literarios observamos claramente en el gráfico expuesto anteriormente como el valor más alto para el exponente es el que ocupa la posición 22 que corresponde al texto *cervantes.txt* con $e = 1,086$, pero contradiciendo la tendencia visible que siguen los textos literarios que es a mayor tamaño, mayor exponente, en el último caso que es el documento más grande de los literarios *castela.txt* con $e = 0,995$ se ha obtenido un valor de (e) menor, es por ello que vamos a ver el ejemplo de dichos textos.

Si tenemos en cuenta que en el texto *cervantes.txt* se narra la conocidísima obra del autor El Quijote, podemos deducir que ha obtenido un mayor valor del exponente en la fórmula de Zipf porque las palabras más frecuentes en este caso nombres propios, etc como Sancho, Quijote, ..., etc. se repiten mucho más en el texto que el resto de palabras, y en el caso del autor Emilio Castelar con su documento *castela.txt*, texto político titulado *Crónica internacional 1890-1898*, es lógico pensar que este documento tenga más variedad en su composición literaria y creativa en el vocabulario, es por ello que si comparamos los dos textos obtenemos conclusiones razonadas respecto al valor obtenido por el exponente de Zipf.

Podemos afirmar siguiendo con la discusión, que la obra *El Quijote*, a pesar de ser una obra narrativa extensa y de gran tamaño trata sobre las aventuras de dos personajes y sobre ellos trata toda la trama de la obra, es por ello que son dos personajes importantes y a los cuales se les menciona en la obra en todo momento, pero en cambio en la obra de Emilio Castelar y su *Crónica Internacional 1890-1898*, aun teniendo mayor tamaño que El Quijote trata: Errores económicos de la gran República sajona.-Diferencias entre los demócratas y los republicanos en América.-Orígenes del proteccionismo anglo-sajón.-La reacción económica en Europa y sus consecuencias.-Correlación entre la guerra por tarifas y la guerra por armas.-Daños traídos por los últimos bills americanos a la producción europea.-Necesidad que tiene América de compenetrar su política y su economía.-Situación de Portugal y España.-Buena ventura de Francia en este período último.-La muerte del rey Guillermo, la desgracia de Parnell y los crímenes nihilistas.-El resultado electoral en Italia.-El Papa y Lavigerie.-Los pietistas germanos y el Emperador.-Estado de Oriente.-Conclusión.

Claramente es visible la gran variedad en el documento de Castelar respecto a la temática de su obra y es por ello una vez más que se obtiene así una conclusión clara respecto a porqué el texto *castela.txt* pese a ser mayor en tamaño que el texto *cervantes.txt*, tiene un valor menor el exponente.

cervantes.txt		Castela.txt	
pala	fr	pala	fr
DON	2988	TODOS	2301
DIJO	2446	TUDO	2268
SANCHO	2152	VIDA	1869
QUIJOTE	2140	AMOR	1747
TUDO	1679	DIOS	1508
BIEN	1481	CORAZON	1400
RESPONDIO	1425	TODAS	1364
SEÑOR	1405	PUES	1335
ESTO	1373	ALMA	1190
QUIEN	1259	SOLO	1175
PUES	1244	AQUEL	1174
TODOS	1175	BIEN	1090
ELLA	1142	MUNDO	1071
MERCED	1024	TODA	1068
SINO	973	AQUELLA	1019
VUESTRA	970	MUY	983
FUE	966	HAN	976
CABALLERO	790	OJOS	974
ESTABA	789	MUERTE	970
SEÑORA	782	QUIEN	946
DECIR	781	MIS	920
AUNQUE	753	CIELO	909
MUY	749	NUESTRA	898
AQUEL	737	NOS	886
AQUI	705	HAY	825
DIOS	699	DIJO	819
CASA	684	SIEMPRE	806
ALLI	652	MUCHO	798
NOS	640	ANGELA	794
LUEGO	631	GUERRA	751
HAY	629	MARGARITA	751
COSA	627	CUANTO	749
LES	627	CONTRA	732
TANTO	623	NADA	725
VER	621	NUESTRO	705

Tabla 21. Ejemplo palabras más frecuentes en texto tipología literario

En este caso, generalizar sobre los literarios o científicos es relativo ya que nos encontramos con casos en los que textos literarios siguen tendencias más cercanas a los textos científicos como el caso de *cervantes.txt*

Cuando todos los textos científicos obtienen un valor $(e) > 1$ siendo $P \cong 100.000$ palabras, los textos literarios aún con textos mayores a 100.000 palabras no obtienen resultados tan altos y no superan el valor de 1 quedándose por debajo, excepto el texto literario en cuestión *cervantes.txt*.

Que los textos literarios obtengan resultados menores para (e) respecto a los científicos significa que cuando un documento tiene más variedad en el vocabulario, el exponente va a disminuir y una consecuencia de dicha variedad gramatical se debe a la gran cantidad de vocabulario de los literarios respecto a los científicos, si observamos en la tabla anterior los textos de similar tamaño (P), los literarios doblan en vocabulario a los científicos. Esta variedad gramatical es la que hace disminuir el exponente y obtener un (e) menor.

Por tanto, que los textos científicos obtengan resultados mayores del exponente (e) respecto a los literarios, significa que cuando un documento tiene menos variedad en el vocabulario (<V) y además como punto a destacar las palabras más frecuentes destacan mucho más que el resto de palabras, esta característica del texto que se da mayoritariamente en los científicos es la que hace aumentar el exponente y obtener un (e) mayor.

En páginas siguientes se estudiará si el valor del exponente depende del tratamiento dado al texto y no de su tipología, en dicho caso se realizarán cuatro pasos de estudio en el que se obtendrán las mismas conclusiones que hemos expuesto ahora y que se relacionan estrechamente a la tipología, como venimos diciendo los textos científicos y literarios tienen unas características gramaticales que los diferencian y parte de dichas características gramaticales se conseguirán reproducir a la hora del tratamiento de los textos, es decir, mas adelante probaremos en que afecta quitar o no las palabras vacías del texto y en qué afecta agrupar las palabras en raíces, para que aumente o disminuya el valor de (e), comprobaremos como en el caso de quitar las palabras vacías a un texto el efecto que se obtiene es que faltan palabras de alta frecuencia , y eso mismo aplicado a la tipología que estamos estudiando es lo mismo que le ocurre a los textos literarios, por tanto se obtendrá un valor de (e) menor. Y en el caso de que no eliminemos las palabras vacías del texto el efecto que se obtiene es que aparecen palabras muy frecuentes, con altas frecuencias respecto al resto de palabras y eso mismo aplicado a la tipología que estamos estudiando es lo mismo que le ocurre a los textos científicos, por tanto se obtendrá un valor de (e) mayor.

Pero en este estudio en el que estamos ahora obtendremos conclusiones respecto a la tipología de los documentos y aún daremos un paso más investigando el comportamiento de (e) en la fórmula de Zipf aplicado a raíces.

Si a la metodología general de investigación y el procedimiento de análisis se le incorpora un nuevo paso como es la agrupación en raíces de una tabla de 358 raíces aplicándole el método de sufijos de Lovins con ligeras adaptaciones al Español y obtenemos en última instancia el valor del exponente de la fórmula de Zipf conseguimos el valor del exponente aplicado a raíces y no a palabras como se ha estudiado anteriormente.

Los resultados obtenidos para el valor de (e) son en el caso de los textos científicos menor respecto al valor obtenido con palabras, y en el caso de los textos literarios el valor de (e) es mayor respecto a las palabras (aunque existen excepciones como el texto *cervantes.txt* que aplicado a raíces obtiene $e=1,041$ menor que aplicado a palabras que obtiene $e=1,086$)

En el siguiente gráfico se compara el valor del exponente de Zipf con textos científicos y literarios aplicado a raíces, los resultados muestran una tendencia interesante en el caso de los textos científicos, el valor del exponente si se aplica a raíces y no en palabras ofrece valores de (e) más bajos que si se obtiene con palabras, y este hecho se puede apreciar más visualmente en los siguientes dos gráficos, y en cambio en el caso de los documentos literarios los resultados ofrecen que el valor de (e) aplicado a raíces no hace disminuir el valor del exponente respecto a las palabras, todo lo contrario mantiene un valor de (e) parecido o incluso mayor a si el proceso se realiza con palabras.

Respecto a los valores más altos obtenidos que se pueden observar en este gráfico concretamente los correspondientes a los textos literarios del 19 al 23 representados en el eje de abscisas corresponde a los texto de: *castela*, *cervantes*, *vari1*, *costa1* y *episogrande*, estos han obtenido un valor muy alto de (e) porque el tamaño de dichos documentos difiere bastante de los de tipo científicos que serían en este caso: *cienti1*, *legall*, *imea2*, *imaa2*, por ejemplo el texto *castela* tiene un tamaño de 450.574 palabras y el texto científico *cienti1* tiene un tamaño de 101.164 palabras, de ahí la gran diferencia en los valores.

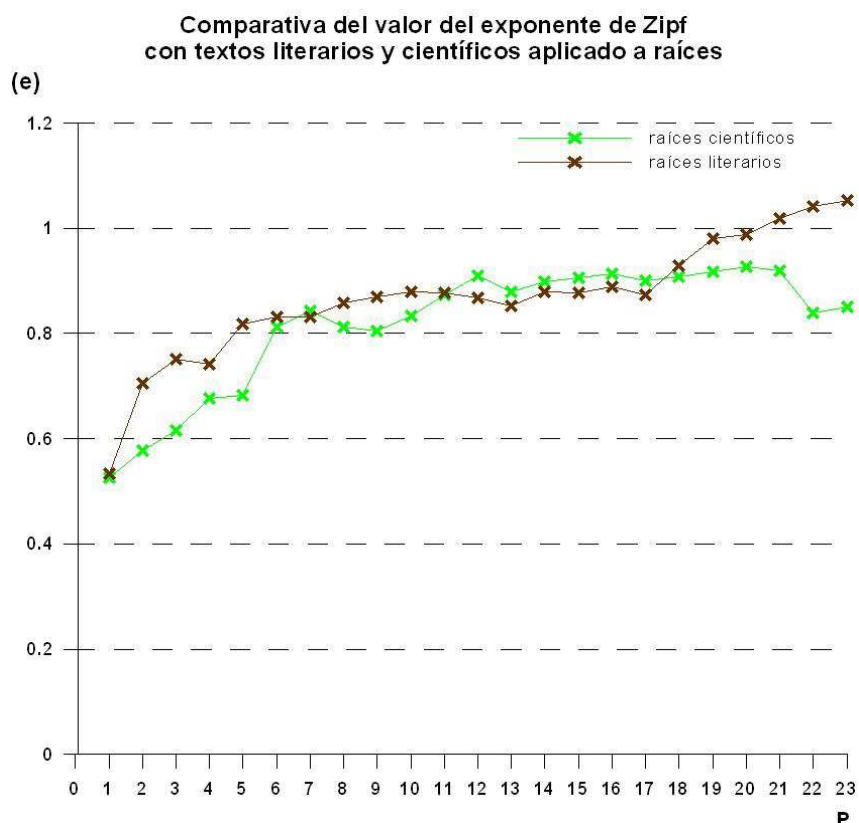


Gráfico 66. Valor del exponente (e) de Zipf con textos literarios y científicos aplicado a raíces

Los dos gráficos siguientes muestran por separado el valor obtenido por el exponente de Zipf en los textos literarios y en los textos científicos comparándolo con palabras y raíces igualmente en cada caso, aquí podrá observarse lo mencionado anteriormente, respecto a que en el caso de los textos literarios la utilización de raíces no parece alterar el valor del exponente, e incluso aumentarlo y en cambio en el caso de los textos

científicos, la aplicación a raíces si parece disminuir más el valor de (e) respecto a si utilizamos palabras.

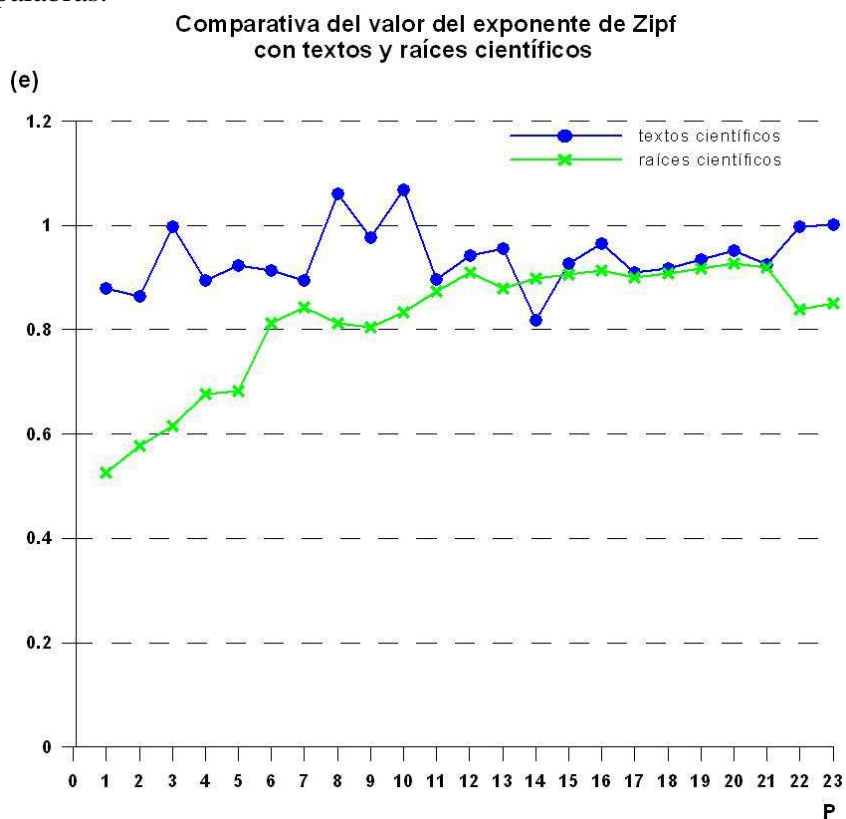


Gráfico 67. Valor del exponente (e) de Zipf con textos científicos aplicado a raíces

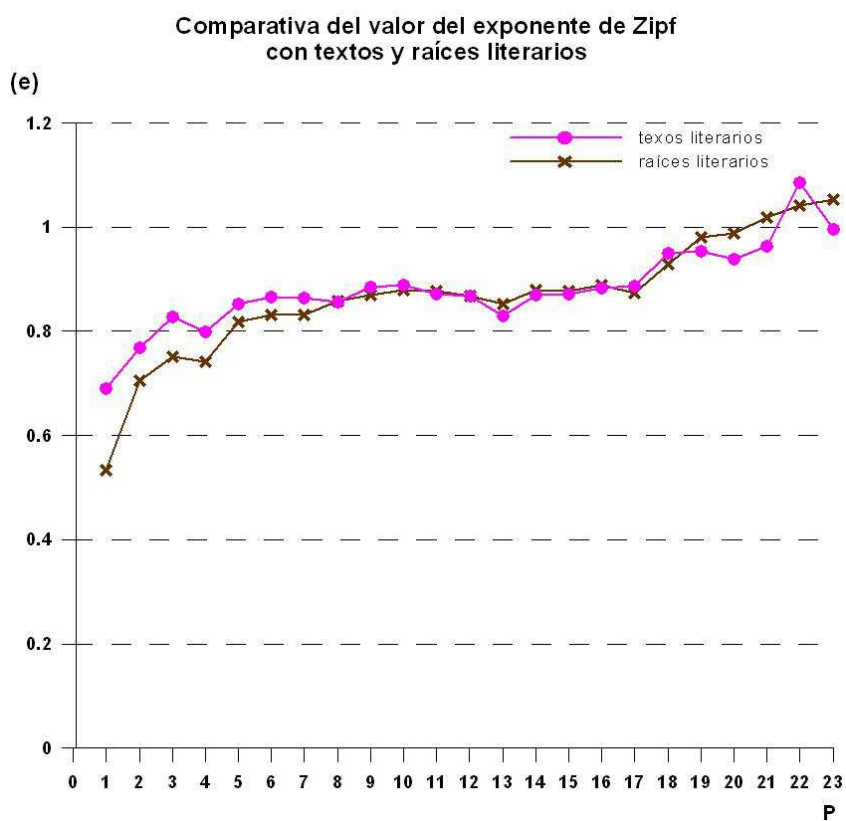


Gráfico 68. Valor del exponente (e) de Zipf con textos literarios aplicado a raíces

Como bien se puede advertir en los textos científicos disminuye más el valor de (e) con raíces respecto a si utilizamos palabras. Esto ocurre porque los textos científicos tienen como característica principal muchas palabras de frecuencias altas y a la hora de agruparse las palabras en raíces, ésta agrupación resulta “mínima”, obteniéndose prácticamente la misma cantidad de raíces que de vocabulario. Recordamos que esto es consecuencia de que los textos científicos tienen menos variantes gramaticales y por tanto tienen menos cantidad de palabras de frecuencias bajas ($fr=1$, $fr=2$).

En cambio, en los textos literarios aumenta más el valor de (e) respecto a los textos científicos si utilizamos raíces. Esto ocurre porque los textos literarios tienen como característica principal muchas palabras de frecuencias bajas, siendo éstas muy abundantes, esto supone que este tipo de textos tienen mucho más vocabulario y más variantes gramaticales, prueba de ello es que los textos literarios doblan en cantidad de vocabulario a los textos científicos. Los textos literarios al agrupar en raíces obtienen un valor superior del exponente (e) , porque las palabras de frecuencias bajas ($fr=1$, $fr=2$) se eliminan, puesto que donde había dos palabras distintas que aparecían sólo una vez cada una quizá tengan la misma raíz y ahora no aparecen como $fr=1$ sino como $fr=2$.

Este argumento que se acaba de exponer se puede comprobar con las experimentaciones realizadas en páginas posteriores donde se analiza el valor de (e) dependiendo del tratamiento dado al documento. En los cuatro experimentos realizados podremos corroborar que lo que ocurre en los textos literarios con las raíces es que el valor de (e) aumenta, sucede lo mismo según el tratamiento dado al texto; si extraemos las raíces de un texto obtenemos un valor de (e) mayor porque provocamos que falten palabras de frecuencias bajas, que es lo que le ocurre en este caso a los textos literarios. Y si al documento en cuestión no le extraemos las raíces, provoca que existan muchas palabras de frecuencias bajas y esto hace disminuir el exponente.

Otra particularidad observada es que en el caso de los textos literarios aplicados a raíces, el valor del exponente crece progresivamente según va aumentando el tamaño de los textos.

Igualmente podemos afirmar respecto a la hipótesis planteada número 2: “El valor del exponente depende del tamaño del texto”

El tamaño del texto afecta al valor del exponente en la fórmula de Zipf, aumentando su valor, es decir cuanto mayor tamaño del texto, el valor del exponente es mayor. Esta afirmación es claramente apreciable en los gráficos anteriores donde el eje de abscisas muestra el tamaño de los textos en modo creciente e igualmente se observa como el valor del exponente decrece si el tamaño es más pequeño y aumenta si el tamaño del texto es mayor.

Resultando que en un tipo homogéneo de textos los pequeños ofrecen valores menores que 1, los textos mayores ofrecen valores mayores que 1, por lo tanto existirá un tamaño de texto intermedio en el que se obtenga un valor del exponente $e=1$, a este valor concreto de P donde se puede obtener $e=1$, es lo que se denomina como *Zipfian Size* (Orlov, 1982)

Debowski (2002) hace mención en su artículo a Orlov (1982), como el precursor de la dependencia de los parámetros con el tamaño del texto, el *Zipfian Size*, combina un modelo matemático con evidencia experimental para mostrar que hay un tamaño de texto para el que el exponente de Zipf vale 1 y a partir de aquí, al aumentar el tamaño, el valor del exponente aumenta ligeramente.

Por tanto el autor Debowski (2002), demuestra experimentalmente y se demuestra en los modelos teóricos, que aumentando el tamaño del texto se alcanza siempre un tamaño donde el exponente vale exactamente 1 y a partir de aquí los textos mayores tienen exponente >1 .

Hasta ahora hemos visto que el valor del exponente en la fórmula de Zipf depende de la tipología del documento analizado porque dichos tipos de documentos tienen unas características que igualmente se pueden observar si aplicamos a los textos un tratamiento u otro (eliminación de palabras vacías, agrupación en raíces, etc.). Hemos visto que también depende del tamaño del texto. A continuación, vamos a corroborar que realmente el valor del exponente de la fórmula de Zipf depende sobre todo del tratamiento dado al texto, ya que podemos forzar a que dicho valor sea mayor o menor, como se demuestra en el estudio siguiente en la hipótesis número tres.

Respecto a la hipótesis planteada número 3: “El valor del exponente depende del tratamiento previo del texto”.

Como se ha explicado en páginas anteriores el tratamiento dado al documento dependerá del analista que puede utilizar una cantidad de palabras vacías y un método distinto de extracción de raíces o lematización, este hecho que a simple vista no se le da importancia, es crucial para obtener unos resultados u otros concernientes al valor del exponente de la fórmula de Zipf. En los ensayos siguientes se demuestra que el valor de los coeficientes en las fórmulas dependen más del tratamiento dado al documento que de la tipología del documento analizado, es decir de si se eliminan más o menos palabras vacías, o si el proceso de extracción de raíces es más exagerado o menos. Entonces que un autor escriba más palabras significativas por página que otro autor, depende más del criterio del analista sobre qué palabras son significativas, que de la forma de escribir del autor, los experimentos con tablas de palabras vacías distintas parecen dar cuenta de esta variación en los coeficientes.

Igualmente observamos que la tabla de palabras vacías es característica del idioma y no de la tipología del documento, y que la tabla de sufijos³² para la extracción de raíces también puede tener un dudoso tratamiento

Así pues, vemos que el valor que obtengamos del exponente de Zipf no es un concepto absoluto asociado al texto, sino que dependerá del tratamiento que realicemos y dependiendo de estos tratamientos del texto el valor del exponente estará alrededor de 1, para corroborar esta afirmación se han realizado diversos ensayos sobre la Ley de Zipf que demuestran que el valor del exponente estará alrededor de 1 dependiendo del tratamiento que demos al texto, es decir dependiendo si en el proceso inicial del documento se incluyen las palabras vacías, si se extraen raíces, etc.

³² En el caso de utilizar el *Stemmer* de Lovins (1968)

Para hacer las comprobaciones se ha analizado un mismo texto concretamente el documento de Larra³³ con $P=138.534$, $V=16.948$ y $P=76.545$, $V=16.837$, realizando cuatro estudios con procedimientos diferentes, los cuales se detallan a continuación:

Estudio 1:

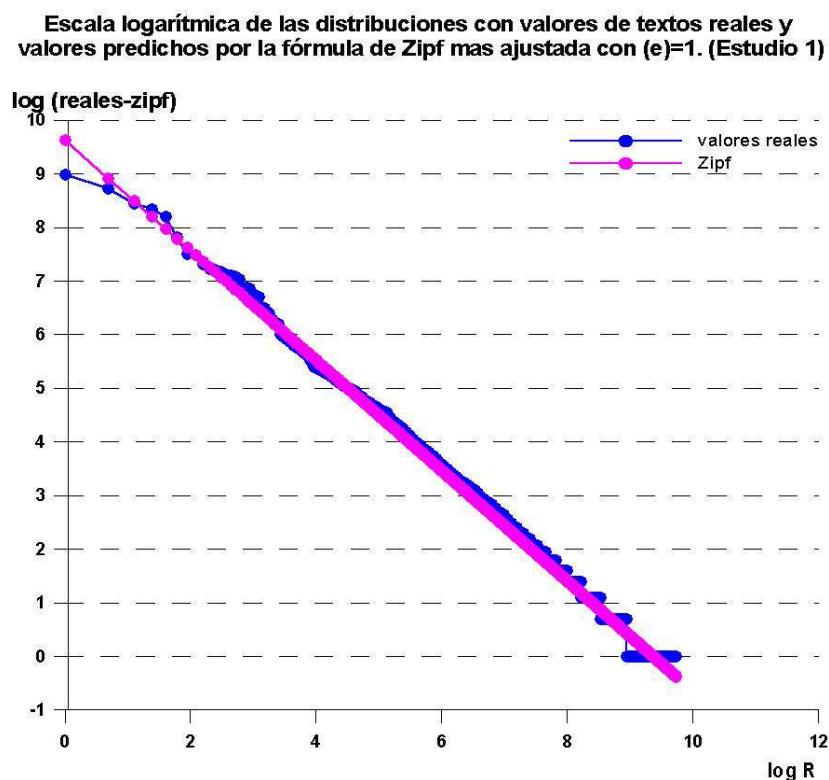
Al documento inicial no le quitamos las palabras vacías y no se extraen las raíces. El texto contiene un total de 1.040.000 caracteres. El valor obtenido del exponente se ha obtenido aplicando la fórmula de Zipf tal cual, es decir en este caso no ha habido ningún ajuste parcial al rango como se realizará en otros estudios.

$P=138.534$

$V=16.948$

El valor obtenido del exponente (e) = 1.02, $K=15264.89$ error=0.0373

Esta es la situación normal, en la que se ajusta muy bien a la fórmula de Zipf con exponente casi igual a 1. Podemos verlo en una gráfica de la siguiente manera:



En la gráfica vemos una línea recta, correspondiente a la fórmula de Zipf perfecta y la línea irregular que son los datos reales del texto del autor Larra, comprobamos que se ajusta bastante bien.

Estudio 2:

³³ Véase Apéndice I

Al documento inicial no le quitamos las palabras vacías pero sí extraemos las raíces con el método de sufijos, este método contiene una tabla con 358 sufijos. Igualmente en este caso, como en el anterior la predicción de Zipf se basa en un ajuste global³⁴ a todos los datos y en consecuencia no se aplica ningún ajuste parcial a ningún tramo del rango.

$P=138.534$

$V=16.948$

$R=8.282$

El valor obtenido del exponente (e) =1.14, $K=24281.14$ error=0.078

El valor superior que obtenemos ahora indica que la gráfica tiene más pendiente, que baja más rápidamente.

En la gráfica se observa que la mayor pendiente se debe esencialmente a que faltan palabras de frecuencia 1, 2,.. las de mayor rango. Esto es lógico que sea así, puesto que donde había dos palabras distintas que aparecían solo una vez cada una, quizá tengan la misma raíz y ahora no aparecen como $fr=1$ sino como $fr=2$.

¿Hasta donde aumenta el exponente? Aquí ha pasado de 1,02 a 1,14. ¿Podría aumentar más? SI. Todo depende de lo que exageremos al agrupar las palabras según sus raíces. Por eso, también hemos realizado un estudio con diferentes y cada vez más agresivos procedimientos para obtener las raíces y con ello conseguimos elevar el valor del exponente tanto como queramos.

Escala logarítmica de las distribuciones con valores de textos reales y valores predichos por la fórmula de Zipf mas ajustada con (e)=1. (Estudio 2)

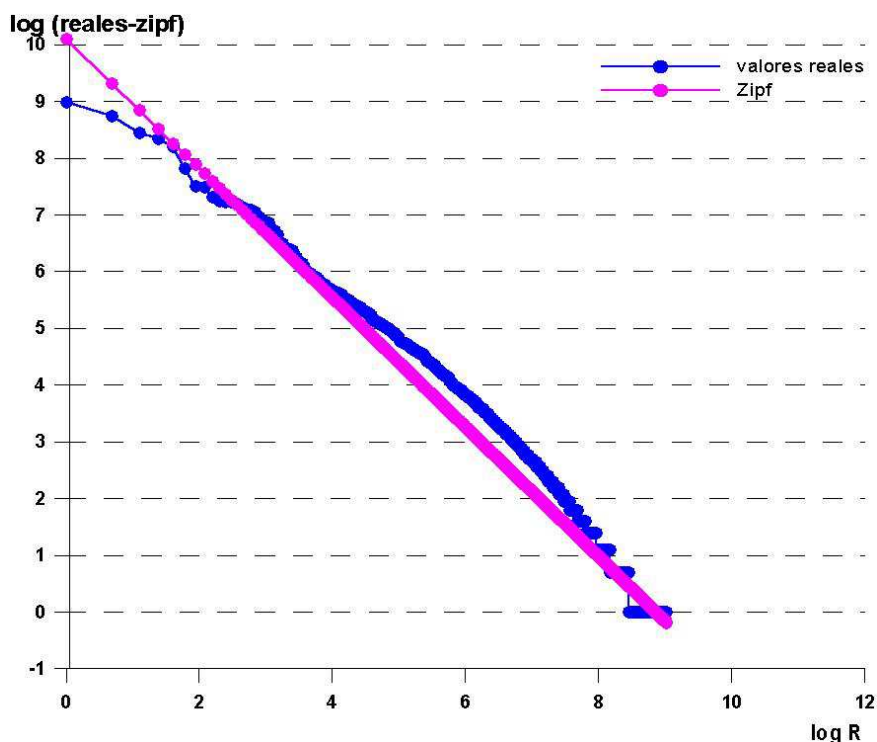


Gráfico 70. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 2

³⁴ En los apartados 6.6 y 6.7 se desarrolla ampliamente un nuevo procedimiento de ajuste parcial de Zipf por tramos.

De todos modos, la parte central de la gráfica sigue teniendo una pendiente cercana a 1. Además de verlo en la gráfica se puede hacer un ajuste parcial de la Ley de Zipf ajustando a algún tramo en el rango, así podemos provocar que se ajuste a un tramo concreto de palabras, por ejemplo en la aplicación³⁵ construida para tal fin se puede especificar un ajuste parcial de la ley de Zipf al tramo de rangos entre 2 y 8 obteniendo en este caso el valor $e=1,02$.

Estudio 3:

Al documento inicial le quitamos las palabras vacías de una tabla de 72 palabras vacías, incluidas las palabras que tengan 1 o 2 letras y no se extraen las raíces.

$P=76.545$

$V=16.837$

El valor obtenido del exponente (e) =0.88, $K=4082.795$ error=0.056

Este valor más pequeño nos dice que la gráfica tiene menos pendiente, que desciende con menos rapidez. Como ahora le hemos quitado las palabras vacías, lo que nos faltan son palabras de alta frecuencia: esto se ve claramente en la gráfica y es el motivo de la menor pendiente.

Si al documento inicial le quitamos las palabras vacías pero en este caso de una tabla de 386 palabras vacías, el valor que obtenemos del exponente es 0.83 todavía más pequeño que el anterior valor 0.88, eso indicará todavía más que faltan palabras de frecuencia alta y por ello la gráfica tiene menos pendiente.

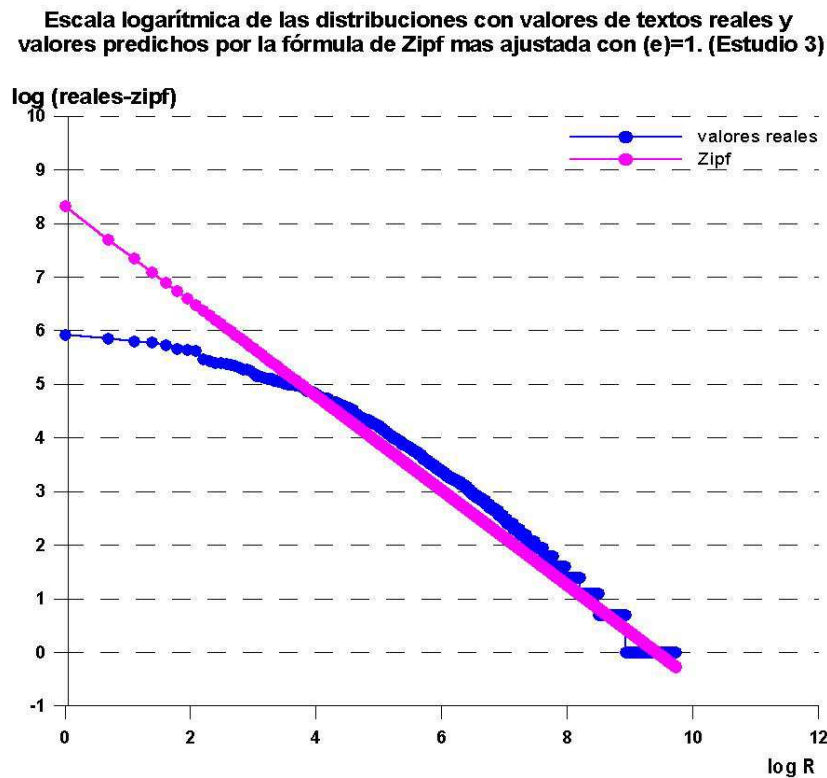


Gráfico 71. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a $(e)=1$. Estudio 3

³⁵ Véase Apéndice II. Base de Datos Topos formulario Analiza Zipf

Igualmente se puede comprobar que si se prescinde de los primeros tramos, haciendo un ajuste parcial al tramo de rangos entre 4 y 10, obtenemos el valor $e=0,99$ que se parece todavía más a 1.

Estudio 4:

Al documento inicial le quitamos las palabras vacías de una tabla de 72 palabras vacías, y sí extraemos las raíces con el método de sufijos de Lovins, este método contiene una tabla con 358 sufijos.

$P=76.545$

$V=16.837$

$R=8.210$

El valor obtenido del exponente (e) =1.01, $K=8636.91$ error=0.12

Ahora parece que sigue siendo aproximadamente igual a 1, pero lo que ocurre es que tenemos los dos efectos anteriores que se compensan: faltan palabras de alta frecuencia, lo que hace disminuir el exponente y faltan palabras de baja frecuencia, lo que hace aumentarlo. El resultado final es que el ajuste es peor

Si al documento inicial le quitamos las palabras vacías pero en este caso de una tabla de 386 palabras vacías, el valor que obtenemos del exponente es 0.97 todavía más pequeño que el anterior valor 1.01, la conclusión es la misma igualmente.

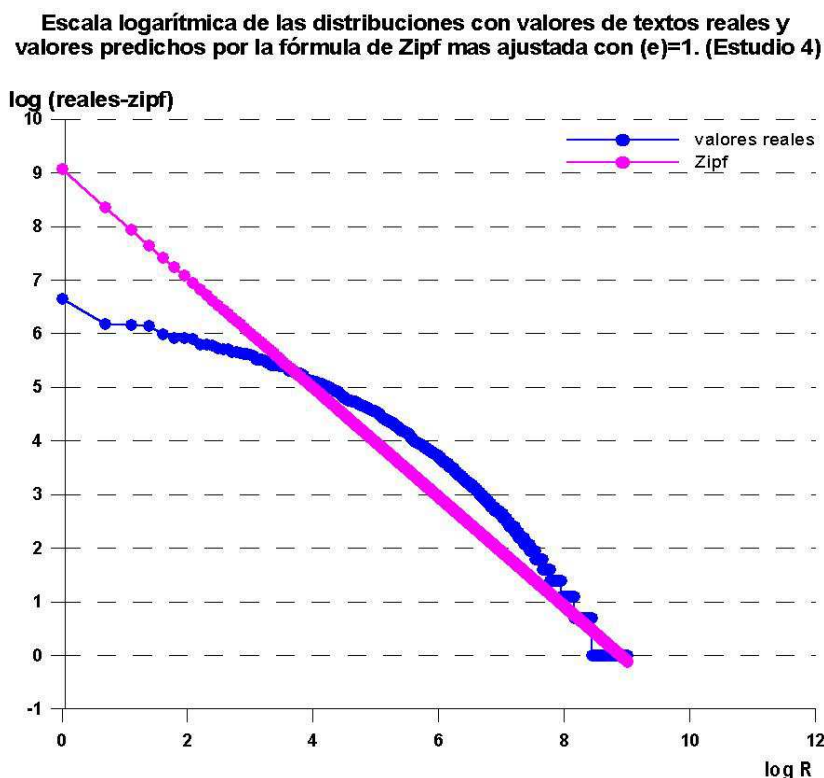


Gráfico 72. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a (e)=1. Estudio 4

Podemos verificar calculando ajustes parciales, como a diferentes tramos corresponden diferentes exponentes, desde los demasiado bajos a la izquierda, hasta los demasiados altos a la derecha.

Parcial rangos entre 0-6 $e=0,4857$
 Parcial rangos entre 2-8 $e=0,7966$
 Parcial rangos entre 4-10 $e=1,2125$

Estos mismos ajustes si se utiliza una tabla de palabras vacías mayor con 386 palabras se obtienen obviamente valores más pequeños del exponente

Parcial rangos entre 0-6 $e=0,4728$
 Parcial rangos entre 2-8 $e=0,7243$
 Parcial rangos entre 4-10 $e=1,1462$

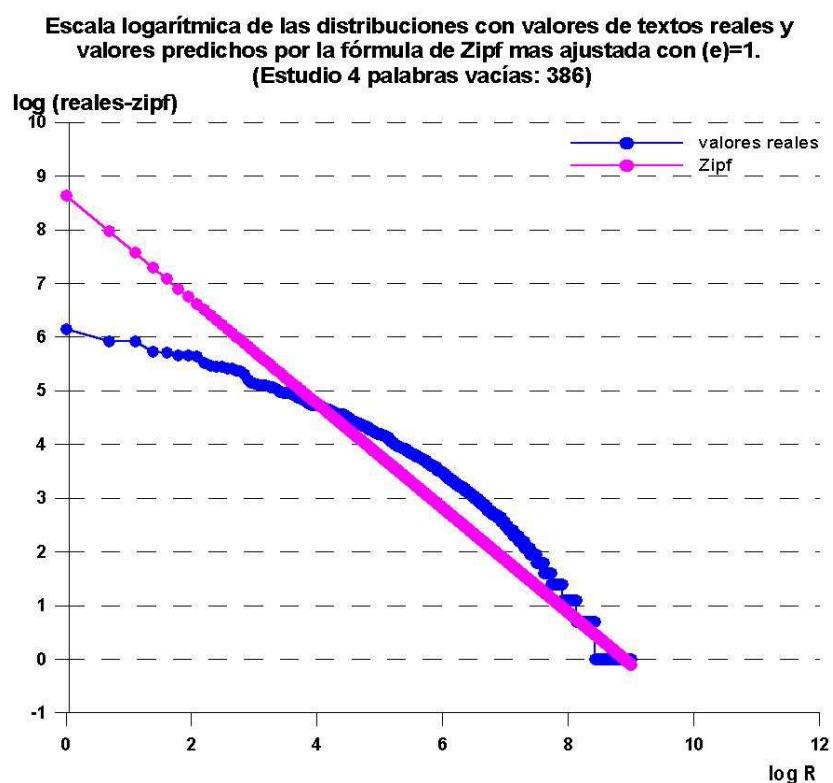


Gráfico 73. Distribución logarítmica de Zipf con valores reales y Zipf ajustada a $(e)=1$ $pv=386$

Vistos los estudios realizados, sugieren que las palabras de bajas frecuencias ($fr=1$, $fr=2$, etc.) son las responsables del aumento del exponente en las fórmulas Zipf-Mandelbrot, entre las palabras de bajas frecuencias estarían los hapaxes que son las palabras que aparecen sólo una vez en el documento y por tanto son consideradas como eventos raros LNRE (Large Number of Rare Events). Asimismo observamos tres partes claramente diferenciadas en el comportamiento de Zipf respecto a la tendencia de los textos reales, la primera parte que claramente se aleja de una tendencia constante son las palabras más frecuentes, la segunda parte que si se ajusta más a una constante son las palabras situadas en rangos intermedios y por último las palabras de bajas frecuencias que tampoco siguen una constante entre las que estarían situadas los hapaxes. Un argumento más extenso sobre estas tres partes diferenciadas en el comportamiento de Zipf respecto a la tendencia de los textos se ofrece en el punto siguiente.

Al respecto Montemurro (2001) apuesta por un modelo matemático para ajustar la Ley de Zipf lo más fielmente posible a los documentos reales estableciendo la existencia de

dos tramos que se alejan de una constante en las gráficas reales de Zipf y proponiendo un modelo matemático que lo representa. Pero lo que nos interesa sobre todo del artículo muy citado actualmente de Montemurro es que el autor apunta al igual que nosotros la evidencia de que el vocabulario en un corpus de texto dado y de diferentes tamaños puede ser dividido en dos partes: una básica en cuya estructura lingüística total cumple la Ley de Zipf-Mandelbrot y una segunda parte que contiene más palabras específicas³⁶ con una función sintáctica menos flexible, estas palabras son las que provocan que la última parte de la curva y en consecuencia las palabras de bajas frecuencias desciendan más rápidamente que el resto de palabras del texto.

Al igual que Montemurro, otros autores como Ferrer Cancho & Ricard (2001) demuestran que según la Ley de Zipf la frecuencia de las palabras como función del rango en tamaño de textos grandes muestra dos valores del exponente ($e \cong 1$ para una primera parte de la función y $e \cong 2$ para una segunda parte claramente diferenciada, demostrando la existencia de dos tramos distintos y esto lo relaciona con conceptos que no pueden expresarse con una sola palabra y por consiguiente necesitan una frase, el autor apuesta por un lenguaje di-gramas, es decir frases con sentido o no.

Esto nos sugiere que debemos hacer una adaptación del ajuste por tramos a la zona donde si que sigue una tendencia constante y por tanto cumple la Ley de Zipf, estos ajustes parciales se aplicarán sobre todo en el capítulo siete, en el cual se realizarán los ajustes a las palabras que se encuentran en la parte central, es decir entre los rangos centrales.

En dicho capítulo se realiza un estudio detallado de la distribución de Zipf aplicado a raíces, una vez visto los resultados obtenidos con las raíces, se ajustará la fórmula de Zipf al tramo central de cada colección de valores, este tramo central correspondería a los rangos entre $V^{0,3}$ y $V^{0,8}$, para realizar el ajuste a los datos centrales se hallará el Punto de Transición o PT (*Transition Point TP*). Sobre el Punto de Transición PT se abordará con más detalle en páginas siguientes, concretamente en el apartado 6.7.4 de este capítulo.

6.6. Ajuste parcial de Zipf por tramos

El ajuste parcial de Zipf por tramos nos va a permitir priorizar y dar más importancia a unas palabras más que otras y esto es posible debido a la propia ambigüedad del estudio de las palabras, ya que igualmente podemos dar más importancia a las palabras distintas, o a todas las palabras que aparecen, o ajustar una función que tiene decimales a unos valores que son enteros por definición (la frecuencia de una palabra es 2 ó 3, pero nunca 2,5), toda esta ambigüedad es lo que nos autoriza a explorar distintos criterios y lo expresaremos por medio de tramos, como por ejemplo podemos decidir como una posibilidad, a complementar con otras, no tener en cuenta las palabras de baja frecuencia. En definitiva a partir de este momento y ya en el capítulo siguiente (Capítulo 7) ya no se analizan los textos en su conjunto sino que se realiza un ajuste por tramos dependiendo de lo que se quiera priorizar.

³⁶ Hapaxes, palabras de $fr=1$

Al igual que muchos autores vienen diciendo hemos podido comprobar que Zipf siempre predice más palabras de frecuencias bajas de las que realmente hay en los textos y debido a la propia ambigüedad del estudio de palabras podemos variar el exponente de Zipf según apliquemos al texto analizado un tratamiento previo de dicho texto más o menos agresivo, es decir dependiendo si el analista elimina más o menos palabras vacías y dependiendo del método de *Stemming* utilizado si éste es más agresivo o menos. Este hecho ha sido ampliamente tratado en el punto 6.5.1 de este capítulo.

Existe cierta arbitrariedad en la obtención de los coeficientes de Zipf y de ello dependerá la pendiente de la distribución de frecuencias que ajustamos, el ajuste que efectuemos dependerá igualmente de cómo valoremos el error.

En los ejemplos llevados a cabo en este estudio se toma como medida de error para el ajuste de Zipf, el error relativo cuadrático promedio³⁷, pero podríamos utilizar otras medidas estandarizadas de error como puede ser el error relativo o el error absoluto y el utilizar unas medidas de error u otras nos permitirá favorecer más a unas palabras que a otras, es decir permitirá variar un poco la inclinación de la curva en la distribución de frecuencias de Zipf permitiendo que se ajuste más o menos a las palabras de frecuencias altas o por el contrario que la curva tenga una mayor pendiente y se ajuste mejor a las palabras de frecuencias bajas.

La medida de error absoluto podemos decir que trata por igual a todas las palabras, es decir que no nos informa de la desviación relativa mayor o menor entre los datos reales y la predicción de Zipf en todos los rangos de frecuencia, en cambio la medida de error relativo podemos decir que si nos informa de la desviación relativa entre los datos, observándose que en las palabras muy frecuentes el error relativo es muy pequeño y en la otra parte de la gráfica en palabras de frecuencias bajas el error relativo obtenido es muy grande, por tanto utilizar el criterio de error relativo beneficia a las palabras de frecuencias bajas disminuyéndose la diferencia entre el ajuste de Zipf y los datos reales.

Estas pequeñas variaciones de la pendiente al modificar el criterio de error utilizado resulta difícil verlo en gráficas reales debido a la pequeña desviación que éstas sufren.

En definitiva afirmamos que las teorías planteadas por muchos autores en la literatura y también analizada en detalle en este estudio sobre la tendencia de Zipf en las palabras de frecuencias bajas, la cual siempre predice más palabras de las que realmente hay en los textos reales, afirmamos que esto es cierto, pero aseverar la cantidad de vocabulario que predice Zipf de más, esto si que sería incorrecto, ya que es muy arbitrario, depende del criterio de error utilizado para el ajuste de Zipf y del tratamiento que se de al texto (eliminación de palabras vacías, método de *Stemming*), obtendremos valores distintos en los coeficientes de Zipf y la curva variará acercándose más a unas zonas de la distribución que a otras.

En conclusión, proporcionar valores fijos a los coeficientes de Zipf o a la cantidad de palabras de frecuencias bajas que predice de más la fórmula de Zipf es muy arbitrario, por tanto no podemos sacar conclusiones cuantitativas ya que depende de las

³⁷ Respecto a los cálculos véase el Estudio complementario 6.8 de este capítulo.

matemáticas utilizadas, del procesado de los documentos y análisis de los mismos y por todo ello los valores obtenidos pueden fluctuar en gran medida.

Según los gráficos siguientes que se exponen a continuación, se ven tres partes diferenciadas en el comportamiento de Zipf respecto a la tendencia en la distribución de frecuencias de los textos reales según su ajuste global y por ello nos lleva a realizar un ajuste parcial a tramos, para forzar que la fórmula de Zipf se acerque más a los datos reales.

Esto lo podemos conseguir igualmente con los estudios realizados anteriormente que demuestran que las predicciones de Zipf siempre obtienen más palabras de frecuencias altas respecto a los textos reales, y esto lo podemos solventar eliminando las palabras vacías, que en consecuencia supone disminuir el valor del exponente. Por otro lado las predicciones de Zipf también siempre obtienen más palabras de frecuencias bajas respecto a los textos reales, y esto lo podemos solventar agrupando en raíces, que en consecuencia supone aumentar el valor del exponente.

Por tanto, lo más llamativo de Zipf que hemos observado en los análisis anteriores es que siempre predice más palabras de frecuencias bajas de las que realmente hay en los textos llevados a estudio y este hecho es lo que vamos a ver a continuación. Es por ello que realizaremos el ajuste parcial de Zipf por tramos. Los gráficos siguientes muestran la tendencia que toman los textos reales agrupados en tramos correspondientes a una escala logarítmica, en contraposición de la predicción que realiza la ley de Zipf, dicho de otro modo más llamativo o sugestivo lo que se está comparando es el proceso de repetición de palabras del autor en qué está influyendo en el proceso de escribir el vocabulario.

Para ello se ha analizado un texto formado por yuxtaposición de varios autores, entre ellos Castelar, Costa, Pérez Galdós, Menéndez Pelayo y Larra³⁸. Dicho texto se compone de un tamaño de $P = 379.995$ y $V = 48.389$. Se representan las gráficas con el logaritmo-frecuencia frente a logaritmo-rango.

El siguiente gráfico como ejemplo de un texto pequeño

³⁸ Ver Apéndice I

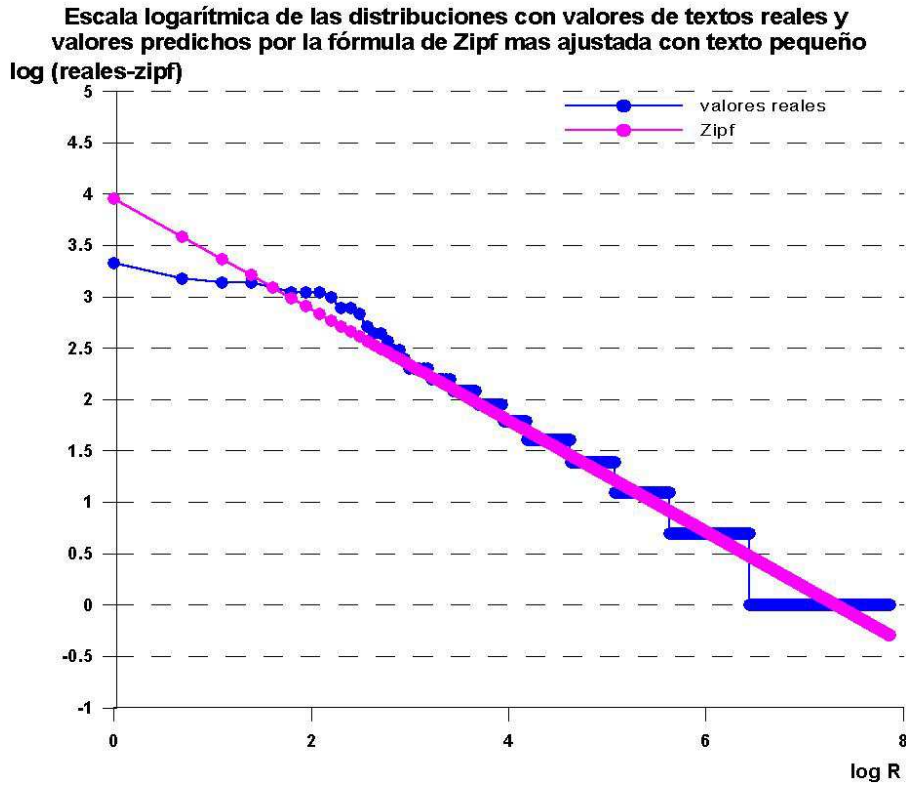


Gráfico 74. Distribución logarítmica de Zipf con valores reales y Zipf ajustada con texto pequeño

El siguiente gráfico como ejemplo de un texto mayor

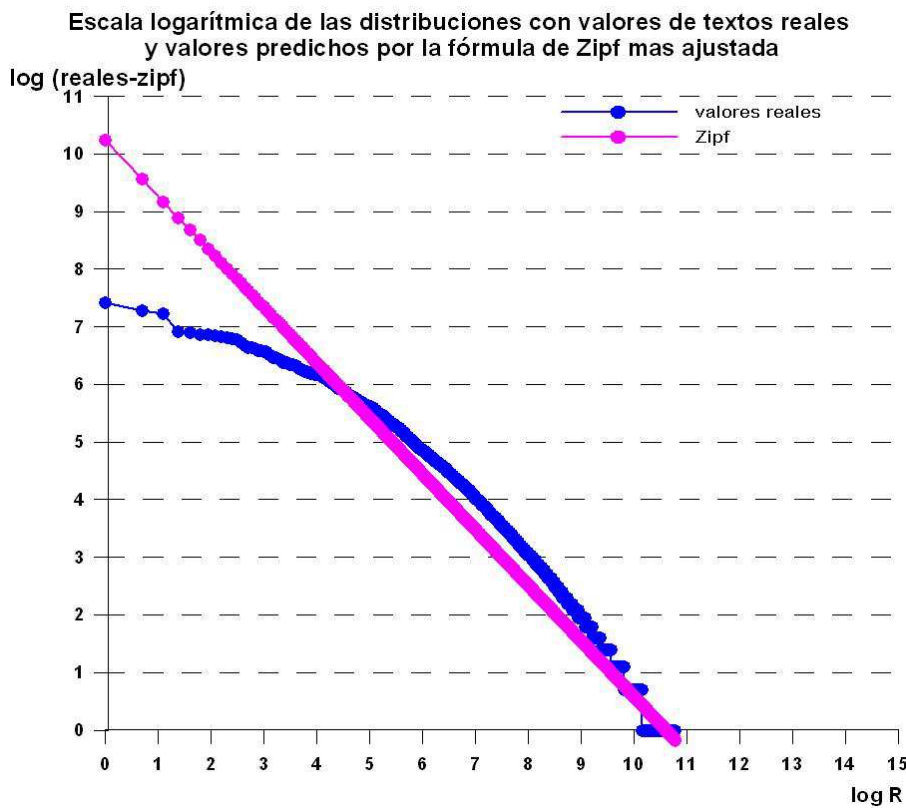


Gráfico 75. Distribución logarítmica de Zipf con valores reales y Zipf ajustada con texto grande

Los gráficos muestran tres partes claramente diferenciadas en primer lugar donde se sitúan las palabras de mayor frecuencia, es decir las palabras de mayor riqueza en el vocabulario, se considera que el total de palabras que conforman esta primera parte es aproximadamente \sqrt{V} , en segundo lugar las palabras de frecuencia intermedia se sitúan por encima de las predicciones de la Ley de Zipf y en tercer y último lugar las palabras de menor frecuencia, éstas abundan menos que lo indicado por Zipf. Esto claramente significa que la ley de Zipf siempre predice un mayor número de estas palabras de las que realmente aparecen en los textos de los autores.

Todo parece indicar que la diferencia entre la distribución teórica de Zipf mejor ajustada y la distribución real de los textos dependerá de cómo se ajuste el exponente (e).

$$Fr = \frac{k}{r^e}$$

Así de este modo, ajustaremos el exponente de la ley teórica de Zipf y daremos valores relativos dependiendo de la tipología de documentos y su tamaño para obtener unos resultados lo más aproximados a la realidad.

Se demuestra, por tanto que cuanto más grande es el documento, dentro de la misma tipología, más disminuye respecto a la fórmula de Zipf, el conjunto de palabras de menor frecuencia, dicho de otro modo o extrapolado a una idea global de la metodología del autor a la hora de escribir, podríamos afirmar que de este modo se observa como en el proceso de repetir palabras del autor, el tamaño del texto influye en el proceso de escribir el vocabulario y utilizar o crear nuevas palabras en dicho documento. O dicho de otro modo el texto crece más que el vocabulario debido al aumento del porcentaje de palabras de frecuencia alta. Este hecho ya ha sido demostrado con la Ley de Heaps desarrollado en el capítulo cinco de esta investigación.

Diversos autores se han visto sorprendidos por esta aparente falta de palabras de baja frecuencia y se han dado diversas explicaciones, todas tienen algo de verdad, pero todas resultan insuficientes

En general tienden a reconocer que hay que dividir el conjunto de rangos en dos tramos y utilizar dos fórmulas de Zipf, con distintos valores de los parámetros. Los motivos que se aducen para este comportamiento están relacionados con la utilización de frases con al menos 2-gramas, 3-gramas, etc. que los autores utilizan en lugar de palabras individuales.

La utilización de di-gramas, ya sean bi-gramas, 3-gramas, 4-gramas, 5-gramas, etc, consiste en grupos de palabras formados por frases que representan conceptos únicos, como pueden ser:

2-gramas: Comunidad Valenciana

3-gramas: Ley de Zipf

4-gramas: Universidad Politécnica de Valencia

5-gramas: Medida de Similitud del Coseno

Varios autores centran sus estudios en la utilización de n-gramas para solventar la falta de palabras de frecuencias bajas respecto a Zipf que presentan los textos analizados con palabras únicas, tratadas individualmente.

Respecto a la utilización de dos tramos y utilizar dos fórmulas de Zipf, como apuntan diversos autores podría ser utilizada, pero en este caso no se va a realizar de este modo, ya que esta modalidad se ha realizado con la Ley de Heaps, concretamente en el capítulo cinco de esta investigación, por esa razón en este caso concreto de Zipf se va a utilizar otro modelo propio llamado Log-% que se desarrollará a continuación.

Uno de los autores que apuesta por un modelo matemático para ajustar la Ley de Zipf lo más fielmente posible a los documentos reales es Montemurro (2001), artículo muy citado, establece la existencia de dos tramos en las gráficas reales de Zipf y propone un modelo matemático que lo representa.

En este caso, el autor Ferrer Cancho & Ricard (2001) apuesta por un lenguaje di-gramas, es decir analizar frases consistentes en grupos de palabras que representan conceptos únicos, este autor muestra experimentalmente la existencia de dos tramos distintos y lo relaciona con conceptos que no pueden expresarse con una sola palabra y por consiguiente necesitan una frase.

Por otro lado los autores Le Quan Ha, Sicilia-García, Ji Ming, Smith (2002) contrastan en su investigación la utilización de palabras individuales (single words) con la utilización de n-gramas (n-grams) en dos idiomas distintos en Inglés y en Mandarín, las conclusiones son interesantes al respecto, los autores llegan a utilizar hasta 5-gramas, para ver la tendencia respecto a la curva de Zipf.

La utilización de di-gramas, ya sean bi-gramas, 3-gramas, 4-gramas, 5-gramas, etc, consiste en grupos de palabras formadas por frases que representan conceptos únicos.

Como se ha manifestado anteriormente, varios autores utilizan conjuntos de palabras: 2-gramas, 3-gramas,..., n-gramas para solventar la falta de palabras de frecuencias bajas respecto a Zipf que presentan los textos analizados con palabras únicas, tratadas individualmente.

(Le Quan Ha, Sicilia-García, Ji Ming, Smith, 2002), en el caso de los textos en Inglés, comparan tres textos de gran tamaño obtenidos del Wall Street Journal de los años 1987, 1988, 1989, de menor tamaño respectivamente, las conclusiones finales son que si se extraen las palabras únicas (single words) y se representan en un gráfico frente a la curva de Zipf mejor ajustada se observa en los tres casos lo que venimos diciendo en nuestra investigación hasta ahora, y es la aparente falta de palabras de baja frecuencia respecto a las predicciones de Zipf.

Cuando los autores analizan dichos textos con n-gramas (n-grams)³⁹ observan que efectivamente se consigue solventar la aparente falta de palabras de bajas frecuencias respecto a las predicciones de Zipf, pero por otro lado se empeora la parte inicial de la

³⁹ Las experimentaciones se realizan con bi-gramas, 3-gramas, 4-gramas y 5-gramas.

curva disminuyendo considerablemente las palabras de frecuencias altas. Es decir utilizar más n-gramas es realmente agrupar varias palabras en una frase, por tanto es como el efecto de quitar las palabras vacías, estamos quitando palabras que “únicas” son muy frecuentes. Esto obviamente produce que el exponente de Zipf se reduzca considerablemente, como apuntan Le Quan Ha, Sicilia-García, Ji Ming, Smith (2002) en su artículo.

Es interesante ver los resultados con el idioma Mandarín, los autores advierten la peculiaridad de este idioma, ya que se basa en sílabas, es decir, caracteres. Una palabra o concepto puede estar formada por una sílaba o carácter, es decir un símbolo, y también otras palabras en el idioma Mandarín se forman con 2 símbolos o más, que se correspondería con la utilización de n-gramas.

En el caso del texto en idioma Mandarín si se extraen las palabras únicas (single words) y se representan en un gráfico frente a la curva de Zipf mejor ajustada, se observa que la falta de palabras de baja frecuencia que venimos observando en todos los casos es aún más acusada y más exagerada que en el idioma Inglés.

El idioma Mandarín respecto al Inglés usando uni-gramas (single words), la pendiente baja más rápidamente, obtiene un valor del exponente (e) mayor, y es muy acusada la falta de palabras de baja frecuencias ($fr=1$, $fr=2$).

Igualmente que ocurre con los textos analizados en el idioma Inglés cuando se utilizan los n-gramas en el idioma Mandarín efectivamente se consigue solventar la acusada falta de palabras de bajas frecuencias respecto a las predicciones de Zipf, pero por otro lado se empeora la parte inicial de la curva, disminuyendo considerablemente las palabras de frecuencias altas, disminuyendo el valor del exponente considerablemente.

En resumen, las palabras al unirse en bi-gramas, 3-gramas,...n-gramas obtienen menores valores de frecuencia que las palabras individualmente (single words), por eso al utilizar n-gramas se reducen bastante las palabras de frecuencias altas, es decir en la parte inicial de la curva se desvía de la de Zipf, en cambio por otro lado los n-gramas incrementan las palabras de frecuencias bajas (consiguiendo así que la parte baja de la curva se ajuste más a la predicción de Zipf).

Simplificando se puede afirmar que existen dos puntos de vista:

- ✓ Establecer un modelo mas complicado que la fórmula de Zipf / Mandelbrot para que explique el comportamiento real en dos tramos
- ✓ Modificar la base de estudio sobre las palabras, añadiéndole otras cosas, frases, para ver que si se cumple la fórmula de Zipf

Pero la cuestión a debatir sería ¿cuales son las frases que se añaden y cuales no?, hasta este momento sabemos que la Ley de Zipf se cumple, pero no del todo ya que predice más palabras de frecuencias bajas de las que realmente hay en los textos.

A partir de este momento no se utilizarán dos tramos y dos fórmulas de Zipf, como apuntan diversos autores, ya que esta modalidad se ha realizado con la Ley de Heaps⁴⁰, sino que presentaremos los datos reales tal como son, pero mediante un modelo propio de presentación desarrollado en esta investigación y que se ha denominado Log-%.

6.7. Modelo Log-% de visualización del efecto de falta de palabras en relación a las predichas por Zipf. Nueva forma de representación.

Creamos un nuevo Modelo denominado Log-% para representar el vocabulario de un texto en tramos iguales en escala logarítmica y en el que quedará representado en cada tramo el porcentaje (%) de la distribución total de frecuencias.

Con este nuevo método representaremos la distribución de frecuencias de un modo muy visual que nos permitirá ver lo que sucede realmente en el texto, mediante un gráfico o dibujo.

6.7.1. Agrupamiento en tramos: Modelo Log-%

Una vez decidido no utilizar dos tramos y dos fórmulas de Zipf para ajustar la Ley de Zipf lo más fielmente posible a la tendencia que siguen los textos reales, incluso poder solventar la aparente falta de palabras de baja frecuencia que tienen los textos reales respecto a las predicciones de Zipf, de entre todas las representaciones posibles, proponemos la siguiente que podemos llamar abreviadamente **Modelo Log-%** vamos a desarrollar un modelo propio denominado **Log-%**, y que consiste en agrupar el vocabulario en tramos iguales en escala logarítmica y asignarle como valor la suma de las frecuencias de las palabras expresado en porcentaje % del total de palabras del texto, Pero vamos a explicar todo esto más detalladamente paso a paso.

La razón de agrupar en tramos es que no tiene ningún sentido preguntar cuantas palabras hay con frecuencia 621, aunque quizá tenga sentido preguntar cuantas hay con frecuencia entre 500 y 600. Si buscamos expresiones que permitan comparar entre textos de distinto tamaño deberíamos plantearlo en términos relativos al vocabulario: ¿cuantas palabras hay con frecuencia menor que el 1% del vocabulario? O quizá dividir el rango de frecuencias en un número fijo de tramos, por ejemplo 10, siendo el primero las palabras con frecuencia 1 a X, el segundo de X+1 a 2X,...

En la aplicación⁴¹ desarrollada para realizar estos cálculos y experimentaciones se permite la exploración de todas estas posibilidades. Y en vista de los comentarios que se han realizado en páginas anteriores respecto a la escasez de palabras de baja frecuencia y a la alta cantidad de palabras muy frecuentes que manifiestan los textos reales, este agrupamiento no va a ser en tramos iguales, para permitir dar más importancia a unas palabras sobre otras.

Por consiguiente se expone paso a paso los valores en eje de abscisas y de ordenadas para la representación del **Modelo Log-%**

⁴⁰ Véase capítulo 5.

⁴¹ Base de Datos TOPOS formulario Analiza Zipf

6.7.2. Modelo Log-% : valores en el eje de abscisas

En el eje de abscisas vamos a representar el conjunto de rangos del total de vocabulario del texto agrupado en 10 tramos iguales en escala logarítmica, esto es; agruparemos en 10 tramos iguales el eje de abscisas con el vocabulario para ver el comportamiento de las palabras más frecuentes y las menos frecuentes en qué tramos quedan representadas, advertiremos que las palabras más frecuentes quedan representadas en los primeros tramos y las menos frecuentes en los tramos finales, y por tanto se observa que existe más palabras en unos tramos que en otros, en consecuencia convertiremos los 10 tramos desiguales en la escala real, en 10 tramos iguales en la escala logarítmica, y de este modo veremos con más detalle los tramos finales donde se sitúan los LNRE, como los hapaxes, y donde estas palabras de menor frecuencia abundan menos que lo indicado por Zipf, siendo la predicción de Zipf mayor en todos los casos.

6.7.2.1. Procedimiento preliminar para la representación del Modelo Log-%.

Agrupamiento en tramos en el rango.

Se introduce el procedimiento inicial con un ejemplo aclaratorio y sencillo de la metodología seguida, imaginemos el texto del autor Larra.txt, el cual una vez procesado por la base de datos Piedras.mdb, se obtiene:

P=78.719 palabras totales en el documento

V=16.449 palabras distintas o vocabulario

En el eje de abscisas representamos rangos de palabras distintas, es decir del vocabulario; por tanto sus valores van a ir desde 1 hasta el total del vocabulario V=16.449, se divide en un número fijo de tramos que establecemos en 10. Si los tramos son iguales, es decir contienen el mismo número de palabras (distintas) los límites entre ellos son los números:

$$\frac{V}{10}, 2 \cdot \frac{V}{10}, 3 \cdot \frac{V}{10}, \dots 10 \cdot \frac{V}{10} = V$$

Podemos hacer las operaciones con una consulta de selección sobre la tabla paladistin, diseñada de la siguiente forma

Campo:	rango	Expr1: Ent([rango]/1644,9)	fr	
Tabla:	paladistin		paladistin	
Total:	Máx	Agrupar por	Suma	
Orden:	Ascendente			
Mostrar:	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Criterios:				
o:				

Figura 10. Procedimiento preliminar para representación del Modelo Log-%

Que produce como resultado la siguiente tabla:

tramosiguales		
MáxDerango	Expr1	SumaDefr
1644	0	50592
3289	1	8846
4934	2	5186
6579	3	3290
8224	4	2580
9869	5	1645
11514	6	1645
13159	7	1645
14804	8	1645
16448	9	1644
16449	10	1

Tabla 22. Resultados del procedimiento preliminar para representación del Modelo Log-%

En este caso desecharemos la última fila con solo una palabra. La tabla contiene en el primer tramo los rangos de 1 hasta 1644, la décima parte del total de rangos; el segundo tramo desde el 1645 hasta el 3289, ...etc. En la última columna tenemos la suma de las frecuencias, lo que significa, el total de palabras que encontramos en el texto pertenecientes a ese tramo; o mejor la parte del texto construida con palabras de ese tramo.

Si expresamos el porcentaje de palabras sobre el total $P=78719$, obtenemos el porcentaje de palabras distintas que se agrupan en cada tramo.

tramosigu			
MáxDerango	Expr1	SumaDefr	tpc
1644	0	50592	64,26911
3289	1	8846	11,23744
4934	2	5186	6,58799
6579	3	3290	4,179423
8224	4	2580	3,277481
9869	5	1645	2,089711
11514	6	1645	2,089711
13159	7	1645	2,089711
14804	8	1645	2,089711
16448	9	1644	2,088441

Tabla 23. Porcentaje de vocabulario agrupadas en cada tramo del Modelo Log-%

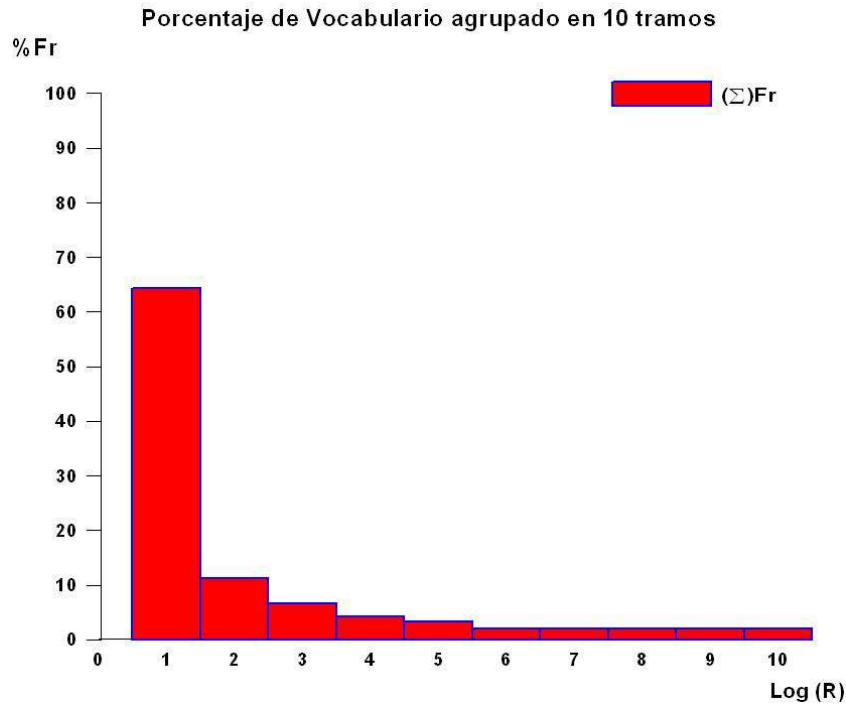


Gráfico 76. Porcentaje de vocabulario agrupado en 10 tramos. Modelo Log-%

Vemos que esta representación no transmite con claridad las características específicas que pueda tener este texto. La única idea que obtenemos es que el 10% de palabras más frecuentes, primera barra de la izquierda, es responsable de más de la mitad del texto. Pero esto ocurre en todos los textos. Esta representación nos informa de lo que ya sabemos de la ley de Zipf.

Lo que buscamos es una representación que nos informe de las “pequeñas” desviaciones de un texto en particular, sobre el comportamiento general de todos los textos. Para ello vamos a equilibrar los tramos: que los primeros sean más pequeños y los últimos más grandes.

6.7.2.2. Procedimiento definitivo para la representación del Modelo Log-%. Agrupamiento en tramos logarítmicos en el rango.

Se divide el eje de abscisas $\log(V)$ en diez partes iguales para definir los tramos por puntos de subdivisión T_1, T_2 , etc. Dicho de otro modo, el primer tramo va a contener las palabras de rango de frecuencias entre 0 (exclusive) y T_1 (inclusive); el segundo entre T_1 (exclusive) y T_2 (inclusive),...así hasta T_{10} . Por ejemplo, si $T_1=2,8$ el primer tramo, entre 0 exclusive y 2,8 inclusive significa las palabras de rango 1 y 2, es decir las dos palabras mas frecuentes.

Los números T_1, T_2^{42} , .. se determinan de modo que:

$$\log(T_1) = \frac{\log(V)}{10}$$

⁴² T_1 =tramo 1; T_2 = tramo 2

$$\log(T2) = \frac{2 \cdot \log(V)}{10}$$

... hasta

$$\log(T10) = \frac{10 \cdot \log(V)}{10} = \log(V)$$

Invirtiendo las fórmulas, pueden expresarse como

$$T1 = V^{\frac{1}{10}}$$

$$T2 = V^{\frac{2}{10}}$$

...

$$T10 = V^{\frac{10}{10}} = V$$

Para este caso la consulta formulada es:

Campo:	rango	Expr1: Ent(Ln([rango])*10/Ln(16449))	fr	
Tabla:	paladistin		paladistin	
Total:	Máx	Agrupar por	Suma	
Orden:	Ascendente			
Mostrar:	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Criterios:				
o:				

Figura 11. Procedimiento definitivo para representación del Modelo Log-%

Y los resultados que se obtienen:

Consulta1		
MáxDerango	Expr1	SumaDefr
2	0	7932
6	1	2028
18	2	2936
48	3	4032
128	4	6254
338	5	8555
893	6	11136
2359	7	12387
6230	8	11956
16448	9	11502

Tabla 24. Resultados del procedimiento definitivo para representación del Modelo Log-%

De esta manera hemos encontrado un compromiso entre los distintos objetivos de la representación:

- 1) el número de tramos es lo suficientemente pequeño para una representación visual comprensible a la primera ojeada. (El número 10 es una elección arbitraria)

- 2) se concede suficiente importancia a las palabras de alta frecuencia que quedan representadas en los primeros tramos.
- 3) las palabras de menor frecuencia quedan adecuadamente representadas en los últimos tramos, aunque la correspondencia no es exacta, el último tramo está formado casi exclusivamente por palabras de frecuencia 1, y unas pocas de frecuencia 2.

Podría forzarse la representación ajustando la escala logarítmica o el número de tramos para que el último coincidiera con las palabras de frecuencia 1 y así, esta representación contendría la parte más útil de la distribución transformada. Pero las complicaciones derivadas de ello anularían la simplicidad del modelo y no sería práctica su utilización para comparar textos.

6.7.3. Modelo Log-%: valores en el eje de ordenadas

En el eje de ordenadas vamos a representar la suma de las frecuencias de las palabras expresado en porcentaje, para establecer el porcentaje de palabras que se hallan en cada tramo.

Por ejemplo, el tramo primero contiene la suma de las frecuencias de las 1.644 palabras más frecuentes, se obtiene $(\Sigma) Fr=50.592$ frecuencias. Esto representa la parte del texto total constituida por repeticiones de estas palabras.

Para obtener una representación con la que comparar textos de distinto tamaño, no expresaremos estas cantidades de manera absoluta sino en tanto por cien de su contribución a formar el texto total, es decir sobre el total de palabras⁴³.

En todo caso, aún antes de convertir los datos en porcentajes, ya vemos en la anterior tabla que los valores quedan del mismo orden de magnitud en todos los tramos. ¿Será cierto que en un texto que cumpla perfectamente la Ley de Zipf, los diez valores en el eje vertical son iguales? si fuera así, el gráfico obtenido nos mostraría visualmente en qué se aparta de la ley este texto concreto.

Si realizamos esta construcción de tramos con muchos ejemplos, la programamos en el formulario de la aplicación⁴⁴ al uso resultaría del siguiente modo:

⁴³ Base de Datos TOPOS formulario Analiza Zipf, tabla reduzipf

⁴⁴ Base de Datos TOPOS formulario Analiza Zipf

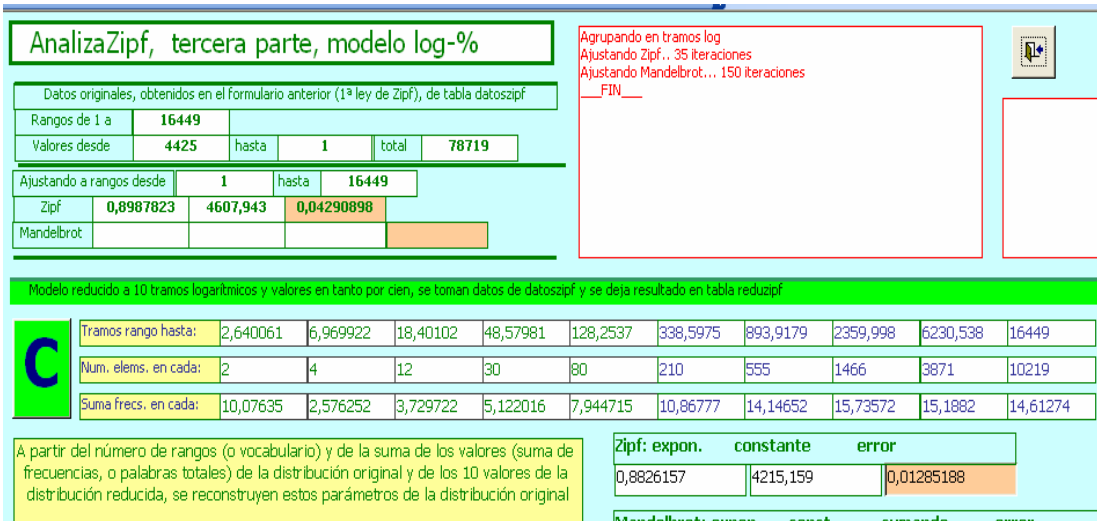


Figura 12. Base de datos del Modelo Log-%

De la imagen anterior se extraen los datos pertenecientes al Modelo reducido a 10 tramos logarítmicos y sus valores en tanto por cien, la suma de frecuencias en cada tramo y obtenemos el correspondiente gráfico del texto Larra.txt:

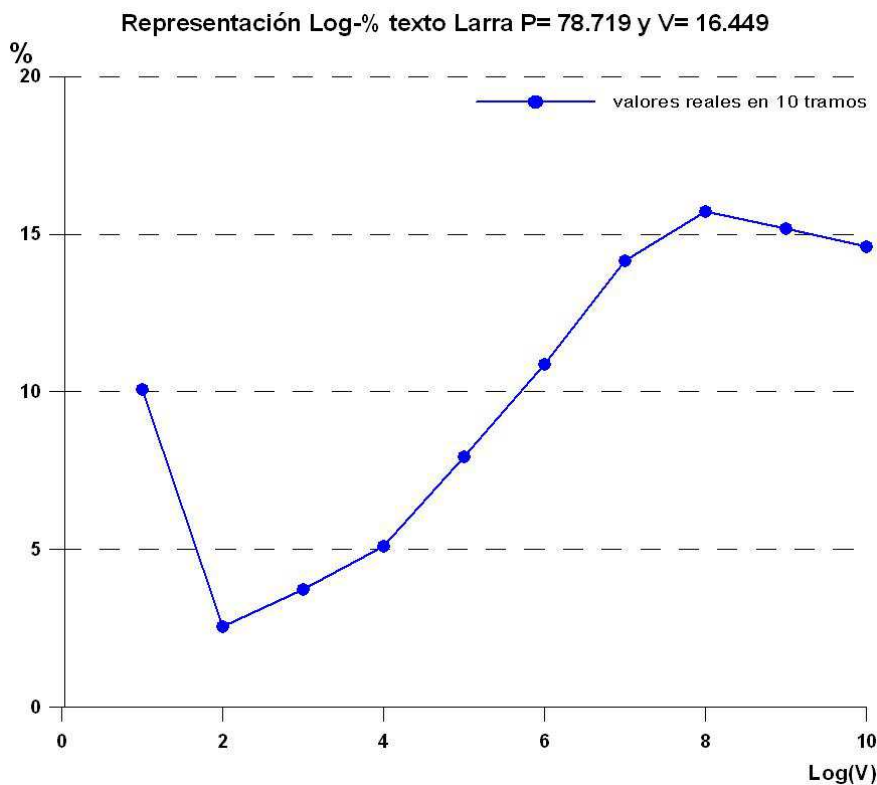


Gráfico 77. Representación Log-% con texto de P=78.719 y V=16.449

Para comparar con hipotéticos textos que cumplan perfectamente la ley de Zipf, en la aplicación⁴⁵ correspondiente se ha desarrollado otro formulario para tal fin, que genera artificialmente un texto sintético ajustado perfectamente a la ley, aunque siempre se encuentran errores de discretización en las frecuencias más altas

⁴⁵ Base de Datos TOPOS formulario TextoSintetico

Al generar un texto sintético existe un error de discretización en las frecuencias más altas y por esa razón aparece el primer y último punto de la gráfica desviado de la tendencia que sigue el resto, este error ocurre por los decimales en los cálculos ya que la aplicación que realiza los cálculos para conseguir un texto perfecto, es decir con un exponente $e=1$, la aplicación incluye decimales y como estamos tratando palabras, éstas lógicamente sólo pueden tener valores enteros, de ahí que al comienzo de la gráfica y en la parte final aparezcan dichos errores tan exagerados.

En el siguiente gráfico comparamos los valores obtenidos en el gráfico anterior de un texto real Larra.txt, con dos textos sintéticos (casi) perfectos, en los textos sintéticos cada uno de ellos con distinto exponente para observar que la inclinación de la recta corresponde a este hecho, al valor del exponente.

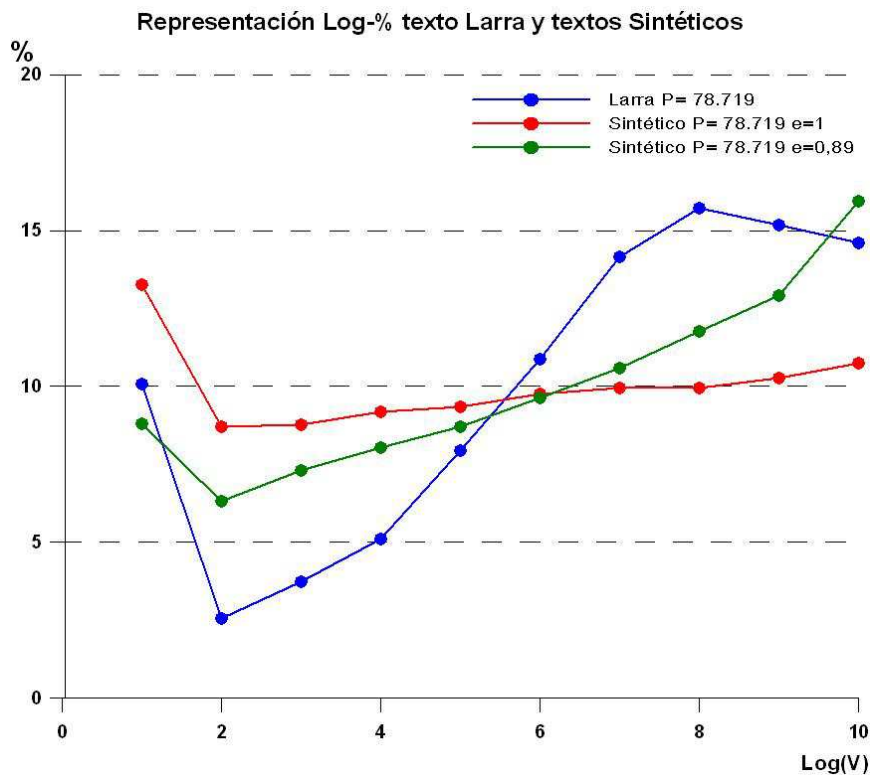


Gráfico 78. Representación Log-% con texto real y sintéticos

La línea coloreada en rojo casi horizontal corresponde al texto sintético de exponente de Zipf igual a 1 (en realidad 0.98): idealmente sería una recta horizontal.

La línea coloreada en verde casi recta pero inclinada es la de un texto sintético con exponente de Zipf igual a 0.89, y la línea coloreada en azul corresponde al texto real Larra.txt, con exponente de Zipf igual a 0,88 similar a la línea verde de un texto sintético, de este modo podemos apreciar la diferencia entre ellos y observar como el texto real no cumple la Ley de Zipf y el texto sintético sí, salvando las excepciones en los extremos de la distribución debidas a que la aplicación realiza los cálculos con decimales y al aplicarlo a palabras estas sólo pueden tener valores enteros.

¿En qué se diferencia nuestro texto de un texto ideal?:

- 1) En las frecuencias altas: tiene menos palabras en las frecuencias altas: tramos 2,3 y 4 y más palabras de frecuencia muy alta.

- 2) En las frecuencias intermedias: parece haber abundancia, tramos 6,7 y 8
- 3) En las frecuencias bajas: hay escasez en las de frecuencia 1 y 2, tramos 9 y 10

En los gráficos siguientes se muestra el texto: *Crónica internacional 1890-1898* del autor: *Emilio Castelar*, subdividido en distintos tamaños, el *CASTELA4* P= 450.574 sería el documento completo y de éste se subdividen tres tamaños más pequeños como el *CASTELA3* P= 367.570, *CASTELA2* P= 244.324 y el más pequeño el texto *CASTELA1* P= 112.466 palabras. Estos gráficos representan los datos reales formando la distribución de frecuencias de Zipf, y distribuyéndolo en los 10 tramos logarítmicos de la representación gráfica del **Modelo Log-%**.

Por tanto los valores representados son valores distribuidos según la teoría de Zipf pero no obtenidos de la fórmula de Zipf.

En estos gráficos vemos como el texto del autor Castelar distribuido según la teoría de Zipf con nuestro Modelo Log-% muestra una relación entre el vocabulario y su frecuencia prácticamente lineal creciente hasta que la curva decae considerablemente obteniendo valores más bajos, recordemos que este Modelo, al representar el eje de ordenadas, en escala logarítmica nos permitirá visualizar lo que ocurre en la parte final de la curva, es decir con las palabras de bajas frecuencias que tanto comentan varios autores como Le Quan Ha, Sicilia-García, Ji Ming, Smith (2002), y de lo que tanto hemos tratado en este Capítulo.

La visión de estos gráficos nos confirma los estudios realizados por dichos autores y los ejemplos realizados en nuestro caso, indicándonos incluso que a mayor tamaño la cantidad de palabras de bajas frecuencias también es mayor, por tanto se aparta todavía más de la predicción de Zipf, como puede apreciarse en la curva cada vez más acusada.

Como hemos mencionado anteriormente para que un texto real se ajuste lo más posible a la Ley de Zipf utilizando nuestro Modelo Log-% como modelo de distribución de Zipf, los datos deben asemejarse lo más posible a una línea recta creciente, y como podemos ver en los gráficos siguientes utilizando un texto real del autor Castelar, esto parece cumplirse pero observamos claramente gracias a esta distribución de Log-% el efecto que producen las palabras de frecuencias bajas consiguiendo así no ajustarse a la predicción de Zipf.

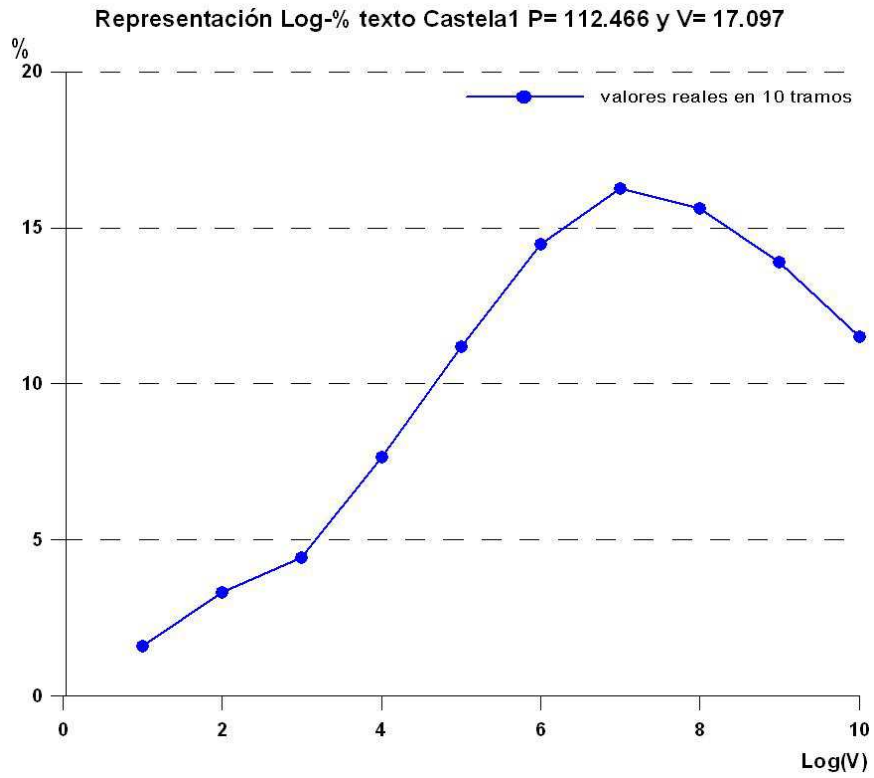


Gráfico 79. Representación Log-% con texto de P=112.466 y V=17.097

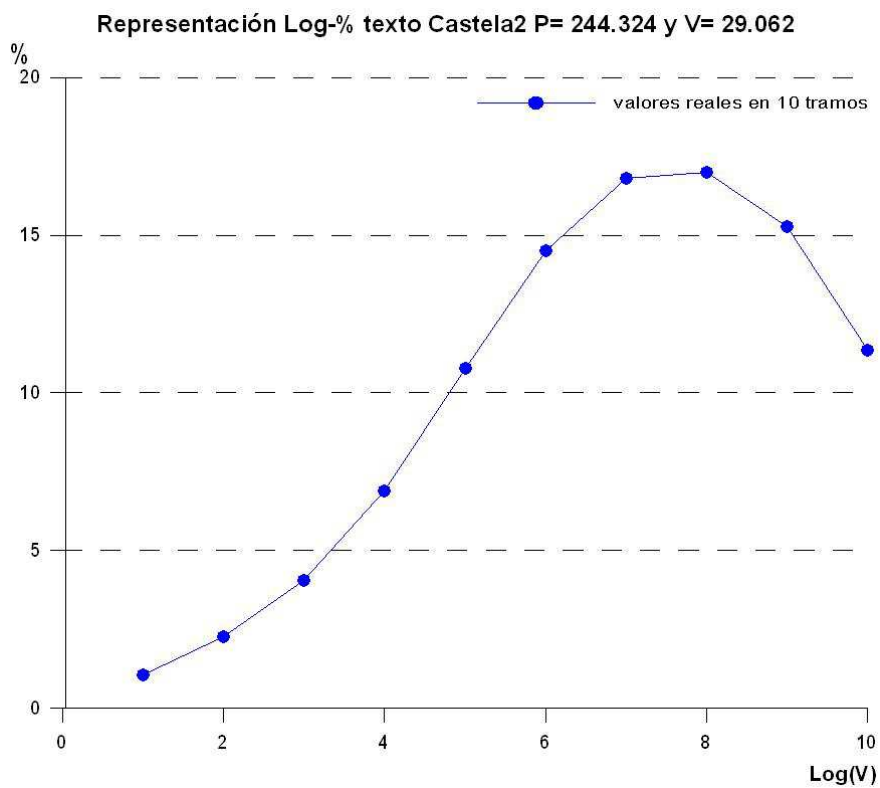


Gráfico 80. Representación Log-% con texto de P=244.324 y V=29.062

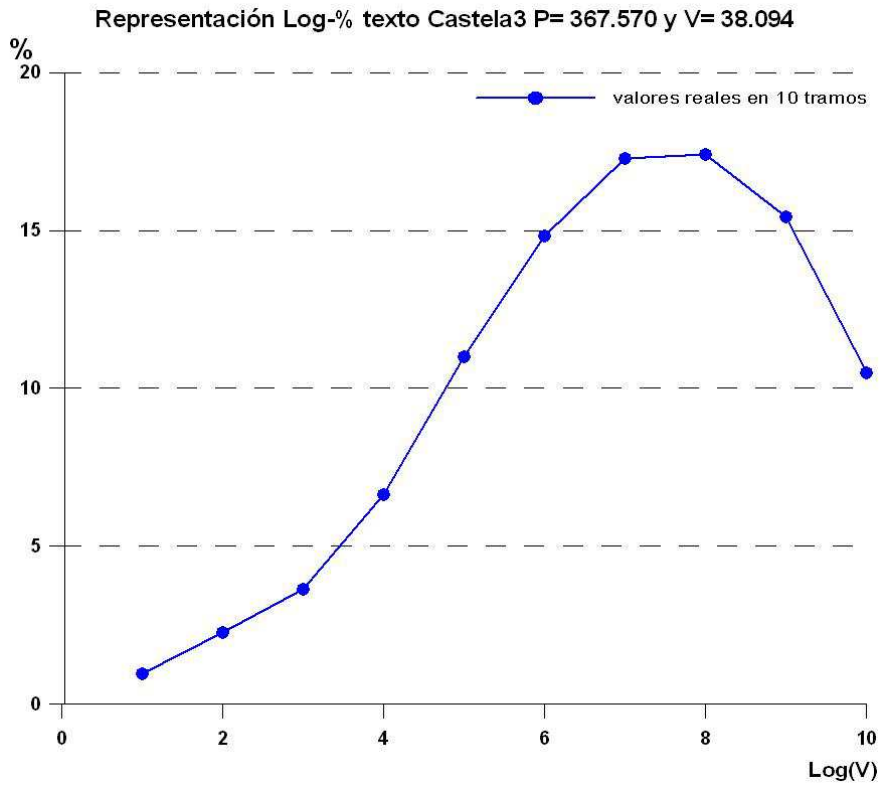


Gráfico 81. Representación Log-% con texto de P=367.570 y V=38.094

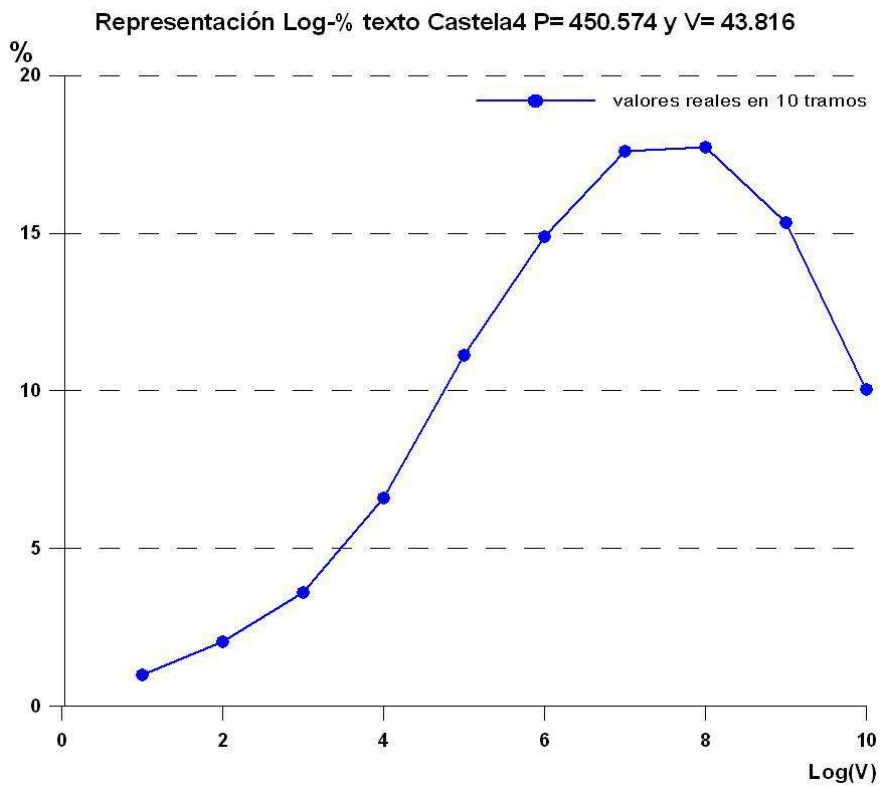


Gráfico 82. Representación Log-% con texto de P=450.574 y V=43.816

Por consiguiente muchos autores han tratado este asunto, dando explicaciones diversas al hecho aparente de que falten palabras de baja frecuencia. ¿Se trata de que la fórmula de Zipf no es la adecuada para representar el lenguaje?, o ¿se trata de que el lenguaje real tiene accidentes que nos hacen verlo como distinto del lenguaje hipotético que sí cumple la fórmula de Zipf? Parece claro que es esto último: no solo el proceso de agrupar en raíces o extraer palabras vacías introduce una arbitrariedad, es que tenemos el caso de los conceptos que no tienen una palabra única, los n-gramas (Si un autor ha querido nombrar el concepto de Comunidad Valenciana solo una vez en un texto, no lo hemos tomado así ya que lo hemos mezclado con comunidad de vecinos y con paella valenciana. Es un accidente el que falte la palabra Comunidad Valenciana para que tenga frecuencia 1. Si se tratase de Galicia no hubiera pasado esto)

Entonces parece que tenemos una pauta natural en la ley de Zipf y tienen un sentido real las desviaciones respecto a ella. Pero, por razones prácticas, debemos quitar las palabras vacías y debemos agrupar en raíces y al hacerlo ya no tenemos el modelo inapelable de la ley de Zipf con exponente 1.

En definitiva, la aplicación de este modelo Log-% proporciona una información visual, más cualitativa que cuantitativa del aspecto general de un texto en relación con las frecuencias de sus palabras, exactamente con la cantidad de palabras de cada frecuencia. De igual modo visto que este modelo en diez tramos Log-% proporciona información esencial y suficiente sobre como es un texto, abre el camino para utilizarlo, tal como se hará en los procesos finales de esta tesis doctoral, como agrupaciones de palabras o de sus frecuencias.

6.7.4. Punto de Transición (Transition Point)

Una vez decidido que debemos hacer una adaptación del ajuste por tramos a la zona donde si que sigue una tendencia constante y por tanto cumple la Ley de Zipf, estos ajustes parciales se aplicarán sobre todo en el capítulo siete, en el cual se realizarán los ajustes a las palabras que se encuentran en la parte central, es decir entre los rangos centrales.

En dicho capítulo se realiza un estudio detallado de la distribución de Zipf aplicado a raíces, una vez visto los resultados obtenidos con las raíces, se ajustará la fórmula de Zipf al tramo central de cada colección de valores, este tramo central correspondería a los rangos entre $V^{0.3}$ y $V^{0.8}$. Esto sugiere que las palabras que caracterizan un texto no sean ni las más frecuentes ni las menos frecuentes, sino las que se encuentran en una frecuencia media de ocurrencia dentro del texto (Luhn, 1958). Para realizar el ajuste a los datos centrales se hallará el punto de transición o PT (*Transition Point TP*).

Según Booth (1967) el punto de transición o PT se puede definir como la frecuencia que divide al vocabulario de un texto en dos: palabras de alta y baja frecuencia, Booth fundamenta el cálculo y el empleo del punto de transición con la siguiente expresión

$$PT = \frac{\sqrt{1 + 8I_1} - 1}{2}$$

Donde I_1 es el número de palabras con frecuencia 1. Se propone el Punto de Transición (PT) como el centro de la zona de palabras de alto contenido informativo. Según Booth (1967) y Urbizagastegui-Alvarado (1999), una vez seleccionados los términos representativos de una parte denominada párrafo virtual, se buscan las oraciones con similitud mayor con el conjunto de palabras de rango próximo al PT y estas constituyen el extracto final de términos. Según Urbizagastegui-Alvarado (1999) existe una vecindad de términos alrededor del punto de transición que describe de manera general el contenido del mismo texto, donde el 25% de las palabras alrededor del PT se encuentran las palabras clave del texto.

Un ejemplo sencillo de obtención del Punto de Transición en un texto de Menéndez Pelayo en el que aplicamos la fórmula:

$$V = 13.342$$

$$\text{Hapaxes} = 7.599 = I_1 = 7.599$$

$$PT = \frac{\sqrt{1 + 8I_1} - 1}{2} = 122.78$$

En este ejemplo, la representación con el Modelo Log-%⁴⁶ obtiene las siguientes subdivisiones:

<i>Palabras</i>	<i>V01</i>	<i>V02</i>	<i>V03</i>	<i>V04</i>	<i>V05</i>	<i>V06</i>	<i>V07</i>	<i>V08</i>	<i>V09</i>	<i>V10</i>
<i>En cada</i>	2	4	11	27	71	183	474	1224	3164	8182
<i>Acumuladas</i>	2	6	17	44	115	298	772	1996	5160	13342

Tabla 25. Ejemplo de las subdivisiones en 10 tramos del Modelo Log-%

Vemos que el PT queda en el tramo V06. Una elección escasa de palabras alrededor del PT podrían ser los tramos V05, V06, V07 que en total tienen 728 palabras, o sea un 5% del vocabulario. Otra elección más abundante sería tomar los tramos desde el V04 al V08. En este caso tendremos 1979 palabras que representan un 15% del vocabulario.

Según manifiestan los autores citados anteriormente, las palabras más significativas tienen valores del rango alrededor del PT, Nuestra clasificación en tramos de rango permite una clasificación sencilla de ellas. Además evita un desplazamiento excesivo hacia las frecuencias altas, al ser los V01, V02, V03 muy pequeños.

6.8. Estudios complementarios: Cálculos fórmulas ajustadas de Zipf y Mandelbrot. Pasos 1-4

La averiguación de los valores de los parámetros tanto del exponente como de los coeficientes en las fórmulas de Zipf y Mandelbrot se ha llevado a cabo mediante un proceso o técnica estándar que nos permitirá obtener un valor verosímil del exponente (e) y coeficientes, es decir un valor verosímil sería realmente un valor inventado, ajustado matemáticamente que permitiría ofrecer una predicción del valor de dichos

⁴⁶ Base de datos TOPOS, formulario AnalizaZipf

parámetros y de este modo se podrán comparar los resultados obtenidos de las predicciones mejor ajustadas y de los datos obtenidos de un texto real.

Para determinar dichos valores de los parámetros en las fórmulas de Zipf y Mandelbrot y conseguir de este modo su ajuste a las frecuencias de las palabras en un texto, el método empleado es un proceso de cuatro pasos que se detalla pormenorizadamente a continuación:

PASO 1

En primer lugar se toma un valor verosímil de los parámetros:

El exponente en la fórmula de Zipf

El exponente y el sumando en la fórmula de Mandelbrot

PASO 2

En segundo lugar se determina por cálculos de compatibilidad con el vocabulario y el total de palabras, el único valor de la constante que es compatible con estos valores de los parámetros, seguidamente se calcula para cada valor del rango, la frecuencia que predice la fórmula.

PASO 3

Una vez hecho esto se comparan los valores predichos con los reales, calculando el error relativo cuadrático promedio:

Error = valor predicho - valor real

Error relativo = error / valor predicho

Error relativo cuadrático = error relativo elevado al cuadrado

Error relativo cuadrático promedio = suma de los errores relativos cuadráticos para cada uno de los rangos, dividida por el total de ellos, es decir, por el vocabulario

PASO 4

En último lugar se modifican el valor de los parámetros y se repiten varias veces los pasos 2) 3) y 4) guiados por un algoritmo matemático de minimización, para detenerse cuando el error sea mínimo. Todo este procedimiento está programado en el formulario AnalizaZipf de TOPOS⁴⁷

Conviene observar que la utilización del error relativo cuadrático promedio es estándar, no estaría justificado hacerlo de otro modo, pero aquí se está considerando que todas las palabras contribuyen por igual a la medida del error. En otras ocasiones se ha decidido dar más importancia a las palabras más frecuentes o a las menos frecuentes, o a las intermedias, con lo que el ajuste tiene un carácter distinto y se obtienen valores distintos de los parámetros.

Con la utilización de esta técnica estándar explicada anteriormente nos permitirá comparar la tendencia de los valores de los datos reales de los textos y los valores de los datos mejor ajustados a las fórmulas de Zipf y Mandelbrot, para estudiar finalmente las leyes de Zipf y Mandelbrot en qué aspecto se ajusta más o menos a la realidad.

⁴⁷ Ver Apéndice II

A continuación se detalla la técnica utilizada para la obtención de las gráficas sintéticas desarrolladas en el apartado 6.3.4 de este capítulo, para explicar el significado de las fórmulas de Zipf y Zipf-Mandelbrot

Como recordatorio denominamos “Sintético” a los valores que se obtienen a partir de cálculos sobre las fórmulas, sin efectuar ningún recuento sobre un texto real, y para ello se ha utilizado la aplicación TOPOS y el formulario SignificadoZipf, este formulario trabaja sobre un hipotético texto de 400 palabras con 50 palabras distintas, datos adecuados para el dibujo de una gráfica. Igualmente pueden cambiarse si se considera necesario.

El sistema fabrica, de una vez, cuatro series de valores de frecuencias. Para Zipf se le dan cuatro valores del exponente. Para Zipf-Mandelbrot se le dan cuatro valores del exponente y otros cuatro del sumando. Los resultados quedan en la tabla “números”, campos y1, y2, y3, y4. Finalmente para la realización de las gráficas se copian una de estas columnas o varias de ellas a una hoja de cálculo y se obtiene así la representación directa de frecuencias frente a rangos.

Para obtener la gráfica sintética de la distribución de Zipf que mejor aproxima a una distribución sintética Zipf-Mandelbrot, con la base de datos TOPOS y el formulario SignificadoZipf obtenemos 4 columnas de frecuencias, ordenadas por rango, correspondiente a una distribución de Zipf-Mandelbrot con sumando distinto de cero que se almacena en la tabla “números”. Seguidamente copiamos a una hoja de cálculo Excel esta columna de frecuencias. Para obtener el exponente de la distribución de Zipf que mejor se ajusta, utilizamos el formulario TOPOS, AnalizaZipf, llevando allí la tabla “números” y la columna deseada. Ajustamos Zipf y copiamos el exponente. Seguidamente volvemos al formulario SignificadoZipf y con el exponente que hemos copiado generamos la serie sintética de frecuencias, que volverá a quedar en la tabla “números”. Para finalmente copiar a Excel esta columna, en paralelo a la anterior y realizamos el gráfico.

6.9. Estudios complementarios: Cómo afecta las palabras vacías y la extracción de raíces al valor del exponente (e) de Mandelbrot.

Como ya sabemos aunque todas las palabras de un texto verifiquen aproximadamente la ley de Zipf con exponente igual a 1, si quitamos algunas de las palabras mas frecuentes, lo que hacemos al suprimir las palabras vacías es destruir la relación entre la frecuencia de la palabra más frecuente y el resto de las frecuencias que permanecen inalteradas.

La representación logarítmica de la gráfica no solo aparece menos inclinada (lo que implica un exponente menor en la fórmula de Zipf) sino que aparece curvada, lo que resulta en un ajuste más defectuoso de cualquier fórmula de Zipf, ya que, sea cualquiera el exponente, siempre es una recta en esta representación.

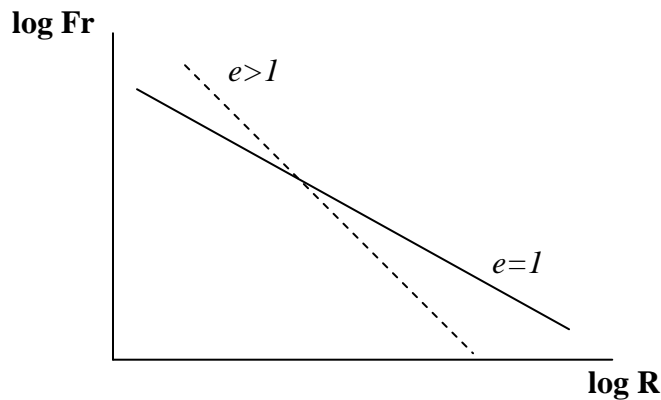


Figura 13. Distribución de Zipf según valor del exponente

Vamos a experimentar a continuación hasta que punto la fórmula de Mandelbrot como generalización de la de Zipf es una adaptación satisfactoria a esta circunstancia

En la siguiente gráfica se muestran los datos correspondientes al texto de larra1.txt después de haber quitado las palabras vacías de una tabla de 72 palabras; los valores obtenidos son $P=76.550$ $V=16.521$ y dando como resultado una vez realizado el ajuste de Mandelbrot

$e=1.035$, $K=16005.01$, $\Sigma=39.706$, $error=0.036$

Comparación logarítmica de las distribuciones con valores de textos reales y la predicción de Mandelbrot con tabla de palabras vacías= 72 p.

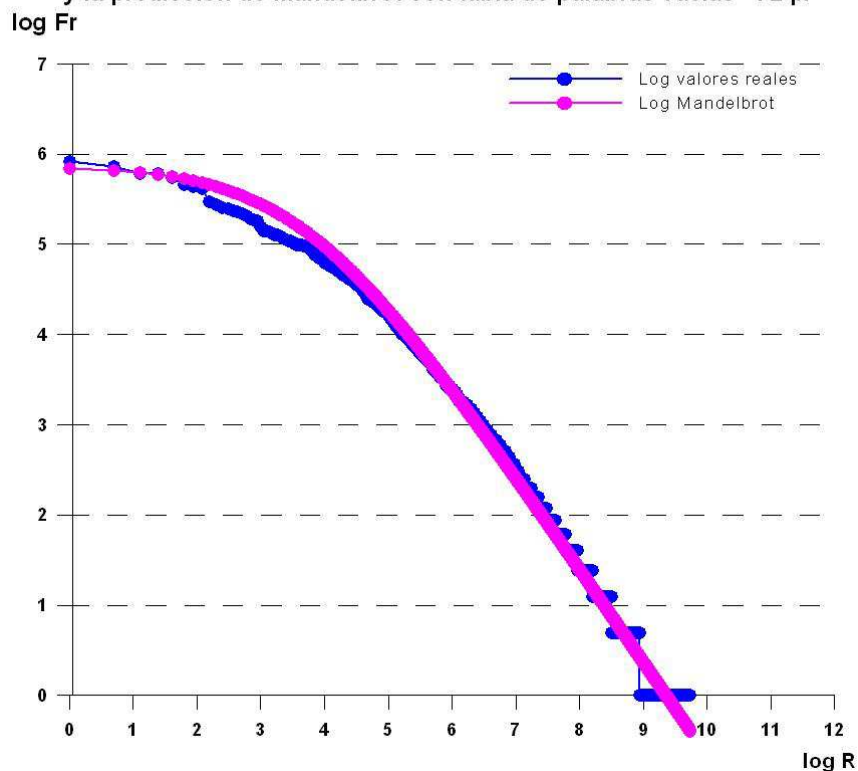


Gráfico 83. Distribución logarítmica de Mandelbrot ajustada con textos reales y $p_v=72$

Podemos considerar el ajuste como bueno en primera instancia, aunque hay una diferencia visible por un escalón en las palabras más frecuentes que impide que la curva regular de la fórmula se adapte completamente a las irregularidades de los datos reales.

Verificamos la hipótesis de que sea debido a posibles palabras vacías que aún no hemos quitado. Revisando las palabras mas frecuentes, encontramos algunas que podrían considerarse vacías y que no figuraban en la lista de palabras a excluir, las incorporamos pasando a una lista de 386 palabras vacías, con la que repetimos todo el proceso.

El análisis resultante ha disminuido el número de palabras totales a 61.434, muchas menos que antes, mientras que el vocabulario queda en 16.375, aproximadamente el mismo que antes. Es decir, se ha modificado la estructura de la colección de frecuencias. Los datos para la fórmula de Mandelbrot que mejor ajusta son:
 $e=0.942$, $K=6698.143$, $\Sigma=30.986$, $error=0.035$

Comparación logarítmica de las distribuciones con valores de textos reales y la predicción de Mandelbrot con tabla de palabras vacías= 386 p.
 log Fr

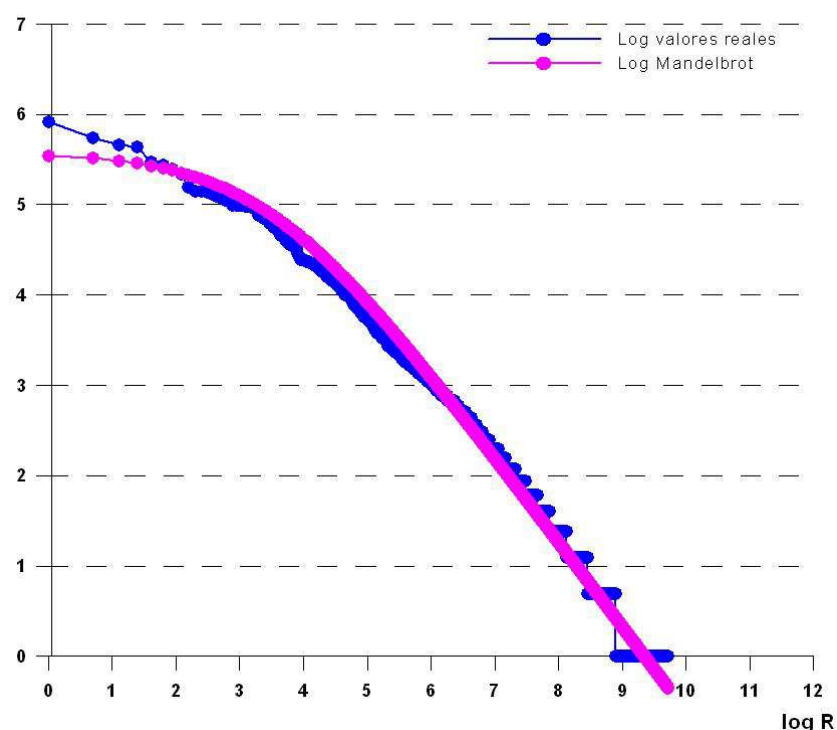


Gráfico 84. Distribución logarítmica de Mandelbrot ajustada con textos reales y pv=386

El tipo de ajuste que observamos es similar al anterior y sigue existiendo una diferencia visible por un escalón en las palabras de mayor frecuencia. Probablemente esto sea cosustancial al proceso y el escalón sea debido a palabras que nadie consideraría como vacías porque tienen un significado preciso, pero que en este texto son funcionalmente vacías por que aparecen en casi todos los documentos. (Podría deberse a que en el texto utilizado larral.txt existe la palabra Enrique que es un nombre propio muy abundante).

Otro asunto a considerar es el de la sustitución del vocabulario por la colección de raíces, que tiene el efecto de retorcer hacia abajo la parte derecha de la gráfica en la representación Log-%, separándola de la línea recta que representa la distribución perfecta de Zipf con exponente igual a 1.

Pero veamos la siguiente gráfica del texto de Larra denominado larra1.txt, habiendo quitado palabras vacías y pasado a raíces con el método de sufijos y utilizando una tabla de 358 sufijos. $P=76.550$ $V=16.521$ y $R=7.841$. Representamos en log-log los valores reales y el mejor ajuste obtenido para Mandelbrot es $e=1.528$, $K=651398.1$, $\Sigma=217.148$ $error=0.034$

Comparación logarítmica de las distribuciones con valores de textos reales y la predicción de Mandelbrot con tabla de sufijos= 358 p.

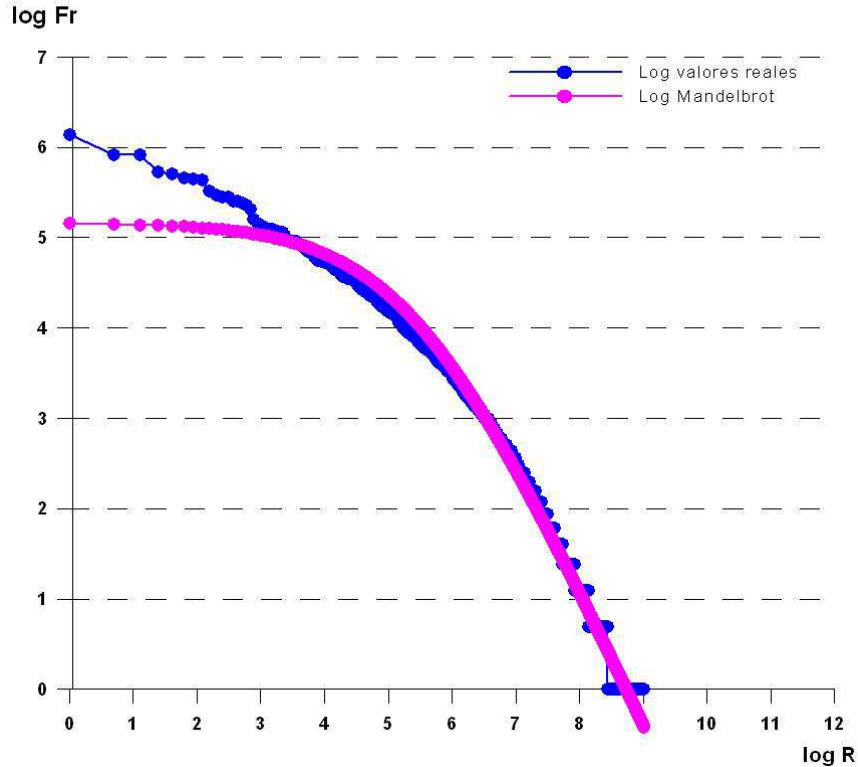


Gráfico 85. Distribución logarítmica de Mandelbrot ajustada con textos reales sufijos=358

Vemos que el modelo de Mandelbrot debido a sus grados de libertad en la elección de exponente y de sumando se adapta al lenguaje real, reproduce la cantidad de palabras de frecuencias bajas. Desde cierto punto de vista esto no es bueno, ya que sabemos que es un accidente la escasez de estas palabras y quizá fuera interesante que el modelo reprodujera algo más esencial y no los accidentes.

Por otra parte, el modelo no es lo suficientemente flexible como para doblarse y no se adapta a las frecuencias más altas como se puede apreciar en el gráfico siguiente.

Como tenemos que trabajar quitando palabras vacías y agrupando las palabras en raíces, es decir, introduciendo accidentes arbitrarios, además de los que ya hay en el lenguaje, en los dos extremos de la curva, tomaremos como modelo “natural” de Mandelbrot el que resulta de ajustar en la parte central los tramos comprendidos entre los rangos 2-8 en el ajuste parcial cuya gráfica es la siguiente y en la cual se obtienen los siguientes valores: $e=1.040$, $K=19266.88$, $\Sigma=75.658$ $error=0.0017$

Comparación logarítmica de las distribuciones con valores de textos reales y la predicción de Mandelbrot con ajuste parcial en los tramos 2-8

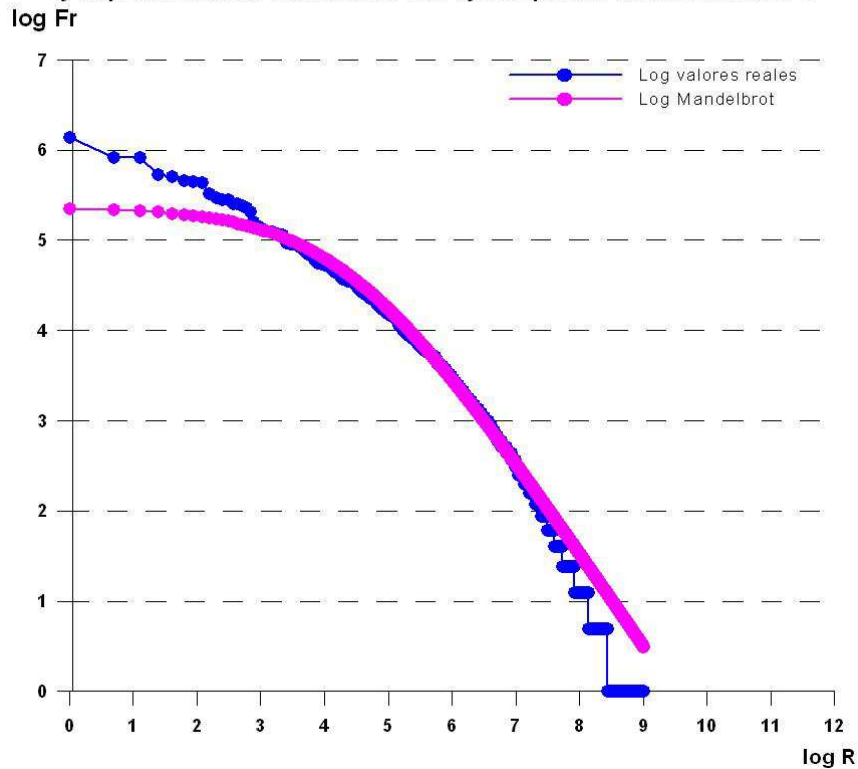


Gráfico 86. Distribución logarítmica de Mandelbrot ajustada a tramos 2-8 con textos reales

El método de *Stemming* utilizado o método para agrupar en raíces es importante para la futura indización y segmentación automática de la información de un texto, aunque ello conlleve no ajustarse al modelo inapelable de la Ley de Zipf y/o Mandelbrot con exponente (e)=1. Pero, como existen distintos métodos de *Stemming*, ¿podemos conseguir con un método u otro que se ajuste lo más posible al exponente 1?, Esta cuestión, justifica el capítulo siguiente en la que se abordará estudios cuantitativos con raíces agrupando a los rangos intermedios

7. Estudios cuantitativos relacionados con raíces. Stemmers

En este capítulo se implementa varios métodos de lematización o Stemming para estudiar su influencia sobre la degradación de la información y el valor de los parámetros de la fórmula de Zipf.

Uno de los ciclos más importantes en la extracción de palabras corresponde al procedimiento de agrupación de las raíces. Este procedimiento realiza una reducción de las palabras a su raíz. El método de *Stemming* o Lematización reduce el número de términos distintos e identifica palabras o términos similares, su finalidad es agrupar palabras que tienen un contenido semántico muy próximo. De este modo los algoritmos conducentes al *Stemming* (proceso de extracción de raíces) agrupan todas las palabras con la misma raíz reduciendo así el número de palabras del índice.

En primer lugar, debemos admitir que la agrupación de palabras según su raíz común no es un proceso unívocamente determinado. Ya en la formación del vocabulario se hace una cierta agrupación: se identifica *Volver* con *volver*, lo que parece inevitable; pero también se identifica *una vuelta del camino* donde el autor ha querido decir una curva, con *a la vuelta del viaje* que significa otra cosa.

Si continuamos agrupando palabras en función de sus primeras letras comunes pensamos que la variación en las últimas letras corresponde a un significado que no es esencial y así quedan agrupadas por el significado esencial. (Esto no es totalmente cierto: un estribillo en poesía no va a ser lo mismo que un estribo para subir a un vehículo)

Por otra parte, los algoritmos de obtención de raíces pueden aplicarse con mayor o menor intensidad, forzando el agrupamiento en distintos grados. Por ejemplo pueden utilizarse tablas más o menos extensas de sufijos o puede obligarse a un mínimo de letras en la raíz, con distintos valores. Estas consideraciones nos hacen ver el proceso de formación del vocabulario esencial de un texto como algo artificial y que admite graduación. En definitiva consiste en variar la granularidad con la que estamos estudiando el texto.

El procedimiento de extracción de raíces se implementará en los sucesivos capítulos con lo cual a partir de ahora es importante aclarar que los estudios cuantitativos no se van a realizar sobre palabras sino sobre las raíces de las palabras. Para ello se utilizará entre otros el Método de Sufijos, el conocido algoritmo de Lovins (1968).

Antes de iniciarnos en este Método de Sufijos como modelo para la extracción de raíces a utilizar en esta investigación se han realizado pruebas y experimentaciones con otro modelo como es el Método de Variedad de Sucesores (Hafer, Weiss, 1974). El Método de Variedad de Sucesores se ha utilizado en un primer momento en esta investigación y ha sido modificado y adaptado en varios aspectos, para que fuese más efectivo con el lenguaje Español debido a sus restricciones léxicas con idiomas distintos del Inglés. Este método está diseñado en su origen para el idioma Inglés y por ello fue necesaria una adaptación práctica del algoritmo para eliminar algunas distorsiones que se producen en su aplicación al Español y así por tanto conseguir aumentar su efectividad.

Según Baeza-Yates, Ribeiro-Neto (1999), este método resulta más complejo que los algoritmos de extracción de raíces por Sufijos como el algoritmo de Lovins (1968).

Debido a la gran adaptación realizada a dicho Método se describe brevemente en qué medida se ha llevado a cabo dicha adaptación para el idioma Español y en qué partes se ha modificado para colaborar con quien estuviera interesado en abordar investigaciones con este algoritmo de extracción de raíces. Aunque cabe señalar que según nuestros estudios y tras el ingente trabajo realizado con la adaptación del Método de Variedad de Sucesores hemos conseguido resultados similares al Método de Sufijos de Lovins (1968) con las adaptaciones al Español.

Así advertimos que los autores Frakes, Baeza-Yates (1992), respecto al Método de Variedad de Sucesores con idiomas distintos del Inglés recomiendan utilizar otros métodos para el Stemming.

En nuestra investigación vamos a ejecutar como método experimental, los siguientes pasos: vamos a tomar un texto fijo, obtendremos la colección de raíces⁴⁸ en el texto, por diversos procedimientos y diversos parámetros, el siguiente paso será ajustar la fórmula de Zipf a cada una y exponer y criticar los resultados.

7.1. Método de Variedad de Sucesores

El Método de Variedad de Sucesores según Frakes, Baeza-Yates (1992) está basado en el análisis estadístico de una colección de documentos, este método como su propio nombre indica utiliza un algoritmo que procesa palabra por palabra buscando las letras que se suceden en una palabra o un término y buscando cuales son los sucesores en toda la colección de palabras.

Básicamente, el algoritmo del Método de Variedad de Sucesores busca los sucesores de cada palabra, es decir compara carácter a carácter cada palabra y extrae la raíz cuando detecta que los caracteres ya no se repiten, es decir en una colección de palabras con la misma raíz cuando en dicha colección cambian las grafías y no se repiten, entonces el algoritmo determina que esa es la raíz y establece el primer pico (peak)⁴⁹. Un ejemplo son las palabras que tienen plural, como por ejemplo *todo/ todos*, el algoritmo fija el primer pico en la raíz *todo*. Un ejemplo sencillo del comportamiento del algoritmo lo tenemos con las palabras siguientes en un texto:

GATO
GATERA
GATO
GIGANTE
GATOS

A partir de aquí el algoritmo busca todos los sucesores de la palabra GATO y les da valores.

G, GA, GAT, GATO, GATOS
2 1 2 1 0

⁴⁸ Para la obtención del vocabulario y las raíces se utiliza la aplicación PIEDRA.mdb y el ajuste y análisis de Zipf con la aplicación TOPOS.mdb

⁴⁹ Peak es un término utilizado por Baeza-Yates y Frakes (1993, p. 134)

El ejemplo muestra que después de la letra G existen dos posibles variaciones en las letras siguiente a la G, estas son: A , I. De este modo, se representan estos valores en un gráfico y se obtiene el primer pico, a partir de ahí, se extrae la raíz, que en este caso sería GAT.

Utilizamos otro ejemplo sencillo del funcionamiento del algoritmo del Método de Variedad de Sucesores a la hora de extraer la raíz de un mismo grupo de palabras.

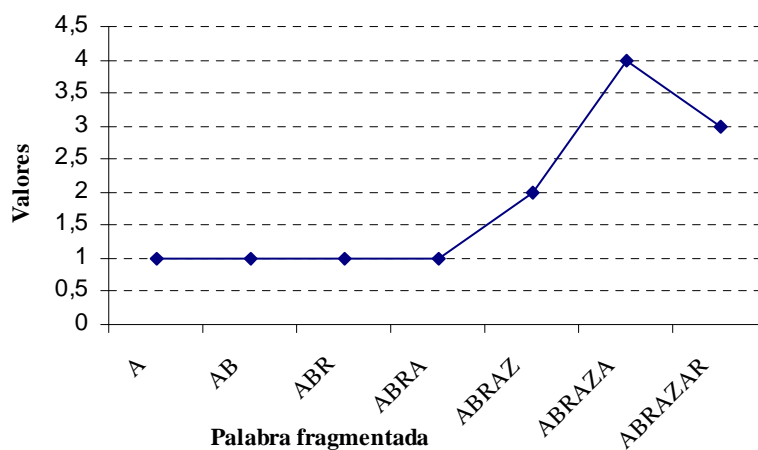


Gráfico 87. Stemmer método de variedad de sucesores

Palabras	Raíz
Abrazada	Abraza
Abrázame	Abraza
Abrazan	Abraza
Abrazar	Abraza
Abrazaré	Abraza
Abrazaron	Abraza
Abrazarte	Abraza
Abrazo	Abrazo

Tabla 26. Ejemplo Stemmer método de variedad de sucesores

Este método presenta algunos problemas que son mejorados, uno de ellos surge cuando algunas palabras tienen prefijos gramaticales como *DES*, *desnutrición*, *desfigurar*, en estos casos, el Método de Variedad de Sucesores cogería como raíz *DES*, no siendo correcto, para estos casos, existe una solución aplicada al idioma Inglés que consiste en que si existen más de 12 palabras distintas que empiezan con el prefijo *DES*, éste se considera un prefijo y se toma la raíz en el segundo pico. En el caso del idioma Español esta solución puede variar a menos palabras, como en nuestro caso en la adaptación realizada al Método de Variedad de Sucesores hemos indicado que deben existir 11 palabras distintas para considerar el prefijo y por tanto tomar la raíz en el segundo pico y no en el primero. Igualmente se han detectado otras complicaciones con verbos, sustantivos, etc.

7.2. Adaptación al Español del Método de Variedad de Sucesores.

El Método de Variedad de Sucesores concebido para aplicar en textos escritos en inglés funciona en todos los idiomas occidentales, siendo el más adaptable, es por ello que para adaptarlo a textos en español se han introducido unas modificaciones importantes y nuevos parámetros al algoritmo para la extracción de las raíces (Rodríguez Luna, 2002).

Entre las modificaciones añadidas se encuentran las siguientes:

Ra: Si incluye la raíz de la anterior palabra, tomarla, en este caso el sistema si localiza varias palabras que tienen la misma raíz que la palabra que la precede, las toma unificando las raíces.

Rp: Quitar plural (si existe singular), en este caso el sistema si encuentra dos palabras con la misma raíz pero con la diferencia de que una de ellas contiene al final una –s, entonces eliminará esta característica y las unificará en una misma raíz. Es decir si es plural y existe el singular, singulariza la palabra antes de contar sucesores.

Rt: Si después de pico hay consonante, tomarla también, el sistema en el caso de varias palabras que unifican su raíz, si alguna de ellas contiene una consonante después del grupo de caracteres que forman la raíz, la consonante que aparezca debe ser tomada también como raíz. La opción “Rt” es muy importante ya que consigue que palabras diferentes no se unifiquen con la raíz de otras que se asemejan. Como por ejemplo las palabras *alma, almas, almacén* las raíces obtenidas serían *alma, almac*, consiguiéndose así que no se unifiquen en la misma raíz palabras distintas.

Rn: Penalizar sucesores de vocal, esta opción se utiliza para nivelar la proporción vocal/consonante en Español, ya que en este idioma está desnivelado.

En Español la distribución de vocales es distinta que en Inglés, ya que normalmente en Español después de una vocal tenemos muchas posibilidades de que aparezca una consonante pero tras una consonante tenemos menos posibilidades de que aparezca una vocal concretamente cinco: a, e, i, o, u. No ocurre lo mismo en el idioma Inglés que existe similar probabilidad de que detrás de una vocal vaya otra vocal u consonante y que detrás de una consonante vaya otra consonante u vocal, Las probabilidades están más equilibradas. Por ello introducimos en el sistema la posibilidad de penalizar los sucesores de vocal con un valor de 1,3. Esto supone que el sistema multiplica el número de sucesores por 1,3 para conseguir en el caso del español que aumenten los sucesores y de este modo la aplicación determine el pico con sentido. Es decir, multiplicamos por 1,3 el número de posibles letras distintas de toda la colección que pueden seguir a la palabra.

Rc: Si la palabra es larga y la raíz coincide, en este caso si la palabra es larga y la raíz coincide con una palabra corta entonces el sistema establece la raíz de dicha palabra larga en el segundo pico.

Entre los parámetros añadidos se encuentran los siguientes:

Número de letras mínimo, en este caso el sistema establece el mínimo de caracteres o letras que debe tener una palabra.

Máximo palabras para Ra, en este caso el sistema determina el máximo de palabras que debe tener el término que se toma por tener la misma raíz que su predecesora.

Penalización vocal/consonante, el sistema necesita este parámetro para nivelar la proporción vocal/consonante con un valor concreto.

Máximo palabras para prefijo, en este caso el sistema establece el máximo de caracteres que debe tener una palabra para que se considere que esta tiene un prefijo.

Letras palabra larga, en este caso el sistema establece el máximo de caracteres que deben tener las palabras que se consideren largas.

Letras palabra corta, en este caso el sistema establece el mínimo de caracteres que deben tener las palabras que se consideren cortas.

Tras modificar el algoritmo original de Método de Variedad de Sucesores con los parámetros y modificaciones anteriormente descritas, los resultados han sido satisfactorios como método de extracción de raíces para textos en español como mostraremos al final de este capítulo donde se evalúan varios Stemmers entre ellos el Método de Variedad de Sucesores con adaptaciones al Español. Otras experimentaciones son las desarrolladas en las conferencias TREC (Text Retrieval Conference), donde se aplicaron diversos mecanismos de *Stemming* para el idioma español, según Harman (1995), que consistieron en la aplicación de los mismos algoritmos de *Stemming* que para el idioma inglés pero con la incorporación de sufijos y reglas para el español.

Utilizaremos el Método de Variedad de Sucesores para establecer comparaciones entre diferentes Stemmers.

En las sucesivas experimentaciones se va a utilizar como método de *Stemming* el Método de Sufijos pero no nos adentraremos en el origen y difusión de este método, sino en su metodología y funcionamiento. Realizando por tanto los estudios cuantitativos sobre las raíces de las palabras.

El algoritmo de Lovins para extraer raíces basado en el método de sufijos puede considerarse un método más abierto a sufrir adaptaciones con otros idiomas como es el caso del español, igualmente el Método de Variedad de Sucesores (Hafer, Weiss, 1974), también resulta fácilmente adaptable a los idiomas occidentales, pero no ocurre lo mismo con otro tipo de algoritmos de extracción de raíces como el de Porter (1980), que está adaptado a la estructura morfológica de las palabras inglesas y está concebido para extraer raíces en el idioma Inglés.

Existen diversos Stemmers para el idioma inglés basados en la eliminación de sufijos derivacionales como Lovins (1968), Dawson (1974), Porter (1980), Paice (1990). En general estos algoritmos no llevan a cabo ningún análisis morfológico sofisticado, sino que se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común.

7.3. Método de Sufijos de Lovins

El algoritmo de Lovins se basa fundamentalmente en una lista de sufijos, cada uno acompañado del número mínimo de letras que deben quedar y a veces, de otra condición suplementaria.

Ejemplos de la lista de sufijos:

-alistically	3	----
-eableness	2	condición E
-entness	2	----
-eature	2	condición Z
-inism	2	condición J
-ery	2	condición E

..., hasta 400 sufijos aproximadamente.

Ejemplos de condiciones sobre la raíz que va a resultar:

Que no termine en e
Que termine en f
Que termine en t o ll
Que no termine en e ni en o
Que no termine en met o en ryst
Que termine en t o dr pero no en tt

..., hasta 28 condiciones

Igualmente se pueden realizar acciones suplementarias sobre palabras duplicadas si la raíz resultante termina en bb, dd, gg, ll, mm, nn, pp, rr, ss, tt, el algoritmo tendría que dejarle sólo una de las letras, es decir dejar sólo b, d, g, l, m, n, p, r, s y t.

Otra de las funciones es la acción de reescritura sobre la raíz, es decir, si la raíz resultante termina en uno de estos sufijos, sustituirla por lo indicado:

Iev	ief
Uct	uc
Umpt	um
Rpt	rb
Urs	ur
Istr	ister

..., tendríamos hasta 33 casos en la acción de reescritura sobre la raíz.

En definitiva la aplicación del algoritmo de Lovins sobre una palabra sigue las siguientes acciones:

- 1) probar todos los sufijos comenzando por los más largos y aplicar si:
 - a. la palabra termina con el sufijo
 - b. el número de letras de la raíz es suficiente
 - c. se verifica la condición
- 2) si la raíz termina en letra duplicada, desdoblarse
- 3) reescribir el final de la raíz si procede

7.4. Adaptación al Español del Método de Sufijos de Lovins

El Método de Sufijos de Lovins (1968) es más abierto y adaptable al español que otros métodos como el de Porter (1980). En nuestro caso hemos adaptado el algoritmo de Lovins (1968) al español pero realizando una adaptación básica, de ahí que existan varias excepciones y parámetros básicos añadidos al método inicial. Para ello se ha creado una lista de sufijos de palabras en español de la cual se ha extraído el final de cada palabra para obtener el sufijo más largo posible de cada palabra, siempre verificando ciertos criterios y creando para ello una lista de excepciones que contiene sufijos de ciertas palabras que puedan resultar una excepción por el tipo de documento que se está estudiando. Igualmente se han incluido parámetros y reglas adicionales para adaptar aún más si cabe dicho algoritmo a nuestro idioma.

Entre las reglas y parámetros que hemos empleado para adaptar el Método de Lovins se ha utilizado el algoritmo en una de las aplicaciones para la obtención de raíces⁵⁰, dichas modificaciones al método original de Lovins y básicas adaptaciones se componen de los siguientes elementos:

- Un valor numérico que indica el número mínimo de letras que debe tener, con carácter general, cualquier raíz.
- Una tabla de sufijos en la que algunos de ellos llevan asociado un valor numérico indicando el mínimo número de letras que deben quedar en la raíz. Se entiende que es un valor superior al general, para ser tenido en cuenta solo en este sufijo.

Esta tabla se ha construido manualmente pero con ciertas técnicas sencillas muy efectivas, basadas en presentar listas de terminaciones de las palabras del texto, ordenadas a la inversa, como si la palabra empezase por su última letra, después la penúltima, etc. En estas listas ya no aparecen los sufijos previamente definidos, con lo que el volumen de trabajo es aceptable.

- Una tabla de excepciones con ciertos sufijos: palabras a las que no se debe aplicar el sufijo, aunque terminen en esas letras. Por ejemplo a METAL no se le debe quitar el sufijo -AL. En cambio sí que podríamos quitárselo a DOCTORAL.

Forma parte del proceso anterior de obtener las palabras del texto, el marcado de los nombres propios. Se hace de una manera semiautomática: la parte automática marca las palabras que siempre aparecen con mayúscula (por ejemplo, no marca "Cuando" que quizá la ha encontrado con mayúscula por ser comienzo de frase, ya que la ha encontrado otra vez como "cuando"). Después de esta parte automática hay facilidades para una revisión manual.

Con todos estos elementos, el algoritmo funciona de la siguiente manera:

1. Se procesan las palabras una a una, descartando los nombres propios.

⁵⁰ Aplicación PIEDRAS, ver apéndice II

2. Para cada palabra se escogen los sufijos que coincidan con el final de la palabra, comenzando por los más largos.
3. Para cada sufijo posible se comprueba: si el número de letras que queda es superior al mínimo general y al mínimo específico, si lo hubiera, para este sufijo; también si la palabra está o no en la lista de excepciones para este sufijo.
4. Si falla alguna de estas condiciones, se descarta el sufijo y se intenta con el siguiente (de menos letras)

En el idioma inglés los algoritmos para la extracción de raíces mediante el método de sufijos eliminan los sufijos y/o los prefijos de las palabras dejando la raíz, estos algoritmos en ocasiones transforman la raíz resultante incluyendo reglas que contienen excepciones y otras reglas, (Harman, 1991) como por ejemplo:

Si una palabra termina en “ies” pero no “eies” o “aies”

Entonces “ies”----- “y”

Si una palabra termina en “es” pero no “aes” o “ees” o “oes”

Entonces “es”----- “e”

Si una palabra termina en “s” pero no “us” o “ss”

Entonces “s”----- “NULO”

La mayoría de los algoritmos utilizan un método de eliminación de sufijos desarrollado por Lovins (1968), el cual se basa en la eliminación de la cadena de caracteres mas larga posible de acuerdo a un conjunto de reglas, este proceso se repite hasta que no se pueden eliminar más caracteres, incluso después de eliminar todos los caracteres posibles de cada palabra, las raíces pueden no combinarse adecuadamente, existiendo excepciones para ello.

Según Frakes, Baeza-Yates (1992), el algoritmo de eliminación de sufijos de Lovins, ha sido utilizado por los algoritmos de Salton (1968), Dawson (1974), Porter (1980), y Paice (1990), según el autor, el algoritmo de Porter es más compacto que el de Lovins, Salton y Dawson, y basándose en la base de la experimentación considera que da un rendimiento en la recuperación de la información comparable a los algoritmos más grandes.

7.5. El Algoritmo de Porter

El algoritmo de Porter, según Frakes, Baeza-Yates (1992), consiste en un conjunto de reglas de condición/acción. Las condiciones se dividen en tres clases:

- ✓ condiciones en la raíz
- ✓ condiciones en los sufijos
- ✓ condiciones en las reglas

Existen varios tipos de condiciones para la raíz

1. La medida: representada por m , de una raíz indica el número de secuencias de vocales (V) y consonantes (C), VC secuencias, así $m=0$, $m=1$, $m=2$, depende del número de secuencias entre vocales y consonantes de cada palabra, por ejemplo:

Medida	Ejemplo
m=0	TR, EE, TREE, Y BY
m=1	TROUBLE, OATS, TREES, IVY
m=2	TROUBLES, PRIVATE, OATEN

Tabla 27. Condiciones para la raíz del algoritmo de Porter.

2. *<X> la raíz finaliza con una letra dada x
3. *v* la raíz contiene una vocal
4. *d la raíz finaliza con una doble consonante
5. *o la raíz finaliza con una consonante-vocal-consonante, (secuencia), donde la última consonante no es w, x o y

Las condiciones en los sufijos básicamente se basan en unos patrones dados o listas de sufijos.

Las condiciones en las reglas son divididas en dos pasos, las reglas son examinadas secuencialmente y sólo una regla de un paso puede aplicarse. El sufijo más largo posible es siempre eliminado debido a la orden de las reglas dentro de un paso, básicamente se basan en las reglas utilizadas. Para ejemplos más concretos de este algoritmo de Porter (Baeza-Yates y Frakes, 1993, p. 139-142) y (Baeza-Yates y Ribeiro-Neto, 1999, p. 433-436).

7.6. Procesos de Evaluación de los Sistemas de Recuperación de Información (SRI)

Según Frakes, Baeza-Yates (1992), un Sistema de Recuperación de Información (SRI) puede ser evaluado por la eficacia en la ejecución, el efectivo almacenamiento de los datos y la efectividad en la recuperación de la información, esta última determinará la relevancia de los documentos recuperados. Igualmente el autor indica la existencia de dos tipos de evaluaciones posibles: la del funcionamiento del sistema y la del funcionamiento de la recuperación, en la que se analiza cómo los documentos recuperados se clasifican de acuerdo a su relevancia con la pregunta efectuada.

Qué documentos son los relevantes de toda la colección posible para un usuario es una labor ardua y complicada para el SRI e incluso para el propio usuario, que puede tener criterios distintos en función de sus necesidades, al igual que para el sistema establecer la relevancia o no de un documento es también un objetivo complejo de alcanzar, ya que incluso un documento puede tener párrafos relevantes y otros no, a esto varios autores lo denominan *relevancia parcial*.

Las primeras evaluaciones de los SRI se realizan a principios de los años cincuenta, los conocidos Proyecto Cranfield, Smart, Medlars, Stairs y las Conferencias TREC⁵¹ sientan las bases de la evaluación de los SRI y confieren unas medidas basadas en la relevancia de los documentos: Precisión y Exhaustividad. Según Valery, Shapiro,

⁵¹ Para información más detallada sobre estos proyectos véase Martínez Méndez, Francisco Javier; Rodríguez Muñoz, José Vicente. (2004). Reflexiones sobre la evaluación de los Sistemas de Recuperación de Información: necesidad, utilidad y viabilidad. Anales de Documentación, nº 7, p. 153-170

Voiskunskii, (1997) estos conceptos fueron utilizados para evaluar la calidad de la recuperación en el sistema. El coeficiente de Exhaustividad (la ratio de la Exhaustividad) se define como una ratio (un porcentaje) del número de documentos relevantes que son recuperados sobre el número total de documentos relevantes en una colección. El coeficiente de Precisión (la ratio de la Precisión) es la (un porcentaje) ratio del número de documentos relevantes recuperados sobre el número total de documentos recuperados. Analíticamente estos coeficientes pueden describirse del siguiente modo:

$$E = \frac{n}{N}(\times 100\%) \qquad P = \frac{n}{M}(\times 100\%)$$

Donde:

- n = número de documentos relevantes recuperados
- N = número de documentos relevantes en la colección
- M = número de documentos recuperados.

Se han desarrollado métodos para evaluar simultáneamente en un sistema de recuperación de información la Exhaustividad y la Precisión. Un método comprende el uso de grafos de Exhaustividad-Precisión-puntos bivariados, donde un eje es para la Exhaustividad y otro para la Precisión. La siguiente figura muestra un ejemplo de tales puntos. Ambos, Exhaustividad y Precisión toman valores entre 1 y 0.

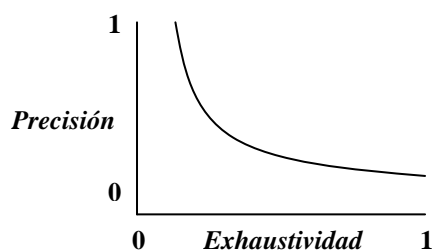


Figura 14. Gráfico Exhaustividad-Precisión

Los puntos de Exhaustividad-Precisión están inversamente relacionados. Esto se explica del siguiente modo, cuando la Precisión sube, la Exhaustividad normalmente baja y viceversa. Una medida de evaluación combinada de Exhaustividad y Precisión, *E*, ha sido desarrollada por Van Rijsbergen (1979) y definida como:

$$E = 1 - [(1 + b^2) P R / (b^2 P + R)]$$

Donde {P = precisión, R = exhaustividad}, y *b* es una medida de la importancia relativa, para un usuario, de Exhaustividad y Precisión. Los investigadores eligen valores de *E* que ellos esperan que reflejarán la Exhaustividad y Precisión que interese al usuario típico. Por ejemplo, si los valores de *b* se encuentran en niveles de 0.50, indica que un usuario estuvo dos veces tan interesado en la Precisión como en la Exhaustividad, y si el valor de *b* fuera 2, nos indica que un usuario estuvo tan interesado en la Exhaustividad como en la Precisión.

Otro punto de vista respecto a los métodos de evaluación de Exhaustividad y Precisión sería la consideración de un nuevo algoritmo de recuperación que pondera sobre un ranking los documentos obtenidos en una consulta, así si tenemos por ejemplo una

respuesta (R) a una consulta (c) de un usuario con los siguientes documentos (Baeza-Yates, Ribeiro-Neto, 1999, p. 76):

$$R_c = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

Consideremos que este nuevo algoritmo devuelve para la consulta c, un ranking de los documentos que se han recuperado. Así se encontrarían ordenados por el ranking los siguientes documentos, de los cuales se señalan con un símbolo circular los documentos que se consideran relevantes en la colección de R_c

- | | | |
|-----------------------|-----------------------|----------------------|
| 1. d ₁₂₃ • | 6. d ₉ • | 10. d ₃₈ |
| 2. d ₈₄ | 7. d ₅₁₁ | 11. d ₄₈ |
| 3. d ₅₆ • | 8. d ₁₂₉ | 12. d ₂₅₀ |
| 4. d ₆ | 9. d ₁₈₇ | 13. d ₁₁₃ |
| 5. d ₈ | 10. d ₂₅ • | 14. d ₃ • |

Tabla 26. Ejemplo Precisión y Exhaustividad

El usuario realiza una consulta a un sistema de información cualquiera y obtiene una respuesta de 15 documentos, de los cuales 10 se consideran relevantes para la consulta de dicho usuario. De este modo el sistema ordena en un ranking de importancia estos documentos obtenidos, así se puede obtener la Precisión y Exhaustividad que tienen para la respuesta de búsqueda dichos documentos relevantes. Si observamos el ranking comenzando por el documento más relevante que se encuentra en la posición número uno, es el documento d₁₂₃ el cual corresponde a un 10% de todos los documentos relevantes que se han dado en la respuesta (R), así pues este documento tiene una Precisión del 100% sobre el 10% de Exhaustividad. En segundo lugar tenemos el documento d₅₆ el cual es el segundo documento más relevante, así tendríamos una Precisión de 66,6% y 20% de Exhaustividad.

Documentos Relevantes	Precisión	Exhasutividad
d ₈₉	100%	0%
d ₁₂₃	100%	10%
d ₅₆	66%	20%
d ₉	50%	30%
d ₂₅	40%	40%
d ₃	33%	50%
d ₅	0%	60%
d ₃₉	0%	80%
d ₄₄	0%	90%
d ₇₁	0%	100%

Tabla 27. Datos de la Precisión y Exhaustividad de los documentos ejemplo

En tercer lugar si se procede con el análisis del ranking generado se puede trazar una curva de la Precisión sobre la Exhaustividad. La Precisión en los niveles de Exhaustividad es más alta del 50% el cual disminuye a 0 porque no todos los documentos relevantes han sido recuperados. El siguiente gráfico muestra los datos que se representan en la tabla comprobando así la curva que se obtiene.

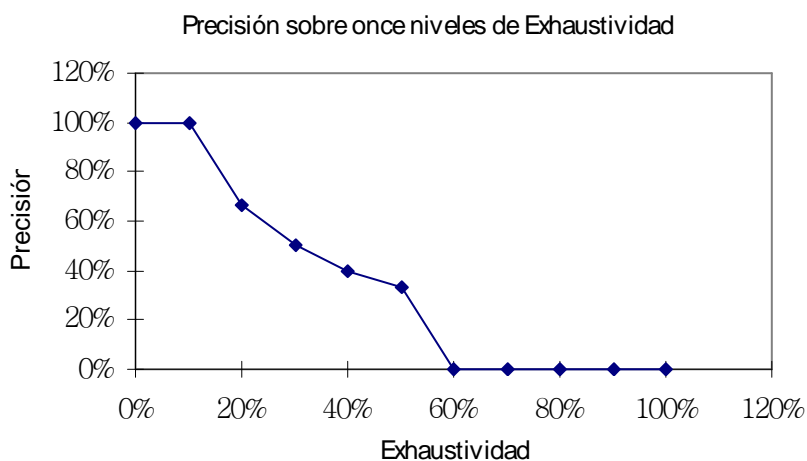


Gráfico 88. Precisión sobre Exhaustividad de los datos ejemplo

Otra línea de trabajo dentro de este mismo ámbito es la evaluación de los motores de búsqueda en la Web. En este campo de trabajo se han llevado a cabo muchos y diferentes procesos de evaluación, algunos de ellos pueden considerarse "superficiales" en tanto que basan sus conclusiones en las características que los propios administradores de estos motores proporcionan a través de sus Webs y no realizan ningún tipo de test o experimentación con los mismos. Otros trabajos si llevan a cabo experimentaciones muchas de ellas relacionadas con el cálculo tradicional de la Precisión de los resultados recuperados y otros van algo más allá, considerando medidas tales como el tamaño del índice de cada motor o el período de actualización de este índice. Es por ello que aunque encontremos una abundante literatura científica se hace precisa una sintetización de la misma y un análisis comparativo de los resultados proporcionados.

De este modo varios autores consideran a diversos motores de búsqueda como los mejores por su flexibilidad, fortaleza de sus búsquedas y el rápido tiempo de respuesta, en este aspecto se pueden observar una amplia diversidad de opiniones, es por ello que es preciso establecer una metodología para el análisis de estos motores de búsqueda de manera que cuando se lleven a cabo nuevos estudios, los resultados procedentes de los mismos no resulten muy dispares.

Chu, Rosenthal (1996) apuestan por la vigencia de los postulados de Lancaster, Fayen (1973), (Tiempo de respuesta, Precisión, Capacidades de búsqueda, Formatos, Documentación e Interface) ya que consideran a los motores de búsqueda como sistemas de recuperación de información caracterizados por un enorme tamaño, por su estructura hipertexto y por su arquitectura distribuida, pero perfectamente susceptibles de ser evaluados con los citados criterios.

Dichos autores proponen el desarrollo de esa metodología que permita evaluar de forma coherente a los distintos motores de búsqueda, en ella se han de considerar cuatro conceptos generales:

Composición de los índices: la composición de los índices afecta de forma muy directa a la calidad de la recuperación de información. Destacan tres componentes importantes: Cobertura, Frecuencia de Actualización y Porción de página web indexada (título, título

y primeros párrafos, página completa, etiquetas <meta>). Las magnitudes de cada motor dependerá del Hardware (Hw) y Software (Sw) dedicado. Por otro lado, el que el índice sea muy extenso no implica calidad y además el poseer un valor muy alto en este parámetro tampoco implica altos niveles en los otros tres.

Capacidades de búsqueda: un motor de búsqueda ha de poseer la posibilidad de utilizar operadores booleanos, búsquedas por expresiones literales, truncamiento de los términos y facilidades de acotar una búsqueda en un determinado campo, con búsquedas por frase literal. Considerándose este conjunto de prestaciones básicas.

Ejecución de la recuperación de información: suele medirse con base a tres parámetros: Precisión, Exhaustividad y Tiempo de Respuesta. Estas medidas, no obstante, incorporan muchas dosis de subjetividad a la hora de determinar cuándo un resultado es relevante o no.

Esfuerzo del usuario: la Documentación y el Interface son elementos a considerar en este apartado y suelen tener un aceptable nivel. De hecho, este parámetro es muy importante porque un usuario no va a hacer uso de un motor si no se encuentra cómodo con su Interface, si no localiza fácilmente la documentación que indica cómo emplearlo y si no la comprende.

A lo largo del desarrollo de este punto se ha introducido y definido conceptos básicos de la evaluación de un sistema de recuperación de información. En definitiva un sistema de recuperación de información típico debe tener los siguientes requerimientos funcionales y no funcionales. Debe permitir a los usuarios añadir, borrar y cambiar documentos en la base de datos. Debe proporcionar a los usuarios la manera de buscar documentos tecleando preguntas, y examinando los documentos recuperados. Un sistema de recuperación de información necesitará soportar grandes bases de datos y recuperar documentos relevantes en respuesta a preguntas interactivamente a menudo en pocos segundos.

Ahora la cuestión es conocer los procesos de evaluación de los Stemmers como es el caso que nos ocupa en este capítulo por tanto se tratará en detalle a continuación.

7.7. Procesos de Evaluación de los Stemmers

Según Salton (1968), los resultados obtenidos en la evaluación de varios métodos de agrupación en raíces “*Stemmers*” indican que el efecto de la agrupación de raíces depende de la naturaleza del vocabulario utilizado. Un vocabulario específico y homogéneo puede presentar diferentes propiedades de combinación que otros tipos de vocabulario

Los distintos métodos de *Stemming* dependen del idioma para el que se desarrollaron es por ello que resulta tarea difícil aplicar algoritmos diseñados para un idioma a textos en otra lengua distinta, por esta razón se han propuesto sistemas elementales de *Stemming* que son independientes del idioma, este es el caso de los n-gramas, (Robertson, Willet, 1999). Respecto al uso de n-gramas, según Figuerola et. al. (2004), una vez descritos diversos algoritmos de normalización de términos con textos en español

consideran desaconsejable su utilización, pues los resultados obtenidos no alcanzan los resultados que se obtienen sin aplicar ningún tipo de normalización.

Los objetivos principales que persigue la extracción de sufijos en la recuperación de información son entre otros, disminuir el tamaño de la tabla de palabras y mejorar la recuperación al unificar palabras con el mismo significado, consecuentemente la evaluación de este tipo de métodos de Stemming es un proceso necesario y definitivo para justificar la calidad de estos métodos en la agrupación de raíces que determinará la obtención final del documento por parte del usuario a la hora de la búsqueda ya sea en una base de datos, en la Web o en SGED: Sistemas de Gestión Electrónica de Documentos, EDMS: Electronic Data Management Systems.

En este sentido los procesos de evaluación (Frakes, Baeza-Yates, 1992), de los métodos de sufijos para el Stemming determinan el rendimiento en la recuperación (no la corrección lingüística) y las medidas estándar de relación Precisión-Exhaustividad.

Respecto a la corrección lingüística tenemos los experimentos de Frakes, Baeza-Yates (1992, p.143), éstos estudian abundantes ejemplos en los que se mide la desviación del *stem* frente a *root*.

Según la autora Gómez Díaz (2005) distingue entre lematización derivativa y lematización flexiva, respecto a la lematización derivativa la define como una agrupación en raíces más agresiva utilizando para ello una lista de sufijos de 230 sufijos y con 3.692 reglas asociadas a ellos, los ejemplos obtenidos son una tabla de 14.577 raíces obtenidas de un total de palabras de 24.414, y la autora define la lematización flexiva como una agrupación en raíces basada en los experimentos de Harman (1991) y Krovetz (1993) la cual no es tan radical como la lematización derivativa, ésta sólo normaliza los plurales y el género y reduce los tiempos verbales a la raíz, utilizando para ello una lista de 88 sufijos y 2.700 reglas asociadas a ellos.

Las conclusiones a las que llega la autora es que la supresión o eliminación de las palabras vacías afectaba positivamente tanto a la Precisión como a la Exhaustividad, tanto si aplicamos la lematización como si no. Esta primera conclusión parece lógica. La segunda conclusión a la que llega la autora es que para el español es más beneficiosa la lematización flexiva (la menos agresiva) suprimiendo antes las palabras vacías.

En nuestro caso realizamos en páginas sucesivas (concretamente en el punto 7.10 de este Capítulo) estudios similares pero evaluando los resultados obtenidos con diversos Stemmers, entre ellos unos son más agresivos que otros y también en nuestras investigaciones utilizamos tablas más grandes de sufijos y de palabras para aplicar estos Stemmers. Por tanto, vamos a evaluar el Método de Variedad de Sucesores con adaptaciones al Español (*MVS*), el Método de Sufijos de Lovins con adaptaciones al Español (*SU*) y el Método denominado Tanto Por Cien (*TCP-%*), Stemmer de creación propia y desarrollado exclusivamente para esta investigación que extrae la raíz de una palabra según el porcentaje de la palabra a obtener, éste puede ser más o menos agresivo dependiendo del porcentaje que le indiquemos.

7.8. Estudio de la distribución de frecuencias de Zipf y Mandelbrot con los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto Por Cien (TCP-%)

Se ha utilizado para este estudio el texto del autor *Miguel de Cervantes Saavedra* y su gran obra *EL INGENIOSO HIDALGO DON QUIJOTE DE LA MANCHA*. Sobre este texto se han realizado los procesos pertinentes de eliminación de palabras vacías, obtención de palabras distintas, extracción de raíces con diversos Stemmers, etc. para posteriormente emplear la distribución de frecuencias de Zipf y Mandelbrot. Todos estos procesos se han llevado a cabo con las aplicaciones creadas al uso⁵².

Es necesario aclarar que la obtención de raíces es un proceso que puede aplicarse con mayor o menor intensidad ya que podemos reducir las raíces según agrupemos el 80% de la palabra o agrupemos el 70% o en su caso los sufijos.

Esta arbitrariedad a la hora de cortar las palabras a su raíz, la realizamos con una de las aplicaciones⁵³ creadas para tal proceso. En dicha aplicación se pueden realizar los diversos experimentos obteniendo las raíces con tres métodos distintos, por el Método de Variedad de Sucesores denominado (*MVS*), por el Método de Sufijos de Lovins denominado (*SU*), y determinando el Porcentaje % de la palabra a obtener denominado (*TCP-%*) (*Tanto Por Cien*).

Con el método de Variedad de Sucesores (*MVS*), explicado ampliamente al comienzo de este capítulo se obtienen las raíces mediante los parámetros y las modificaciones añadidas al método adaptado al idioma Español.

El Método de Sufijos de Lovins (*SU*) consiste en quitar los sufijos como se detalla al comienzo de este capítulo, pero realizamos una adaptación básica del método de Lovins (1968) al español, de ahí que existan varias excepciones y parámetros básicos añadidos al método inicial explicadas igualmente en páginas anteriores.

El Método de cortar un Porcentaje de las palabras (*TCP-%*) depende de dos parámetros, el % de letras de la palabra que se pretende dejar y el número mínimo de letras que se exige que queden. Así por mostrar un ejemplo aclaratorio: TCP 60% 7 54% , significa por ejemplo si se aplica a una palabra de 10 letras, por un lado intentamos quitarle las últimas 4 letras, para que se quede en su 60%, pero por otro lado exigimos que queden como mínimo 7 letras, luego solo se eliminan 3.

Una vez se ejecuta el Stemmer se obtienen las raíces resultantes y se compara con el vocabulario, este sería el valor del segundo porcentaje del ejemplo mostrado, es decir, este determinaría a que % se ha reducido, dicho de otro modo este segundo porcentaje indicaría como de drástico es el método. Con lo cual significa que de las raíces obtenidas qué % del vocabulario representan. Aunque cada uno de los porcentajes mostrados son medidas distintas, se demuestra que están relacionadas ya que cuando uno disminuye el otro también disminuye.

⁵² Véase Apéndice II.

⁵³ Base de datos PIEDRAS, Véase Apéndice II.

Para ilustrar las diferencias o similitudes entre distintos Stemmers en la distribución de frecuencias de Zipf y Mandelbrot aplicamos todos ellos a un mismo ejemplo:

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo:
Aplicación PIEDRAS.mdb, formulario PIEDRA
PreProceso: separar documentos: Cada punto y aparte + Mínimo 800 caracteres
AnalizarTexto: colección de 72 palabras vacías
Nº letras para palabra corta que no se toma=2
Guión separa palabras=SI
Números son palabras=NO
Quitar palabras vacías=SI
Total: Documentos=2.258
Palabras=280.501
Palabras distintas=28.097
Raíces Método Variedad de Sucesores (MVS)=11.521
Reducción de: 41%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20

Su distribución de frecuencias en diagrama log(rango)-log(frecuencia) es la siguiente:

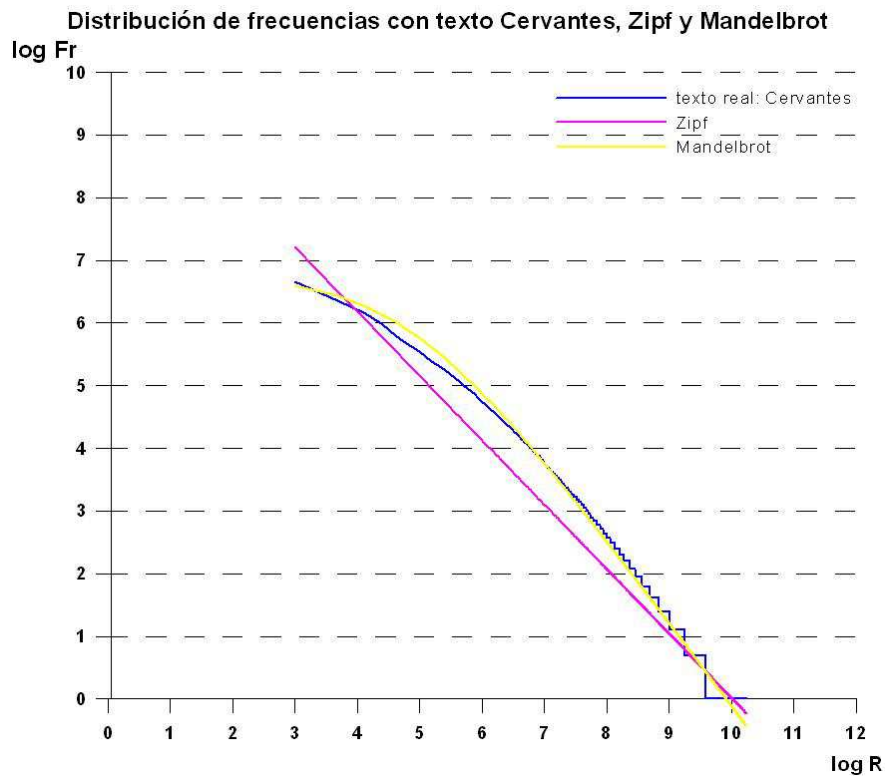


Gráfico 89. Distribución de frecuencias texto Cervantes, Zipf y Mandelbrot

El exponente de Zipf es 1,02, en la gráfica su representación es la línea recta coloreada en rosa que se ajusta bastante bien como tendencia general, aunque infravalora las frecuencias centrales. La curva de Mandelbrot (exponente 1,33 sumando 125,3) se ajusta mejor a todo el recorrido.

Si aplicamos el mismo ejemplo con raíces volviendo a aplicación PIEDRAS.mdb para ejecutar el Método de Variedad de Sucesores (MVS) con los siguientes parámetros:

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo:
 Aplicación PIEDRAS.mdb, formulario PIEDRA
 ExtraerRaícesVS: si incluye la raíz de la anterior palabra, tomarla=SI
 Quitar plural si existe singular=SI
 Si después del pico hay consonante, tomarla=SI
 Nº letras de la Palabra corta igual a su raíz=3
 Penalización vocal/consonante=1,3
 Cantidad de palabras para prefijo=11
Total: Raíces Método Variedad de Sucesores (MVS)=11.521
 Reducción de: 41%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20

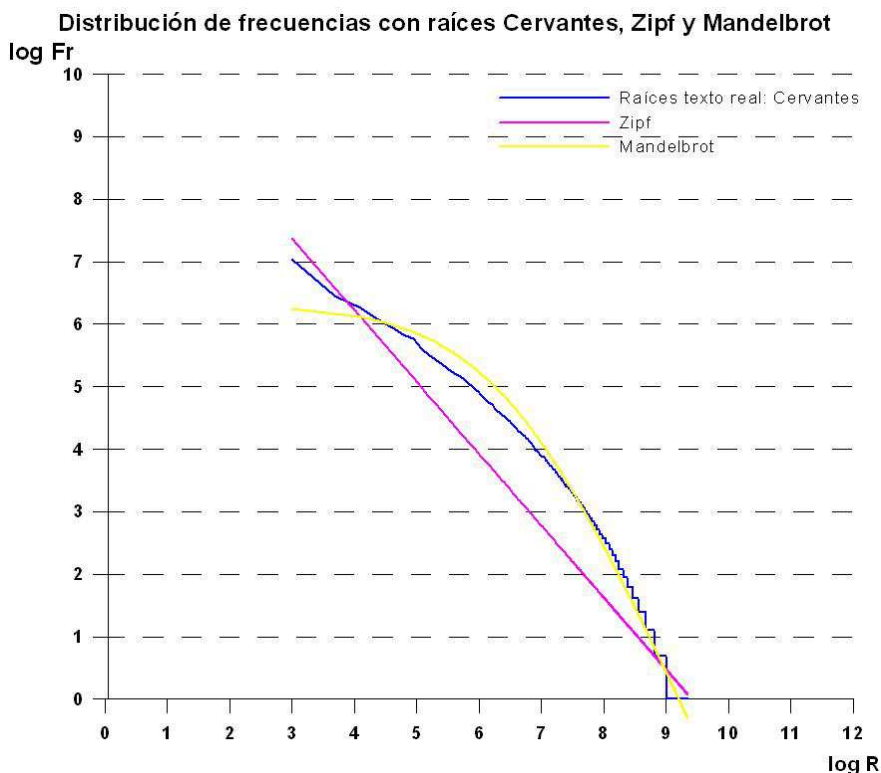


Gráfico 90. Distribución de frecuencias con raíces texto Cervantes, Zipf y Mandelbrot

La línea de datos del texto de Cervantes es la que termina con escalones y está coloreada en azul; la recta de Zipf tiene de pendiente (exponente de Zipf) 1,14,

representa una tendencia global, pero la curva real es decir del texto objeto de estudio, sólo tiene esta pendiente en su tramo central; la curva de Mandelbrot, que se ajusta bastante mejor, tiene un exponente de 2,28 y un sumando de 674,4.

CONCLUSIONES:

Comparando esta gráfica que muestra la distribución de frecuencias de Zipf y Mandelbrot con raíces y la gráfica anterior que muestra la distribución de frecuencias de Zipf y Mandelbrot con la de todas las palabras del texto Cervantes observamos las siguientes diferencias:

La representación de las frecuencias ha adquirido una fuerte curvatura llevando hacia abajo las de rangos elevados (esto significa que ha disminuido relativamente la cantidad de palabras/raíces de frecuencia baja, lo que es consecuente con el proceso de agrupar palabras según sus raíces)

Esta curvatura ha motivado un incremento considerable del exponente de Mandelbrot, pasando a ser 2,28 ya no es similar al de Zipf, recuérdese que ya se dio una explicación de la relación entre el exponente de Mandelbrot y la curvatura al estudiar las gráficas sintéticas.

A continuación se presentan las gráficas que muestran la distribución de frecuencias de Zipf y Mandelbrot con raíces obtenidas por varios Stemmers, en los cuales se especifican con detalle los parámetros.

Texto analizado:	Cervantes.txt de 3.139 K
Tratamiento previo:	Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaicesSufijos:	Mínimo número de caracteres que ha de tener cualquier raíz = 3 Tabla de sufijos=358
Total:	Raíces Método Sufijos (SU 3)=11.604 Reducción de: 41%
Para dibujar gráfica:	Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot:	Zipf (e)= 1,17 Mandelbrot (e)= 1,90 (Σ)= 225,67

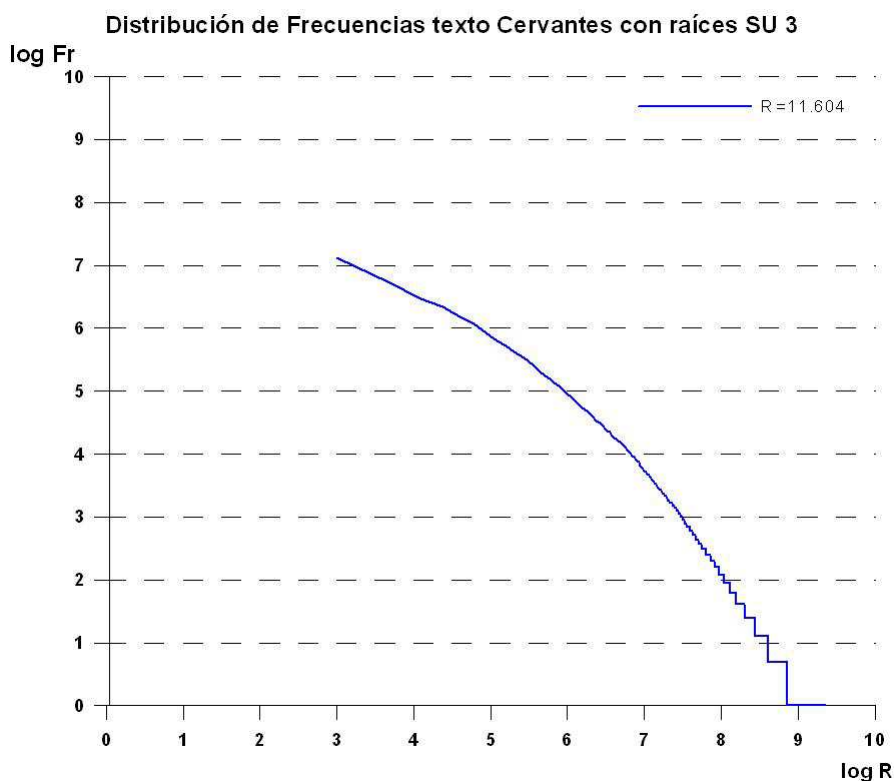


Gráfico 91. Distribución de Frecuencias texto Cervantes con raíces SU 3

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Sufijos de Lovins con adaptaciones al Español y además le incluimos a dicho algoritmo que el mínimo número de caracteres que ha de tener la raíz es de 3 caracteres, así por tanto se ha obtenido un total de 11.604 raíces, lo que supone un 41% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,17$ y para Mandelbrot de $e=1,90$ y $\Sigma=225,67$. Es necesario recordar que la gráfica dibujada utiliza un método denominado *Esponjar Gráfica* que realiza una selección de los datos, a intervalos regulares, copiándose 1 dato de cada 20, lo que reduce bastante los valores dibujados en la gráfica dando una visión general de los mismos.

Texto analizado:	Cervantes.txt de 3.139 K
Tratamiento previo:	Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaicesSufijos:	Mínimo número de caracteres que ha de tener cualquier raíz = 5 Tabla de sufijos=358
Total:	Raíces Método Sufijos (SU 5)=15.500 Reducción de: 55%
Para dibujar gráfica:	Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot:	Zipf (e)= 1,11 Mandelbrot (e)= 1,82 (Σ)= 340,23

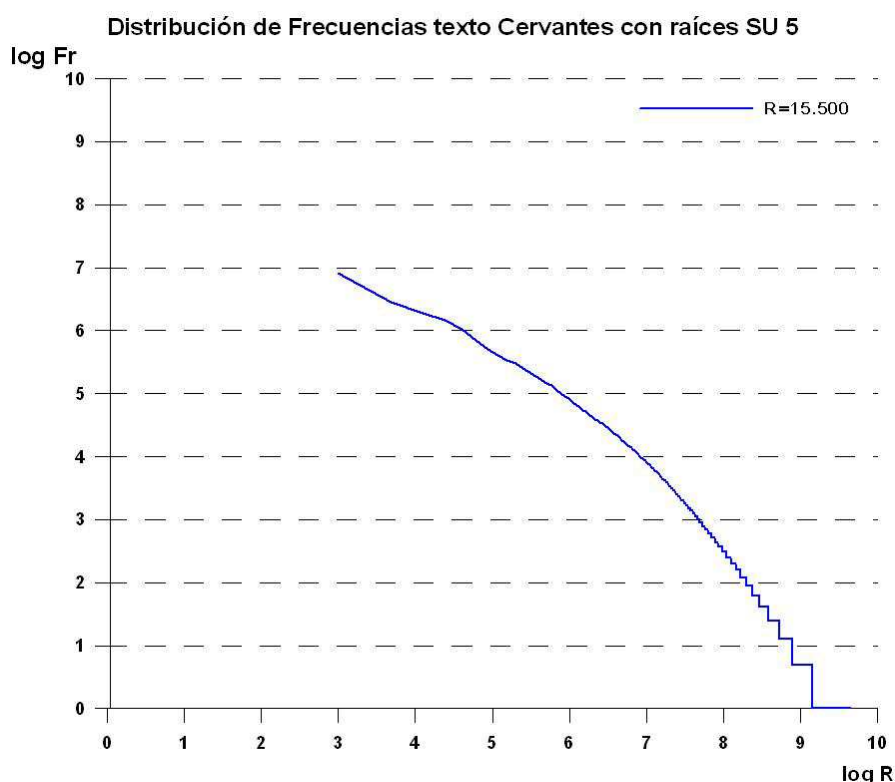


Gráfico 92. Distribución de Frecuencias texto Cervantes con raíces SU 5

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Sufijos de Lovins con adaptaciones al Español y además le incluimos a dicho algoritmo que el mínimo número de caracteres que ha de tener la raíz es de 5 caracteres, así por tanto se ha obtenido un total de 15.500 raíces, lo que supone un 55% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,11$ y para Mandelbrot de $e=1,82$ y $\Sigma=340,23$. Es necesario recordar que la gráfica dibujada utiliza un método denominado *Esponjar Gráfica* que realiza una selección de los datos, a intervalos regulares, copiándose 1 dato de cada 20, lo que reduce bastante los valores dibujados en la gráfica dando una visión general de los mismos. Este método es menos agresivo que el anterior ya que obligamos a que la raíz tenga como mínimo 5 caracteres y no 3 como en el ejemplo primero, por ello el porcentaje reducido es mayor un 55% respecto a un 41% en el caso anterior y las raíces obtenidas obviamente al ser un método menos agresivo se obtiene mayor número de raíces en total 15.500 respecto a las raíces obtenidas en el ejemplo anterior 11.604, lo que observamos al respecto de los valores de Zipf y Mandelbrot es que utilizando el Método de Sufijos de Lovins con adaptaciones al español si dicho método es más agresivo aumentan ligeramente los valores del exponente en ambos casos.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:

Aplicación PIEDRAS.mdb, formulario PIEDRA

ExtraerRaicesVS: si incluye la raíz de la anterior palabra, tomarla=SI

Quitar plural si existe singular=SI

Si después del pico hay consonante, tomarla =SI

Nº letras de la Palabra corta igual a su raíz=3
Penalización vocal/consonante=1,3
Cantidad de palabras para prefijo=11
Total: Raíces Método Variedad de Sucesores (MVS)=11.521
Reducción de: 41%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot: Zipf (e)= 1,14
Mandelbrot (e)= 2,28 (Σ)= 674,45

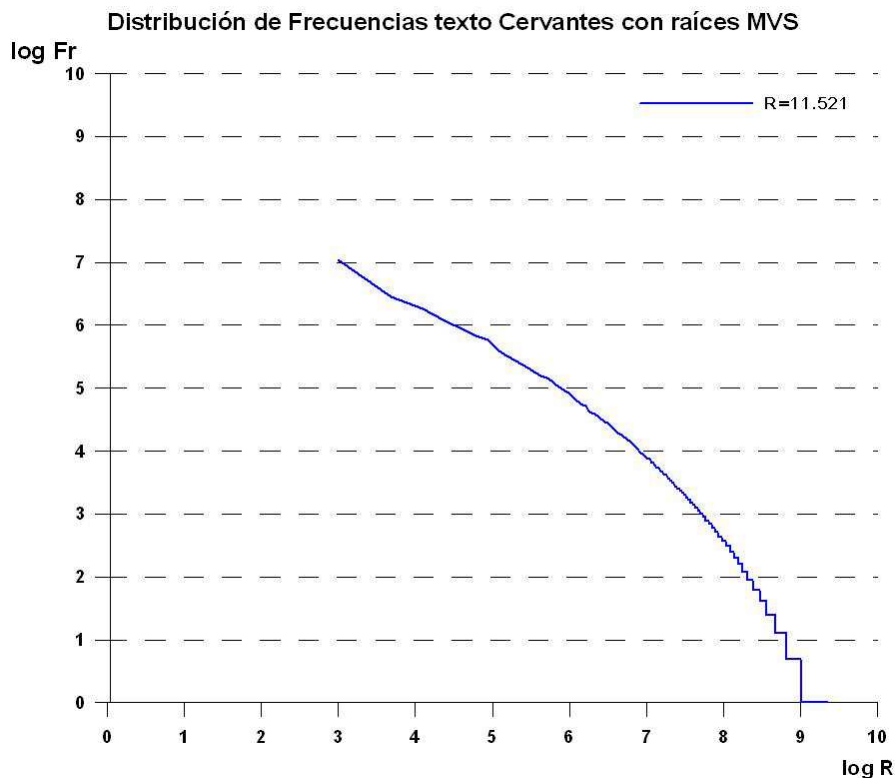


Gráfico 93. Distribución de Frecuencias texto Cervantes con raíces MVS

CONCLUSIÓN:

En este caso hemos utilizado el Método de Variedad de Sucesores con adaptaciones al Español, obteniéndose un total de 11.521 raíces, lo que supone un 41% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,14$ y para Mandelbrot de $e=2,28$ y $\Sigma=674,45$. En este caso también se utiliza el sistema para dibujar la gráfica de *Esponjar Gráfica* explicado anteriormente. Este método obtiene un total de raíces similar al primero que hemos utilizado el *SU 3*, que si recordamos resultaba un método más agresivo de agrupación que el *SU 5*, por tanto observamos que el *MVS* obtiene resultados similares que el *SU*, efectuándose una reducción del 41% del vocabulario en ambos casos. Respecto a los valores de Zipf y Mandelbrot, en el caso del exponente de Zipf el *MVS* obtiene un valor de $e=1,14$ muy similar al valor de $e=1,17$ obtenido con el método de *SU 3*, pero en cambio el valor del exponente de Mandelbrot

es mucho mayor en el caso del MVS $e=2,28$ respecto al valor de $e=1,90$ obtenido con el método SU 3.

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo:
 Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaícesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 80%
 Mínimo número de caracteres que ha de tener cualquier raíz=6
Total: Raíces Método Sufijos (TCP 80% 6)=19.879
 Reducción de: 70%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot: Zipf (e)= 1,07
 Mandelbrot (e)= 1,62 (Σ)= 285,88

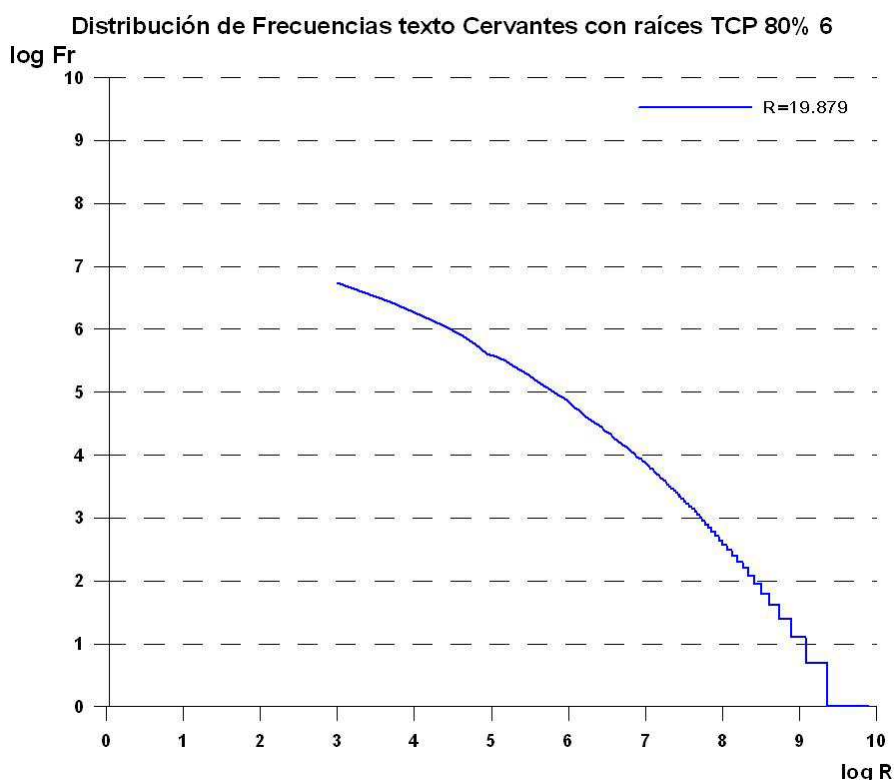


Gráfico 94. Distribución de Frecuencias texto Cervantes con raíces TCP 80% 6

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que el número de letras de la palabra que tomamos como raíz= 80% y el mínimo número de caracteres que ha de tener cualquier raíz=6, así por tanto se ha obtenido un total de 19.879 raíces, lo que supone un 70% de reducción respecto al

vocabulario, además de obtener unos valores para Zipf de $e=1,07$ y para Mandelbrot de $e=1,62$ y $\Sigma=285,88$. Este es el caso menos agresivo de todos los vistos hasta ahora y según las conclusiones obtenidas hasta el momento, un método menos agresivo indica que los valores del exponente de Zipf y Mandelbrot disminuyen ligeramente.

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo:
Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaicesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 60%
Mínimo número de caracteres que ha de tener cualquier raíz=5
Total: Raíces Método Sufijos (TCP 60% 5)=12.331
Reducción de: 43%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot: Zipf (e)= 1,14
Mandelbrot (e)= 2,04 (Σ)= 422,15

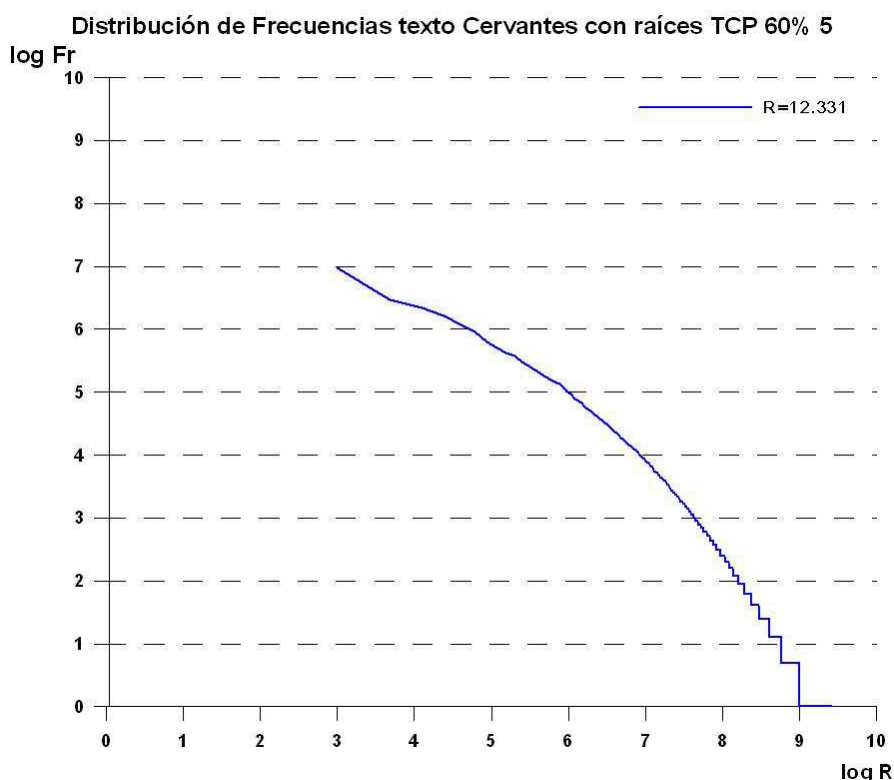


Gráfico 95. Distribución de Frecuencias texto Cervantes con raíces TCP 60% 5

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz=

60% y el mínimo número de caracteres que ha de tener cualquier raíz=5, así se ha obtenido un total de 12.331 raíces, lo que supone un 43% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,14$ y para Mandelbrot de $e=2,04$ y $\Sigma=422,15$. Este ejemplo es mucho más agresivo a la hora de la agrupación de raíces que el anterior, por tanto y según las conclusiones obtenidas hasta el momento, un método más agresivo indica que los valores del exponente de Zipf y Mandelbrot aumentan ligeramente. Se aprecia una similitud entre la cantidad de raíces obtenidas y los valores de los exponentes de Zipf y Mandelbrot con el método visto anteriormente de *MVS*.

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo:
 Aplicación PIEDRAS.mdb, formulario PIEDRA
 ExtraerRaícesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 40%
 Mínimo número de caracteres que ha de tener cualquier raíz=4
Total: Raíces Método Sufijos (TCP 40% 4)=5.369
 Reducción de: 19%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot: Zipf (e)= 1,28
 Mandelbrot (e)= 3,38 (Σ)= 724,36

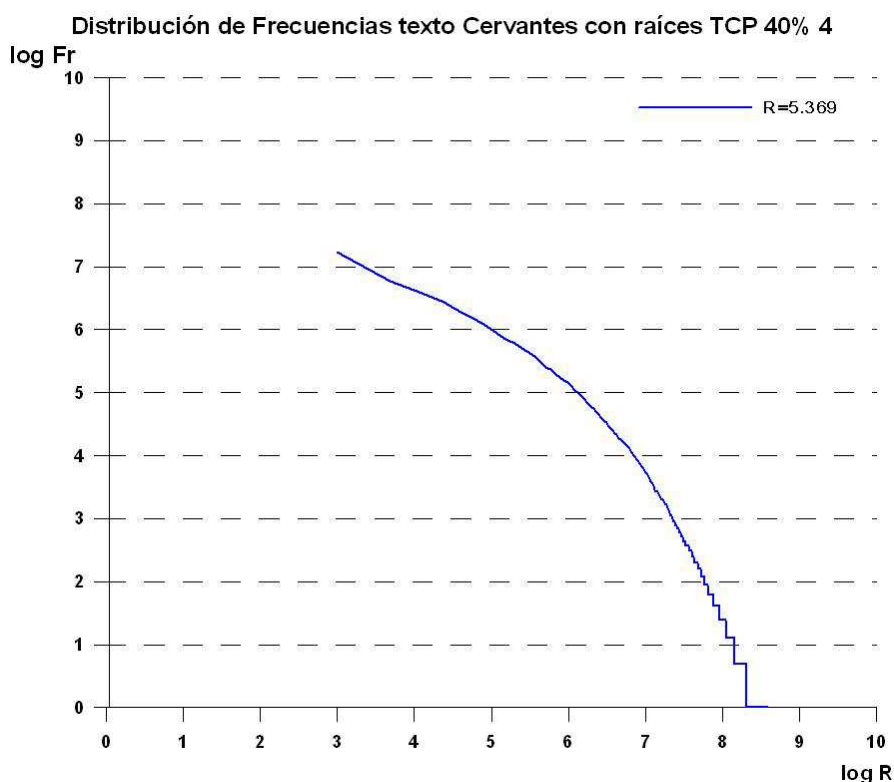


Gráfico 96. Distribución de Frecuencias texto Cervantes con raíces TCP 40% 4

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz=40% y el mínimo número de caracteres que ha de tener cualquier raíz=4, así por tanto se ha obtenido un total de 5.369 raíces, lo que supone un 19% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,28$ y para Mandelbrot de $e=3,38$ y $\Sigma=724,36$. Este ejemplo es mucho más agresivo a la hora de la agrupación de raíces que el anterior, por tanto y según las conclusiones obtenidas hasta el momento, un método más agresivo indica que los valores del exponente de Zipf y Mandelbrot aumentan ligeramente. Es revelador el aumento significativo del exponente de Mandelbrot que aumenta exageradamente cuando la agrupación en raíces es más agresiva y por tanto se obtienen menos raíces, respecto al valor del exponente de Zipf mantiene ligeramente su valor respecto a los demás Stemmers.

Texto analizado:	Cervantes.txt de 3.139 K
Tratamiento previo:	Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaicesTCP:	Tanto por cien del número de letras de la palabra que tomamos como raíz= 30% Mínimo número de caracteres que ha de tener cualquier raíz=3
Total:	Raíces Método Sufijos (TCP 30% 3)=1.761 Reducción de: 6%
Para dibujar gráfica:	Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20
Zipf y Mandelbrot:	Zipf (e)= 1,51 Mandelbrot (e)= 5,63 (Σ)= 579,98

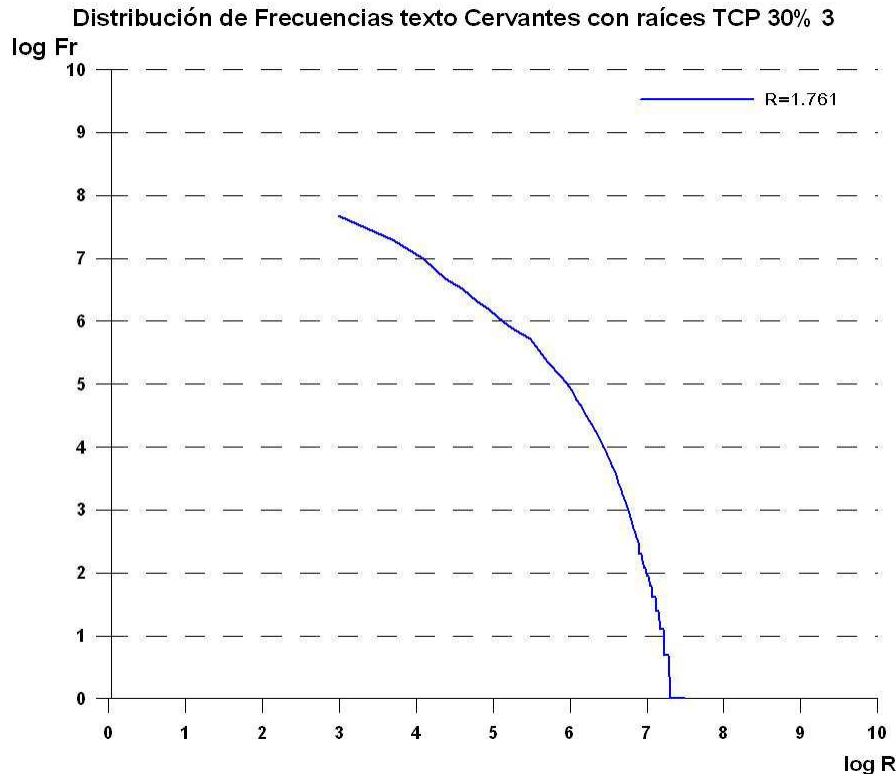


Gráfico 97. Distribución de Frecuencias texto Cervantes con raíces TCP 30% 3

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que el número de letras de la palabra que tomamos como raíz= 30% y el mínimo número de caracteres que ha de tener cualquier raíz=3, así por tanto se ha obtenido un total de 1.761 raíces, lo que supone un 6% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de $e=1,51$ y para Mandelbrot de $e=5,63$ y $\Sigma=579,98$. Este ejemplo, mucho más agresivo a la hora de la agrupación de raíces que el anterior, confirma que con un método más agresivo los valores del exponente de Zipf y Mandelbrot aumentan ligeramente. Salvo en el caso del exponente de Mandelbrot que como hemos dicho en el ejemplo anterior el aumento es significativo aumentando exageradamente cuando la agrupación en raíces es más agresiva y por tanto se obtienen menos raíces.

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo: Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaícesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 20%
 Mínimo número de caracteres que ha de tener cualquier raíz=2
Total: Raíces Método Sufijos (TCP 20% 2)=262
 Reducción de: 0%
Para dibujar gráfica: Aplicación TOPOS.mdb, formulario ESPONJARGRAFICA. Método Esponjar Gráfica: para

llevar a Excel sólo una selección de los datos, extraída a intervalos regulares. Se copian sólo 1 dato de cada 20

Zipf y Mandelbrot: Zipf (e)= 2,10
Mandelbrot (e)= 6,70 (Σ)= 74,18

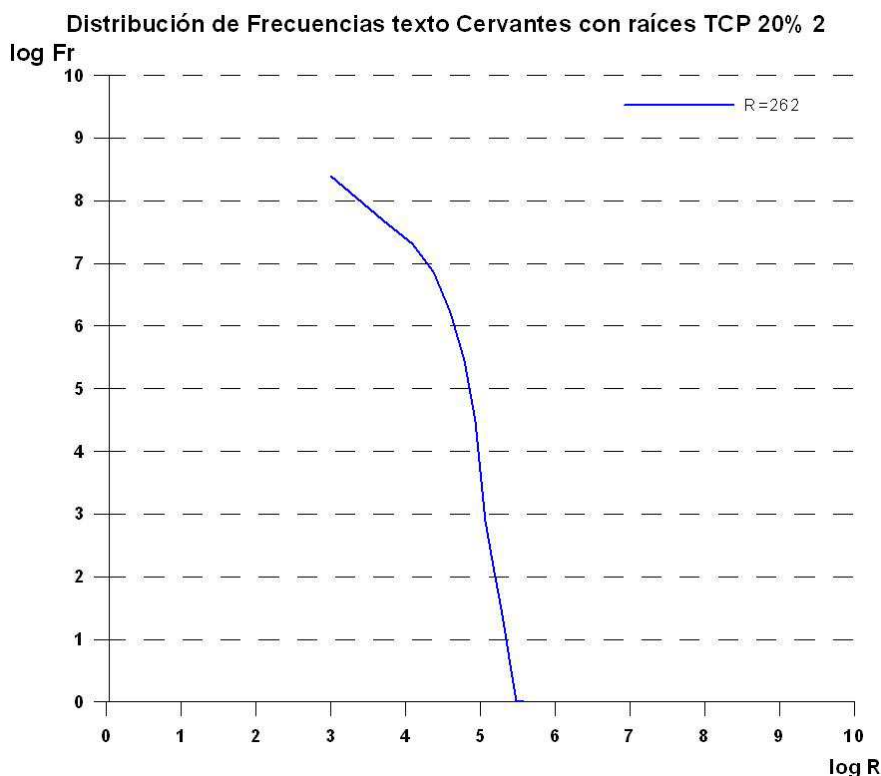


Gráfico 98. Distribución de Frecuencias texto Cervantes con raíces TCP 20% 2

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que el número de letras de la palabra que tomamos como raíz= 20% y el mínimo número de caracteres que ha de tener cualquier raíz=2, así por tanto se ha obtenido un total de 262 raíces, lo que supone un 0% de reducción respecto al vocabulario, además de obtener unos valores para Zipf de e=2,10 y para Mandelbrot de e=6,70 y Σ =74,18. Este ejemplo, el más agresivo a la hora de la agrupación de raíces de todos los ejemplos mostrados confirma que con un método más agresivo los valores del exponente de Zipf y Mandelbrot aumentan ligeramente. Salvo en el caso del exponente de Mandelbrot que como hemos dicho en el ejemplo anterior el aumento es significativo aumentando exageradamente cuando la agrupación en raíces es más agresiva y por tanto se obtienen menos raíces.

Por tanto los resultados obtenidos en la distribución de frecuencias de Zipf y Mandelbrot con los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto por Cien (TCP), sobre el texto Cervantes.txt, de 280.501 palabras, se detalla a continuación ordenado según las raíces obtenidas:

Stemmer	Vocabulario (raíces)	Zipf (e)	Mandelbrot (e)	Mandelbrot (Σ)
Palabras	28.097	1,02	1,33	125,3
TCP 80% letras, mínimo 6	19.879	1,07	1,62	285,88
Sufijos, raíz mínimo 5 letras	15.500	1,11	1,82	340,23
TCP 60% letras, mínimo 5	12.331	1,14	2,04	422,15
Sufijos, raíz mínimo 3 letras	11.604	1,17	1,90	225,67
Variedad de sucesores	11.521	1,14	2,28	674,45
TCP 40% letras, mínimo 4	5.369	1,28	3,38	724,36
TCP 30% letras, mínimo 3	1.761	1,51	5,63	579,98
TCP 20% letras, mínimo 2	262	2,10	6,70	74,18

Tabla 28. Valores obtenidos en la distribución de frecuencias texto Cervantes con diversos Stemmers

El siguiente gráfico es un cuadro comparativo de los gráficos de páginas anteriores que constituyen los resultados obtenidos con varios Stemmers.

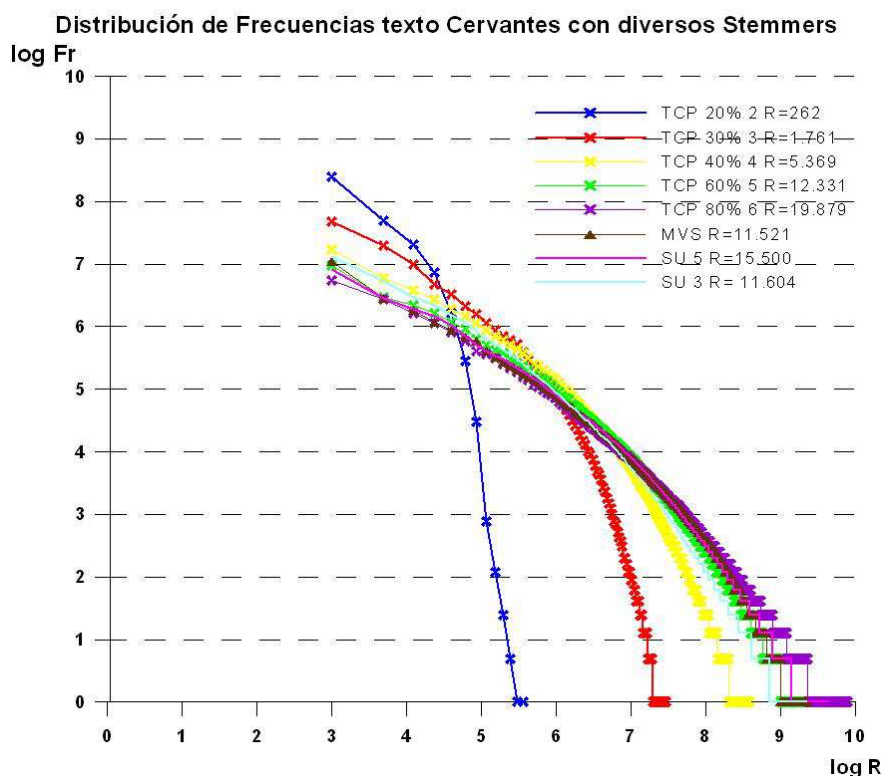


Gráfico 99. Distribución de Frecuencias texto Cervantes con diversos Stemmers

Según los resultados de la tabla anterior y este gráfico sugiere que el exponente de Zipf mantiene cierta regularidad a través de procesos más o menos agresivos de extracción de raíces. Sobre todo si tenemos en cuenta que el exponente de Zipf está

calculado sobre todos los valores y no ajustado a tramos como realizaremos a continuación.

Podemos sacar las siguientes conclusiones:

- ✓ El exponente de Zipf calculado sobre todos los valores de un texto fijo (*Cervantes*) mantiene cierta regularidad con una progresión monótona con Stemmers más o menos agresivos.
- ✓ Es irrelevante el método de Stemming seguido para obtener las raíces: la tabla está ordenada según número de raíces obtenidas y vemos una progresión monótona en el exponente de Zipf a pesar de que los métodos aparecen entremezclados. Incluso con el método de dejar la palabra reducida a un tanto por cien fijo de sus letras, que es bastante artificial.
- ✓ Si exageramos más la agrupación en raíces, el valor del exponente aumenta ligeramente, como ya se apuntaba en el anterior capítulo seis, al agrupar en raíces faltan palabras de frecuencias bajas y esto hace aumentar el exponente. A menos raíces resultantes mayor valor del exponente (siempre dentro del intervalo 1 a 2), o dicho de otro modo si el Stemmer es más agresivo: menos raíces resultantes, se obtiene mayor valor del exponente
- ✓ El sumando de la fórmula de Mandelbrot también crece al agrupar más las palabras indicando que aumenta la curvatura en cualquier representación gráfica, es decir cada vez se cumple menos la Ley de Zipf ya que se aleja de la distribución de Zipf. (el último valor de 74,18 debe relacionarse con el valor tan alto 6,70 del exponente; este caso, en el que las palabras se han reducido a solo dos letras debe considerarse como un caso muy extremo donde las cosas quedan ya algo distorsionadas)
- ✓ La aplicación de diferentes Stemmers afecta sólo a los tramos inicial y final de la pendiente en la distribución de frecuencias de Zipf y Mandelbrot.
- ✓ ¿Qué ocurre si calculamos el exponente de Zipf sólo a la parte central, es decir a la distribución de frecuencias alrededor del PT (Transition Point)?, mostrará el exponente de Zipf una progresión monótona como hemos visto hasta ahora o por el contrario se va a demostrar la independencia de la pendiente en este tramo central.

A continuación vamos a corroborar que el exponente de Zipf calculado sobre las frecuencias alrededor del PT, es decir las frecuencias centrales de un texto fijo mantiene su valor aproximadamente 1, sin desviaciones significativas, por tanto podemos asegurar que si ajustamos Zipf a las raíces de la parte central de la distribución de frecuencias éste no depende del procedimiento (Stemmer) utilizado sino que es independiente. Efectivamente comprobaremos que utilizar un Stemmer u otro afecta sólo a los tramos inicial y final donde se sitúan las palabras de frecuencias altas y bajas. Si observamos la representación log-log en el penúltimo de los casos, la distribución de frecuencias de las raíces obtenidas del texto *Cervantes* mediante la utilización del Stemmer TCP 30% 3. En este caso se dibujan todos los datos del texto, es decir no se

utiliza la opción de Esponjar Gráfica que permite sólo una selección de los datos, extraída a intervalos regulares en el cual se copian sólo 1 dato de cada 20, sino que el siguiente gráficos muestra todos los datos.

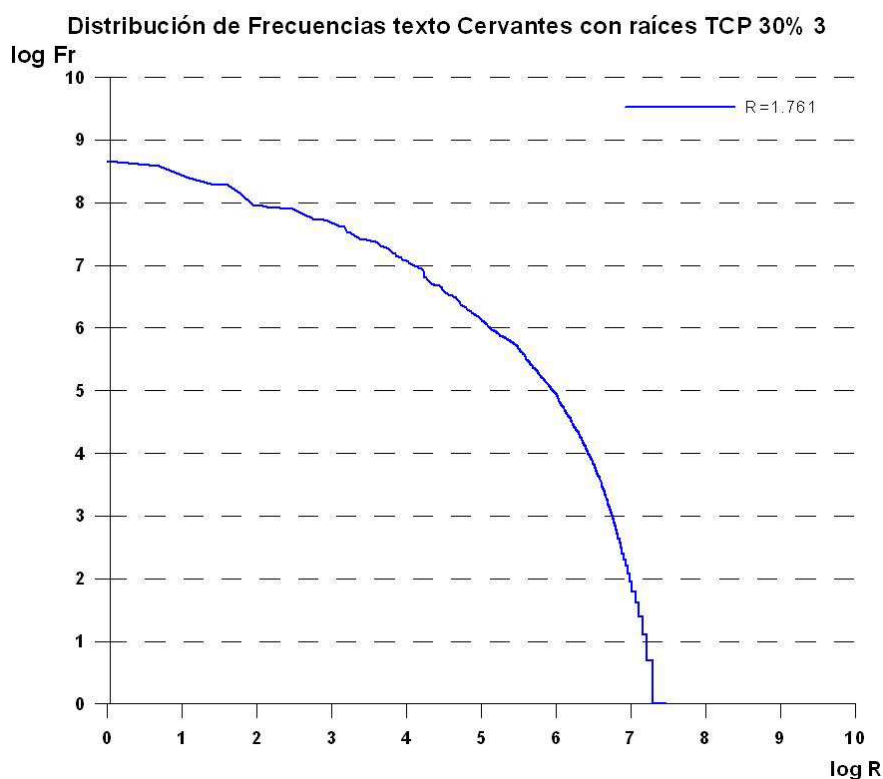


Gráfico 100. Distribución de Frecuencias texto Cervantes con raíces TCP 30% 3

Claramente podemos distinguir aquí tres tramos:

- ✓ El primero corresponde a palabras de muy alta frecuencia que en su mayoría cumplen una función sintáctica en el texto, las palabras vacías (artículos, preposiciones, expresiones comunes del lenguaje, nombres propios) y cuya abundancia responde más a estadística que a la fórmula de Zipf
- ✓ El tramo central, que en la representación log-log aparece bastante rectilíneo es el que sí cumple la ley de Zipf. De la consideración de esta gráfica y muchas otras similares que no reproducimos puede deducirse que el tramo central está formado aproximadamente por los rangos cuyo logaritmo está entre $0,4 * v$ y $0,7 * v$.
- ✓ El último en los rangos superiores, delata el fenómeno ampliamente comentado en la literatura, de la aparente falta de palabras. Sea cual sea su causa o explicación hay que admitir que este tramo de rangos no cumple la ley de Zipf.

Visto que la distribución de frecuencias de las raíces manifiesta tres tramos claramente diferenciados en los que el valor del exponente de Zipf tiene una ligera dependencia con la intensidad de extracción de raíces: a menos raíces resultantes, mayor valor del exponente (siempre dentro del intervalo 1 a 2) y que el valor del exponente de Zipf muestra una progresión monótona a pesar del Stemmer utilizado vamos a tomar los

tramos centrales equivalente aproximadamente a las frecuencias próximas al Punto de Transición (PT), a los que ajustaremos la distribución de Zipf con la división en tramos del Modelo Log-% de creación propia y ampliamente estudiado en el Capítulo seis.

7.9. Estudio de la distribución de frecuencias de Zipf ajustado a las frecuencias próximas al PT (Transition Point) con los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto por Cien (TCP)

Una vez visto los resultados obtenidos con las raíces vamos a ajustar la fórmula de Zipf al tramo central de cada colección de valores, este tramo central correspondería a los rangos entre $V^{0,4}$ y $V^{0,7}$, Hallando así el Punto de Transición o PT (*Transition Point TP*).

Así pues, aprovechando la división en tramos que hemos creado con el Modelo Log-%⁵⁴, vamos a tomar los tramos centrales equivalente aproximadamente a las frecuencias próximas al Punto de Transición (PT).

De las anteriores gráficas se desprende que el exponente de Zipf calculado sobre todos los valores de la distribución de frecuencias del texto de Cervantes mantiene cierta regularidad a través de procesos más o menos agresivos de extracción de raíces.

A continuación reproducimos todas las gráficas en forma log-log de los valores de textos reales y de la predicción según la fórmula de Zipf ajustada al tramo central, para constatar la independencia de la pendiente de este tramo y que la diferente agrupación de las palabras afecta solo a los tramos inicial y final.

Repetimos entonces el estudio haciendo ajustes parciales de la fórmula de Zipf, sólo a las frecuencias de las raíces alrededor del PT; ya que el resto de autores de la literatura utiliza la fórmula de Zipf para predecir frecuencias en todos los rangos.

Siguiendo con el mismo ejemplo del autor *Miguel de Cervantes Saavedra* y su gran obra *EL INGENIOSO HIDALGO DON QUIJOTE DE LA MANCHA*.

Texto analizado:	Cervantes.txt de 3.139 K
Tratamiento previo:	
Aplicación PIEDRAS.mdb, formulario PIEDRA	
ExtraerRaicesSufijos:	Mínimo número de caracteres que ha de tener cualquier raíz = 3 Tabla de sufijos=358
Total:	Raíces Método Sufijos (SU 3)=11.604 Reducción de: 41%
Para dibujar gráfica:	Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.
Zipf:	Zipf (e)= 0,89

⁵⁴ Véase Capítulo 6

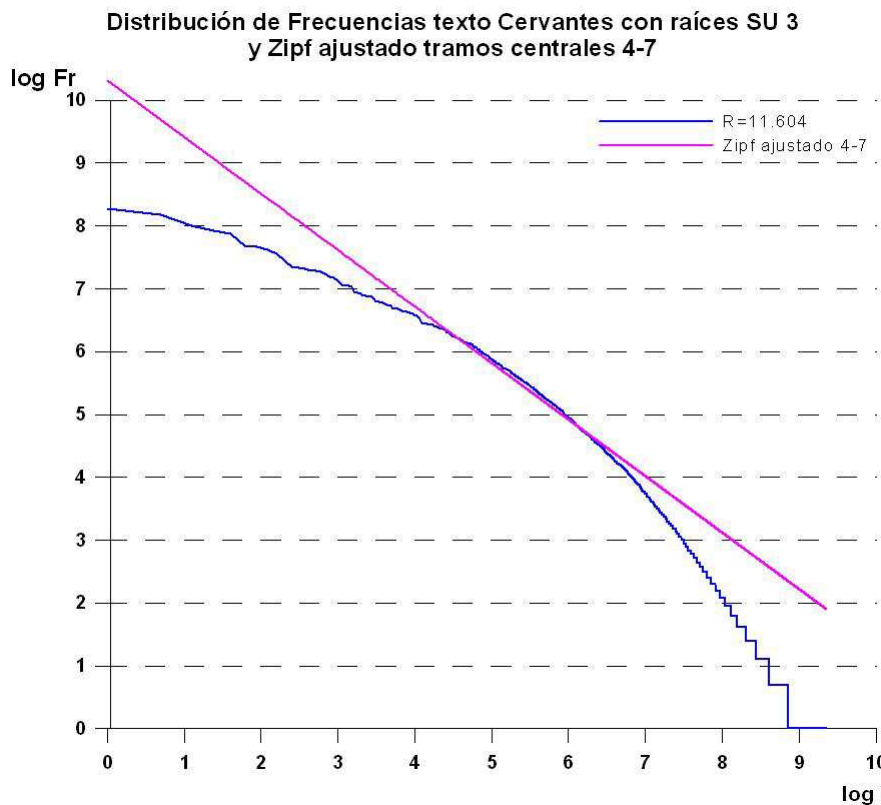


Gráfico 101. Distribución de frecuencias texto Cervantes con raíces SU 3 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Sufijos de Lovins con adaptaciones al Español y además le incluimos a dicho algoritmo que el mínimo número de caracteres que ha de tener la raíz es de 3 caracteres, así por tanto se ha obtenido un total de 11.604 raíces, lo que supone un 41% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de $e=0,89$.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:
 Aplicación PIEDRAS.mdb, formulario PIEDRA
 ExtraerRaicesSufijos: Mínimo número de caracteres que ha de tener cualquier raíz = 5
 Tabla de sufijos=358

Total: Raíces Método Sufijos (SU 5)=15.500
 Reducción de: 55%

Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,80

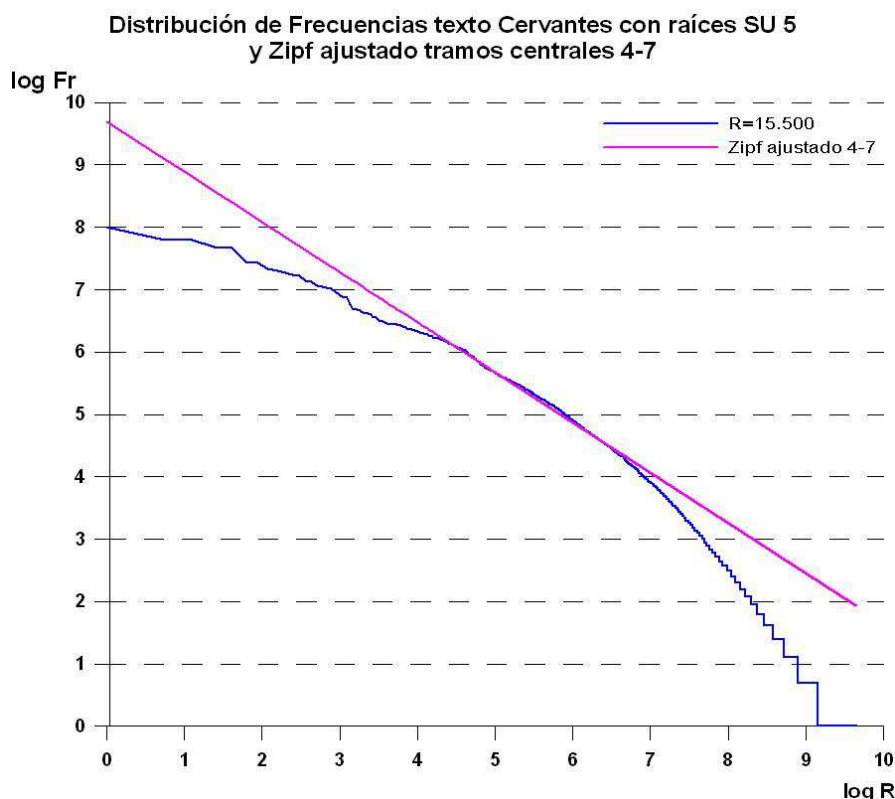


Gráfico 102. Distribución de frecuencias texto Cervantes con raíces SU 5 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Sufijos de Lovins con adaptaciones al Español y además le incluimos a dicho algoritmo que el mínimo número de caracteres que ha de tener la raíz es de 5 caracteres, así por tanto se ha obtenido un total de 15.500 raíces, lo que supone un 55% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de $e=0,80$. Este método es menos agresivo que el anterior ya que obligamos a que la raíz tenga como mínimo 5 caracteres y no 3 como en el ejemplo primero, por ello el porcentaje reducido es mayor un 55% respecto a un 41% en el caso anterior y las raíces obtenidas obviamente al ser un método menos agresivo se obtiene mayor número de raíces en total 15.500 respecto a las raíces obtenidas en el ejemplo anterior 11.604, lo que observamos al respecto del valor del exponente de Zipf es que utilizando el Método de Sufijos de Lovins con adaptaciones al español si dicho método es más agresivo aumentan ligeramente los valores del exponente de Zipf.

Texto analizado:	Cervantes.txt de 3.139 K
Tratamiento previo:	
Aplicación PIEDRAS.mdb, formulario PIEDRA	
ExtraerRaícesVS:	si incluye la raíz de la anterior palabra, tomarla=SI Quitar plural si existe singular=SI Si después del pico hay consonante, tomarla =SI Nº letras de la Palabra corta igual a su raíz=3 Penalización vocal/consonante=1,3 Cantidad de palabras para prefijo=11
Total:	Raíces Método Variedad de Sucesores (MVS)=11.521 Reducción de: 41%

Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,80

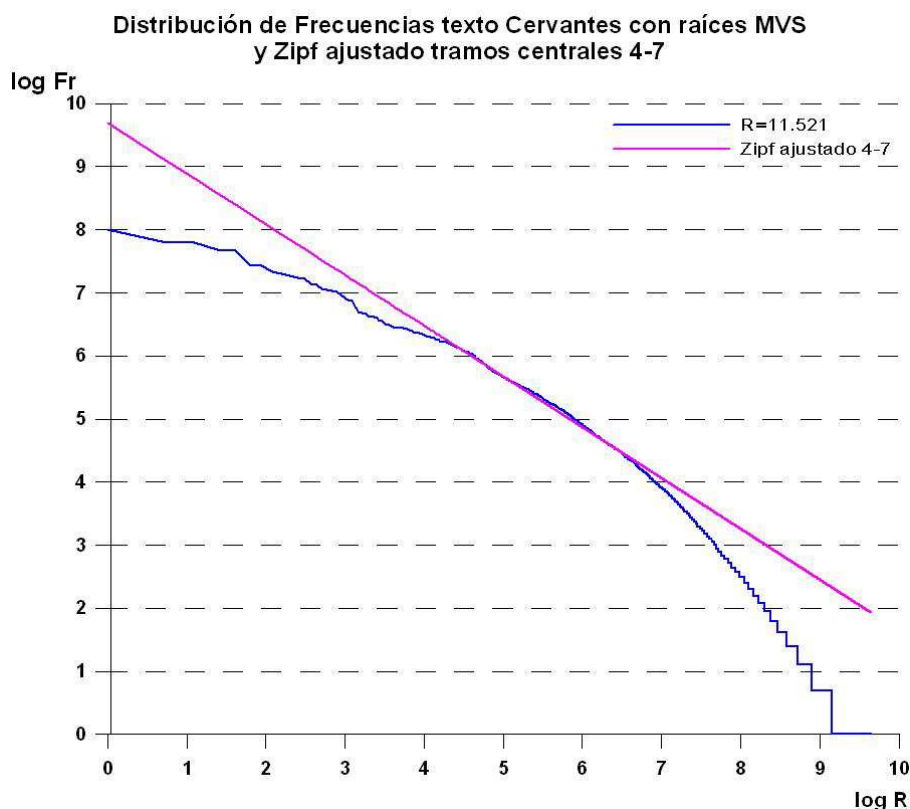


Gráfico 103. Distribución de frecuencias texto Cervantes con raíces MVS y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso hemos utilizado el Método de Variedad de Sucesores con adaptaciones al Español, así por tanto se ha obtenido un total de 11.521 raíces, lo que supone un 41% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de $e=0,80$.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:

Aplicación PIEDRAS.mdb, formulario PIEDRA

ExtraerRaicesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 80%

Mínimo número de caracteres que ha de tener cualquier raíz=6

Total: Raíces Método Sufijos (TCP 80% 6)=19.879

Reducción de: 70%

Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,83

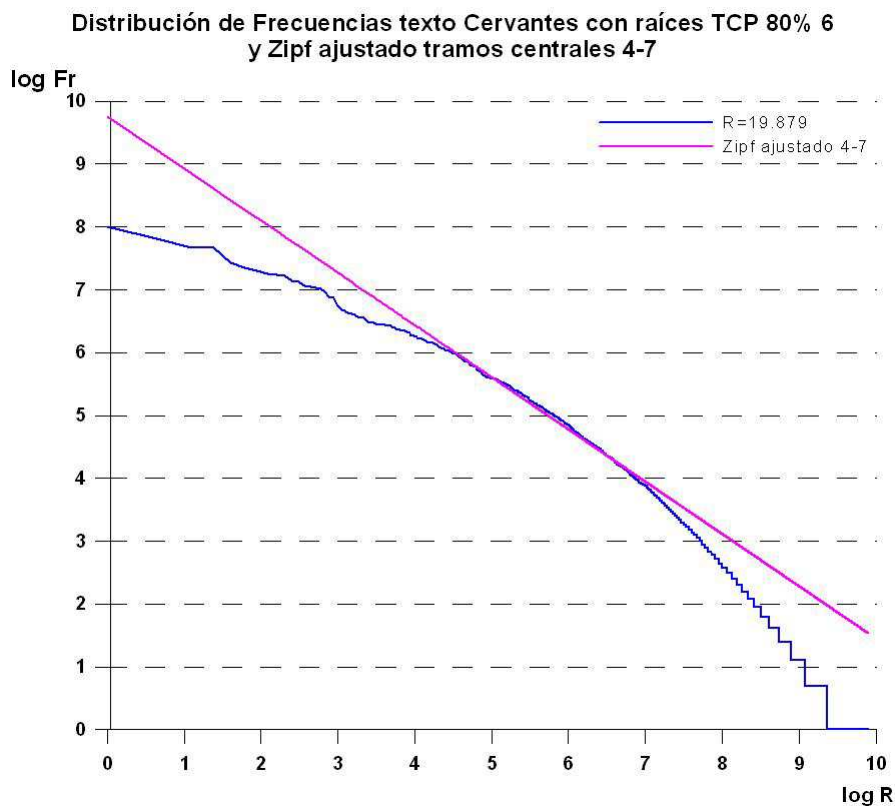


Gráfico 104. Distribución de frecuencias texto Cervantes con raíces TCP 80% 6 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso hemos utilizado el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz= 80% y el mínimo número de caracteres que ha de tener cualquier raíz=6, así por tanto se ha obtenido un total de 19.879 raíces, lo que supone un 70% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de $e=0,83$. Este es el caso menos agresivo de todos los vistos hasta ahora y el exponente mantiene su valor respecto a los casos anteriores.

Texto analizado: Cervantes.txt de 3.139 K
Tratamiento previo: Aplicación PIEDRAS.mdb, formulario PIEDRA
ExtraerRaicesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 60%
 Mínimo número de caracteres que ha de tener cualquier raíz=5
Total: Raíces Método Sufijos (TCP 60% 5)=12.331
 Reducción de: 43%
Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logy y logz, de este modo se

dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,79

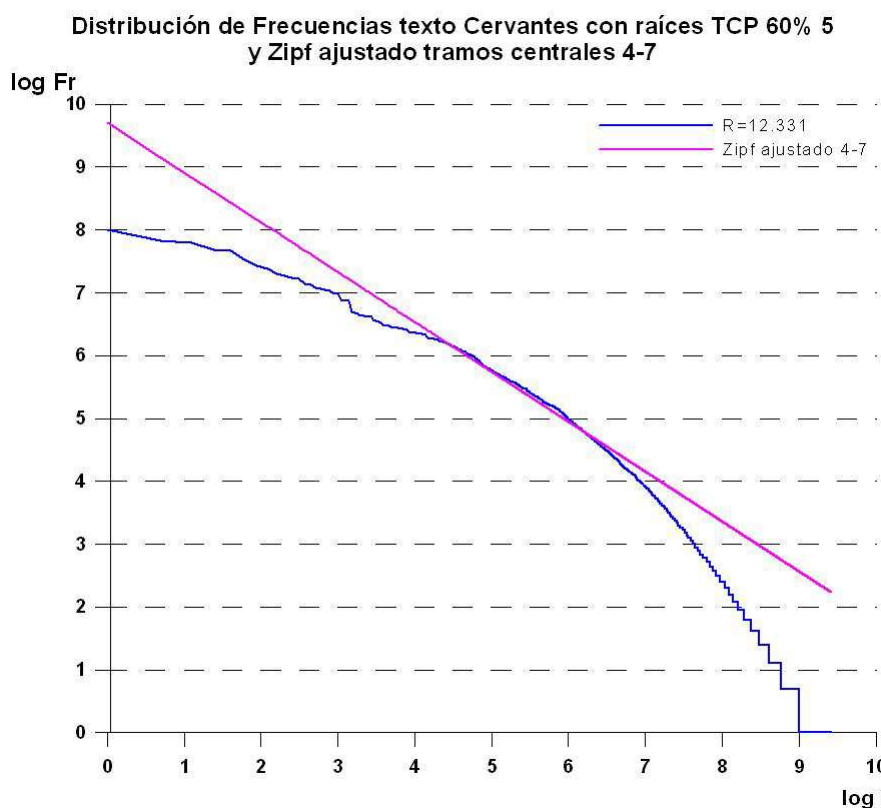


Gráfico 105. Distribución de frecuencias texto Cervantes con raíces TCP 60% 5 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz= 60% y el mínimo número de caracteres que ha de tener cualquier raíz=5, así por tanto se ha obtenido un total de 12.331 raíces, lo que supone un 43% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de e=0,79. Este ejemplo es mucho más agresivo a la hora de la agrupación de raíces que el anterior, y constatamos la independencia del exponente respecto al Stemmer utilizado y su grado de agrupación.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:

Aplicación PIEDRAS.mdb, formulario PIEDRA

ExtraerRaicesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 40%
Mínimo número de caracteres que ha de tener cualquier raíz=4

Total: Raíces Método Sufijos (TCP 40% 4)=5.369
Reducción de: 19%

Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,72

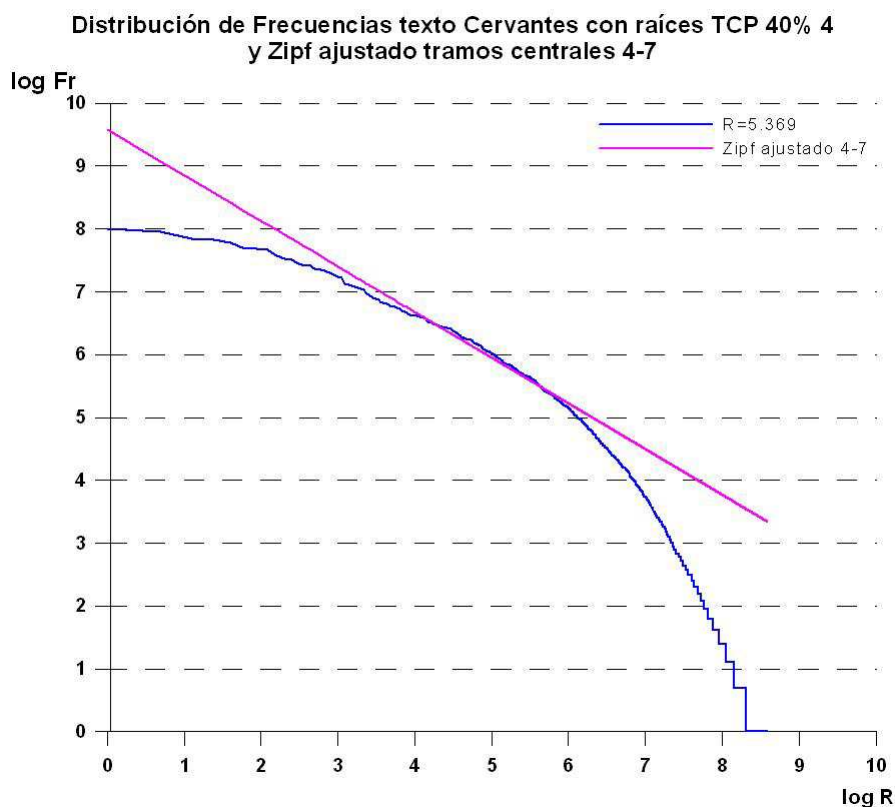


Gráfico 106. Distribución de frecuencias texto Cervantes con raíces TCP 40% 4 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz= 40% y el mínimo número de caracteres que ha de tener cualquier raíz=4, así por tanto se ha obtenido un total de 5.369 raíces, lo que supone un 19% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de $e=0,72$. Este ejemplo es mucho más agresivo a la hora de la agrupación de raíces que el anterior, por tanto y según las conclusiones obtenidas hasta el momento, observamos que disminuye ligeramente respecto al anterior, no siendo significativas las desviaciones del exponente.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:

Aplicación PIEDRAS.mdb, formulario PIEDRA

ExtraerRaicesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 30%

Mínimo número de caracteres que ha de tener cualquier raíz=3

Total: Raíces Método Sufijos (TCP 30% 3)=1.761
 Reducción de: 6%
 Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.
 Zipf: Zipf (e)= 0,84

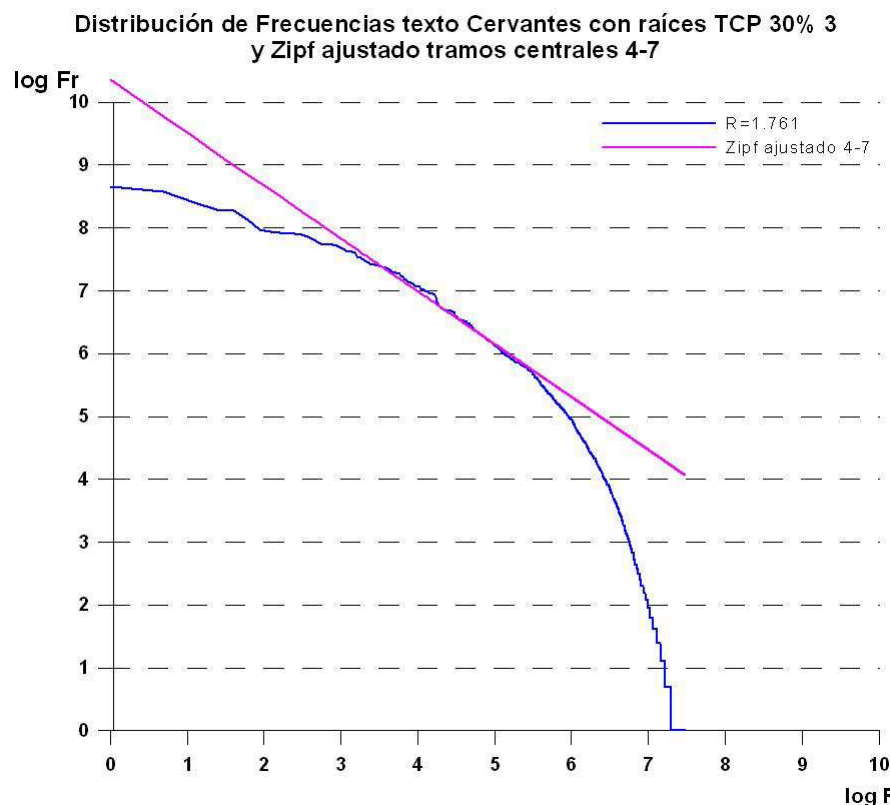


Gráfico 107. Distribución de frecuencias texto Cervantes con raíces TCP 30% 3 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz= 30% y el mínimo número de caracteres que ha de tener cualquier raíz=3, así por tanto se ha obtenido un total de 1.761 raíces, lo que supone un 6% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de e=0,84. Este ejemplo, mucho más agresivo a la hora de la agrupación de raíces que el anterior, altera las conclusiones vistas hasta ahora ya que en vez de seguir la tendencia de que con un método más agresivo el valor del exponente de Zipf disminuye ligeramente cuando Zipf se ha ajustado al tramo central, lo que ocurre es el caso contrario y aumenta el valor notablemente.

Texto analizado: Cervantes.txt de 3.139 K

Tratamiento previo:
 Aplicación PIEDRAS.mdb, formulario PIEDRA

ExtraerRaícesTCP: Tanto por cien del número de letras de la palabra que tomamos como raíz= 20%
 Mínimo número de caracteres que ha de tener cualquier raíz=2

Total: Raíces Método Sufijos (TCP 20% 2)=262
 Reducción de: 0%

Para dibujar gráfica: Aplicación TOPOS.mdb, tabla datoszipf, se exportan los datos a Excel de las columnas logv y logz, de este modo se dibujan en el gráfico todos los datos de la distribución de raíces.

Zipf: Zipf (e)= 0,81

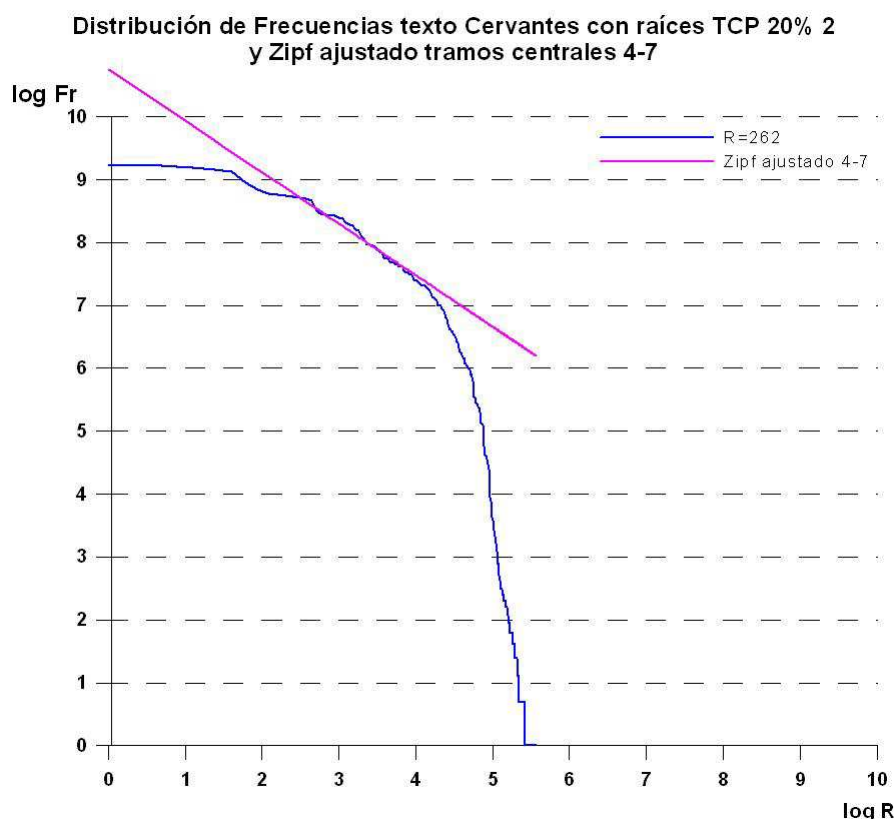


Gráfico 108. Distribución de frecuencias texto Cervantes con raíces TCP 20% 2 y Zipf ajustado tramos 4-7

CONCLUSIÓN:

En este caso volvemos a utilizar el algoritmo de Tanto por Cien TCP-% incluyendo entre las restricciones que del número de letras de la palabra que tomamos como raíz= 20% y el mínimo número de caracteres que ha de tener cualquier raíz=2, así por tanto se ha obtenido un total de 262 raíces, lo que supone un 0% de reducción respecto al vocabulario, además de obtener el valor para Zipf ajustado al tramo central de e=0,81. Este ejemplo, el más agresivo a la hora de la agrupación de raíces de todos los ejemplos mostrados igual que el ejemplo anterior altera las conclusiones vistas hasta ahora ya que en vez de seguir la tendencia de que con un método más agresivo el valor del exponente de Zipf disminuye ligeramente cuando Zipf se ha ajustado al tramo central, lo que ocurre es el caso contrario y aumenta el valor notablemente. Igualmente

observamos claramente como en el gráfico la distribución de frecuencias queda distorsionada respecto a las anteriores

Por tanto los resultados obtenidos en la distribución de frecuencias de Zipf ajustado a las frecuencias próximas al PT (Transition Point) con los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto por Cien (TCP) sobre el texto Cervantes.txt, de 280.501 palabras, se detalla a continuación ordenado según las raíces obtenidas:

<i>Stemmer</i>	<i>Vocabulario (raíces)</i>	<i>Zipf (e)</i>
Palabras	28.097	1,02
TCP 80% letras, mínimo 6	19.879	0,83
Sufijos, raíz mínimo 5 letras	15.500	0,80
TCP 60% letras, mínimo 5	12.331	0,79
Sufijos, raíz mínimo 3 letras	11.604	0,89
Variedad de sucesores	11.521	0,80
TCP 40% letras, mínimo 4	5.369	0,72
TCP 30% letras, mínimo 3	1.761	0,84
TCP 20% letras, mínimo 2	262	0,81

Tabla 29. Valores obtenidos en la distribución de frecuencias de Zipf ajustada al PT (Punto de Transición) texto Cervantes con diversos Stemmers

Podemos sacar las siguientes conclusiones:

- ✓ El exponente de Zipf mantiene su valor de aproximadamente 1, con independencia de la granularidad con la que estamos observando el texto. Las desviaciones alrededor de 1 que se observan en la tabla no son significativas. Por ello se demuestra que la utilización de varios Stemmers no afecta al valor del exponente de Zipf
- ✓ Igualmente constatamos la independencia de la pendiente de la fórmula de Zipf ajustada al tramo central alrededor del PT y que la diferente utilización de varios Stemmers afecta sólo a los tramos inicial y final donde se sitúan las palabras de frecuencias altas y bajas.

En el siguiente gráfico se muestran los valores obtenidos del exponente de Zipf con distintos Stemmers aplicados a distintos textos concretamente 23 textos distintos, todos ellos de tipo literario y de tamaños diversos. Con este gráfico demostramos que si calculamos el exponente de Zipf con todos los valores de la distribución de frecuencias y no sólo a los centrales el valor de (e) obtenido con palabras y con raíces es distinto. Y por otro lado demostramos que el exponente (e) de Zipf calculado sobre todos los valores de la distribución de frecuencias con varios textos diferentes aumenta con el tamaño del texto. En sentido general, con todos los Stemmers o con palabras ocurre que: (e) es mayor cuando (P) es mayor y viceversa.

Siendo, (e) = exponente de Zipf, (P) = tamaño del texto

La conclusión referida en páginas anteriores también es claramente observable en este gráfico, ya que a menos raíces resultantes, mayor valor del exponente; es decir el

Stemmer MVS y SU3 son más agresivos obteniendo menos raíces que el Stemmer TCP806, por tanto los valores de los exponentes en cada caso son mucho mayores.

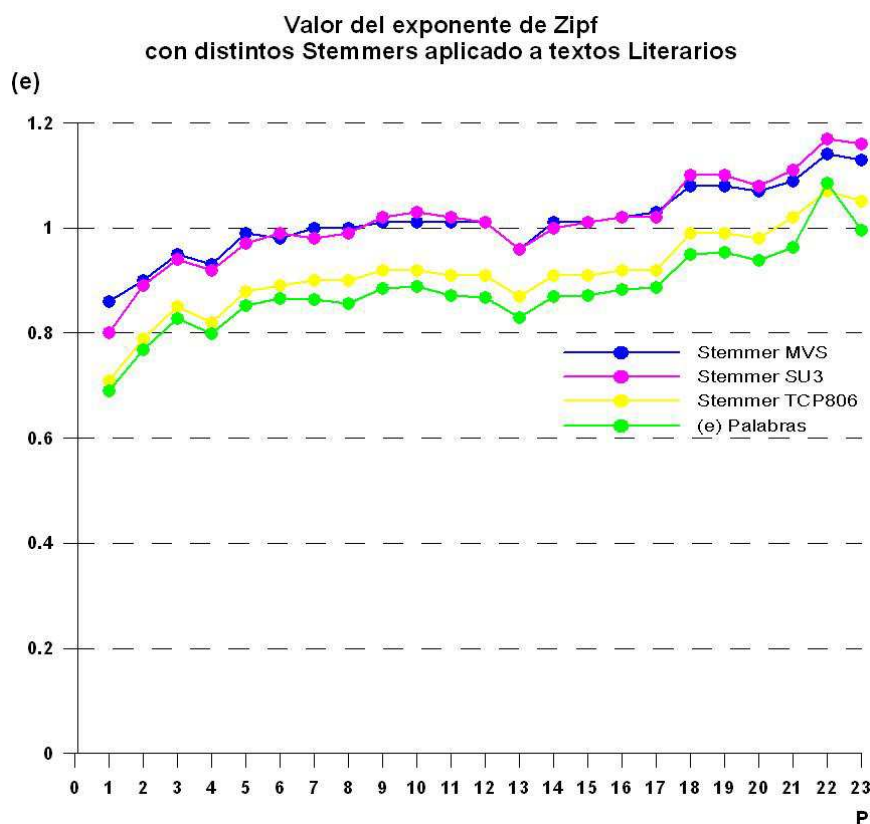


Gráfico 109. Valor del exponente de Zipf con distintos Stemmers aplicado a textos literarios

Leyenda gráfico:

Eje de abscisas (P)	Texto Literario	P literarios	(e) Zipf palabras	(e) Zipf Stemmer MVS	(e) Zipf Stemmer SU3	(e) Zipf Stemmer TCP806
1	gg1	6821	0,69	0,86	0,80	0,71
2	seis	25696	0,77	0,90	0,89	0,79
3	ksoti2	28279	0,83	0,95	0,94	0,85
4	ksoti1	28582	0,80	0,93	0,92	0,82
5	erudi1	53940	0,85	0,99	0,97	0,88
6	ksoti	56854	0,87	0,98	0,99	0,89
7	dos	73790	0,86	1,00	0,98	0,90
8	castea	74939	0,86	1,00	0,99	0,90
9	larra1	76550	0,88	1,01	1,02	0,92
10	episo1	81398	0,89	1,01	1,03	0,92
11	casted	86666	0,87	1,01	1,02	0,91
12	castec	89556	0,87	1,01	1,01	0,91
13	costac	91714	0,83	0,96	0,96	0,87
14	costaa	93140	0,87	1,01	1,00	0,91
15	casteb	93270	0,87	1,01	1,01	0,91
16	costab	94283	0,88	1,02	1,02	0,92
17	vari	95045	0,89	1,03	1,02	0,92

<i>Eje de abscisas (P)</i>	<i>Texto Literario</i>	<i>P literarios</i>	<i>(e) Zipf palabras</i>	<i>(e) Zipf Stemmer MVS</i>	<i>(e) Zipf Stemmer SU3</i>	<i>(e) Zipf Stemmer TCP806</i>
18	castee	106079	0,95	1,08	1,10	0,99
19	episogrande	204757	0,95	1,08	1,10	0,99
20	costa1	279141	0,94	1,07	1,08	0,98
21	vari1	380054	0,96	1,09	1,11	1,02
22	cervantes	385213	1,09	1,14	1,17	1,07
23	castela	450574	1,00	1,13	1,16	1,05

Tabla 30. Leyenda gráfico núm. 109

Podemos sacar las siguientes conclusiones:

- ✓ En líneas generales, ya se realice el estudio con un mismo texto o diferentes si comparamos si afecta el Stemmer al valor del exponente (e) de Zipf, a menos raíces (Stemmer más agresivo) obtendremos un valor mayor de (e).
- ✓ La diferente utilización de los Stemmers afecta sólo a los tramos inicial y final (palabras de frecuencias altas y bajas) de la distribución de frecuencias de Zipf.
- ✓ El aplicar un Stemmer necesariamente provoca este efecto de falta de palabras de frecuencias bajas, pero lo provoca el Stemmer no lo provoca el autor, no es consecuencia ni del lenguaje, ni del autor, ni del texto.

7.10. Evaluación de los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto por Cien (TCP), Precisión-Exhaustividad.

En páginas anteriores se comentaban las evaluaciones con diferentes Stemmers de la autora Gómez Díaz (2005), la cual indicaba sus conclusiones al respecto, según la autora la supresión o eliminación de las palabras vacías afectaba positivamente tanto a la Precisión como a la Exhaustividad, tanto si aplicamos la lematización como si no. Esta primera conclusión parece lógica. La segunda conclusión a la que llega la autora es que para el español es más beneficiosa la lematización flexiva (la menos agresiva) suprimiendo antes las palabras vacías.

Los autores Figuerola et. al. (2004), analizan diversos algoritmos de normalización para evaluar su eficacia, explorando así las posibilidades y efecto del Stemming en la Recuperación de Información en Español, tras los experimentos realizados concluyen que la normalización de términos produce mejoras en la Recuperación, igualmente esta primera conclusión parece lógica, la segunda conclusión a la que llegan los autores es que el uso de n-gramas parece desaconsejable e igualmente sugieren que los algoritmos más complejos que incluyen conocimiento lingüístico, (lematización flexiva y derivativa) no alcanzan los resultados que se obtienen sin aplicar ningún tipo de normalización, por tanto no superan los resultados conseguidos con un simple stemmer (reduce los plurales a singular), mucho más fácil de implementar.

En esta investigación vamos a realizar la evaluación de los Stemmers: Método de Variedad de Sucesores (MVS), Método de Sufijos (SU) y Método de extracción Tanto

por Cien (TCP-%), utilizaremos para ello las medidas de evaluación universalmente conocida Precisión-Exhaustividad.

Se ha utilizado para el estudio de la Precisión y Exhaustividad, el texto del autor *Miguel de Cervantes Saavedra* y su gran obra *EL INGENIOSO HIDALGO DON QUIJOTE DE LA MANCHA*. Sobre este texto se han aplicado los diferentes Stemmers que hemos desarrollado para posteriormente realizar las búsquedas en la aplicación⁵⁵ implementada para la Recuperación de Información desarrollada con el modelo vectorial, previamente analizado el texto con los distintos procesos de Stemming, la obtención de las similitudes, etc.

Realizaremos dos tipos de búsquedas una con palabras clave, es decir más corta, suponiendo que las consultas reales de los usuarios tienden a ser muy breves y otro tipo de búsqueda, más larga donde se utiliza una frase o expresión en texto libre.

El siguiente gráfico representa a la búsqueda: *MOLINOS Y GIGANTES*, con esta búsqueda más corta y utilizando dos palabras clave evitamos el ruido documental, así por lógica la Precisión y Exhaustividad será más alta que si realizamos una ecuación de búsqueda en texto libre. El gráfico muestra claramente cómo el Stemmer *TCP806* y *MVS*, que son menos agresivos en la agrupación en raíces ofrecen mayor grado de Precisión y Exhaustividad y en cambio observamos cómo con el Stemmer *SU* y el *TCP303* muestran menor Precisión y Exhaustividad, inclusive el *TCP303* el cual es un Stemmer de creación propia muy agresivo y forzado a agrupar el 30% de letras de la palabra y que además exige que el número mínimo de letras que se queden sean 3.

⁵⁵ Base de datos ARENA.mdb, Véase Apéndice II.

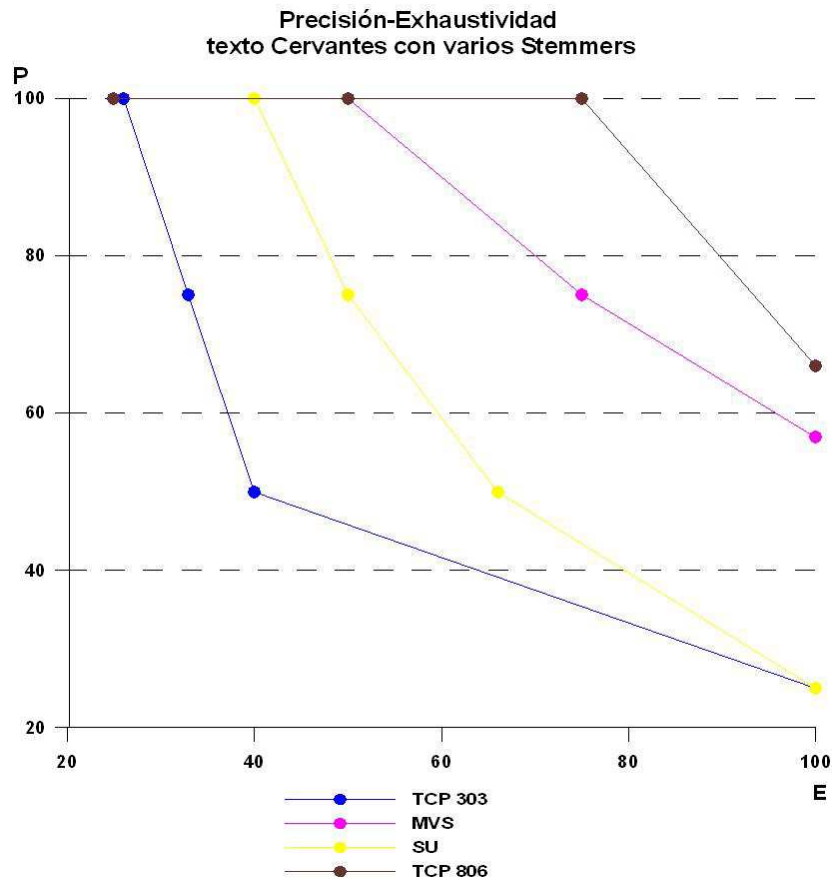


Gráfico 110. Precisión-Exhaustividad texto Cervantes con varios Stemmers

El siguiente gráfico representa a la búsqueda: *LA LUCHA DE DON QUIJOTE CONTRA LOS MOLINOS*, en este caso realizamos una ecuación de búsqueda más larga que la anterior y en texto libre. El gráfico P-E muestra claramente cómo el Stemmer *TCP806* y *MVS* que son menos agresivos en la agrupación en raíces siguen ofreciendo mayor grado de Precisión y Exhaustividad y en cambio observamos cómo con el Stemmer *SU* muestran menor Precisión y Exhaustividad. En este caso podemos observar que el Stemmer *TCP303* se ha eliminado del gráfico debido a que la P-E obtenida es 0%, por tanto no es significativo dibujarlo en la gráfica pero si indica la desfavorable Precisión y Exhaustividad obtenida.

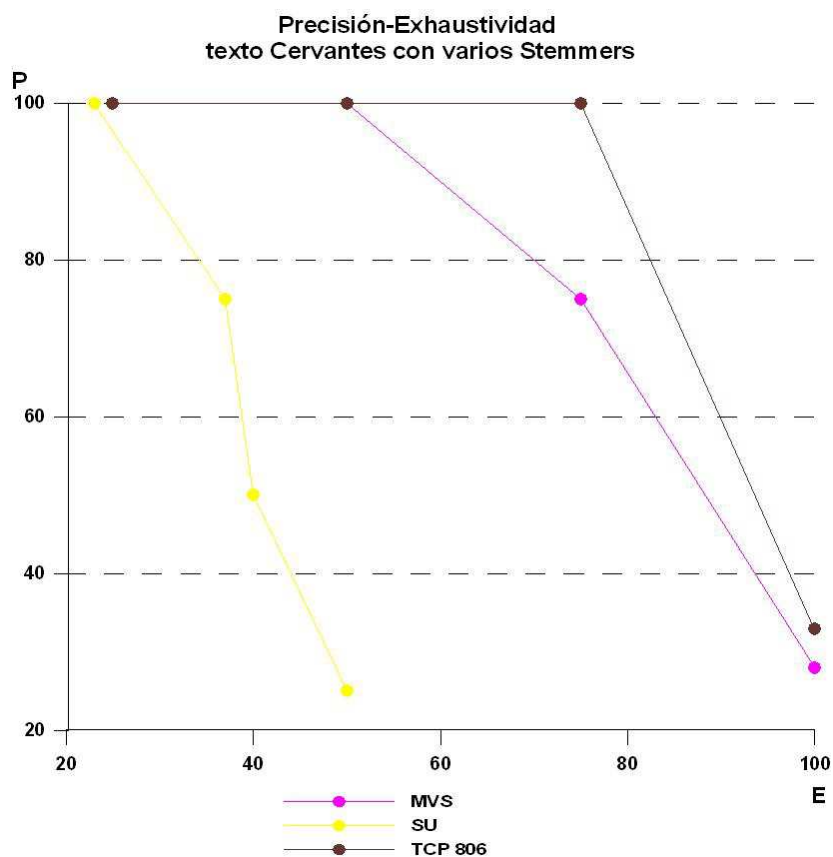


Gráfico 111. Precisión-Exhaustividad texto Cervantes con varios Stemmers

En líneas generales y tras varios ejemplos realizados confirmamos que la Precisión y Exhaustividad en la Recuperación de Información mejora si se utilizan Stemmers poco agresivos, es decir Stemmers que realizan operaciones no muy drásticas en la lematización. Los Stemmers que reducen la raíz de la palabra a tres o cuatro caracteres (lematización más agresiva) reducen el éxito en la Recuperación. Igualmente utilizar en las búsquedas palabras cortas amplía el éxito en la Recuperación de Información frente a la utilización de frases. Podemos por tanto ratificar las conclusiones ofrecidas por la autora Gómez Díaz (2005).

8. Introducción a las palabras asociadas. La relación semántica y conceptual de los términos

En este capítulo se implementa una aplicación informática que construye la colección de palabras asociadas definidas a partir de un texto y cierta granularidad de sus fragmentos y que muestra la cantidad de palabras y las posibilidades de su reducción y de su selección. Se establecen criterios en base a la cantidad de pares de palabras asociadas para seleccionar términos que describen la temática de cada fragmento del texto.

8.1. La similitud entre palabras

La similitud entre palabras es un aspecto trascendental para esta investigación que se aborda ampliamente en este capítulo. Se considera que dos palabras están asociadas cuando éstas aparecen conjuntamente en el mismo documento. Si estas aparecen asociadas en pocos documentos, no tienen relevancia alguna, pero por el contrario si aparecen asociadas sucesivamente en varios documentos si que tienen una especial relevancia.

Dos palabras coocurren cuando aparecen simultáneamente en el mismo documento. Por tanto dos palabras estarán más asociadas entre sí cuanto mayor sea la coocurrencia entre ellas. Este hecho es muy significativo ya que se tendrán en cuenta las parejas de palabras que tengan un grado de cohesión muy alto, entendiéndose por cohesión, la intensidad de las asociaciones entre palabras. Estos casos determinan una mayor relevancia de los documentos que por el contrario sucede cuando dos palabras asociadas coocurren en menor medida en un texto, es decir, en el que el grado de cohesión es menor. Lingüistas como Harris el maestro de Chomsky sostiene que la noción de la aparición conjunta es la clave del análisis de la significación de los textos.

Por este motivo, en este capítulo se estudia la coocurrencia de palabras asociadas o pares de palabras en cada uno de los documentos más pequeños llevados a estudio y además se estudia el grado de cohesión que los pares de palabras tienen en cada documento.

Igualmente para los autores Callon, Courtial, y Herve (1995), el método de las palabras asociadas se basa en el cómputo de las apariciones conjuntas de palabras que definen el índice para los diferentes documentos de un fichero. Cuantas más palabras coocurren con frecuencia en textos diferentes, más se refuerza las relaciones entre estos textos y las conexiones entre dichos temas.

Por tanto, el fundamento metodológico del análisis de las palabras asociadas es la noción de aparición conjunta de palabras clave en los documentos que han sido reunidos para constituir un fichero. Desde un punto de vista metodológico se trata de definir un índice (o varios) para medir la intensidad relativa de estas apariciones conjuntas y para llegar a representaciones simplificadas de las redes a las que dan forma. Según afirman Callon, Courtial, y Herve (1995).

Para medir la intensidad relativa de las apariciones conjuntas de las palabras en los documentos se tiene en cuenta las frecuencias de las dos palabras consideradas y se

utiliza el denominado índice de equivalencia o de asociación, el cual mide la intensidad de la asociación entre dos palabras i y j realizada sobre el conjunto de documentos del fichero. Se obtiene el valor 1 cuando la presencia de i acarrea automáticamente la presencia de j , y viceversa, es decir, cuando las dos palabras están siempre juntas. Por el contrario es igual a 0 cuando la mera presencia de una de las dos palabras excluye la de la otra. Así llamaremos índice de equivalencia (E_{ij}) al coeficiente cuyo valor viene dado por la fórmula siguiente:

$$E_{ij} = \frac{C_{ij}^2}{c_i + c_j}$$

Donde, en un documento grande o fichero F dividido en documentos pequeños n .

C_i = número de apariciones de la palabra clave i en la totalidad de documentos n

C_j = número de apariciones de la palabra clave j en la totalidad de documentos n .

C_{ij} = número de apariciones conjuntas de las palabras i y j en la totalidad de documentos n .

Otros autores Ruiz-Baños y Contreras-Cortés (1998) afirman que C_i son los documentos que contienen la palabra i , y C_j son los documentos que contienen la palabra j . Obviamente no es lo mismo el número de veces que una palabra aparece en la totalidad de los documentos; esto sería su frecuencia, y otra muy distinta son los documentos que contienen la palabra en cuestión. (sería la frecuencia en los documentos en los que aparece dicha palabra).

El cálculo de todos los coeficientes entre todos los pares de palabras posibles genera un número de relaciones importantes, pero sería vano pretender visualizarlo. Por eso se utilizan algoritmos para identificar clusters que reúnan las palabras que están frecuentemente asociadas a otras, es decir, entre las cuales los índices de equivalencia son altos. Un método para construir estos clusters puede consistir en no tomar en consideración más que las relaciones existentes y en ordenar los clusters en función de la fuerza de las asociaciones entre las palabras que los constituyen (Callon, Courtial, y Herve, 1995). Existen algoritmos que seleccionan entre todas las palabras del cluster aquella que es la más central y que se retiene para asignar un nombre al cluster.

Los clusters tienen dos parámetros cuantitativos: la densidad y la centralidad, según Callon, Courtial, y Herve (1995), nos ofrecen nociones de centralidad y densidad en los clusters: *la centralidad* se ocupa en un cluster de la intensidad de sus relaciones con otros clusters. La medición de la centralidad permite ordenar los diferentes clusters procedentes de un fichero por orden de centralidad creciente.

La densidad pretende caracterizar la intensidad de las relaciones que unen las palabras que componen un cluster determinado. Los clusters pueden ser colocados por orden de densidad creciente.

Según Ruiz-Baños y Contreras-Cortés (1998), la centralidad o índice de cohesión externa es la suma de los índices de equivalencia de todos los enlaces externos que posee un tema con otros. Y el concepto de densidad o índice de cohesión interna lo define como la intensidad de las asociaciones internas de un tema y representa el grado de desarrollo que posee.

8.2. Las palabras asociadas

Las palabras asociadas o también conocidas en la literatura como coocurrencias son una parte importante en esta investigación, autores como Evert (2005) se refiere a las palabras asociadas o coocurrencias de dos palabras como *Word pairs*.

Evert (2005) distingue entre coocurrencias relacionales y posicionales. Las coocurrencias posicionales son las palabras que coocurren dentro de una cierta distancia la una de la otra. Esta distancia es denominada como el intervalo colocacional. Una ventaja de las coocurrencias posicionales es que ellas son directamente observables en un corpus o texto. Por otro lado las coocurrencias relacionales están basadas en una interpretación lingüística de un corpus observable. Esto supone que la identificación de coocurrencias relacionales en algunas cantidades de texto sustanciales requiere pre-procesamiento automático lingüístico.

Para construir el Sistema de Indización y Segmentación Automática que se ha desarrollado definiremos un mapa temático del texto, para construir este mapa temático se utiliza principalmente y entre otros un método de análisis del corpus o documento en cuestión mediante la obtención de las palabras asociadas.

Los *términos* que finalmente formarán parte de dicho mapa temático o Sistema de Indización y Segmentación Automática serán por tanto una combinación de raíces obtenidas del texto y coocurrencias que superen tanto el umbral de frecuencia en los documentos como además del umbral de significación. Por tanto la importancia de las palabras asociadas se calculará por el grado de asociación de cada pareja para finalmente escoger las que superen los umbrales determinados según la fórmula de Dice (1945). En base a la asociación de las palabras revelaremos la estructura de las relaciones existentes en el texto que servirán para establecer los clusters.

8.3. Colocaciones (Collocations)

El primero en utilizar el término *colocación* fue Firth (1957), que acuñó este vocablo para designar la frecuencia de aparición de una unidad léxica con respecto a otra, frecuencia en principio estadísticamente demostrable. Según Firth “*collocations are actual words in habitual company*”. El interés de Firth por la colocación fue seguido por sus discípulos que la dotaron de identidad lingüística.

Otros autores como Halliday (1961) define las colocaciones como: *coocurrencias estadísticamente superior a lo esperable*.

Básicamente entendemos como colocaciones la combinación de palabras en la que una palabra exige la presencia de otra para expresar un sentido léxico.

Según la autora Travalia (2006) las colocaciones gramaticales en español son las combinaciones como “consistir en” y “carecer de”, que contienen un verbo más un complemento preposicional de régimen, éstas se han identificado tradicionalmente como colocaciones gramaticales en español, estas combinaciones no se ajustan al concepto básico de la colocación: dos elementos que co-aparecen en el discurso de

forma frecuente, sin presentar una fijación completa, la autora propone nuevos tipos de colocaciones gramaticales en español que sí se adaptan a este concepto.

En las investigaciones sobre colocaciones se da sustancialmente más protagonismo a las colocaciones léxicas que a las gramaticales. Por lo que respecta al español, la mayoría de los trabajos se ocupan preferentemente de las primeras, y se limita a mencionar que existen también colocaciones gramaticales.

Según el autor Koike (2001) discrepa con la distinción entre colocaciones léxicas y gramaticales que establecía Benson (1986) argumentando que la frontera entre las dos clases de colocaciones no es tan nítida.

Las colocaciones léxicas se caracterizan por ser dos unidades léxicas que aparecen juntas en el discurso. La co-aparición frecuente posiblemente sea el rasgo principal de estas unidades, en cambio las colocaciones gramaticales como “consistir en” y “carecer de”, contienen un verbo más un complemento preposicional de régimen.

Según la autora una colocación es la combinación frecuente, pero no obligatoria, de dos elementos. Por lo tanto, las combinaciones que tradicionalmente se han denominado colocaciones gramaticales no se deben considerar como tales, dado que no representan casos de coocurrencia frecuente. “carecer” y “de”, por ejemplo, no son dos elementos que se suelen usar juntos, sino dos elementos que deben co-aparecer de forma obligatoria. En este sentido “carecer de” y otras construcciones similares como “referirse a” se asemejan más a construcciones fijas de la lengua.

La autora opina que son colocaciones gramaticales solo los casos en los que la coaparición del verbo y la preposición no es obligatoria, lo cual implica que el verbo puede utilizarse solo.

Entre los diferentes sentidos que se han otorgado al término *colocación*, Alonso Ramos (1993) destaca los de combinaciones usuales o probables de dos palabras o combinaciones en las que una unidad exige la aparición de la otra. Irsula (1992), define las colocaciones como:

“...combinaciones frecuentes y preferentes de dos o más palabras, que se unen en el seno de una frase para expresar determinados acontecimientos en situaciones comunicativas establecidas...”

Por su parte, Cruse (1986) las considera: *“secuencias de unidades léxicas que co-ocurren habitualmente pero que, no obstante, son completamente transparentes en el sentido de que cada constituyente léxico es también un constituyente semántico”*

Independientemente de la definición que se ofrezca de las colocaciones, casi todos los lingüistas coinciden en que este fenómeno lingüístico se caracteriza sobre todo por contener dos elementos léxicos que co-aparecen con frecuencia.

La autora Travalia (2006) propone una nueva clase de colocaciones que no se caracterizan por la co-aparición frecuente de sus constituyentes, aunque sí reúnen un vínculo léxico entre elementos, transparencia semántica, composicionalidad formal y restricciones combinatorias, esta nueva clase de colocaciones las denomina

colocaciones implícitas y son la base de otras combinaciones frecuentes, sin aparecer ellas mismas de forma habitual en el discurso. Dicho de otro modo las colocaciones implícitas corresponden a la estructura profunda (Chomsky 1957, 1965) de una serie de combinaciones. Mientras que estas combinaciones son frecuentes en el discurso, las colocaciones implícitas que están detrás de ellas no aparecen en el discurso. Según la autora las colocaciones implícitas se contraponen a las explícitas, estas últimas incluyen a su vez, las dos clases de colocaciones tradicionales, las léxicas y las gramaticales. Respecto a este tipo de colocaciones implícitas no las tendremos en cuenta debido a que para nuestro estudio nos basaremos básicamente en las colocaciones léxicas.

8.4. Granularidad

Los documentos textuales o en línea se componen de pedazos de información con contenido textual y semántico que se convierten en objetos informativos que tendrán que ser procesados y estructurados en un corpus para su posterior recuperación. Analizar la naturaleza de dichos objetos informativos y determinar la granularidad, fijará el nivel de descomposición o grado en el que pueden ser divididos los contenidos del documento a tratar.

El nivel mínimo de granularidad de un contenido es aquel grado de descomposición en el que una determinada información sigue manteniendo su significación comunicativa, o el grado en el que físicamente no puede seguir descomponiéndose.

Por ejemplo, habrá casos en los que un mismo texto pueda ser utilizado en contextos diferentes o que por el contrario, se encuentre vinculado de forma persistente a un contexto determinado.

En definitiva, la granularidad en almacenamiento de datos se refiere a la especificidad a la que se define un nivel de detalle en una tabla, es decir si hablamos de una jerarquía la granularidad empieza por la parte más alta de la jerarquía siendo la granularidad mínima, el nivel más bajo.

8.5. Estimación de las parejas de raíces en un texto

Para realizar una estimación de la cantidad de parejas de raíces en un texto o documento tendremos en cuenta que un documento está formado por P que es la cantidad total de palabras, por V el vocabulario o palabras distintas y por R las raíces obtenidas mediante un *Stemmer* cualquiera. Para estimar la cantidad total de parejas de raíces en un texto tenemos que para cada documento caracterizado por un número V correspondiente al número de palabras distintas del texto podría establecerse que el número de combinaciones de parejas es $V \cdot V$.

Si se desea desechar las parejas formadas por repeticiones de la misma palabra, la expresión sería $V \cdot (V - 1)$. Y si además no se quieren considerar como distintas las parejas formadas con las mismas palabras pero en orden inverso, la expresión se ha de dividir entre 2, resultando

$$\frac{V \cdot (V - 1)}{2} \quad [1]$$

Si se parte de la hipótesis de que en un texto las raíces constituyen el 30% del número total de palabras distintas del texto (V) se puede considerar que el número de raíces del texto será $R = 0,3 \cdot V$.

Aplicando la expresión [1] obtenida anteriormente, el número de parejas de raíces se puede determinar con la expresión:

$$\frac{(0,3 \cdot V) \cdot (0,3 \cdot V - 1)}{2} \quad [2]$$

Despreciando el factor no cuadrático ya que los textos tienen una cantidad de palabras considerable (V es un número elevado) se podría decir de forma aproximada que el número de parejas de raíces es:

$$Rp = 0,045 \cdot V^2 \quad [3]$$

Por tanto se puede concluir que el número de parejas de raíces (Rp) es proporcional al cuadrado de V , siendo el coeficiente de proporcionalidad un número muy pequeño, inferior a 1 (del orden de 0,045).

Para tener una estimación verosímil de lo que puede significar esta cantidad en relación al tamaño de un texto expresado como número total de palabras, aplicamos la ley de Heaps y su función potencial.

Unos valores típicos para la ley de Heaps puede ser 25 para el coeficiente y 0,6 para el exponente, lo que llevado a la anterior fórmula [3]:

$$0,04 \cdot (25 \cdot P^{0,6}) \cdot (25 \cdot P^{1,2}) = 25 \cdot P^{1,2}$$

Esto equivale a una cantidad considerable de parejas de palabras; véanse algunos casos en la siguiente tabla

<i>Palabras totales en el texto</i>	<i>Rp teóricas (Parejas de raíces)</i>
10.000	Aprox. 1,5 millones
20.000	Aprox. 3.6 millones
30.000	Aprox. 5,8 millones
40.000	Aprox. 8,3 millones

Tabla 31. Cantidad de parejas de raíces teóricas en los documentos

8.5.1. Contando parejas de raíces asociadas

Hasta aquí, hemos enunciado un problema de simple combinatoria; vamos a incluir ahora los hechos derivados de la forma de distribución de las palabras en los textos.

Lo que queremos contar no es el total de parejas teóricamente posibles, sino de las que aparecen, las que se pueden calificar como asociadas. Para que merezca llamarse asociada una pareja de palabras o raíces debe aparecer en el texto con cierta proximidad entre ellas, más veces que las que las leyes de la estadística sugieren, para el caso de que ambas palabras se comporten como independientes.

Para medir la proximidad entre la aparición en el texto de dos palabras o raíces utilizamos su pertenencia conjunta a un documento. De este modo introducimos una nueva variable que es la cantidad de documentos en que hemos dividido el texto, o quizá mejor el tamaño de estos documentos. Si los documentos son pequeños, resulta más difícil que dos palabras pertenezcan conjuntamente a uno de ellos. Un mismo texto puede estar dividido en unos pocos documentos grandes o en muchos documentos pequeños. Podemos llamar a esta propiedad y de hecho se le conoce como la granularidad de la división del texto en documentos.

Para una revisión inicial vamos a fijar arbitrariamente el tamaño de cada uno de los documentos en que dividimos un texto grande: lo establecemos en 300 caracteres (esta cantidad es aproximada, ya que siempre forzamos a que el final de un documento coincida con un punto y aparte). Al dividir un texto grande en documentos más pequeños de $\cong 300$ caracteres, estamos considerando documentos de unas 70 palabras, en los textos literarios, si bien en los textos científicos o legales son menos de 70 palabras, debido a que normalmente dichos textos tienen palabras más largas que los literarios.

Como hemos mencionado anteriormente lo que queremos contar no es el total de parejas teóricamente posibles, sino de las que aparecen, las que se pueden calificar como asociadas, y para ello vamos a considerar su pertenencia conjunta a un documento, por tanto en un primer análisis no se va a establecer ningún criterio estadístico o semántico de momento, bastará que al menos dos raíces aparezcan conjuntamente en un documento para que las seleccionemos como pareja. Las llamaremos *raíces pre-asociadas*.

Cuando decimos que vamos a contarlas es porque ahora se trata de un trabajo experimental no de una deducción teórica. Los valores que obtengamos dependerán de cómo los autores han desarrollado sus pautas de agrupar palabras.

En cinco ejemplos de textos variados obtenemos los siguientes datos:

<i>Texto</i>	<i>Palabras totales</i>	<i>Raíces</i>	<i>Promedio palabras en un doc.</i>	<i>Rp teóricas (Parejas de raíces)</i>	<i>Raíces pre-asociadas</i>
Azorín	130.370	10.874	71	58,3 M	2,4 M
Castela	407.290	16.113	73	129,6 M	8,1 M
Cienti1	100.753	11.334	55	63,8 M	1,0 M
Legal1	94.808	6.039	35	18,0 M	0,7 M
Quijote	158.679	9.312	65	43,2 M	2,6 M

Tabla 32. Cantidad de parejas de raíces teóricas y pre-asociadas en varios documentos

La conclusión es que no todas las parejas posibles llegan a tener una existencia real por que nunca coinciden ambas palabras en un mismo documento.

Por tanto, podemos afirmar que el número teórico de parejas posibles debe dividirse por 20 o más, para estimar el número de raíces pre-asociadas.

8.5.2. La granularidad del texto y las parejas de raíces asociadas encontradas.

Si los documentos son pequeños resulta más difícil que dos palabras pertenezcan conjuntamente a uno de ellos. Un mismo texto puede estar dividido en unos pocos documentos grandes o en muchos documentos pequeños. Podemos llamar a esta propiedad la granularidad de la división del texto en documentos.

Se presentan a continuación unas gráficas que muestran el número de parejas de raíces pre-asociadas dependiendo de la granularidad. Se observa que en todos los casos crece de acuerdo a una ley potencial de exponente cercano a 0,6.

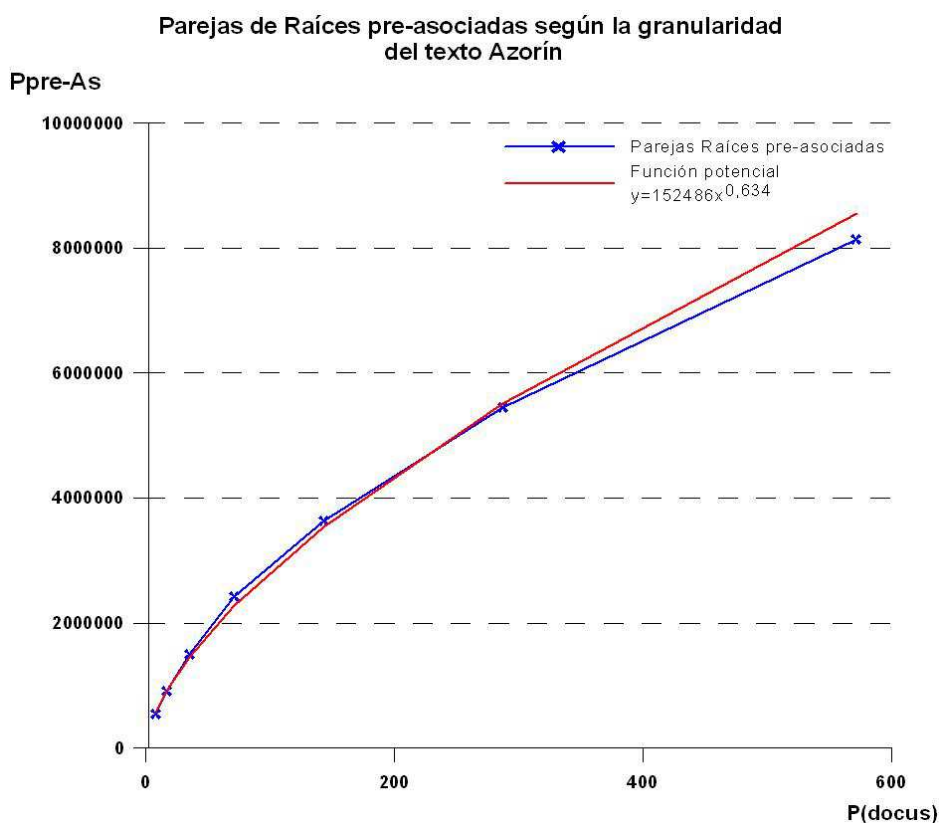


Gráfico 112. Parejas de raíces pre-asociadas según granularidad texto Azorín

Se puede concluir que la relación de la granularidad y el número de parejas de raíces pre-asociadas sigue una función potencial con un exponente aproximadamente entre 0,5.

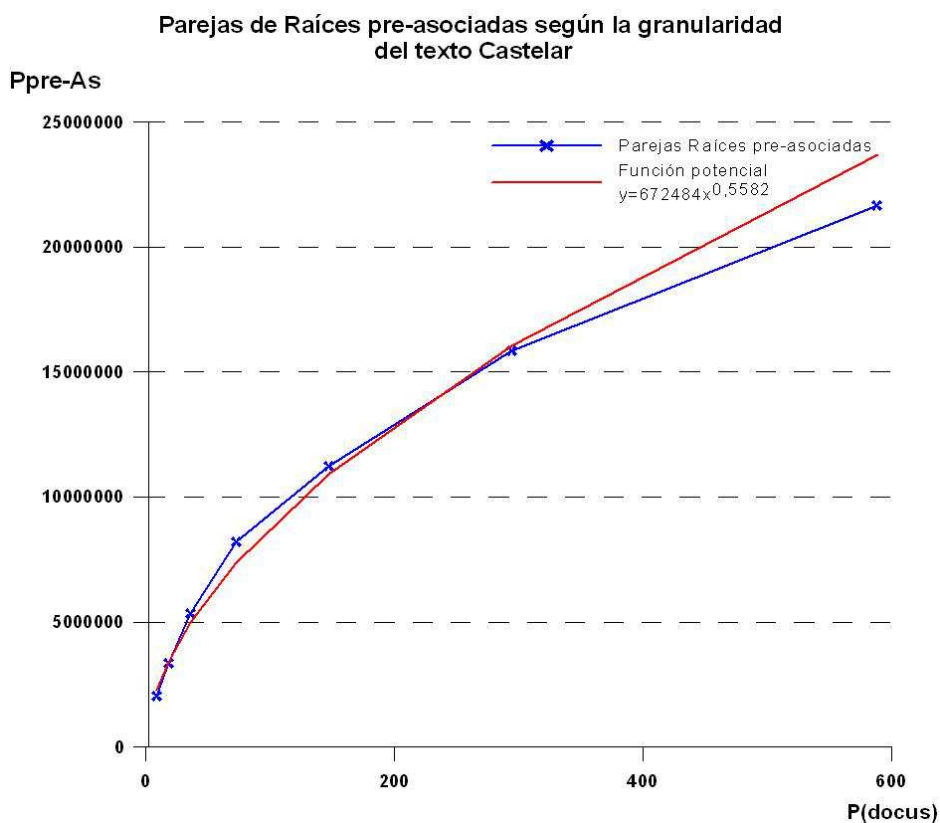


Gráfico 113. Parejas de raíces pre-asociadas según granularidad texto Castelar

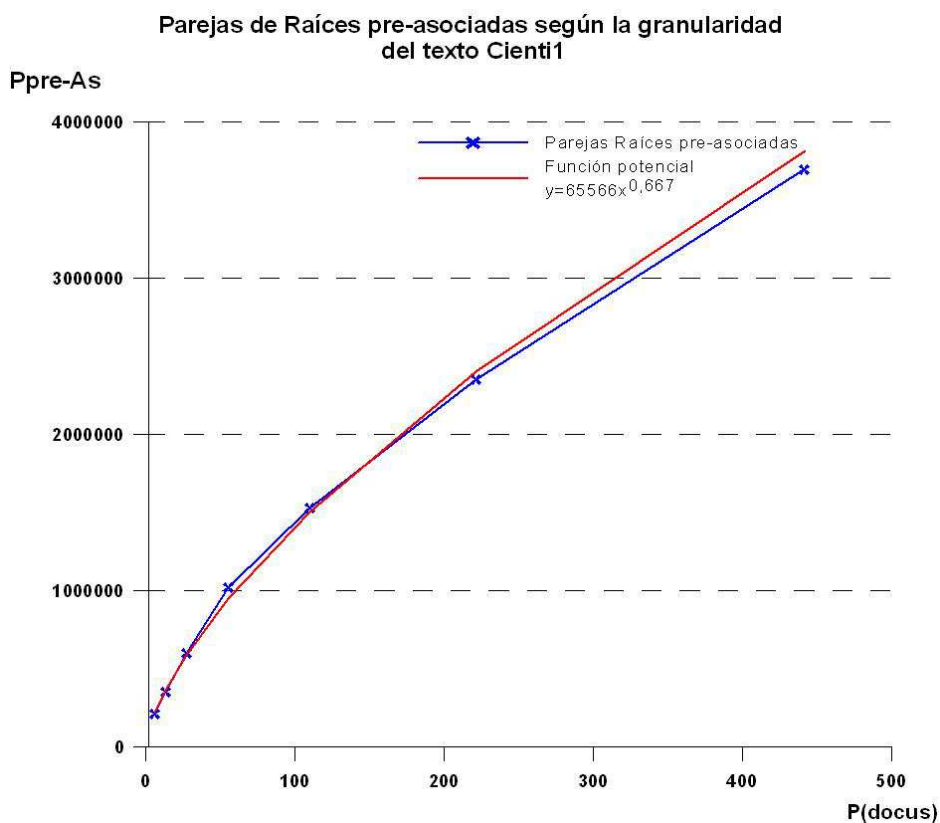


Gráfico 114. Parejas de raíces pre-asociadas según granularidad texto Cient1

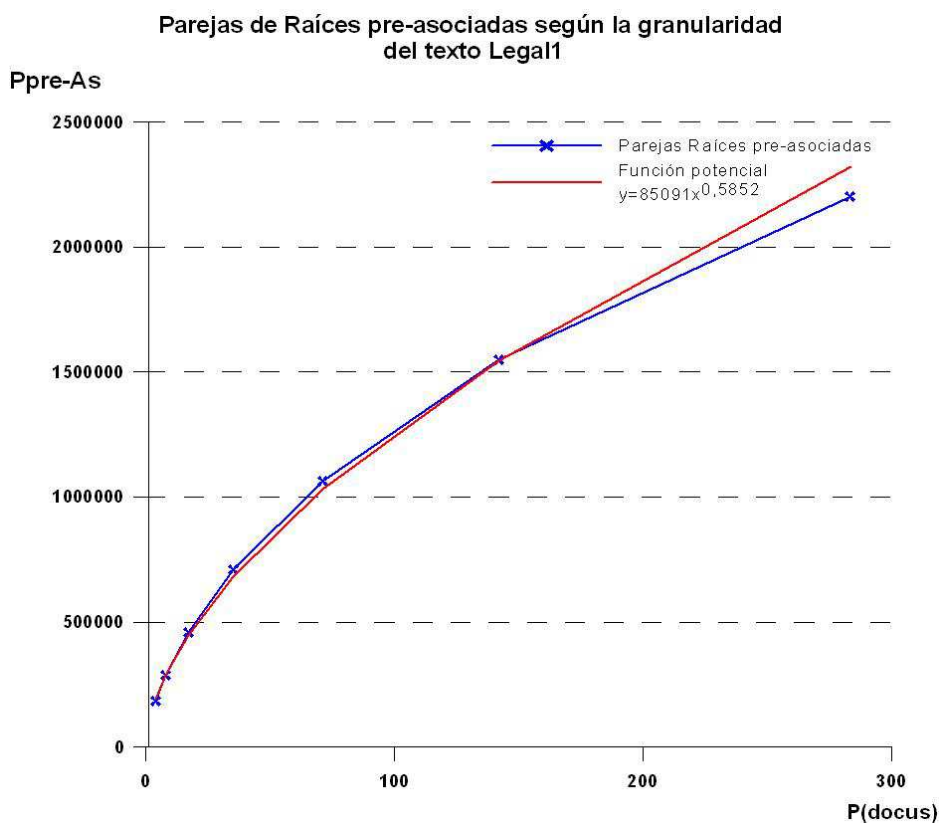


Gráfico 115. Parejas de raíces pre-asociadas según granularidad texto Legal1

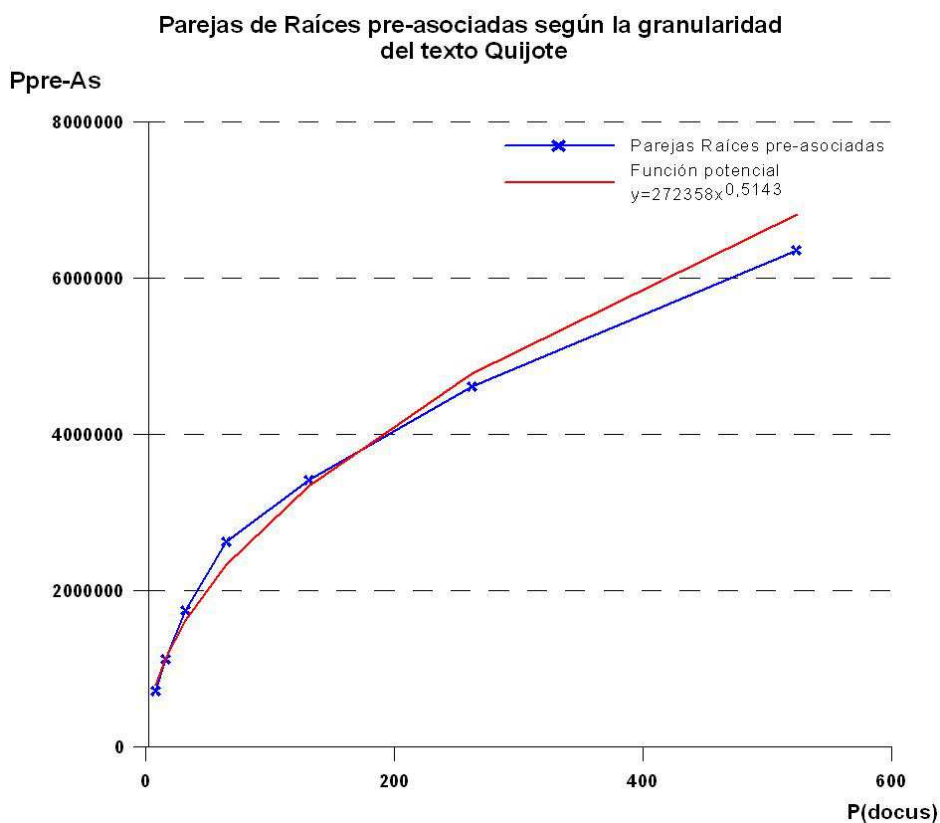


Gráfico 116. Parejas de raíces pre-asociadas según granularidad texto Quijote

Los gráficos muestran en un primer tramo que cuando los documentos son más pequeños hay menos raíces pre-asociadas y en un segundo tramo que aumenta el número de parejas con documentos más grandes.

Obviamente las gráficas muestran cómo la granularidad de la división del texto en documentos es determinante para obtener más o menos cantidad de parejas de raíces. Por tanto si decidimos restringir el sentido de las raíces pre-asociadas volviendo a dividir el texto en documentos más pequeños, por ejemplo a la mitad, el número de parejas que podemos esperar es la siguiente fórmula potencial:

$$PA_{mitad} = K \cdot (0,5 \cdot TD_{antes})^{0,6} = K \cdot 0,5^{0,6} TD_{antes}^{0,6} = 0,5^{0,6} \cdot (K \cdot TD_{antes}^{0,6}) = 0,66 \cdot PA_{antes}$$

Por tanto, cambiando el tamaño de los documentos a la mitad, el número de raíces pre-asociadas se reduce a dos tercios de las que había.

Es importante aclarar que las gráficas anteriores y lo mencionado anteriormente solo se refiere a documentos suficientemente pequeños. Si el tamaño de los documentos es muy grande, no representa ninguna información el que dos palabras aparezcan en un documento. Entonces deja de tener sentido el concepto de *pre-asociadas* y simplemente estaríamos estimando la cantidad total de parejas de raíces en un texto, acercándose la cantidad a la formulación [1]:

$$\frac{R \cdot (R - 1)}{2}$$

8.5.3. Selección de parejas de raíces asociadas que aportan información

A partir de ahora vamos a fijar arbitrariamente el tamaño de cada uno de los documentos en que dividimos un texto grande, y lo establecemos a $\cong 300$ caracteres (esta cantidad es aproximada, ya que siempre forzamos a que el final de un documento coincida con un punto y aparte). Al dividir un texto grande en documentos más pequeños de $\cong 300$ caracteres estamos considerando documentos de unas 70 palabras, en los textos literarios, si bien como hemos mencionado anteriormente en los textos científicos o legales son menos de 70 palabras.

Hasta ahora no se había establecido ningún criterio estadístico o semántico pero a partir de este momento se exigen dos condiciones de tipo estadístico a las parejas pre-asociadas para considerarlas asociadas:

En primer lugar, que su coincidencia no sea fruto del azar. Las palabras muy frecuentes, es posible que aparezcan ambas en algún documento debido a su misma abundancia y no a que tengan relación una con otra.

Respecto a las palabras asociadas con el criterio de superar la esperanza estadística, si el texto se ha dividido en D documentos y la palabra aa aparece en fa documentos, y la palabra bb en fb documentos, dado un documento específico x , la probabilidad de que contenga aa es $\frac{fa}{D}$ y de que contenga bb $\frac{fb}{D}$. La probabilidad de que contenga

simultáneamente aa y bb es la probabilidad compuesta o producto de las probabilidades, que será $\frac{fa \cdot fb}{D \cdot D}$. Por tanto, el número esperado de documentos donde aparecen ambas palabras:

$$f_{ab} = D \cdot \frac{fa \cdot fb}{D \cdot D} = \frac{fa \cdot fb}{D}$$

Diremos que un par de palabras superan el criterio estadístico cuando aparecen conjuntamente en más de estos documentos.

Por tanto consideraremos las raíces asociadas estadísticamente que coincidan en más documentos de los que sería verosímil estadísticamente, es decir las que superan el criterio estadístico.

En segundo lugar sólo tomaremos las parejas que coincidan como mínimo en dos documentos (criterio de repetición). Nuestro objetivo final es agrupar documentos de tema parecido; una pareja que aparezca en un solo documento puede estar muy relacionada, pero es inútil para nuestro propósito.

Consideraremos que una pareja de raíces es asociada si:

- pre-asociada
- supera el criterio estadístico
- criterio de repetición

Completando una tabla ya presentada anteriormente:

Texto	Palabras totales	Raíces	Promedio palabras en un doc.	Rp teóricas (Parejas de raíces)	Raíces pre-asociadas	Raíces asociadas Crit. Est. Crit. Rep.
Azorín	130.370	10.874	71	58,3 M	2,4 M	0,7 M
Castela	407.290	16.113	73	129,6 M	8,1 M	3,2 M
Cient1	100.753	11.334	55	63,8 M	1,0 M	0,2 M
Legal1	94.808	6.039	35	18,0 M	0,7 M	0,2 M
Quijote	158.679	9.312	65	43,2 M	2,6 M	0,8 M

Tabla 33. Cantidad de parejas de raíces teóricas, pre-asociadas y asociadas con criterio estadístico y de repetición en varios documentos

Cuando aplicamos el criterio estadístico y el criterio de repetición observamos grandes diferencias entre unos autores y otros. Aquí no encontraremos una gran uniformidad ante los textos estamos considerando un hecho que depende de las características del texto, de la forma de escribir del autor. Para tener una guía aproximada podemos enunciar lo siguiente:

El número de parejas de raíces asociadas es dos órdenes de magnitud menor que el número teórico de parejas posibles. (Análogo a dividir por 100). Recordemos que esta observación se refiere a documentos de $\cong 300$ caracteres. Si no fuera así, deberíamos

considerar también que cambiando el tamaño de los documentos a la mitad, el número de raíces pre-asociadas se reduce a dos tercios de las que había.

8.6. Relación del número de parejas de raíces asociadas y el tamaño del texto

A partir de este momento para tener menos palabras descartaremos a las pre-asociadas y trataremos de determinar cual es la tasa de crecimiento de las palabras que sí nos interesan para el estudio, las palabras asociadas. Es obvio que al crecer el texto crecerá el número de raíces asociadas.

Como se menciona en páginas anteriores, unos valores típicos para la Ley de Heaps puede ser la fórmula $25 \cdot P^{1,2}$ como expresión de cuántas parejas hay en función del tamaño del texto o número total de palabras P.

Pero esta fórmula ha sido deducida observando sólo un tamaño en cada texto y aunque su deducción fuera totalmente correcta, la inestabilidad de la fórmula potencial tantas veces mencionada hace que otras expresiones puedan ser también aproximadamente correctas.

Como ilustración tenemos la gráfica de tres funciones potenciales con diferentes parámetros y en ella observamos como a pesar de tener diferentes parámetros las funciones potenciales obtenidas son similares:

$$PA = 0,2 \cdot P^{1,2}$$

$$PA = 8,5 \cdot P^{0,9}$$

$$PA = 95 \cdot P^{0,7}$$

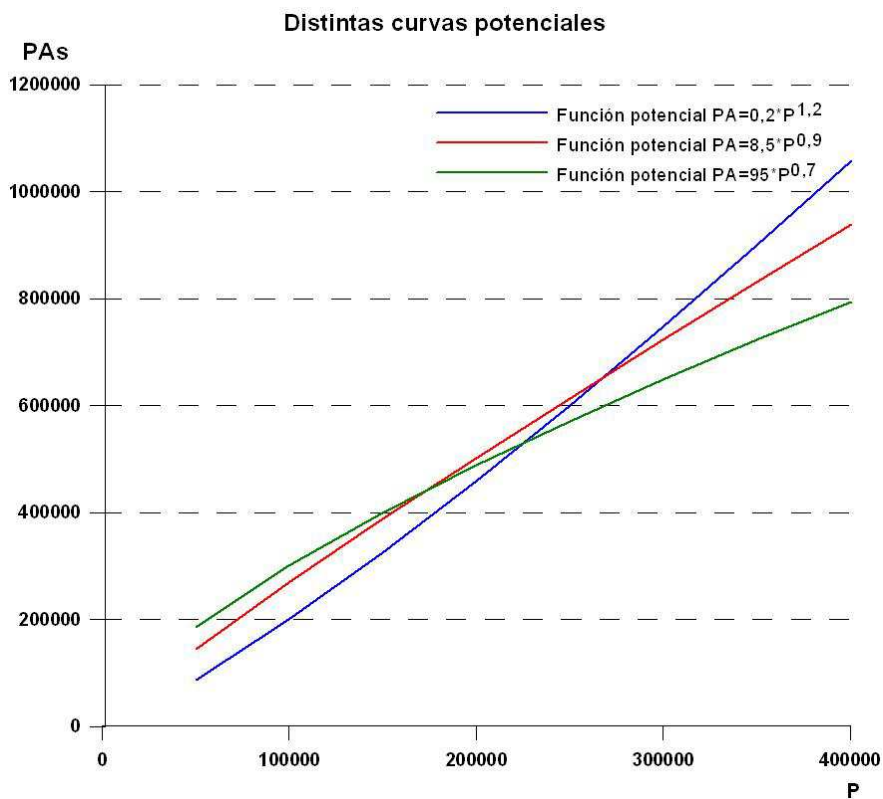


Gráfico 117. Funciones potenciales con diferentes parámetros

Como decíamos en capítulos anteriores, concretamente en el capítulo cinco de esta tesis doctoral, la función potencial es muy inestable y dicha inestabilidad de los coeficientes resultan parecidos, pero no del todo. Los valores de los coeficientes a pesar de ser distintos pueden estar generando funciones que son parecidas entre sí.

Se hace necesaria una confrontación con datos de textos reales y distintos tamaños de textos de similares características para decidir cuál de estas formas de crecimiento es la verdadera. Especialmente hay que determinar el valor adecuado del exponente.

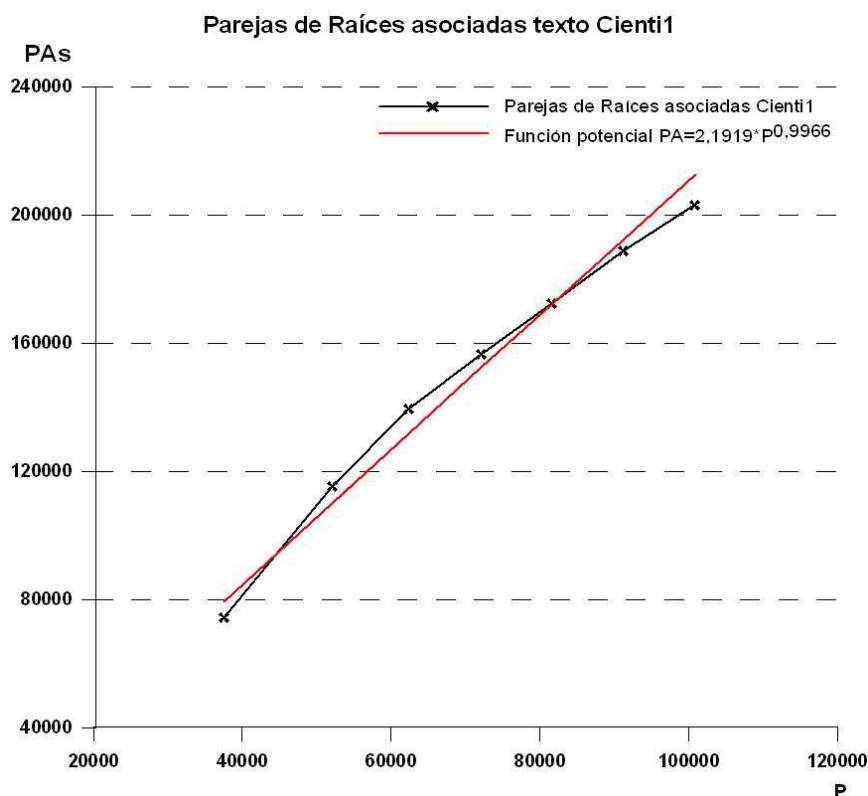


Gráfico 118. Parejas de raíces asociadas y función potencial obtenida texto Cient1

El texto legal que se representa en la gráfica recordemos que está formado por una yuxtaposición sucesiva de textos distintos, esto supone una variedad de vocabulario bastante alta lo que supone que el valor del exponente en la fórmula potencial aumente hasta un valor de 1,02, como puede apreciarse en el siguiente gráfico.

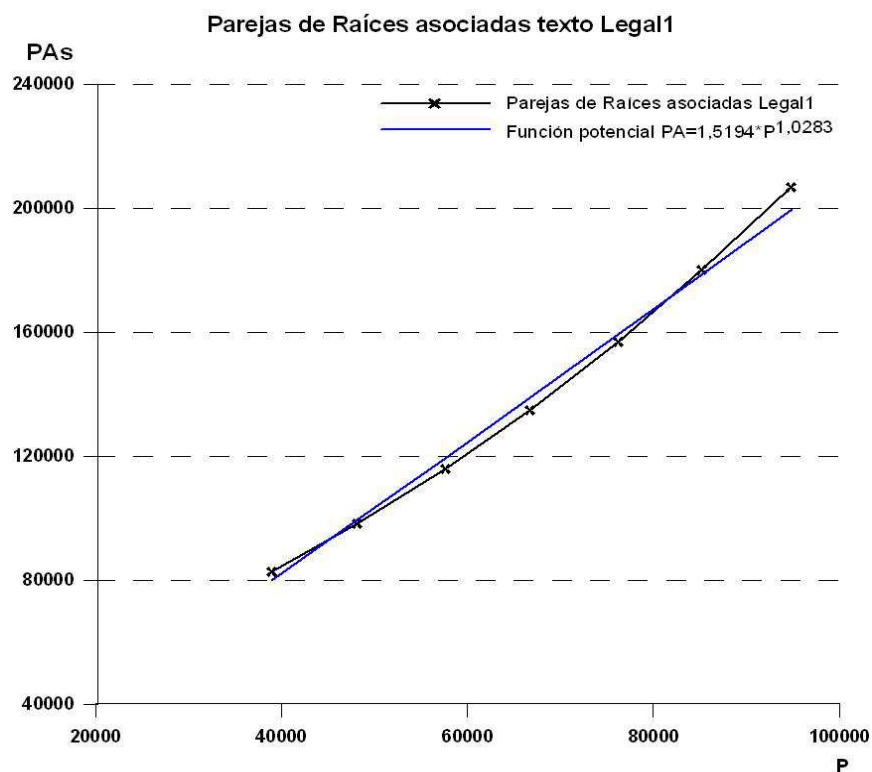


Gráfico 119. Parejas de raíces asociadas y función potencial obtenida texto Legal1

Observamos cómo la tendencia del exponente en la fórmula potencial es diferente según los textos.

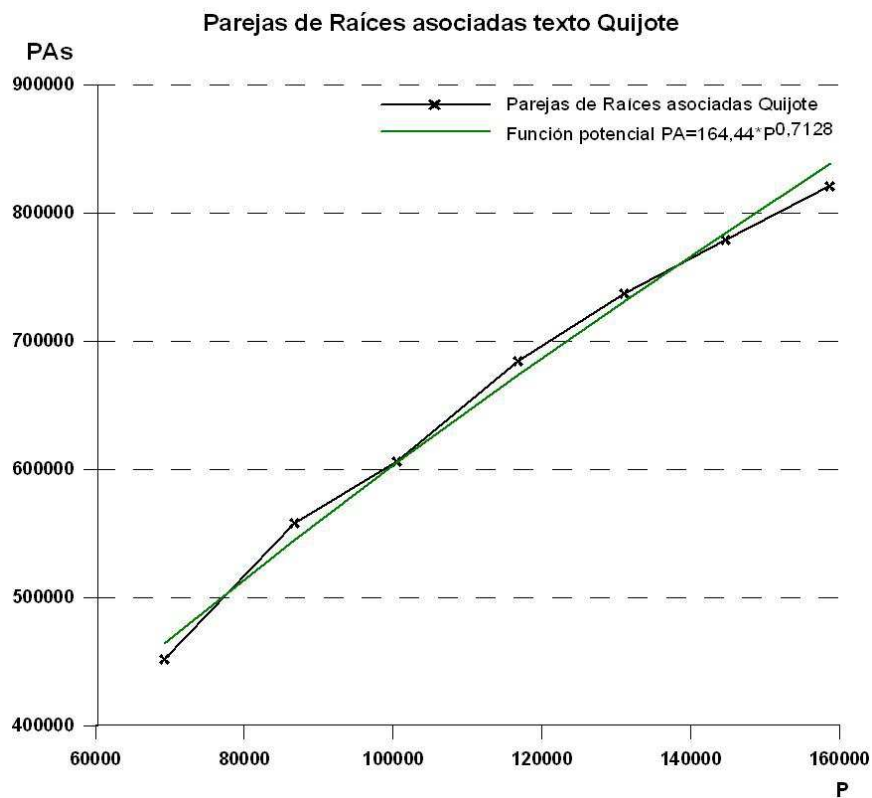


Gráfico 120. Parejas de raíces asociadas y función potencial obtenida texto Quijote

Vemos aquí reflejada la disparidad de situaciones en los textos: Científicos y legales están formados por la sucesión de pequeños fragmentos que pueden tratar asuntos distintos y por eso hacen crecer el exponente a un valor mayor que 1, el número de raíces asociadas. Por otra parte, las aventuras del Quijote insiste en los mismos temas y por eso presenta un exponente menor.

Sin embargo, la situación no es tan distinta de unos a otros ya que se ve compensada por el mayor coeficiente en el texto del Quijote. Podemos dar una fórmula de compromiso entre todas ellas muy imperfecta, solo para estimar el orden de magnitud:

La cantidad de raíces asociadas es aproximadamente igual a $10 \cdot P^{0,9}$ (10 veces el total de palabras elevado a 0,9) con esta fórmula podemos obtener las palabras asociadas de un texto general, pero eso sí, teniendo en cuenta que para una estimación más precisa, si el texto aborda insistentemente el mismo asunto recomendamos bajar el exponente y aumentar el coeficiente; mientras que si se trata de yuxtaposición sucesiva de asuntos distintos proceder a la inversa.

En definitiva, podemos enunciar las siguientes reglas:

Si consideramos palabras asociadas a las obtenidas por recuentos puramente combinatorios o de sentido estadístico:

Su número crece con el tamaño del texto siguiendo una ley potencial, en resumidas cuentas, podemos afirmar que el número de palabras asociadas con criterios combinatorios o estadísticos sigue una ley potencial con exponente mayor que el de Heaps. Esta última conclusión es razonable si atendemos a que el número de parejas tiene alguna relación con el cuadrado del número de raíces.

En conclusión, los exponentes de una fórmula potencial que mide la cantidad de parejas de palabras en relación al tamaño del texto se rigen por una regla básica: El exponente no puede ser menor que el de Heaps, y el número de parejas debe crecer como mínimo tan rápido como crece el vocabulario y como máximo según el cuadrado del vocabulario $2e$ (doble del de Heaps).

8.7. El espectro de similitudes

En relación a la similitud y la frecuencia de las parejas de raíces, las parejas de raíces asociadas presentan distintos grados de similitud entre sus componentes. Utilizando una de las fórmulas para medirla: concretamente la *fórmula de Dice*, (Dice, 1945) que es una de las más sencillas sabemos que su valor es un número decimal entre 0 y 1. Podemos sospechar que la mayoría de las parejas tendrán similitud pequeña, próxima a 0, y sólo unas pocas tendrán similitud alta cercana a 1.

Tratamos de establecer cuantitativamente la distribución de los valores de la similitud. Como era de esperar, la mayoría de las parejas tienen similitud muy baja. La siguiente gráfica que corresponde al texto de Azorín muestra la cantidad de parejas encontradas en cada uno de los tramos de similitud 0 a 0,1, 0,1 a 0,2 etc.

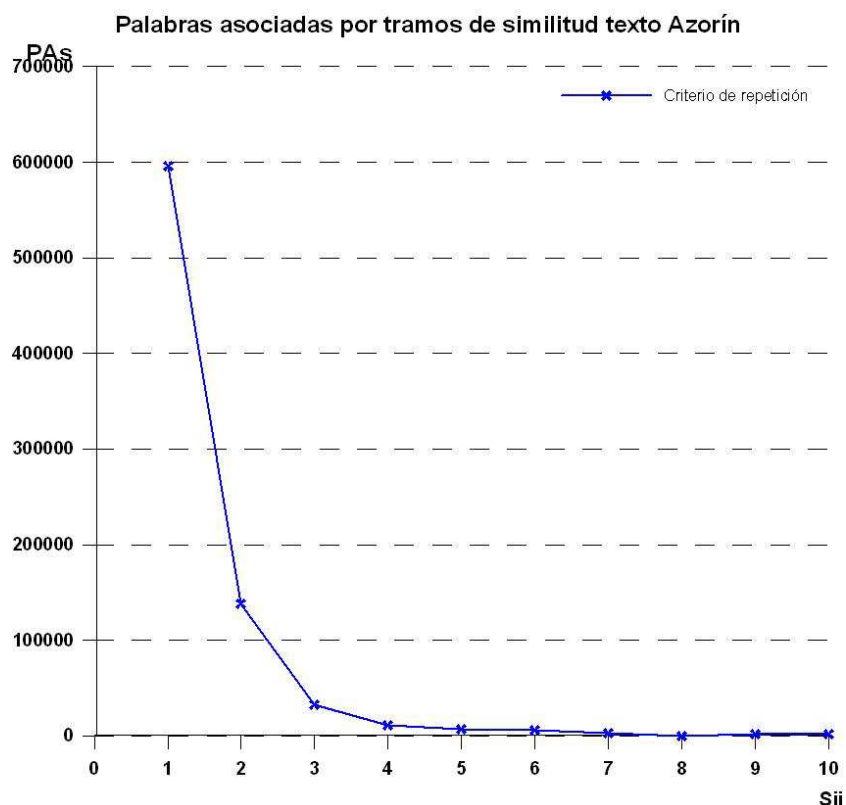


Gráfico 121. Espectro de similitudes texto Azorín

De igual modo, en el siguiente gráfico vemos el espectro de similitudes, en el eje de abscisas se detalla los diferentes tramos de similitud desde el primer tramo de similitud de 0-0,1 hasta el último tramo de similitud de 0,9-1, en el eje de ordenadas la cantidad de palabras asociadas, los tres casos corresponden al texto *episo1.txt* de $P = 81.398$ palabras y $R = 7.561$ raíces, la línea coloreada en azul corresponde a la similitud de las pre-asociadas, la línea coloreada en rosa corresponde a la similitud de las parejas que superan el criterio estadístico y la línea coloreada en amarillo corresponde a la similitud de las parejas que muestran un criterio de repetición, se observa como en los dos primeros casos la cantidad de parejas con similitudes muy bajas es mucho mayor que en caso del criterio de repetición.

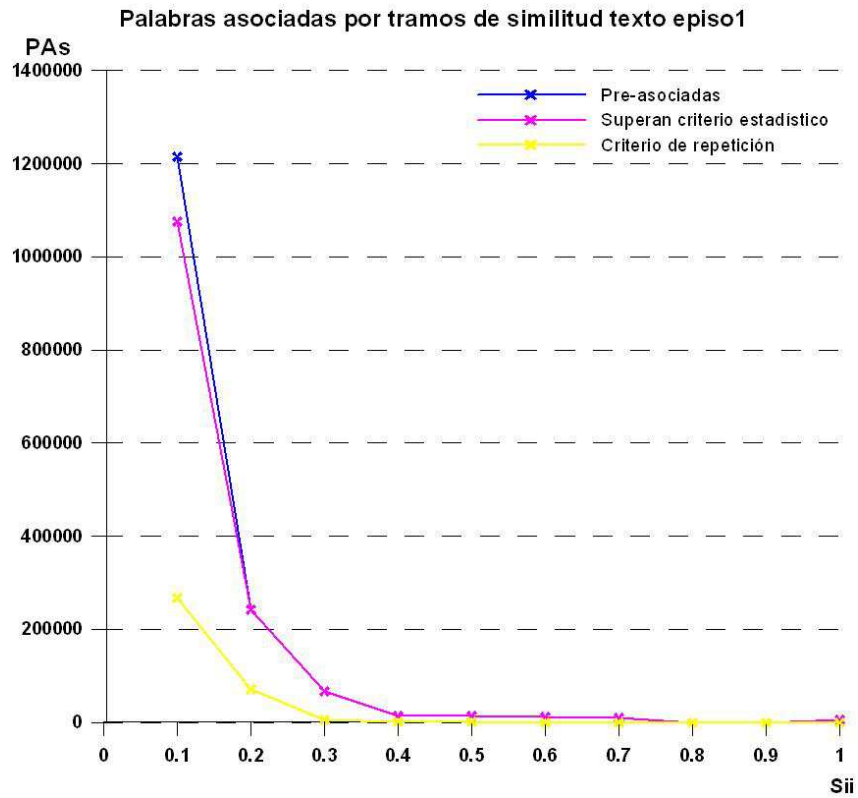


Gráfico 122. Espectro de similitudes texto episo1

Como la distribución está fuertemente concentrada en los primeros tramos pasamos a una representación logarítmica, en este caso para el texto de Cervantes y concretamente su obra el Quijote

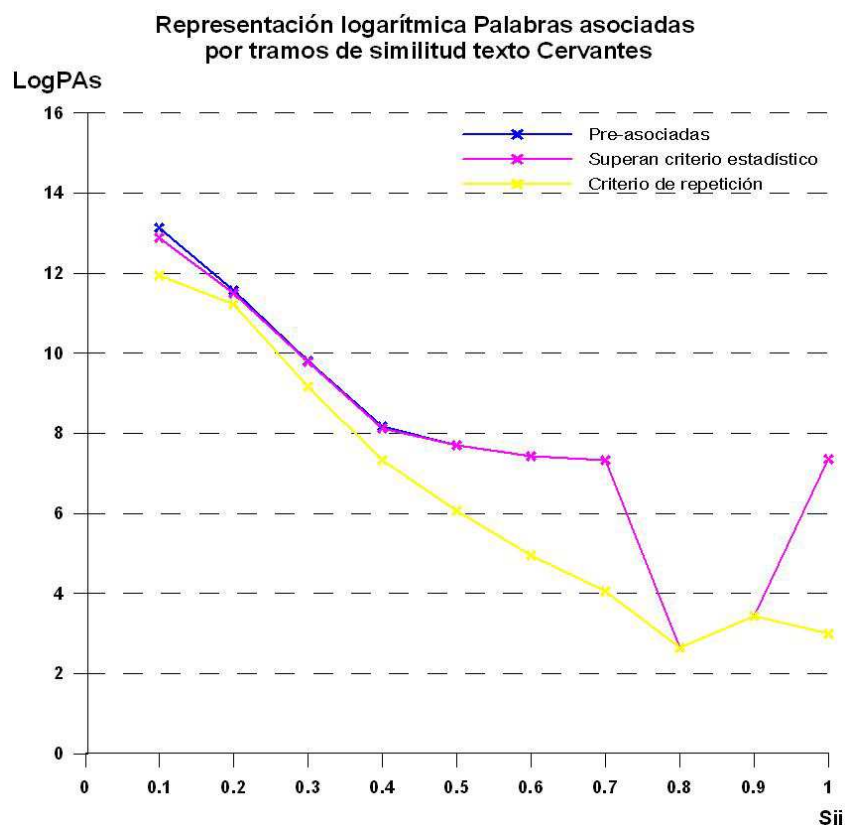


Gráfico 123. Representación logarítmica del espectro de similitudes texto Cervantes

Para mejorar la visibilidad, usamos escala logarítmica para el número de parejas

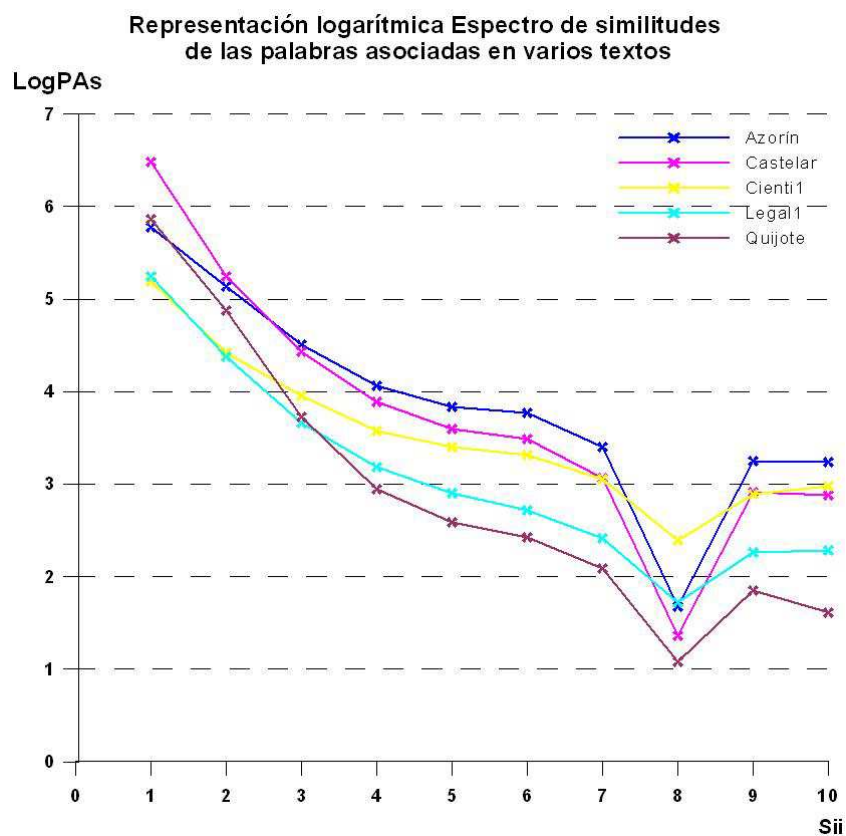


Gráfico 124. Representación logarítmica del espectro de similitudes varios textos

Resulta notable la semejanza de las cinco gráficas a pesar de corresponder a textos cualitativa y cuantitativamente diferentes.

A la vista de ellas establecemos las siguientes conclusiones:

Las parejas de raíces asociadas conforman dos poblaciones distintas: una que se concentra en los valores inferiores de la similitud y va teniendo menos abundancia conforme pasamos a valores más altos; otra población que corresponde a parejas con valores muy próximos a 1. Estas parejas de palabras con similitud cercana a 1, son palabras que suelen ir unidas y que son equivalentes a una sola palabra, son las llamadas “colocaciones” o también conocidas como “collocations” (Evert, 2005), según el autor las colocaciones son palabras clave con similitud superior a lo que resultaría si fueran independientes. Firth (1957) acuñó este vocablo para designar la frecuencia de aparición de una unidad léxica con respecto a otra, frecuencia en principio estadísticamente demostrable. Otros autores como Halliday (1961) define las colocaciones como: “coocurrencias estadísticamente superior a lo esperable”.

Como nota aclaratoria de carácter técnico, en los recuentos efectuados para obtener el espectro de similitudes se ha observado que la aplicación del criterio estadístico hace descartar parejas de raíces exclusivamente de muy baja similitud. Dado que finalmente todas serán descartadas por su poco interés y considerando que el coste computacional de aplicar este criterio es elevado, en lo sucesivo prescindiremos de su aplicación sin que ello influya en los resultados.

Otra conclusión es que el criterio de repetición afecta por igual a todos los tramos de similitud. Su utilización es razonable respecto a nuestro objetivo ya que si queremos obtener grupos de documentos de temática similar, ¿que valor tiene una información que afecte solo a un documento? Esto nos lleva a insistir en criterios similares a este para reducir el número de palabras asociadas a considerar.

En el proceso siguiente que vamos a desarrollar trataremos de identificar previamente las colocaciones para tratarlas como si fueran una sola palabra. Utilizaremos el nombre de “*términos*” para referirnos al conjunto resultante formado tanto por raíces como por di-palabras o di-gramas, resultantes de las colocaciones.

9. Segmentación automática del texto. Identificación de cambios temáticos

En este capítulo se implementa un método de segmentación automática del texto que está basado únicamente en procedimientos cuantitativos y que utiliza las fórmulas obtenidas anteriormente sobre crecimiento del vocabulario para dividir el texto en fragmentos y determinar una estructura en niveles y subniveles de forma jerárquica.

9.1. Segmentación Automática del Texto (Text Segmentation)

Para el procesamiento de textos al igual que en la elaboración de resúmenes automáticos, se utiliza una técnica conocida como Segmentación de textos para la elaboración de resúmenes automáticamente y la elaboración de clasificaciones, de este modo en el caso de una búsqueda obtener los segmentos o pasajes más relacionados con la consulta de un usuario en lugar del documento completo. La tarea de detectar automáticamente los cambios temáticos que se producen en un documento es una tarea difícil pero útil.

A esta técnica se le conoce en la literatura como Segmentación de textos (*Text Segmentation*) o Segmentación por tópicos.

Hasta ahora en nuestras investigaciones hemos abordado los textos desde un enfoque sincrónico hemos estudiado diferentes características de éstos escogiendo para ello partes aleatorias y no sucesivas de los textos o documentos. Pero a partir de este momento y para llevar a cabo nuestro objetivo de la Segmentación del texto abordaremos los documentos desde un enfoque diacrónico. Estudiaremos el texto completo como un todo constituido de partes disjuntas y correlativas que van desde el comienzo del texto, cuando el autor comienza a escribir hasta el final del mismo.

Una vez aclarado esto, el proceso a seguir para obtener la Segmentación del texto será dividir el documento o texto completo en tramos disjuntos y correlativos de los cuales cada tramo corresponderá a una temática distinta y dichos tramos o documentos se agruparán en ventanas deslizantes que formarán grupos de documentos.

Con frecuencia, un documento contiene varios temas éstos se definen como fragmentos de texto formados por palabras, oraciones o párrafos.

El proceso automático para identificar en un texto los temas o tópicos que lo forman, es lo que se conoce como Segmentación del texto (*Text Segmentation*) y en las investigaciones realizadas sobre la Segmentación se emplea la cohesión léxica.

La cohesión léxica es un parámetro cuantitativo que pretende caracterizar la intensidad de las relaciones que unen las palabras que componen un cluster determinado. Los clusters pueden ser colocados por orden de densidad creciente. Según Halliday y Hasan (1976), la cohesión léxica es una propiedad semántica del discurso referida a las relaciones de sentido que existe entre las unidades textuales en el texto.

Entre los elementos que indican relación de sentido se distinguen la repetición o reiteración léxica, la paráfrasis, la elipsis y otras.

La cohesión léxica es un elemento muy útil y eficaz para detectar los cambios temáticos en un texto, porque las unidades textuales se agrupan en clusters calculando la cohesión, la cual es el promedio de las similitudes entre todas las parejas de palabras.

Existen métodos desarrollados para la Segmentación de textos que se basan en la cohesión léxica como el propuesto por Hearst (1997) y su algoritmo denominado TextTiling. Este algoritmo divide textos explicativos en unidades de discurso de múltiples párrafos. Al contrario que muchos modelos de discurso que asumen una Segmentación jerárquica de éste, el autor determinó representar el texto en una secuencia lineal de segmentos o temas.

Otro método desarrollado para la Segmentación de textos basado en la cohesión léxica es el propuesto por Heinone (1998) que a diferencia de Hearst propuso un método que emplea una ventana que recorre todo el texto y determina para cada párrafo, el párrafo más similar dentro de la ventana. Esta se formará por una cantidad de párrafos superiores e inferiores al que se analiza. Es un método útil para controlar la longitud (en cantidad de palabras) de los segmentos. Aunque este método logra determinar una correspondencia óptima entre la longitud de los segmentos que se obtienen, la longitud deseada para estos y el valor de similitud asociado a cada párrafo tiene el inconveniente de que el vector de cohesión del documento asocia cada párrafo con el valor de similitud más alto en su ventana, pero no considera que este valor puede corresponderse con un párrafo superior o inferior a él.

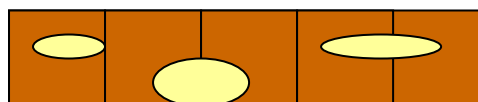
El objetivo principal de la Segmentación en esta Tesis doctoral es detectar los cambios temáticos que se producen en un documento, para establecer su estructura temática y obtener la indización automática de cada una de sus partes. De este modo, se obtiene la categorización del texto o documento utilizando la enumeración de sus partes temáticas a modo de niveles o estructura arbórea.

9.2. Estudio diacrónico del texto y segmentación del texto

Los cálculos realizados hasta ahora en todos los capítulos anteriores pueden considerarse sincrónicos en el sentido de que se han tomado y mezclado resultados en fragmentos de texto situados en cualquier posición.

Documento

**Estudio Sincrónico
utilizado en capítulos anteriores**



Pero cualquier texto tiene una naturaleza secuencial: aunque sea un texto formado por yuxtaposición de fragmentos de distintos autores, escritos en distintos momentos, su lectura exige comenzar por el principio y atravesarlo secuencialmente.

**Estudio Diacrónico
utilizado por Heaps**

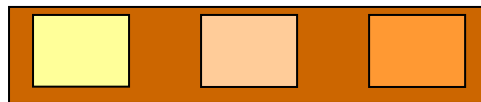
Documento



El enfoque diacrónico es el utilizado por Heaps ya que considera los fragmentos del texto como partes cada vez mayores del texto original.

**Estudio Diacrónico
utilizado en esta investigación**

Documento



El estudio diacrónico del texto utilizado en esta investigación consiste en tramos disjuntos que son los documentos y ventanas deslizantes que son grupos de documentos.

Desde este punto de vista diacrónico se divide el texto en fragmentos y se cuenta el vocabulario de cada uno, obtenemos así una sucesión de valores ordenados. Cada uno de ellos es un valor tomado por la variable aleatoria cuya media es el vocabulario promedio, dependiendo del tamaño del fragmento y quizá predicho por la ley de Heaps, la cual se ha comprobado que sigue una distribución normal.

Pero los sucesivos valores no son independientes: la intención del autor al continuar desarrollando un tema y las exigencias del lenguaje hace que exista una relación, no matemática sino de carácter aleatorio, entre el vocabulario de un fragmento y el de los que le preceden. Esta relación está abierta mientras el autor escribe el texto, pero queda fijada, en uno de sus infinitos valores posibles, en cuanto el texto ha quedado escrito. Hace visible así una huella característica del texto.

Para ilustrarlo con un ejemplo, tomamos un texto de Cervantes formado por el Quijote seguido de las Novelas Ejemplares, con un total de 3.214.045 bytes, $P = 280.471$ palabras y $V = 27.999$, lo dividimos en fragmentos sucesivos de tamaño fijo y contamos el vocabulario⁵⁶ en cada uno de ellos y cuantas palabras nuevas tiene, tomando medias móviles para suavizar la curva obtenemos la siguiente figura:

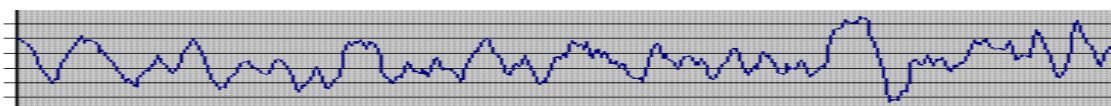


Figura 15. Novedad de palabras en texto de Cervantes

Acudiendo al texto podemos identificar cada tramo de esta figura: el primer tramo descendente corresponde a las primeras aventuras de Don Quijote cuando es armado caballero, los molinos etc; el tramo ascendente que le sigue contiene razonamientos y cuentos de pastores; el tramo descendente corresponde otra vez a aventuras (los yangüeses, la venta, los borregos, el muerto, los batanes, el barbero, los galeotes); el

⁵⁶ Aplicación RENOS, Véase Apéndice II

otro tramo ascendente, interrumpido por un pico corresponde a los razonamientos que Don Quijote hace en Sierra Morena y al cuento del pastor Cardenio, etc.

Esta representación no depende de una elección particular del tamaño de los fragmentos o de la media móvil. Comparamos distintas figuras obtenidas con distintos parámetros para ver que hay más o menos detalles dependiendo de la granularidad pero subyace una pauta que depende del texto y no del procedimiento.

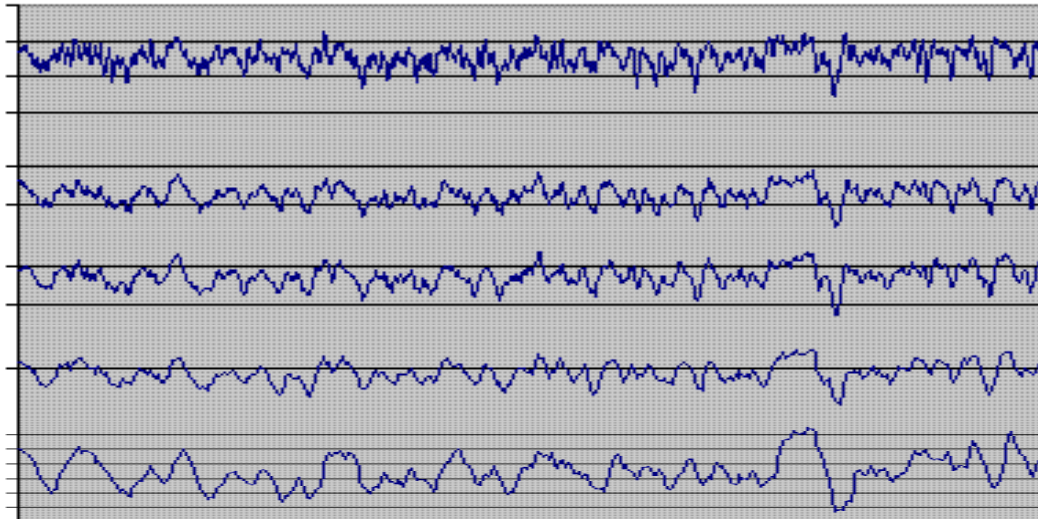


Figura 16. Novedad de palabras en tramos de distintos tamaños

Comparamos ahora las figuras obtenidas para distintos tamaños de los tramos es decir, distinta granularidad con 600, 1.000, 3.000 y 10.000 palabras; la magnitud representada es el cociente entre el vocabulario real de cada tramo y el del promedio de todos los tramos del mismo tamaño

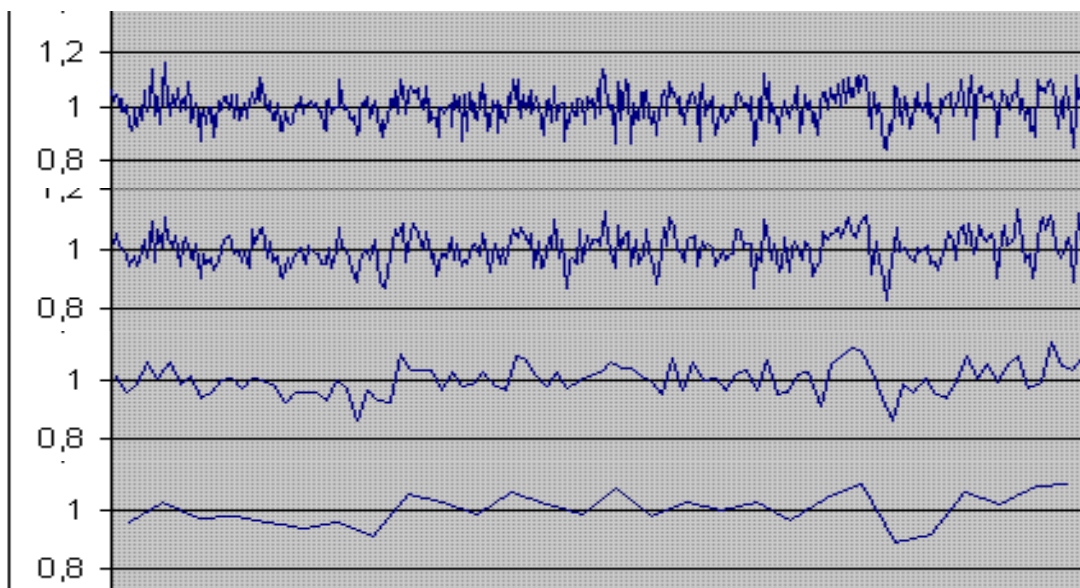


Figura 17. Novedad de palabras en tramos de distintos tamaños: 600, 1.000, 3.000, 10.000 palabras

De la misma figura ampliamos el tramo final, para ver que a distintas escalas, la constitución es similar, es decir a menor escala muestra la misma tendencia. Por tanto, podríamos afirmar que sigue una estructura fractal.

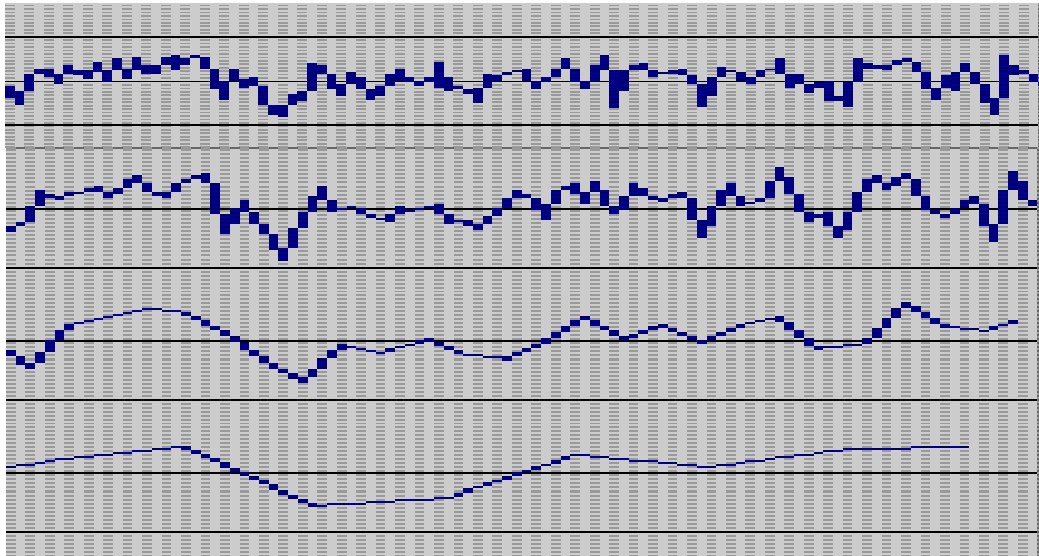


Figura 18. Ampliación tramo final de figura 17

Como otra forma de visualización vamos a partir de la misma subdivisión inicial en tramos de 600 palabras, valores divididos por el promedio en todos los tramos y suavizar la curva tomando medias móviles que agrupen los tramos de 3 en 3 y luego de 10 en 10

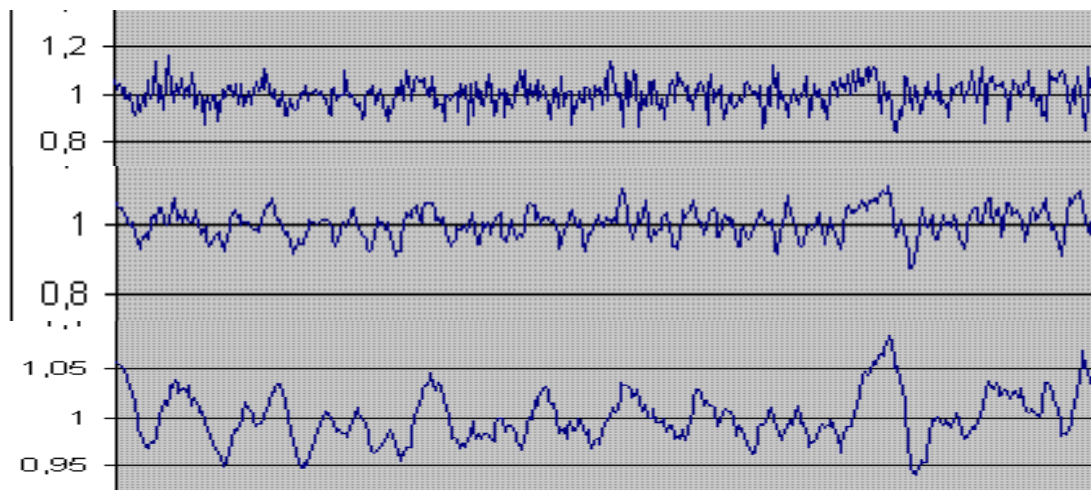


Figura 19. Novedad de palabras (medias móviles)

Otra posibilidad es tomar los tramos solapados parcialmente. En la siguiente comparación de dos figuras tenemos la inicial en tramos disjuntos de 600 palabras cada uno y la obtenida con tramos de 1.000 palabras en los que las 300 últimas son también las 300 primeras del tramo siguiente, de este modo obtendremos la cantidad de palabras nuevas respecto a las anteriores.

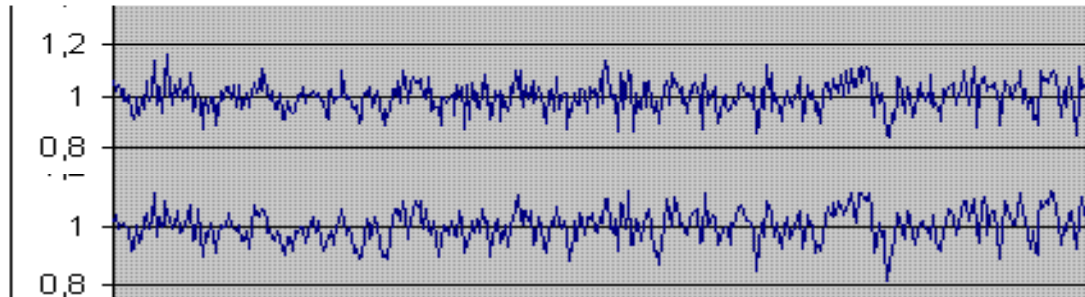


Figura 20. Novedad de palabras en tramos solapados

Se comprueba de este modo la identidad esencial de las figuras aunque hayan sido obtenidas para tramos de distintos tamaños y de distintos puntos de comienzo. Lo que representan es por tanto, una propiedad del texto que se manifiesta al ser observado en su desarrollo secuencial, diacrónico.

9.2.1. Granularidad de los documentos

En lugar de tomar fragmentos sucesivos de tamaño fijo lo adecuado es tomar la subdivisión en documentos, tal como se hizo inicialmente con el texto. Esto requiere disponer de una expresión bastante exacta de la fórmula de Heaps ya que para comparar el vocabulario de dos documentos de distinto tamaño hay que utilizar su relación al vocabulario hipotético que correspondería a cada tamaño.

Se muestra a continuación un texto formado por la concatenación de 6 textos diferentes, Larra, Galdós, Menéndez Pelayo, Legislación sobre bachiller, Costa y colección de resúmenes. Total de 269.693 palabras, vocabulario de 38.884, dividido en documentos por salto de línea después de 1.000 palabras o más, total 1.831 documentos

Primero vemos la figura del cociente entre el vocabulario real de cada documento y el hipotético que le correspondería según la fórmula (triple) de Heaps

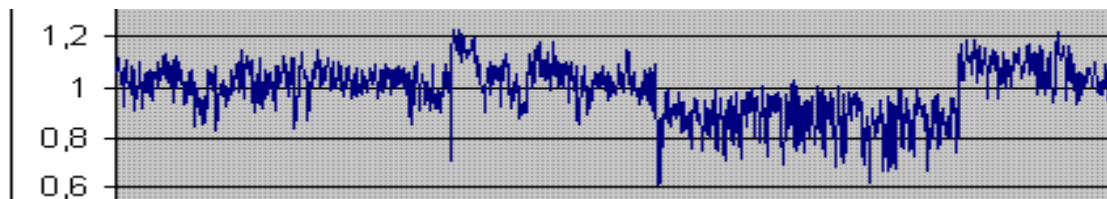


Figura 21. Superávit de vocabulario en tramos

Las distintas partes que constituyen este texto tienen su reflejo en forma de tramos a distinta altura oscilando el vocabulario real en torno a un promedio distinto para cada parte.

9.3. Novedad de las palabras en un texto

El tratamiento diacrónico nos permite observar las palabras nuevas que se introducen en un documento, las que no aparecen en los documentos anteriores. Su valor oscila más o en menos alrededor de un valor característico. Por ejemplo, estudiando cada documento respecto a los 20 documentos que le preceden podemos hacer:

PP = número de palabras en total en los 20 documentos precedentes
 VP = el vocabulario hipotético que predice la fórmula de Heaps
 PT = número de palabras en los 21 documentos, precedentes y actual
 VT = su vocabulario hipotético

Entonces, $VT - VP$ = el número de palabras distintas y que no están en los documentos precedentes, que podemos esperar en el documento actual.

Por tanto, obtenemos la Novedad de las palabras de cada tramo definiendo un tramo de texto como una secuencia de documentos consecutivos. El tramo se hace deslizar a lo largo del texto calculando en cada posición, se completa el tramo con los documentos inmediatamente anteriores, se cuentan los vocabularios de ambos para detectar las palabras nuevas que aparecen en el tramo, se compara su valor con el obtenido del mismo modo pero utilizando la triple fórmula de Heaps en lugar de contar.

También definimos como Novedad Inversa al mismo proceso pero recorriendo el texto del final al principio y Novedad Actual sería la diferencia entre la novedad y la novedad inversa. Valores altos de la novedad actual indican que aquí aparecen muchas palabras nuevas y en los documentos sucesivos aparecen menos. Igualmente definimos como Evolución de la Novedad la diferencia entre la novedad actual de un tramo y la del tramo anterior en el deslizamiento. Valores altos indican que empieza un subtexto de distinta temática.

Contando efectivamente las palabras con estas características que aparecen en el documento y viendo su desviación respecto a las hipotéticas descubrimos cuales son los documentos que introducen como novedad, respecto a sus precedentes, mayor cantidad de palabras. En la figura siguiente vemos el cociente o tasa de aparición de novedad en las palabras entre estas dos magnitudes.

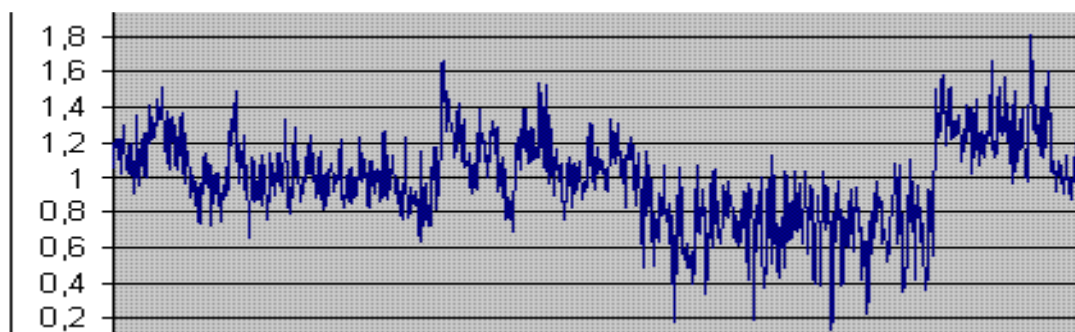


Figura 22. Tasa de aparición de novedad de palabras

Es de esperar que los puntos más altos de esta figura indiquen comienzo del desarrollo de una parte temáticamente distinta pero esto queda enmascarado por los distintos niveles de amplitud relativa del vocabulario que corresponden a cada parte del texto. No obstante, pueden hacerse algunas operaciones para que destaquen más claramente.

En primer lugar, invertimos el recorrido es decir, entresacamos las palabras de un documento que no aparecen en los 20 documentos siguientes y las mostramos relativamente a las que podrían esperarse según la fórmula de Heaps. La figura es parecida pero distinta a la anterior; las mostramos conjuntamente:

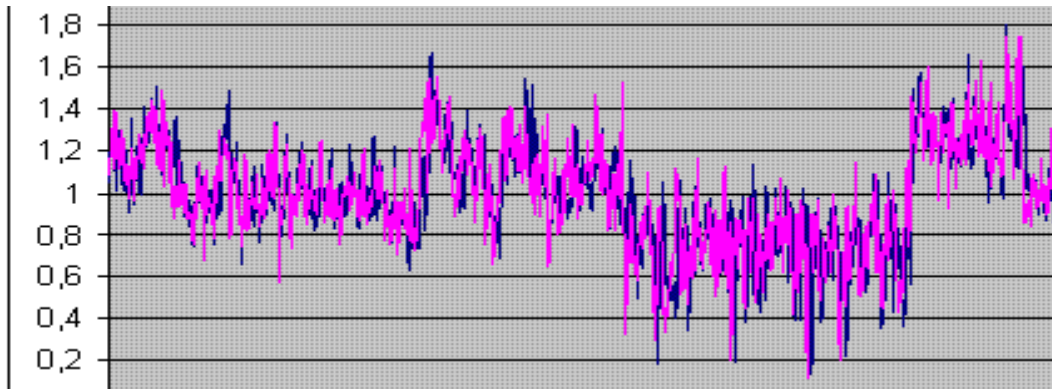


Figura 23. Tasa de novedad de palabras (recorrido inverso)

Los puntos en que ambas figuras alcanzan máximo local corresponden a documentos cuyo vocabulario tiene novedades tanto respecto a los que le preceden como a los que le siguen. Esto corresponde generalmente a párrafos enumerativos: en el texto de legislación sobre bachiller abundan ya que contiene las listas de temas de cada asignatura; en el texto de Cervantes destaca el episodio de la batalla con los borregos, donde Don Quijote enumera citándolos por sus nombres, todos los caballeros que participan.

Para efectos del presente trabajo lo que hay que detectar son los documentos iniciales de un nuevo asunto, para dividir todo el texto en partes temáticas. Estos serán documentos que incorporen novedades en el vocabulario respecto a los documentos precedentes pero cuyo vocabulario persista en los siguientes. Es decir, que presenten un máximo en la figura en orden directo a la vez que un mínimo en la del orden inverso. Para ponerlos de manifiesto representamos la diferencia entre ambas:

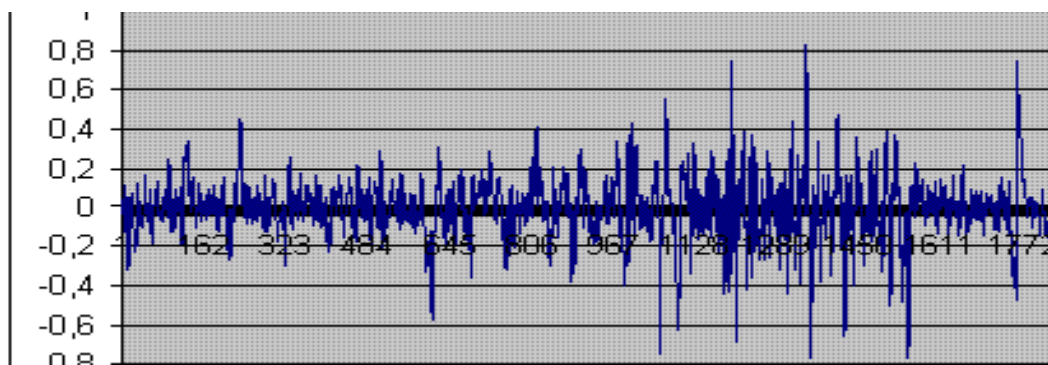


Figura 24. Puntos iniciales de asunto

Aquí los puntos más altos indican documento inicial de una parte del texto que aborda otra temática distinta. Pero aún podemos reforzar su efecto visual, filtrando algunos efectos que puedan ser debidos a otras causas. Aprovechamos la idea de que el documento anterior a un documento inicial de parte es el final de la parte anterior, y si leemos todo el texto del final al principio es este el que presenta las novedades persistentes en el vocabulario. Como eso es lo que representa la figura que se ha restado aparece como un mínimo local.

Reuniendo todo, la separación entre partes distintas temáticamente viene señalada por un mínimo seguido inmediatamente por un máximo. Este tipo de puntos en una figura

se ponen de manifiesto representando la diferencia de cada valor con el anterior (lo que en definitiva es una expresión del crecimiento brusco que se tiene en este punto).

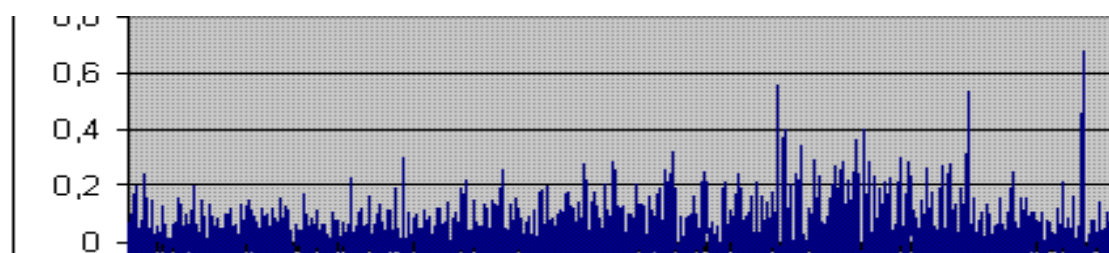


Figura 25. Puntos iniciales de asunto (en forma de valores positivos)

Aquí ya podemos prescindir de los mínimos carentes de significado. Los puntos más altos marcan cambio de tema. En concreto para este ejemplo, los tres más altos situados a la derecha, separan los fragmentos de bachiller, Costa y artículos. Dentro del tramo de bachiller se encuentran muchos puntos altos ya que es una enumeración de temarios de asignaturas. Y de manera parecida puede encontrarse una interpretación a cada uno de los otros puntos altos.

9.4. Aplicación a la Segmentación Automática del texto

Al comienzo de este capítulo se estudia el texto desde un enfoque diacrónico, compuesto por tramos sucesivos disjuntos que son los documentos, compuestos por ventanas deslizantes que forman grupos de documentos. De este modo se divide el texto en fragmentos sucesivos y se cuenta el vocabulario de cada uno de ellos, obteniendo así una sucesión de valores ordenados. Una vez hemos dividido el texto en fragmentos pequeños y ordenados, es cuando se realiza la Segmentación automática del texto en niveles temáticos y cada uno de los niveles corresponderá a cada uno de los picos detectados en las figuras anteriores, que se corresponden con la novedad de las palabras. Se segmenta el texto por los 6 puntos de mayor variación de novedad (siempre que los tramos obtenidos sean suficientemente grandes). Y en cada una de estas partes temáticas se congregarán documentos diferentes del texto origen, que en ocasiones pueden ser sucesivos o no, pero tendrán en común la misma temática. El proceso concluye cuando se forma la estructura temática del texto subdividido en partes o niveles. Para el cual se utilizan parámetros que miden la relevancia de las palabras y técnicas de clustering según las similitudes, para encontrar las palabras características de cada subdivisión

10. Sistema de Indización y Segmentación automática del texto: MALLOV.

10.1. Sistema de Indización y Segmentación Automática MALLOV

Se ha desarrollado un Sistema de Indización y Segmentación Automática denominado MALLOV, en el cual se aplican las técnicas y modelos desarrollados a lo largo de los capítulos que conforman esta tesis doctoral, El sistema MALLOV es capaz de realizar la estructura temática del texto o documento en cuestión y por supuesto la indización automática de cada una de sus partes temáticas en las que ha sido dividido.

El fin de MALLOV es comprobar que los resultados obtenidos en los capítulos anteriores, utilizados conjuntamente consiguen un funcionamiento razonable. Para obtenerlo es importante no añadir las potentes herramientas ya conocidas⁵⁷ sino mantenernos estrictamente en aplicar los métodos objeto de este estudio.

La Segmentación del Texto (*Text Segmentation*) es un proceso trascendental para elaborar resúmenes automáticos, aunque esa no sea la cuestión que abordamos en esta investigación, sino más bien nuestro objetivo es enumerar las partes temáticas en las que se divide el texto. Es decir obtendremos la categorización del texto o documento utilizando la enumeración de sus partes temáticas a modo de niveles o estructura arbórea.

Para realizar esta estructura arbórea en la que obtendremos los diferentes temas que coexisten en un documento, en primer lugar hemos de decidir cómo vamos a representar el árbol jerárquico y sus niveles y subniveles, y para ello se decide crear la siguiente estructura:

Nivel 0, el conjunto de palabras que describen el contenido general del documento.

Nivel 1, el conjunto de palabras que representan el contenido de los distintos capítulos del documento.

Nivel 2, el conjunto de palabras que detallan el contenido de pequeños grupos de documentos

Nivel 3, sería el conjunto de palabras que relatan el contenido de los párrafos del documento.

Y así sucesivamente con las demás subdivisiones. El conjunto de términos de todas ellas formarán el vocabulario de un documento.

Por consiguiente, para desarrollar esta estructura, realizamos un cálculo teórico de verosimilitud con fórmulas hipotéticas que proporcionarán un orden de magnitud de los posibles resultados respecto a la cantidad de niveles a decidir en la segmentación, aunque finalmente no se utilice el resultado de este cálculo teórico de verosimilitud sino que los niveles se formarán dependiendo del tamaño de los documentos y de las palabras que aparezcan.

⁵⁷ Véase capítulo 3

Para el cálculo teórico de verosimilitud, si tenemos un total de v niveles en el árbol jerárquico, y cada nivel está formado por g palabras que representan el texto global (nivel 0), los grupos de documentos (nivel 1), contenido de un documento (nivel 2), o que representan un párrafo (nivel $v-1$), y a su vez estos niveles tienen d subdivisiones.

Si no se repiten las palabras en cada uno de estos niveles, mediante la siguiente fórmula hipotética se halla el número total de palabras en cada nivel. En los cuales el total de palabras es igual al vocabulario.

$$g + dg + d^2g + \dots + d^{v-1}g = g(d^0 + d^1 + \dots + d^{v-1}) = \frac{gd(d^{v-1} - 1)}{d - 1}$$

Para calcular el número total de niveles que se formarán automáticamente dependiendo del tamaño del vocabulario del texto a analizar, es preciso primeramente decidir arbitrariamente un número que consideremos adecuado para la cantidad de subdivisiones que deben constar en la creación de dicho árbol jerárquico, como por ejemplo el 7, este número es la reflexión de que, si se considera en la Clasificación Decimal Universal (CDU) que diez clasificaciones son suficientes, y si por el contrario es lógico pensar que 4 ó 5 son muy escasas, entonces consideraremos que 7 es un número adecuado para plasmar la cantidad de subdivisiones.

El penúltimo nivel es igual a la cantidad de documentos totales que forman el documento grande y original, a esta cantidad la llamamos D ,

$$7^{v-2} = D$$

La fórmula hipotética que proporciona un orden de magnitud de los posibles niveles es la siguiente:

$$v = 2 + \frac{\ln D}{\ln 7}$$

Una vez elaboradas las fórmulas se aplican a varios ejemplos para ver su comportamiento,

Ejemplo 1:

Texto	Episodios Nacionales, Autor: Galdós
Documentos	1102
Vocabulario	17.103 palabras

$$\frac{g7(7^{v-1})}{7-1} = V \qquad g = \frac{6V}{7(7^{v-1} - 1)}$$

Con esta fórmula hipotética hallamos el número de términos que compondrán cada nivel.

Una vez realizamos los cálculos teóricos de verosimilitud obtenemos un valor de $v = 5,60$ y un valor de $g = 0,87$. Ante el valor de v , sólo es posible un número entero de

niveles, así que redondeamos este valor hacia arriba con lo que finalmente tendremos $v = 6$, lo que significa que decidimos tomar 6 niveles, y del mismo modo ante el valor de g , sólo es posible un número entero de palabras o términos, así que redondeamos este valor hacia arriba, con lo que tendremos $g = 1$, justamente en este ejemplo no podríamos redondear hacia abajo, si no nos encontraríamos con 0 palabras en cada nivel, y evidentemente no podríamos hacer ningún estudio ni experimentación posible.

Una vez obtenido estos valores, observamos que resulta escasa la cantidad de palabras para cada nivel, ya que con sólo una palabra por nivel no determina la temática del texto, documentos, párrafos etc. Así que rebajamos los valores de v y g y una vez confirmados los nuevos valores obtenidos, $v = 5$ y por consiguiente $g = 6,1 \cong 6$ palabras en cada grupo.

Esquema hipotético Segmentación

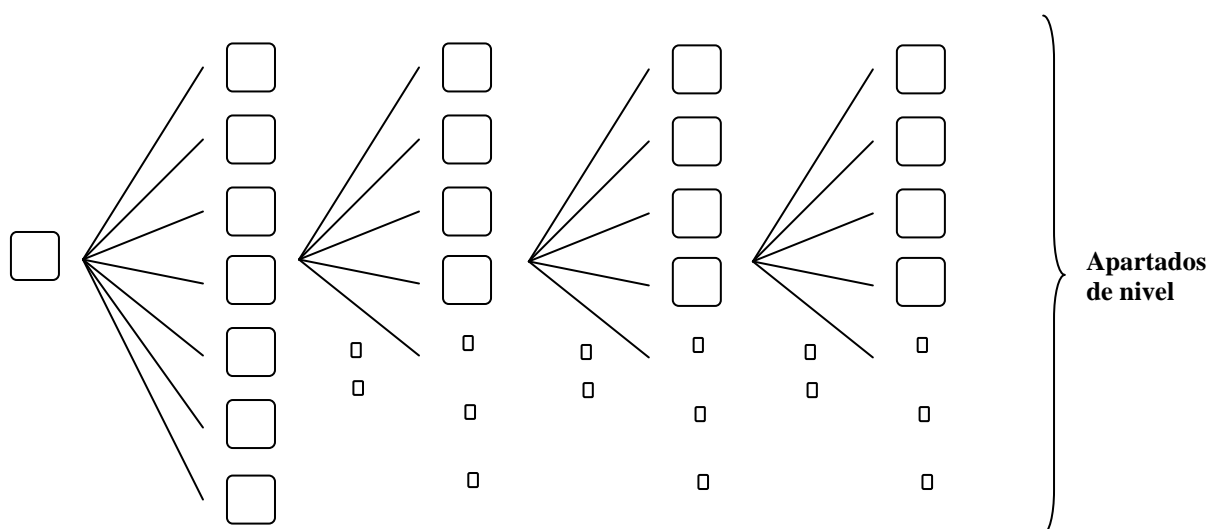


Figura 26. Niveles en Sistema de Indización y Segmentación automática MALLOV

- Nivel0:* 6 palabras que representan el contenido general del documento.
- Nivel1:* 7 grandes apartados de nivel1, con un total de 42 palabras
- Nivel2:* 49 grandes apartados de nivel2, con un total de 294 palabras.
- Nivel3:* 343 apartados de nivel3, correspondientes a documentos contenidos en el documento general o pequeños grupos de documentos, con un total de 2058 palabras.
- Nivel5:* 2401 apartados pertenecientes a párrafos, en promedio son la mitad del documento, con un total de 14.406 palabras, se aproxima al total de vocabulario del texto pero no exactamente.

Si tomamos otro ejemplo:

Ejemplo 2:

Texto	Episodios Nacionales, Autor: Galdós
Documentos	3000
Vocabulario	10.290 palabras

Para aplicar el número de niveles aplicamos la fórmula hipotética:

$$v = 2 + \frac{\ln 3000}{\ln 7} = 6,1 \cong 6$$

Para obtener la cantidad de palabras en cada nivel, aplicamos la fórmula hipotética:

$$g = \frac{6 \cdot 10.290}{7(7^{5-1} - 1)} = 0,5$$

En este caso observamos que 0,5 palabras por cada grupo no son suficientes, así que decidimos tomar 4 niveles, y los resultados tampoco son muy favorables, ya que obtendríamos un total de 25 palabras por grupo, y esta cantidad serían demasiadas. Finalmente tomamos 5 niveles y 4 palabras por grupo.

Obviamente la finalidad de esta representación arbórea, sus niveles y el redondear los resultados obtenidos mediante estas formulaciones intuitivas es conseguir una cantidad de niveles acordes al tamaño del documento o texto que se está analizando y también obtener una cantidad de términos de indización acordes igualmente al tamaño del documento.

Es por ello que aunque nuestra estructura de temas del texto la hallamos dividido en siete niveles un documento puede verse representado sólo en cuatro, seis o tres niveles indistintamente, y siempre dependiendo del tamaño y cantidad de temas distintos encontrados en el documento en cuestión.

Una vez representadas las partes temáticas del texto en sus niveles correspondientes mediante las palabras indizadas, estos se agrupan en bloques distribuidos jerárquicamente según se desglose el documento en cuestión, es decir, primeramente existirá un bloque inicial que describa el contenido global de todo el documento, y en este bloque habrá una cantidad inicial de palabras o descriptores, seguidamente este bloque inicial se subdividirá en varios bloques, los cuales corresponden a distintas partes del documento total, cada uno de estos también contendrá una serie de palabras que describa el contenido y así sucesivamente hasta poder formar las divisiones necesarias y llegar a describir cada párrafo del documento en cuestión.

Los términos que contiene el primer bloque serán las palabras de mayor frecuencia en el documento, y así sucesivamente las palabras que describen los contenidos de varias partes del documento total sean las de menor frecuencia, etc.

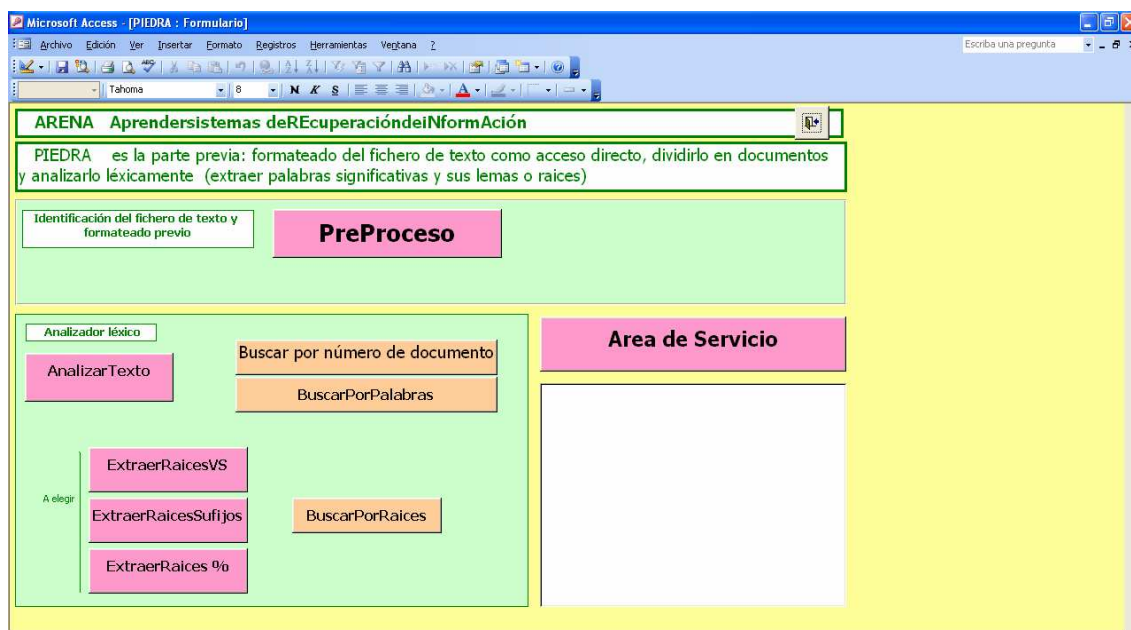
En definitiva, todas las experimentaciones realizadas hasta ahora con los textos se han llevado a cabo para formar una estructura coherente respecto al tamaño del documento y que se puedan seleccionar los términos que definen mejor el contenido general del documento, pero no sólo eso, sino que también se pueda seleccionar los que mejor definen el contenido de cada uno de los capítulos o subdocumentos en los que se divide el documento grande de partida u original, y así sucesivamente hasta seleccionar de igual modo los términos más característicos de los párrafos, etc. Es decir indizaremos automáticamente la granularidad de los documentos objeto de análisis.

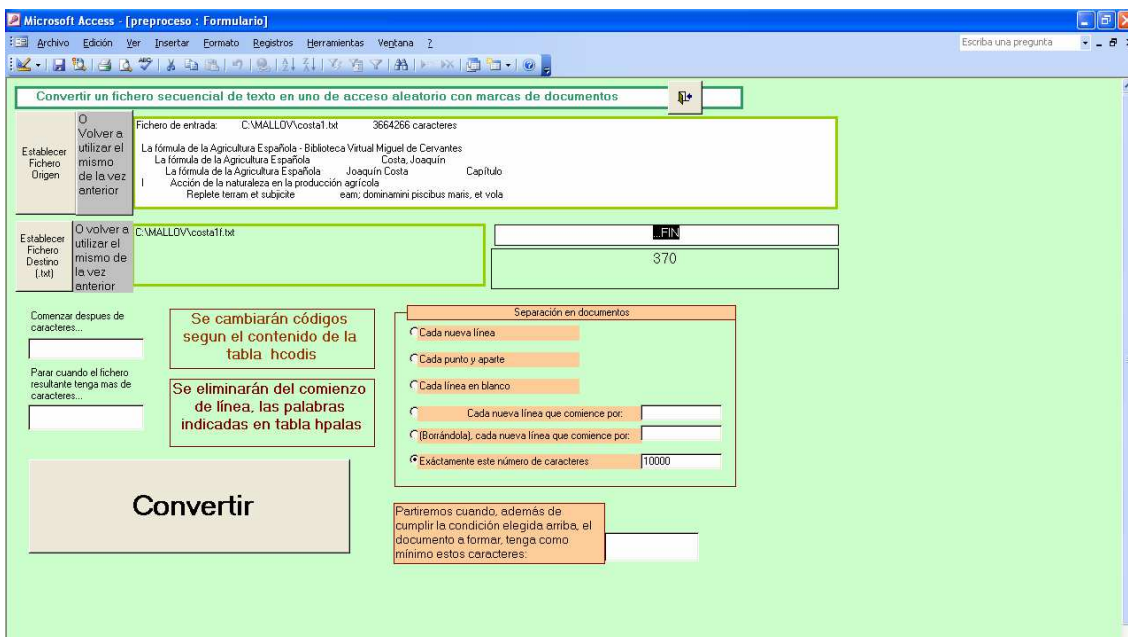
10.1.1. Sistema de Indización y Segmentación Automática MALLOV: Procedimiento básico

El procedimiento para implementar el Sistema de Indización y Segmentación Automática MALLOV ha supuesto el estudio y perfeccionamiento de los métodos cuantitativos y leyes clásicas en Recuperación de Información, como son los modelos relativos al proceso de repetición de palabras (Zipf, 1949), (Mandelbrot, 1953) y al proceso de creación de vocabulario (Heaps, 1978). Se realiza una crítica de las circunstancias de aplicación de los modelos y se estudia la estabilidad de los parámetros de manera experimental mediante recuentos en textos y sus fragmentos. Se establecen recomendaciones a priori para los valores de sus parámetros, dependiendo de las circunstancias de aplicación y del tipo de texto analizado. Se observa el comportamiento de los parámetros de las fórmulas para vislumbrar una relación directa con la tipología de texto analizado. Se propone un nuevo modelo (Log-%) para la visualización de la distribución de frecuencias de las palabras de un texto.

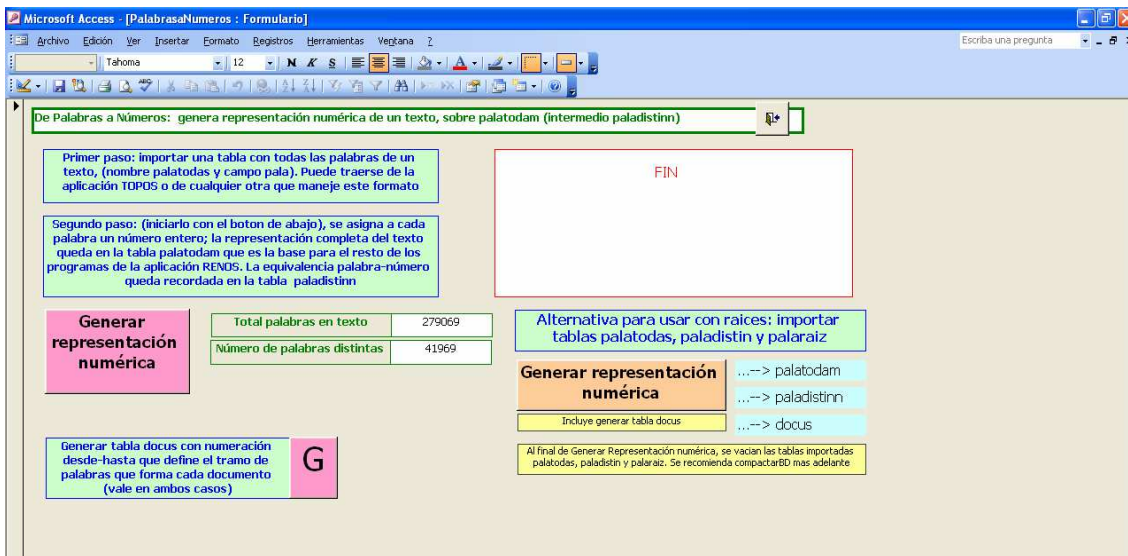
El objetivo final es detectar los cambios temáticos que se producen en un documento, para establecer su estructura temática y obtener la indización automática de cada una de sus partes. De este modo, se obtiene la categorización del texto o documento utilizando la enumeración de sus partes temáticas a modo de niveles o estructura arbórea.

El procedimiento para la implementación del sistema ha consistido primeramente en convertir los ficheros de textos en ficheros de acceso aleatorio con formato fijo para compatibilidad entre los distintos programas utilizados; dividirlos en una serie de documentos o fragmentos pequeños, según distintos criterios; formar la colección ordenada de palabras que los componen, prescindiendo de palabras vacías; obtener el vocabulario o colección de palabras distintas; sustituir el vocabulario por la colección de raíces mediante la aplicación de un Stemmer; identificar saltos con posibles cambios semánticos para dividir el texto en partes; identificar palabras importantes para la caracterización e indización de cada parte y buscar pares de palabras asociadas para reforzar el valor semántico.

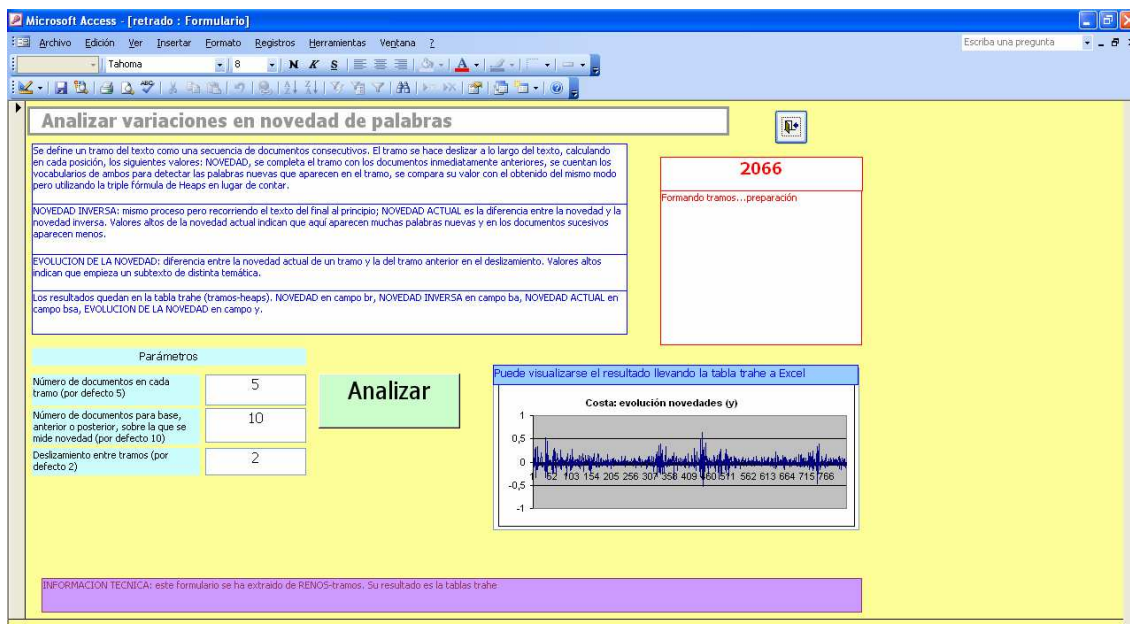




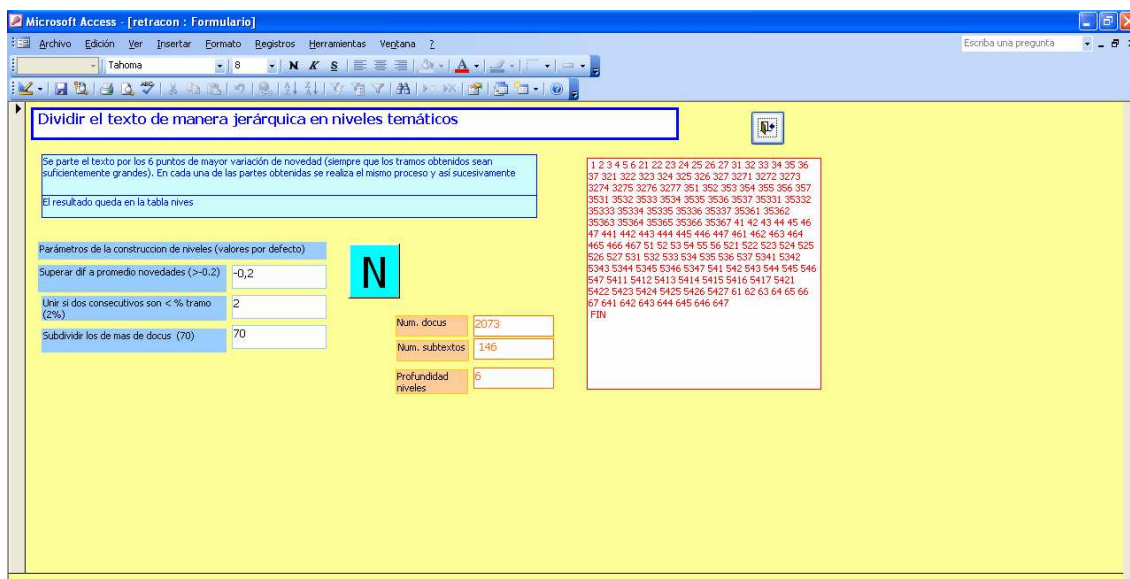
Una vez los textos se han desgranado en distintas tablas con las palabras, vocabulario, raíces, parejas de palabras asociadas, etc., se genera la representación numérica de un texto a partir de sus raíces, de este modo el sistema trabajará con números lo cual acelera los procesos de cálculo.



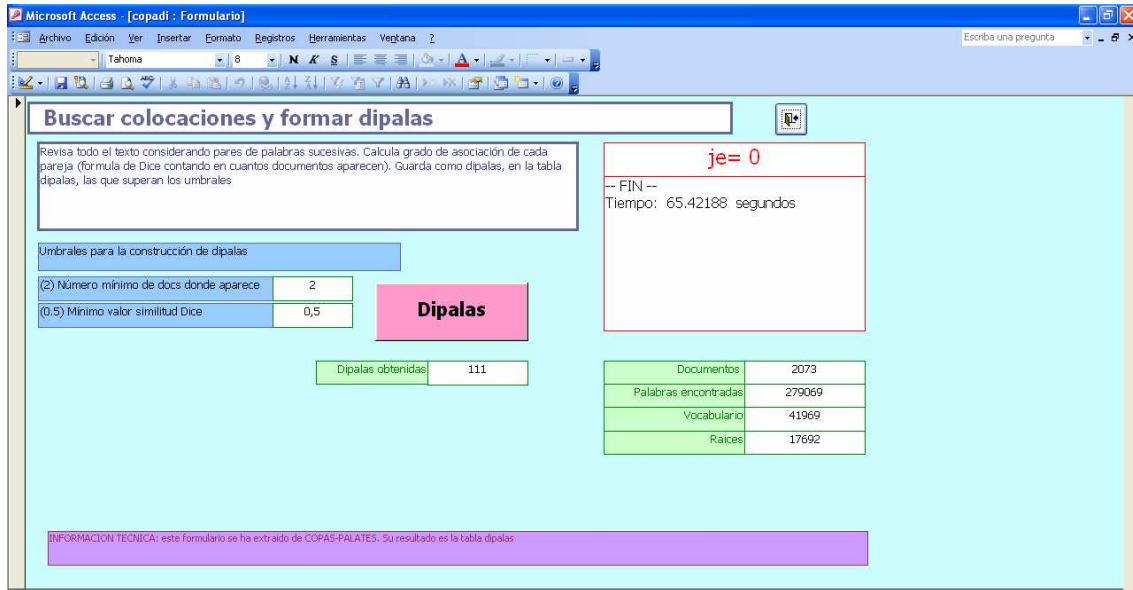
Otro paso importante es detectar las variaciones de novedad, las palabras nuevas en el vocabulario respecto a los documentos precedentes, pero cuyo vocabulario persista en los siguientes, para dividir el texto en partes temáticas automáticamente (*Text Segmentation*).



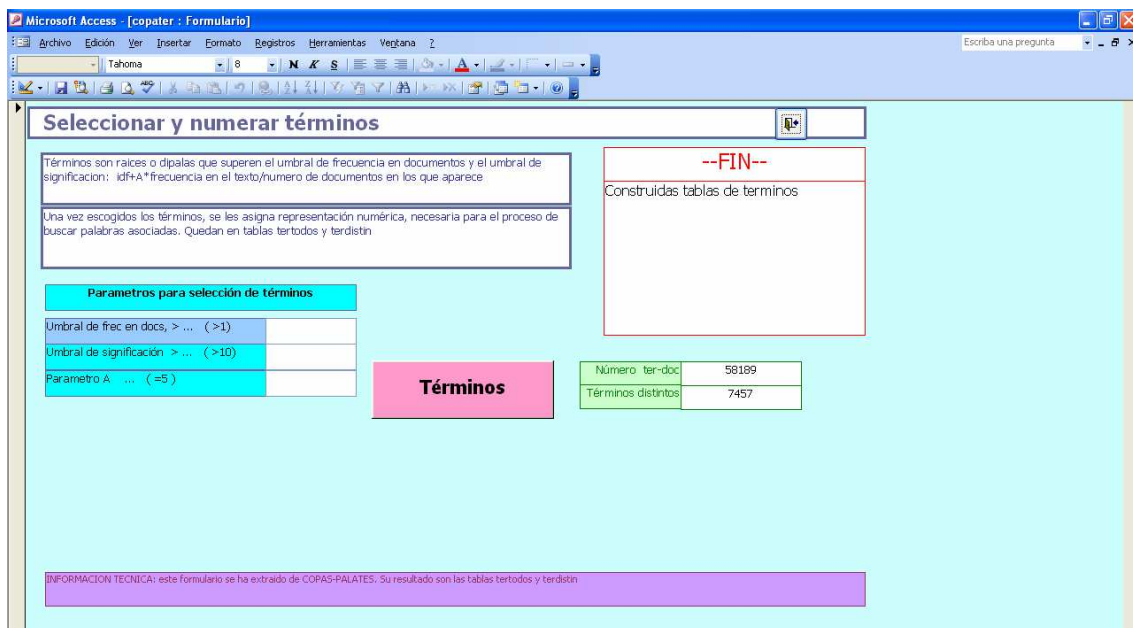
A continuación se parte el texto por los 6 puntos de mayor variación de novedad (siempre que los tramos obtenidos sean suficientemente grandes). En cada una de las partes obtenidas se realiza el mismo proceso y así sucesivamente. Dividimos el texto semánticamente.



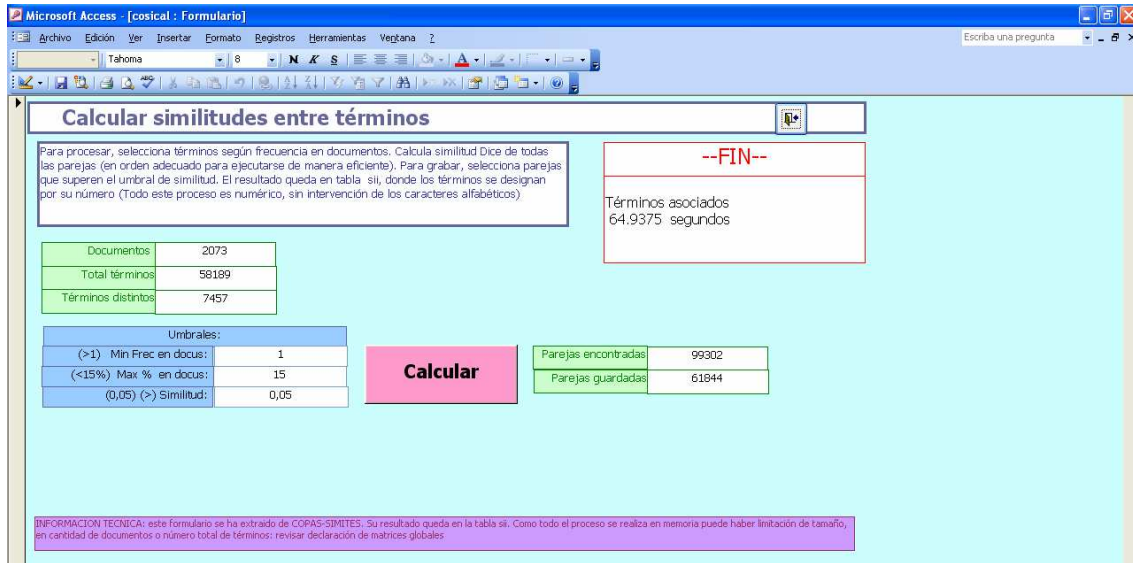
La importancia de las palabras asociadas, colocaciones o di-gramas es tal que se calculará el grado de asociación de cada pareja para finalmente escoger las que superen los umbrales determinados según la fórmula de Dice (1945). Una vez identificadas previamente las colocaciones se tratan como si fueran una sola palabra, uniéndolas con un guión, por ej. (Comunidad-Valenciana). En base a la asociación de las palabras revelaremos la estructura de las relaciones existentes en el texto que servirán para establecer los clusters. Utilizaremos el nombre de “*términos*” para referirnos al conjunto resultante, formado tanto por raíces como por di-gramas, resultantes de las colocaciones.



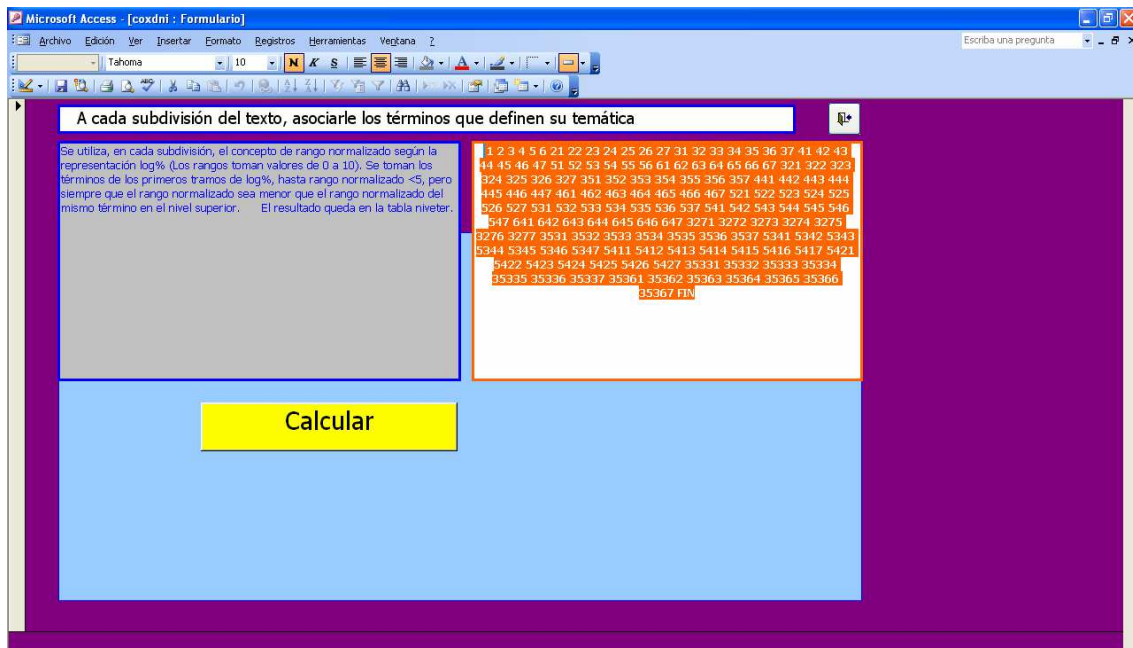
Seguidamente se divide el texto en las partes temáticas (*Text Segmentation*), y para ello se seleccionan los “*Términos*”, que formarán parte de la Indización y Segmentación Automática,



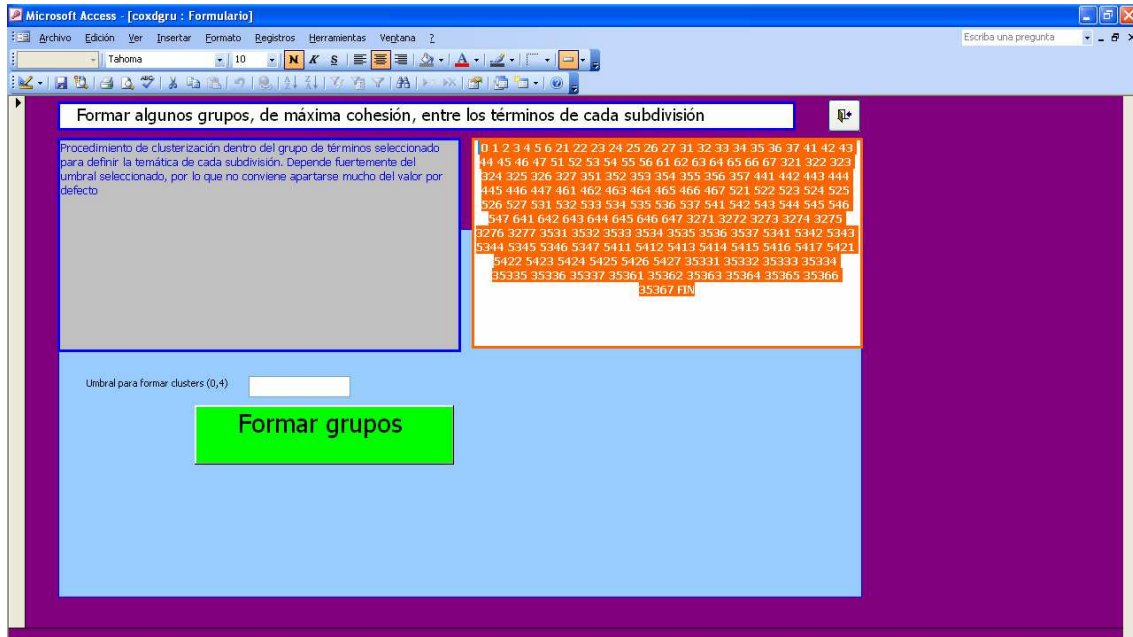
Los *términos* que finalmente formarán parte de dicho mapa temático o Sistema de Indización y Segmentación Automática serán por tanto una combinación de raíces obtenidas del texto y coocurrencias que superen el umbral de frecuencia en documentos además del umbral de significación. Finalmente se obtiene la matriz de similitudes y se colocan automáticamente los términos de indización en cada nivel de Segmentación utilizando la representación Log-% desarrollada en esta Tesis doctoral.



El siguiente paso es la construcción de Clusters o grupos de términos calculando la máxima cohesión, la cual pretende caracterizar la intensidad de las relaciones que unen las palabras que componen un cluster determinado. Los clusters pueden ser colocados por orden de cohesión creciente.



Seguidamente se procede a la clusterización dentro del grupo de términos seleccionado para definir la temática de cada subdivisión, esto dependerá fuertemente del umbral seleccionado, por lo que no conviene apartarse mucho del valor por defecto.



Una vez constituidas las partes temáticas del texto en sus niveles correspondientes con los términos indizados, estos se agrupan en bloques distribuidos jerárquicamente según se desglose el documento en cuestión. El bloque inicial describe el contenido global de todo el documento con una cantidad inicial de términos. Seguidamente este bloque inicial se subdivide en varios bloques, los cuales corresponden a distintas partes del documento total, cada uno de estos también contiene una serie de términos que describe el contenido y así sucesivamente hasta poder formar las divisiones necesarias y llegar a describir cada párrafo del documento en cuestión

Finalmente se ejecuta la Indización y Segmentación del texto, se muestra el resultado de los clusters mediante la estructura arbórea, los términos indizados se muestran en mayúsculas los que se corresponden con los clusters formados y en minúsculas los términos complementarios que son más importantes en ese nivel y no en el nivel anterior.

Microsoft Access - [coveni : Formulario]

Archivo Edición Ver Insertar Formato Registros Herramientas Ventana 2

Tahoma 8

EsquemaTexto

La distribución del texto en niveles y los terminos que definen su tematica está en las tablas nives, niveter y terdistin

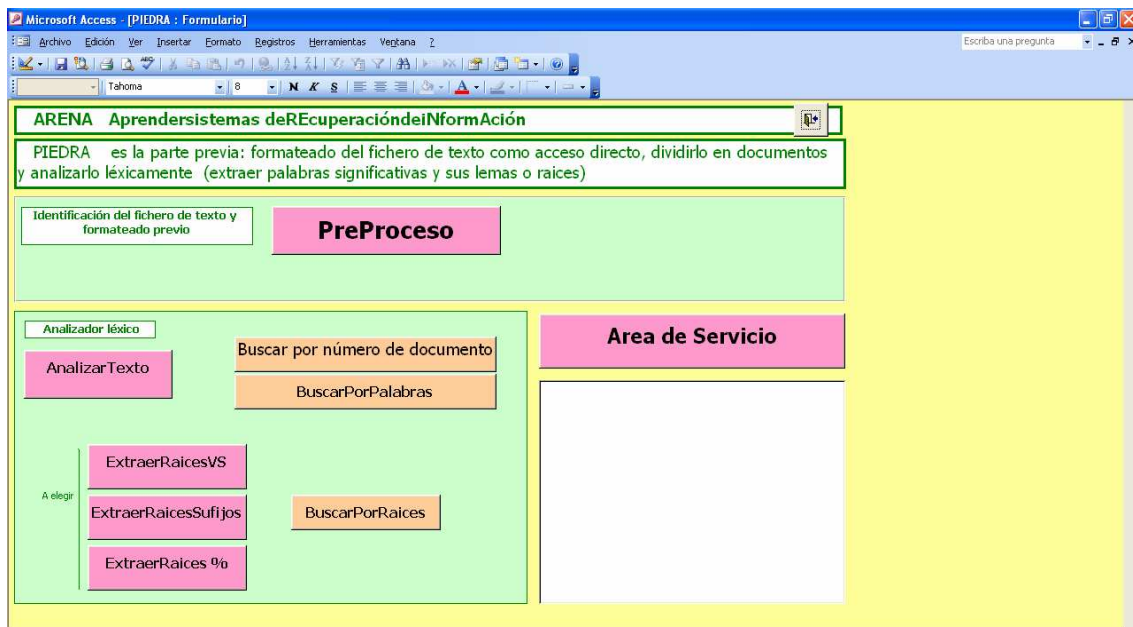
Documentos		
Desde	Hasta	Total
0	1	2073

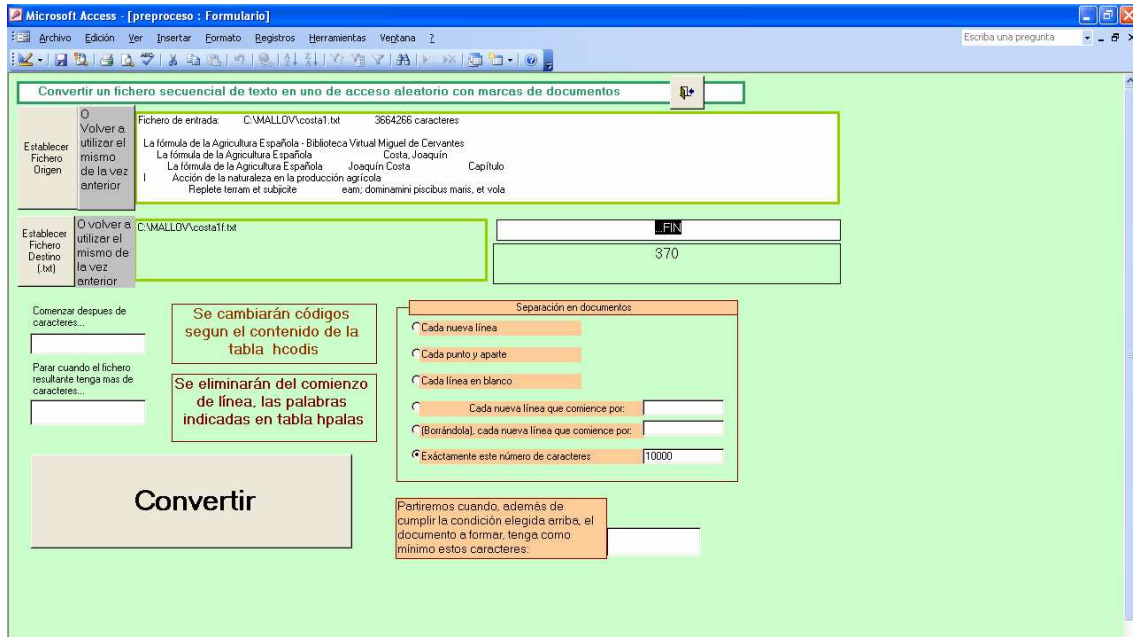
TODO MUESTRA PUEBLO AGUA SEÑA ARBOL
CULTO POLITICO AGRICULTOR DERECHO
NUMISIO COSTA PLANTO LEY RIO FRUTO
PAG PATRIA RIEGO CANALS

	Desde	Hasta	Total	Subdivisiones
1	1	33	33	0
PLANTO ANIMALES CULTO HUEVA PECES VEGETALES DOMESTICA ARBUSTO VEGETA AGUA VERGEL OLIVO RIO ARADO ROCA COSTUMBRE VIENTO AREN POZOS ARTESIANOS LAGO CHARCA				
2	33	129	97	7
(OASIS SAHARA) AGRICULTOR PLANTO CULTO VEGETA OLIVO ARADO ARTIFICIALES SUD PATRIA CEREALES TRIGO INDIGENA PALMA DESIERTO LLUVIAS ROCA VIENTO AREN POZO ESTADOS_UNIDOS FERROCARRIL SUBTERRANEA HECTOLITRO AMERICANA ZONA				
3	129	855	727	7
(OLIGARQUIA CACIQUES OLIGARCA) (OPOSICION OPOSITOR) AGRICULTOR LEY TIPO GOBERNAR LEVES ALIMENTO IGNORO PATRIA SOBERANIA REGISTRA EDUCA ESCUELAS LECCION SEQUE PAG REY ESTUDIOS GOBIERNOS DERECHO ELECTORAL CONTESTA METODO CRISIS POLITICO TITULO MAESTRA VOTO ESCUELA NIÑO				
4	855	1115	261	7
(CANALS RIEGO SOBRRARRE CANO TAMARITTE) AGUA RIO EMIGRA SEQUE HAMBRE ACEQUIA NIEVA REGADA FERROCARRIL VAPOR CONSTRUIR GOBIERNOS POLITICO REGAR EBRO HIDRAULICO ZONA COMISO PESETA ARAGONES MORET LITERA ESERA				
5	1115	1873	759	6
(PRUDENCIA PALLINA) (NUMISIO THEODOSIO) LEY RIO GOBERNAR LEVES ENFERMA IGNORO CESAR ESCLAVO VINO COSTUMBRE VIENTO AIRE PAG DIOSA REY GRATO INDIVIDUAL CODIGO DERECHO MAXIMA PREGUNTO LEGISLA ACEPTA REPUBLICA TITULO RESTAURO SUR APROBO MAESTRA ESTATUA NIÑO				
6	1873	2073	201	7
(FRUTO PLANTO ARBOL) CULTO INBIERTO COSECHA ENCINAS COLONO CASTAÑO CEREALES TRIGO PROPIETARIO RECOLECCION HECTAREA METRO HORTALIZAS ALMENDRO HUERTA EXPORTA MELOCOTONES ALBARICOQUES FINCA PESETA MANZANA				
-	-	-	-	-

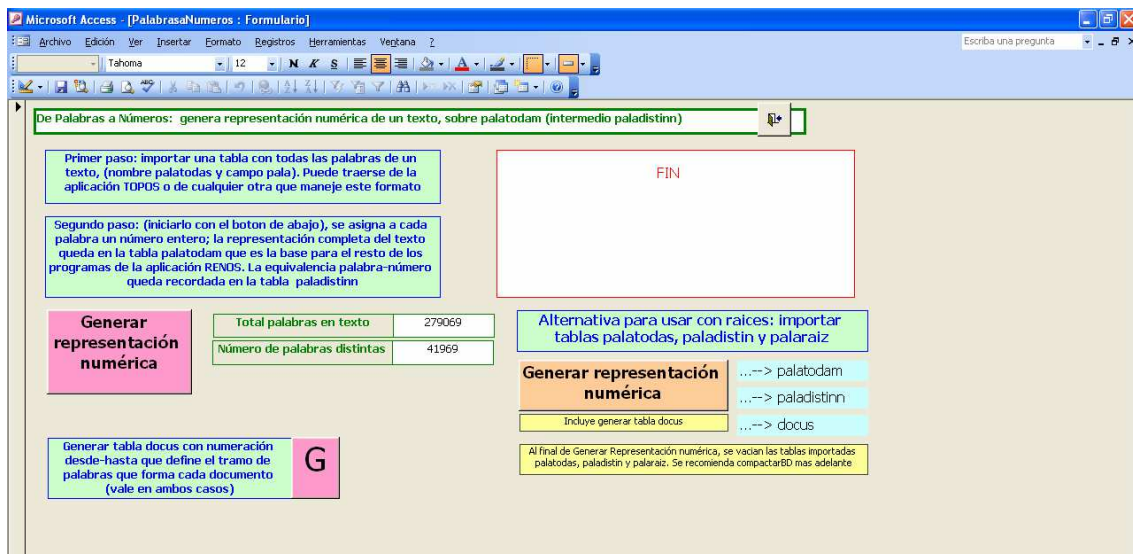
10.1.2. Sistema de Indización y Segmentación Automática MALLOV: Procedimiento completo

Para la implementación del Sistema de Indización y Segmentación Automática MALLOV se realiza en primer lugar la conversión de los ficheros de textos en ficheros de acceso aleatorio con formato fijo para realizar el procesado del texto, del cual extraemos las palabras y el vocabulario, para posteriormente mediante la utilización de un Stemmer extraer las raíces del vocabulario. Estas palabras se almacenan en las tablas correspondientes denominadas *palatodas*: todas las palabras del texto, *paladistin*: palabras distintas o vocabulario, *palaraiíz*: raíces obtenidas del vocabulario, *docus*: número de bloques en los que se ha dividido el texto, número de documentos totales en los que se ha dividido el texto y longitud en caracteres de cada uno de ellos. La longitud de cada documento permite calcular medidas relativas, ya que no es lo mismo que una palabra aparezca x veces en un documento pequeño que en uno grande.





A continuación se genera la representación numérica de un texto a partir de sus raíces, de este modo el sistema trabajará con números lo cual acelerará los procesos de cálculo.



En caso de textos de más de 40.000 palabras, para cada uno de los tres rangos de tamaño, se generan varios subtextos extraídos aleatoriamente del texto total; se cuenta el vocabulario de cada uno de ellos y se ajusta una función potencial a los valores de (palabras, vocabulario). Aquí se entiende por vocabulario el conjunto de raíces distintas, no palabras distintas, lo que influye en los valores de los coeficientes y exponentes obtenidos.

Para este proceso se utiliza de manera esencial la representación numérica del texto, donde quedan parámetros básicos del texto, su representación numérica y sobre todo, los parámetros para la triple fórmula de Heaps que se utiliza en las fases siguientes. El resultado se guarda en la tabla heaps3.

id	x
188	253732
189	32000
190	-32000
191	0
192	3000
193	25373
194	2,028297
195	0,8384579
196	16,11601
197	0,5855477
198	50,55013
199	0,4730992

Tabla 34. Parámetros para la triple fórmula de Heaps

Microsoft Access - [rehe3h3 : Formulario]

Archivo Edición Ver Insertar Formato Registros Herramientas Ventana 2

Tahoma 8

Triple fórmula de HEAPS

Recomendable solo para textos de más de 40000 palabras

Para cada uno de los tres rangos de tamaño, se generan varios subtextos extraídos aleatoriamente del texto total; se cuenta el vocabulario de cada uno de ellos y se ajusta una función potencial a los valores de (palabras, vocabulario). Aquí se entiende por vocabulario el conjunto de raíces distintas, no palabras distintas, lo que influye en los valores de los coeficientes y exponentes obtenidos.

Para este proceso se utiliza de manera esencial la representación numérica del texto. El resultado se guarda en la tabla heaps3

Ajustando triple Heaps
 Desde 100 hasta 3000
 100 390 680 970 1260 1550 1840 2130 2420 2710 3000
 Desde 3000 hasta 27906
 3000 5490 7961 10471 12962 15453 17943 20434 22924 25415 27906
 Desde 27906 hasta 279069
 27906 53022 78138 103254 128371 153487 178603 203720 228836
 253952 279069

Rangos de tamaños		Fórmulas		
desde	hasta	Coefficiente	Exponente	Error
100	3000	1,979438	0,8307182	0,003633913
3000	27906	23,8583	0,5300764	0,002037342
27906	279069	57,02121	0,4564165	0,00139324

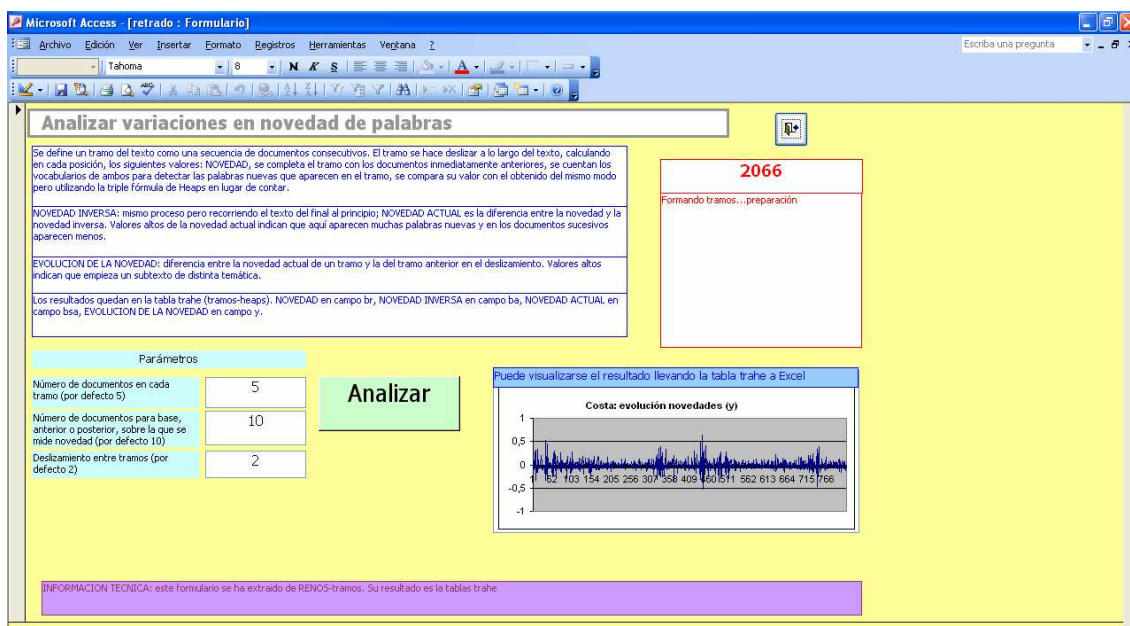
INFORMACIÓN TÉCNICA: este formulario se ha extraído de RENOS-HEAPS3. Su resultado queda en la tabla heaps3

Seguidamente se identifican saltos con posibles cambios temáticos para dividir el texto en partes; y para ello se detectan las variaciones de novedad, las palabras nuevas en el vocabulario respecto a los documentos precedentes, pero cuyo vocabulario persista en los siguientes.

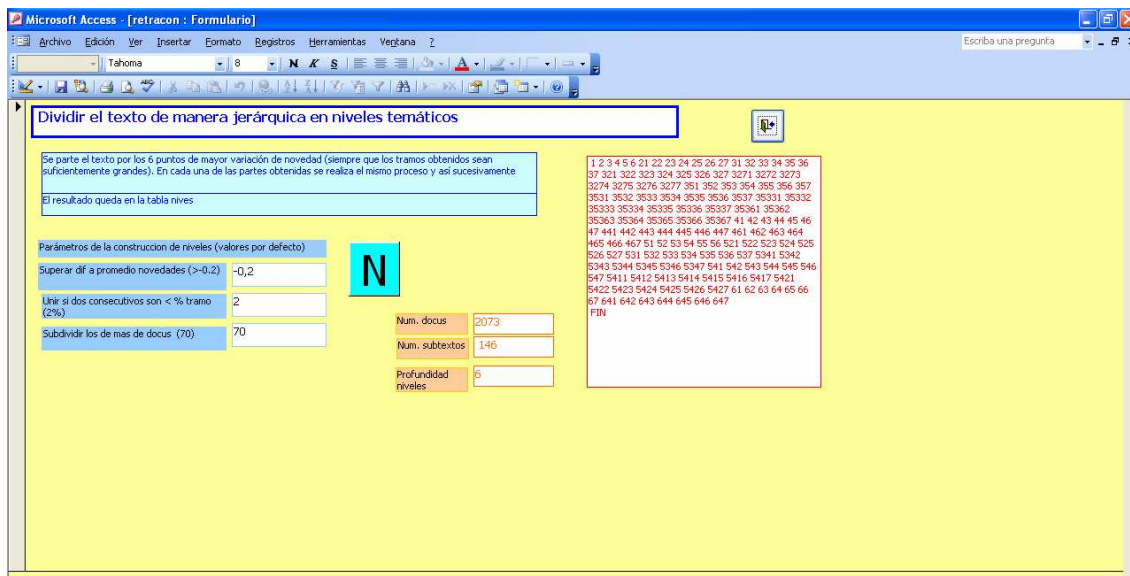
Para llevarlo a cabo, se define un tramo del texto como una secuencia de documentos consecutivos, el tramo se hace deslizar a lo largo del texto, calculando en cada posición, los siguientes valores: *Novedad*, se completa el tramo con los documentos inmediatamente anteriores, se cuenta el vocabulario de ambos para detectar las palabras nuevas que aparecen en el tramo, se compara su valor con el obtenido del mismo modo pero utilizando la triple fórmula de Heaps en lugar de contar.

Para hallar la *Novedad Inversa* se utiliza el mismo proceso pero recorriendo el texto del final al principio y para hallar la *Novedad Actual* es la diferencia entre la *Novedad* y la *Novedad Inversa*. Valores altos de la *Novedad Actual* indican que aquí aparecen muchas palabras nuevas y en los documentos sucesivos aparecen menos.

La *Evolución de la Novedad* es la diferencia entre la *Novedad Actual* de un tramo y la del tramo anterior en el deslizamiento. Valores altos indican que empieza un subtexto de distinta temática.



A continuación se parte el texto por los 6 puntos de mayor variación de novedad (siempre que los tramos obtenidos sean suficientemente grandes). En cada una de las partes obtenidas se realiza el mismo proceso y así sucesivamente. Dividimos el texto semánticamente.



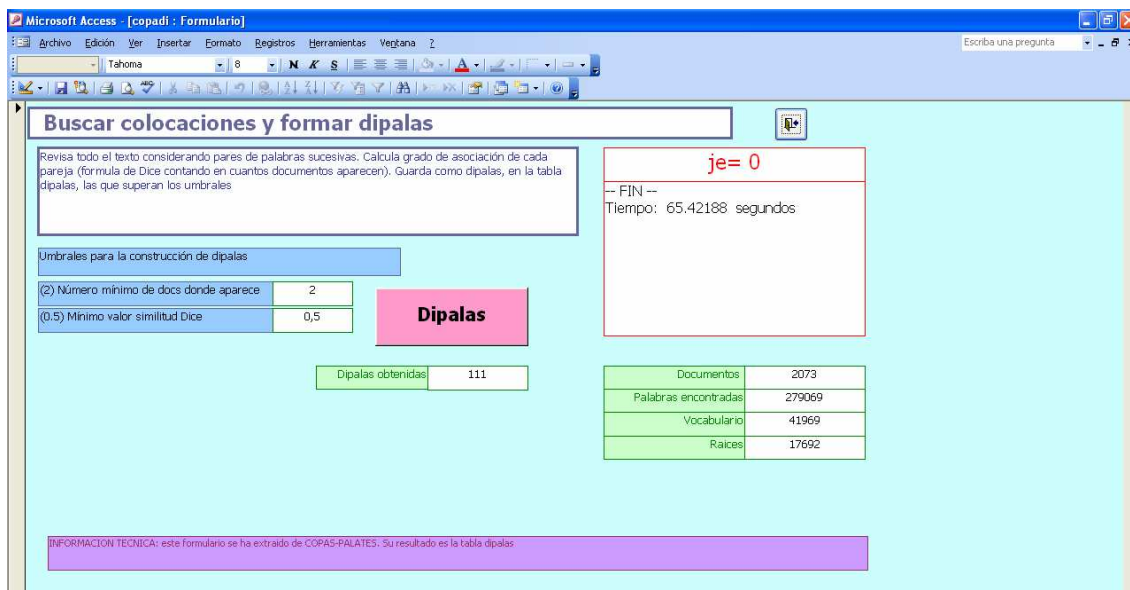
Se localizan las colocaciones y se construye la tabla *di-palas* para albergar las colocaciones o di-gramas encontradas, Se obtiene la frecuencia de aparición en el texto, el valor de similitud según Dice (1945)⁵⁸ y la frecuencia de la raíz en documentos

⁵⁸ El valor de la similitud entre palabras formulada por Dice (1945), ofrece valores entre 0-1, siendo 1 el valor de similitud más alto para dos palabras

distintos. Las colocaciones se extraen de textos con granularidad más pequeña y no en documentos de 300 caracteres como decíamos en el capítulo 8, es por ello que se obtienen pocas di-gramas o colocaciones. Para localizar las colocaciones se calcula las similitudes de todas las palabras del texto junto con las palabras que le siguen, y se comprueba cuales son colocaciones importantes para la temática del texto, para escoger estas colocaciones importantes se unen con un guión y de este modo pasan a ser una palabra como por ejemplo: *Comunidad-Valenciana*.

pala1	pala2	fr	Dice	frd
acido	carbonico	6	0,5217391	3
aeso	isona	3	0,75	3
altamira	buylla	5	0,7692308	5
alto	aragon	123	0,5245203	95
alvarez	ossorio	5	0,625	3
amigables	componedores	3	0,8571429	2
ammiano	marcelino	3	0,8571429	3
animadas	moleculas	3	0,8571429	3
arterial	hidraulico	7	0,6086956	7
asa	comerlo	3	0,6666667	3
biblioteca	virtual	77	0,9112426	76
cabaña	asa	3	0,5454546	3
carnero	prendido	3	0,75	3
cavite	santiago	7	0,56	7
cedulas	hipotecarias	7	0,56	5
comerlo	indio	3	0,6	3
comprometer	arbitros	3	0,6	2
conejar	gallinero	3	0,75	3
correos	telegrafos	3	0,6666667	3
corruptio	optimi	3	1	2
costa	joaquin	155	0,5081967	75
cuencas	hidrograficas	7	0,5185185	7
cursus	publicus	5	0,7142857	4
deidades	gentilicas	4	0,5333334	4
deliciosa	visualidad	3	0,75	3
derriba	machetazo	3	0,6666667	3
didymo	veriniano	3	0,75	3
encajonan	oprimen	6	0,8571429	6
encina	trufera	7	0,5185185	4
encorvado	bestia	3	0,6	3
esencias	resinosas	3	0,6	2
esparto	halfa	3	0,5454546	2
estados	unidos	63	0,6395939	49
eusebio	hieronymo	7	0,7777778	6
explicita	implicitamente	3	0,6	3

Tabla 35. Colocaciones y valor de similitud de Dice



Seguidamente se divide el texto en las partes temáticas (*Text Segmentation*), y para ello se seleccionan los “*Términos*”, que formarán parte de la Indización y Segmentación Automática, los cuales para ser considerados Términos deben cumplir las siguientes premisas:

- ser raíces importantes
- ó ser di-gramas, colocaciones seleccionadas y por consiguiente unidas con un guión
- cumplir el umbral de significación

Para cumplir el umbral de significación los nuevos Términos se basan en la siguiente formulación:

$$idf + A \cdot \frac{fr}{frd}$$

Donde,

A es un parámetro variable

fr es la frecuencia en el documento de la palabra

frd es el número de documentos en los que aparece la raíz

Los nuevos términos pasan a dos tablas denominadas *terdistin* y *tertodos* donde se ubica la nueva numeración que le corresponde a cada término (*nter*), la raíz obtenida de la palabra (*pa*), la frecuencia de la raíz en los documentos distintos (*frd*), el término (*pala*), la frecuencia del término en el documento (*fr*), el valor obtenido del *Idf* (*Inverse Document Frequency*), (*idf*), el rango normalizado (*ran*) y el rango normalizado de cada término en cada nivel (*iin*).

nter	pa	frd	fr	pala	idf	ran	iin
1	agricultura	243	457	agricultura	1,906948	7,590802	7,202044
2	costa_joaquin	75	155	costa_joaquin	3,082521	6,704071	7,216196
3	joaquin	112	205	joaquin	2,681511	8,276586	6,95595
4	costa	202	405	costa	2,091742	8,702365	7,580373
5	eam	3	4	eam	6,301397	8,264945	5,89

nter	pa	frd	fr	pala	idf	ran	iin
6	qua	11	12	qua	5,002114	7,793891	5,314404
7	vob	6	7	vobis	5,60825	8,15261	7,779784
8	herbam	2	2	herbam	6,706862	9,747356	7,730711
9	genesis	9	9	genesis	5,202785	7,859041	6,7841
10	introd	10	11	introduccion	5,097425	8,90564	7,290061
11	frut	232	428	fruto	1,953272	4,40518	7,219972
12	res	8	8	resinas	5,320568	7,861403	4,64456
13	fibr	8	9	fibra	5,320568	9,06395	6,846287
14	textil	4	4	textiles	6,013715	9,304595	8,236356
15	calor	78	131	calor	3,043301	6,711368	7,229645
16	plant	247	443	planto	1,890621	7,594487	7,659745
17	animal	38	51	animales	3,762423	7,183078	6,802033
18	metamorfose	9	9	metamorfosea	5,202785	8,221792	6,874507
19	inert	8	8	inerte	5,320568	8,703928	7,597758
20	muscul	14	17	musculo	4,760952	7,286857	6,811967
21	nerv	10	10	nervio	5,097425	7,86376	6,680719
22	chisp	7	7	chispa	5,454099	8,28406	6,402618
23	electr	20	28	electrica	4,404277	8,223573	6,956427
24	sex	7	7	sexo	5,454099	6,718624	7,242977
25	embr	5	5	embrion	5,790572	8,225349	8,106361
26	calienta	7	7	calienta	5,454099	8,751415	7,524792
27	rede	7	7	redes	5,454099	9,065768	6,982869
28	sav	13	14	savia	4,83506	8,266312	4,884295
29	canales	151	331	canales	2,38273	8,705489	7,256193
30	laborator	6	6	laboratorio	5,60825	9,136222	6,903362
31	labra	15	18	labra	4,691959	8,707048	7,269296
32	hues	10	11	hueso	5,097425	7,296519	6,832202
33	glut	2	2	gluten	6,706862	9,74754	7,759398
34	gras	10	13	grasa	5,097425	8,708607	7,237701
35	almid	3	3	almidon	6,301397	9,068492	7,008901

Tabla 36. Términos de indización

Selecionar y numerar términos

Términos son raíces o dipalas que superen el umbral de frecuencia en documentos y el umbral de significación: $idf+A \times \text{frecuencia en el texto/numero de documentos en los que aparece}$

Una vez escogidos los términos, se les asigna representación numérica, necesaria para el proceso de buscar palabras asociadas. Quedan en tablas tertodos y terdistin

Parámetros para selección de términos

Umbral de frec en docs, > ... (>1)	
Umbral de significación > ... (>10)	
Parámetro A ... (=5)	

Términos

Número ter-doc	58189
Términos distintos	7457

Construidas tablas de terminos

--FIN--

INFORMACION TECNICA: este formulario se ha extraido de COPAS-PALATES. Su resultado son las tablas tertodos y terdistin

Seguidamente se calcula la matriz de similitudes, se seleccionan términos según frecuencia en documentos, se calcula la similitud de Dice de todas las parejas (en orden adecuado para ejecutarse de manera eficiente). De las parejas encontradas se seleccionan las parejas que superen el umbral de similitud. El resultado queda en tabla *sii* (*similitud*), donde los términos se designan por su número. Todo este proceso es numérico, sin intervención de los caracteres alfabéticos.

Calcular similitudes entre términos

Para procesar, selecciona términos según frecuencia en documentos. Calcula similitud Dice de todas las parejas (en orden adecuado para ejecutarse de manera eficiente). Para grabar, selecciona parejas que superen el umbral de similitud. El resultado queda en tabla *sii*, donde los términos se designan por su número (Todo este proceso es numérico, sin intervención de los caracteres alfabéticos)

Documentos	2073
Total términos	58189
Términos distintos	7457

Umbral:

(>1) Min Frec en docs:	1
(<15%) Max % en docs:	15
(0,05) (>) Similitud:	0,05

Calcular

Parejas encontradas	99302
Parejas guardadas	61844

Términos asociados
64.9375 segundos

--FIN--

INFORMACION TECNICA: este formulario se ha extraido de COPAS-SMITES. Su resultado queda en la tabla *sii*. Como todo el proceso se realiza en memoria puede haber limitación de tamaño, en cantidad de documentos o número total de términos: revisar declaración de matrices globales

El siguiente paso es la construcción de Clusters o grupos de Términos, para ello dentro de cada nivel se agrupan los clusters de un modo sencillo. Dentro de cada nivel obtenemos una cantidad de Términos y agrupamos dichos Términos calculando la máxima cohesión.

Se utiliza, en cada subdivisión, el concepto de rango normalizado según la representación Log%. Para calcular el valor del rango normalizado, una vez situados los términos en cada nivel, colocamos los nuevos términos o parejas (di-gramas) en el

esquema de niveles utilizando la representación Log-% que se desarrolla ampliamente en el Capítulo 6, los rangos normalizados toman valores entre 0-10, en esta representación se distribuirán todos los Términos desde el rango 0 – 0,5 que serán los Términos que conformarán el Nivel 1, y así sucesivamente. De cada Término que sale en los rangos normalizados de 0-5 cogemos sólo los que son importantes en ese tramo del documento y luego vemos esos mismos términos en el siguiente subnivel. Pero ya no cogemos esos términos sino los que quedan más a la derecha (Rangos normalizados de (5-...)). Se almacenan en la tabla *niveter* la cual contiene el nivel, el Término, el rango normalizado, que se calcula con la siguiente formulación:

$$10 \cdot \frac{\log r}{\log V}$$

Y el rango normalizado relativo el cual es el cociente del rango normalizado y el nivel inmediato superior (más grande), el resultado debe ser menor que 1 para coger el Término, ya que si fuera mayor que 1 significa que dicho Término es importante en el nivel superior y entonces no nos interesa para el nivel en el que estamos, es decir, el nivel inmediato.

El resultado queda en la tabla *niveter*:

ni	nter	ran	ranrela
0	1	2,005242	2,005242
0	4	2,780976	2,780976
0	11	2,576931	2,576931
0	16	2,459017	2,459017
0	29	3,234751	3,234751
0	45	3,030706	3,030706
0	47	2,177759	2,177759
0	52	1,551467	1,551467
0	130	3,102934	3,102934
0	157	1,229509	1,229509
0	173	2,953493	2,953493
0	216	0	0
0	361	0,7757335	0,7757335
0	519	2,870555	2,870555
0	653	3,352665	3,352665
0	1024	2,683597	2,683597
0	1385	2,327201	2,327201
0	3707	3,170782	3,170782
0	6288	1,801198	1,801198
0	6319	3,29526	3,29526
1	15	4,135928	0,9180832
1	16	0	0
1	17	4,40995	0,7655097

Tabla 37. Términos de indización en cada nivel

La densidad pretende caracterizar la intensidad de las relaciones que unen las palabras que componen un cluster determinado. Los clusters pueden ser colocados por orden de densidad creciente.

Según Ruiz-Baños y Contreras-Cortés (1998) la centralidad o índice de cohesión externa es la suma de los índices de equivalencia de todos los enlaces externos que posee el tema con otros. Y el concepto de densidad o índice de cohesión interna lo define como la intensidad de las asociaciones internas de un tema y representa el grado de desarrollo que posee.

Dentro de los futuros trabajos de esta Tesis Doctoral se contempla el aplicar la centralidad a los clusters.

Finalmente se realiza la Indización y Segmentación del texto, se muestra el resultado de los clusters mediante la estructura arbórea, los Términos indizados se muestran en mayúsculas los que se corresponden con los clusters formados y en minúsculas los Términos complementarios que son más importantes en ese nivel y no en el nivel anterior.

La distribución del texto en niveles y los Términos que definen su temática se almacena en las tablas *nives*, *niveter* y *terdistin*, las cuales se pueden consultar en páginas anteriores.

La distribución del texto en niveles y los términos que definen su temática está en las tablas *nives*, *niveter* y *terdistin*

Desde	Hasta	Total	Subdivisiones
1	33	33	0
33	129	97	7
129	855	727	7
855	1115	261	7
1115	1873	759	6
1873	2073	201	7
-	-	-	-

Documentos

Desde	Hasta	Total
1	2073	2073

EsquemaTexto

PLANTO ANIMALES CULTO HUEVA PECES VEGETALES DOMESTICA ARBUSTO VEGETA AGUA VERGEL OLIVO RIO ARADO ROCA COSTUMBRE VIENTO AREN POZOS_ARTESIANOS LAGO CHARCA

(OASIS SAHARA)
AGRICULTOR PLANTO CULTO VEGETA OLIVO ARADO ARTIFICIALES SUD PATRIA CEREALES TRIGO INDIGENA PALMA DESIERTO LLUVIAS ROCA VIENTO AREN POZO ESTADOS_UNIDOS FERROCARRIL SUBTERRANEA HECTOLITRO AMERICANA ZONA

(OLIGARQUIA CACIQUES OLIGARCA) (OPOSICION OPOSITOR)
AGRICULTOR LEY TIPO GOBERNAR LEYES ALIMENTO IGNORO PATRIA SOBERANIA REGISTRA EDUCA ESCUELAS LECCION SEQUE PAG REY ESTUDIOS GOBIERNOS DERECHO ELECTORAL CONTESTA METODO CRISIS POLITICO TITULO MAESTRA VOTO ESCUELA NIÑO

(CANALS RIEGO SOBRARBE CAÑO TAMARITE)
AGUA RIO EMIGRA SEQUE HAMBRE ACEQUIA NIEVA REGADA FERR OCARRIL VAPOR CONSTRUIR GOBIERNOS POLITICO REGAR EBRO HIDRAULICO ZONA COMISO PESETA ARAGONES MORET LITERA ESERA

(PRUDENCIA PAULINA) (NUMISIO THEODOSIO)
LEY RIO GOBERNAR LEYES ENFERMA IGNORO CESAR ESCLAVO VINO COSTUMBRE VIENTO AIRE PAG DIOSA REY GRATO INDIVIDUAL CODIGO DERECHO MAXIMA PREGUNTO LEGISLA ACEPTA REPUBLICA TITULO RESTALURO SUR APROBO MAESTRA ESTATUA NIÑO

(FRUTO PLANTO ARBOL)
CULTO INERTO COSECHA ENCINAS COLONO CASTAÑO CEREALES TRIGO PROPIETARIO RECOLECCION HECTAREA METRO HORTALIZAS ALMENDRO HUERTA EXPORTA MELOCOTONES ALBARICOQUES FINCA PESETA MANZANA

11. Discusión de Resultados

Se enumeran los resultados siguiendo con la estructura y nomenclatura habitual utilizada en las hipótesis y objetivos y que también se manejará en las conclusiones, sin duda esta estructura creemos ayudará al lector a ubicar los resultados obtenidos de cada una de las experimentaciones que se han realizado en cada capítulo a lo largo de esta tesis doctoral. Recordamos que la nomenclatura se compone en primer lugar del número del capítulo correspondiente y en segundo lugar del número de resultado dentro de dicho capítulo.

Respecto a los resultados obtenidos de la aplicación del modelo de crecimiento del vocabulario aportando críticas y una nueva versión de la Ley de Heaps:

- 5.1** Puede definirse un procedimiento riguroso para calcular los parámetros con gran precisión (utilizando promedios de vocabulario en muchos fragmentos). De esta forma los parámetros obtenidos constituyen una característica del texto y no del procedimiento.
- 5.2** Aplicada la fórmula de Heaps (con los parámetros obtenidos con gran precisión) se observan discrepancias sistemáticas: no es un modelo totalmente adecuado. Sin embargo, estas discrepancias se atenúan enormemente si se utiliza una doble fórmula de Heaps, o sea dos juegos de parámetros para dos rangos de valores.
- 5.3** Distintas formas de cálculo sincrónico y promediado de los parámetros de Heaps a un mismo texto conducen a resultados distintos. Esta anomalía viene motivada por la distinta elección de los fragmentos de tamaño intermedio para realizar el cálculo. (Método sincrónico, método que es independiente de las sucesiones temáticas del texto). Su efecto sobre el modelo de crecimiento no es muy importante ya que distintas funciones potenciales, sobre un cierto intervalo de valores, pueden proporcionar resultados muy parecidos. Por otra parte con el método sincrónico quedan unificadas las distintas formas de cálculo.
- 5.4** Existe una relación directa entre los valores de los parámetros de Heaps y las características propias de la tipología de los textos (literarios, científicos, etc); puede ponerse de manifiesto claramente con una representación gráfica⁵⁹.
- 5.5** Admitiendo poca precisión en los parámetros de Heaps, resultan muy predecibles: es posible dar reglas sencillas para obtenerlos sin apenas cálculos. Es decir, sin hacer cálculos encontramos la función potencial que mejor sirve para la interpolación.

⁵⁹ Véase figura 9, pág. 119.

- 5.6** Aplicada la fórmula de Heaps en un rango de valores superior al texto de donde se ha obtenido, se observan discrepancias sistemáticas. Pueden corregirse cambiando el modelo de Heaps por otras fórmulas similares que atenúen el crecimiento.

Respecto a los resultados obtenidos de la aplicación del modelo de distribución de frecuencias de Zipf para aportar nuevas fórmulas sobre el crecimiento del vocabulario en los textos:

- 6.1** Respecto a la interpretación de los parámetros de la fórmula de Zipf-Mandelbrot: en igualdad de otras condiciones, a mayor valor del exponente, mayor es la frecuencia de las palabras más frecuentes y menor la de las menos frecuentes.
- 6.2** El sumando de Mandelbrot predice menores frecuencias de las palabras más frecuentes, lo que quedaría compensado con un aumento del exponente; en este caso son las frecuencias intermedias las que aumentan.
- 6.3** Se demuestra con un amplio estudio que el valor del exponente (e) en las fórmulas de Zipf y Mandelbrot depende estrechamente de la tipología del texto (literario, científico, etc.) del tamaño del texto y del tratamiento o analizador léxico que se utilice en el texto.
- 6.4** Respecto a la desviación de los textos reales en la distribución de frecuencias derivada de la fórmula de Zipf-Mandelbrot a la que quedan ajustados, de forma experimental hemos demostrado con carácter muy general, es decir para todos los tipos de texto, que para frecuencias altas la predicción de Zipf queda por encima del texto real, para frecuencias intermedias, las discrepancias se invierten y para frecuencias bajas la predicción de Zipf-Mandelbrot queda por encima de los textos reales.
- 6.5** Se proporciona una forma visual (Modelo Log-%) de comprobar las cantidades relativas de palabras de distintos tipos, en cuanto a frecuencias presentes en un texto. Esta visualización muestra el cumplimiento en líneas generales de la Ley de Zipf, pero también su incumplimiento parcial, obteniendo así valoraciones realistas.

Respecto a los resultados obtenidos de la aplicación de un método de lematización de las palabras:

- 7.1** Desarrollamos un método de Lematización o Stemming denominado TCP-%, que extrae la raíz de una palabra según el porcentaje de la palabra a obtener; éste puede ser más o menos agresivo dependiendo del porcentaje que le indiquemos para así realizar las diferentes experimentaciones con los textos.

- 7.2** Realizamos modificaciones amplias a diversos Stemmers, como el método de variedad de Sucesores (Hafer, Weiss, 1974) y la adaptación al Español del Método de Sufijos de Lovins (1968).
- 7.3** Para exponente (e) de Zipf calculado sobre todos los valores de un texto, se obtienen distintos valores pero que son parecidos entre sí, independientemente del Stemmer utilizado y de si éstos son más agresivos o menos.
- 7.4** Afinando un poco más, tras varias experimentaciones realizadas comprobamos que a pesar del resultado anterior (7.3) si el Stemmer es más agresivo (menos raíces resultantes) se obtiene mayor valor del exponente (e) de Zipf y lo mismo le pasa al \sum de la fórmula de Mandelbrot.

Respecto los resultados obtenidos del estudio cuantitativo del concepto de palabras asociadas:

- 8.1** Se demuestra que sí afecta la granularidad del texto en la cantidad total de pares de palabras encontradas, incluso aumenta el valor del exponente (e) de la función potencial respecto a palabras únicas.
- 8.2** Evidentemente con textos más grandes se obtienen más palabras asociadas. El número de palabras asociadas crece con el tamaño del texto siguiendo una función potencial con exponente (e) superior a Heaps.
- 8.3** Aplicando adecuadamente un umbral de similitud de Dice (1945), el número de parejas resultantes tiende a ser independiente del tamaño del texto.
- 8.4** Aplicando adecuadamente un umbral, el espectro de similitudes muestra gráficamente la existencia de colocaciones o colocations.

Respecto a los resultados obtenidos de la implementación de un método de segmentación automática del texto visualizando las fórmulas obtenidas anteriormente sobre crecimiento del vocabulario:

- 9.1** Se demuestra que las discrepancias del texto respecto a los parámetros de la Ley de Heaps contribuye para obtener algo práctico: la novedad de las palabras en un texto. Con la fórmula de Heaps predecimos cuantas palabras debería haber y contamos las que realmente hay, así obtenemos la diferencia y obtenemos la novedad de palabras de un texto.
- 9.2** Una vez implementado el Sistema de Segmentación funciona adecuadamente respecto a nuestras expectativas.

Respecto los resultados obtenidos de la implementación de un prototipo de Sistema de Indización y Segmentación Automática:

- 10.1** El resultado final es un prototipo de Sistema de Indización y Segmentación Automática que dibuja un esquema en niveles jerárquico con los Términos temáticos obtenidos automáticamente del texto y el cual compila todos los resultados objeto de estudio consiguiendo un funcionamiento razonable sin añadir las potentes herramientas lingüísticas ya conocidas (véase capítulo 3.2) y utilizando exclusivamente técnicas numéricas o cuantitativas.

12. Conclusiones

Esta tesis doctoral se fundamenta en la realización de diversos estudios cuyos resultados convergen finalmente en el último capítulo de esta investigación, en el que se desarrolla un prototipo de Sistema de Indización y Segmentación Automática para textos en Español. Sin cada uno de estos estudios no habría sido posible la consecución final del software de dicho Sistema.

De este trabajo de investigación pueden extraerse diversas conclusiones que pueden dividirse en una conclusión general y en conclusiones específicas sobre cada uno de los estudios de los métodos cuantitativos y leyes clásicas en recuperación de información que se han examinado con sumo detalle. Se ha hecho un recorrido transversal por varios de los modelos y leyes clásicas en recuperación de información, como son los modelos relativos al proceso de repetición de palabras (Zipf, 1949), (Mandelbrot, 1953) y al proceso de creación de vocabulario (Heaps, 1978), que contribuyen a los métodos algorítmicos, tratando de encontrar mejoras mediante modificaciones o ampliaciones significativas. Todo ello trabajando siempre con textos relativamente largos y en idioma Español, para dar solución a los problemas específicos que en este contexto puedan plantearse.

La conclusión general alcanzada es que hemos perfeccionado métodos cuantitativos que pueden ser utilizados en distintos problemas o aplicaciones del Tratamiento de la Información, en particular lo hemos aplicado a desarrollar un Sistema de Indización y Segmentación Automática comprobando que los resultados son aceptables aún sin utilizar los métodos lingüísticos.

Tras la demostración teórica y la implementación de un prototipo de software de un Sistema de Indización y Segmentación Automática, a grandes rasgos éste realiza las siguientes funciones: procesa el texto, genera la representación numérica de las palabras de un texto para agilizar el proceso de las bases de datos, detecta las variaciones de novedad de las palabras, divide el texto en partes temáticas automáticamente, se buscan las *colocations* y se construye la tabla de éstas, selecciona los Términos que formarán parte del sistema, los Términos seleccionados cumplen las siguientes premisas: son raíces o di-gramas con alta frecuencia en el texto, cumplen el umbral de significación y tienen valores altos de similitud según Dice (1945), se obtiene la matriz de similitudes, se colocan los Términos en el esquema de niveles utilizando la representación Log-% con rango normalizado (0-10), se agrupan los Términos en grupos o clusters calculando la máxima cohesión y finalmente se obtiene la indización y segmentación automática.

En definitiva este Sistema compila todos los resultados objeto de estudio consiguiendo un funcionamiento razonable sin añadir las potentes herramientas lingüísticas ya conocidas (véase capítulo 3.2) y utilizando exclusivamente técnicas numéricas o cuantitativas.

Recordamos que la nomenclatura se compone en primer lugar del número del capítulo correspondiente y en segundo lugar del número de conclusión dentro de dicho capítulo.

Respecto a las conclusiones específicas sobre el modelo de crecimiento del vocabulario aportando críticas y una nueva versión de la Ley de Heaps:

- 5.1 Utilizando los valores obtenidos de forma elemental para los parámetros de Heaps, su variabilidad impide la deducción de características de un fragmento del texto.
- 5.2 Los parámetros obtenidos con gran precisión (promedio de vocabularios de muchos fragmentos y doble o triple fórmula de Heaps permiten en su aplicación a un fragmento del texto, interpretar la discrepancia en un vocabulario como una característica propia del fragmento.
- 5.3 Las características propias de un fragmento obtenidas de este modo, tienen relación con propiedades estructurales o semánticas del texto.
- 5.4 Aprovechamos las discrepancias del texto respecto a los parámetros de la Ley de Heaps para obtener algo práctico (Novedad de las palabras).

Respecto a las conclusiones específicas sobre el modelo de distribución de frecuencias de Zipf para aportar nuevas fórmulas sobre el crecimiento del vocabulario en los textos:

- 6.1 Los parámetros resultantes de ajustar la fórmula de Zipf- Mandelbrot a un texto proporcionan información cuantitativa sobre los distintos tipos semánticos de palabras en relación con su frecuencia en el texto.
- 6.2 Los textos reales manifiestan una tendencia estable a apartarse de las predicciones teóricas de Zipf-Mandelbrot, por lo que la información obtenida de ella es utilizable con la modificación adecuada.
- 6.3 De una forma sencilla y automática pueden extraerse tramos de rangos de palabras en la cantidad y con las características deseadas para la posterior construcción de un sistema de tratamiento de la información.

Respecto a las conclusiones específicas sobre el proceso de un método de lematización de las palabras:

- 7.1 Demostramos fehacientemente que la utilización de lematizadores o Stemmers sobre un texto sólo afecta a los tramos inicial y final de la pendiente en la distribución de frecuencias de Zipf y Mandelbrot, es decir sólo afecta a las palabras de frecuencias altas y bajas y no a las intermedias.
- 7.2 Realizando un ajuste parcial a la distribución de frecuencias de Zipf por tramos a la parte central Punto de Transición (PT) (Booth, 1967) queda demostrado que si ajustamos Zipf a las raíces de la parte central de la distribución de frecuencias éste es independiente del Stemmer utilizado.

Respecto a las conclusiones específicas sobre el estudio cuantitativo del concepto de palabras asociadas con vistas a extraer conclusiones semánticas:

- 8.1** A través del espectro de similitudes podemos localizar las collocations, lo cual es imprescindible para el manejo de cualquier sistema de tratamiento de la información.
- 8.2** Transformadas las collocations en palabras ordinarias y rehecho espectro de similitudes. Disponemos de una potente herramienta de términos en base a la cohesión de clusters.

Respecto a las conclusiones específicas para implementar un método de segmentación automática del texto visualizando las fórmulas obtenidas anteriormente sobre crecimiento del vocabulario:

- 9.1** Aplicando la forma perfeccionada de la Ley de Heaps (triple función potencial) a fragmentos tomados con el criterio diacrónico y comparado con el recuento real de palabras, la discrepancia encontrada, el exceso encontrado a veces nos indica la cantidad de palabras nuevas.
- 9.2** Perfeccionando este método de distintas maneras se obtiene un procedimiento efectivo de Segmentación Automática.

Respecto a las conclusiones específicas para implementar un prototipo de Sistema de Indización y Segmentación Automática:

- 10.1** Todos los resultados de los capítulos cinco a nueve de esta tesis doctoral convergen en el Sistema de Indización y Segmentación Automática desarrollado.

BIBLIOGRAFÍA

ADRIANO DE JESÚS HOLANDA. Et. al. (2003). Basic word statistics for information retrieval: thesaurus as a complex network. *Workshop em Tecnologia da Informaçao e da Linguagem Humana*, Brasil: October 12th.

ALONSO RAMOS, M. (1993). Las funciones léxicas en el modelo lexicográfico de I. Mel'cuk. Tesis doctoral, Madrid: UNED.

ARANO, Silvia. "La ontología: una zona de interacción entre la Lingüística y la Documentación" [en-línea]. *Hipertext.net*, núm. 2, 2003. <<http://www.upf.edu/hipertextnet/numero-2/ontologia.html>> [ref. de 08 de noviembre 2012].

BAAYEN, R. HARALD. (1991). A stochastic process for word frequency distributions. *Annual Meeting of the ACL. Proceedings of the 29th annual meeting on Association for Computational Linguistics*. California, p. 271-278

BAAYEN, R. HARALD. (2001). Word frequency distributions, Dordrecht: Kluwer Academic Publishers.

BAEZA-YATES, R.; RIBEIRO-NETO, BERTHIER. (1999). Modern Information Retrieval. Harlow: Addison-Wesley.

BELKIN, N. J., BRUCE C.W. (1987). Retrieval techniques, *Annual of Information Science and Technology*, vol. 22, p. 109-145

BELY, N. Et. al. (1970). *Procedures d'analyse sémantique appliqués a la documentation scientifique*, París: Gauthier Villar.

BENSON, M. Et.al. (1986). BBI Combinatory Dictionary of English: A Guide to Word Combinations, Amsterdam, John Benjamins

BERNERS-LEE, T. (1999). Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor. San Francisco: Harper

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. (2001). The semantic web. *Scientific American*, May, vol. 284, n. 5, p. 34-43

BLOOMFIELD, M. (1970). Evaluation of Indexing. *Special Libraries*. vol. 61, n. 8, p. 429

BOOTH, A.D. (1967). A Law of Occurrences for Words of Low Frequency. *Information & Control*, vol. 10, n. 4, p. 386-393

BRADFORD, S. C. (1948). *Documentation*. London: Crosby Lockwood.

BRONLET, PH. & AUSLOOS, M. (2003). Generalized (m, k)-Zipf law for fractional Brownian motion-like time series with or without effect of an additional linear trend. *International Journal of Modern Physics C*, vol. 3 n. 14 p. 351-365

CALLON, M; COURTIAL, J-P; HERVE, R. (1995). *Cienciometría: el estudio cuantitativo de la actividad científica*. Gijón: Trea.

CARPENTIER, ELISABETH. (1982). Histoire et informatique: Recherches sur le vocabulaire des biographies royales françaises. *Catris de Civilisation Medievale*. vol. 25, p. 3-30

CHOMSKY, N. (1957). *Syntactic Structures*. The Hague, Mouton.

CHOMSKY, N. (1965). *Aspects of the theory of Syntax*. Cambridge: M.I.T.Press.

CHU, HETING.; ROSENTHAL, MARILYN. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. *ASIS Annual Conference Proceedings*. October 19-24.

CHURCH, KENNETH W. AND HANKS, PATRICK (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol.16 n. 1, p. 22-29

CLEVERDON, C.W. (1967). The Cranfield Test on Index Language Devices. *ASLIB Proceedings*, vol. 19, n. 6, p. 173-194

CODINA, L. Y PEDRAZA-JIMÉNEZ, R (2011). Tesauros y ontologías en sistemas de información documental. *El profesional de la información*, septiembre-octubre, vol. 20, n. 5, p. 555-563

CRUSE, D.A. (1986). *Lexical Semantics*. Cambridge, Cambridge UP.

CURRÁS, EMILIA. (1998). *Tesauros: manual de construcción y uso*. Madrid : Kaher II.

DAWSON, J. (1974). Suffix Removal and Word Conflation. *ALLC Bulletin*, vol. 2, n. 3, p. 33-46

DEBOWSKI, LUKASZ (2002). Zipf's law against the text size: a half-rational model. *Glottometrics*, vol. 4

DEERWESTER, S., Et. al. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. 41, p. 391-407.

DENISOV, S. (1997). Fractal binary sequences: Tsallis thermodynamics and the Zipf law. *Physics Letters A*, vol. 235 p. 447-451

DICE, LEE R. (1945). Measures of the amount of ecologic association between species. *Ecology*, vol. 26, p. 297-302

- DUNNING, TED (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Association for Computational Linguistics*, vol. 19 n.1, p. 61-76
- EVERT, STEFAN (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations, Universität Stuttgart: Institut für maschinelle Sprachverarbeitung.
- EVERT S, BARONI M. (2007). ZipfR: Word frequency distributions, *Proc 45th Ann Meeting of the Association for Computational Linguistics*. p. 29-32
- FERRER CANCHO, RAMON & RICARD V. SOLÉ. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, vol. 8, p. 165-173
- FERRER I CANCHO R. (2005). The variation of Zipf's law in human language. *The European Physical Journal B*. vol. 44, p. 249-257
- FIGUEROLA, CARLOS G.; ZAZO, ÁNGEL F.; RODRÍGUEZ VÁZQUEZ DE ALDANA, EMILIO, Et. al. (2004). La Recuperación de Información en español y la normalización de términos. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia artificial*, vol. 22, p.135-145
- FIGUEROLA, CARLOS G., RODRÍGUEZ VÁZQUEZ DEL ALDANA, EMILIO, ZAZO, ÁNGEL F., ALONSO BERROCAL, JOSÉ LUIS. (2006). Encontrar documentos a través de las palabras. *Nuestras palabras: entre el léxico y la traducción*, coord. Por María Teresa Fuentes Morán, Jesús Torres del Rey, p. 147-174
- FIRTH, J. R. (1957). A synopsis of linguistic theory 1930–55. *Studies in linguistic analysis*, Oxford.: The Philological Society, p. 1–32
- FOLTZ, P. W. (1990) Using Latent Semantic Indexing for Information Filtering. En R. B. Allen (Ed.) *Proceedings of the Conference on Office Information Systems*, vol. 40-47. Cambridge, MA,: MIT Press.
- FOLTZ, P. W. (1996) Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*. vol. 28, n. 2, p. 197-202
- FONDIN, H. (1977). La structure et le vocabulaire de l'analyse documentaire: contribution pour une mise au point. *Documentaliste*, vol. 14, n. 2
- FRAKES, W. B.; BAEZA-YATES, R. Eds. (1992). *Information Retrieval: Data Structures and Algorithms*. New York: Prentice Hall, Englewood Cliffs.
- GARCÍA GUTIÉRREZ, ANTONIO LUIS (1984). *Lingüística Documental*. Barcelona: Mitre.
- GARCÍA MARCO, F.J. (2007). Ontologías y organización del conocimiento: retos y oportunidades para el profesional de la información. *El profesional de la información*, noviembre-diciembre, vol. 16, n. 6, p. 541-550

GIL LEIVA, I., RODRÍGUEZ MUÑOZ, J.V. (1996). Tendencias en los sistemas de indización automática. Estudio evolutivo. *Revista Española de Documentación Científica*, vol. 19, n. 3, p. 273-291

GIL LEIVA, I., RODRÍGUEZ MUÑOZ, J.V. (1996). El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos. *Revista General de Información y Documentación*, Madrid: Servicio de publicaciones de la Universidad Complutense, vol. 6, n. 2.

GIL LEIVA, I., RODRÍGUEZ MUÑOZ, J.V. (1997). Análisis de los descriptores de diferentes áreas del conocimiento indizadas en las bases de datos del CSIC. *Revista Española de Documentación Científica*, vol. 2, n. 20, p. 150-160

GIL URDICIAIN, B. (1996). Manual de lenguajes documentales. Madrid: Noesis.

GÓMEZ DÍAZ, R. (2005). La lematización en español: una aplicación para la recuperación de información. Gijón: Trea

HAFER, M.; WEISS, S. (1974). Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval*, vol. 10, n. 11-12 p. 371-385

HALLIDAY M.A.K. (1961). Categories of the theory of grammar. *Word*, vol. 17, n. 3, p. 241-292

HALLIDAY M.A.K.; HASAN, R. (1976). Cohesion in English. New York: Longman Group

HARMAN D. (1991). How effective is suffixing?. *Journal of the American Society for Information Science*, vol. 42, n.1, p. 7-15

HARMAN D. (1995). Overview of the fourth text retrieval conference (TREC-4). In D.K. Harman, editor, *The Fourth Text Retrieval Conference*, vol. 4, n. 1-24, Gaithersburg, Maryland: National Institute of Standards and Technology (NIST), Defense Advanced Research Projects Agency (DARPA)

HEAPS, J. (1978). Information Retrieval. Computational and Theoretical Aspects. Academic Press.

HEARST, M.A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, vol. 23, n. 1, p. 33-64

HEINONEN, O. (1998). Optimal Multi-paragraph Text Segmentation by Dynamic Programming. Helsinki: University of Helsinki.

IGLESIA APARICIO, J. Y MONJE JIMÉNEZ, M.T. (2012). *Gestión de la información en la web 2.0*. [curso on-line]. Fundación Germán Sánchez Ruipérez. Centro Internacional de Tecnologías Avanzadas

- IRSULA, J. (1992). Colocaciones sustantivo-verbo. Wotjak, G. (ed), Estudios de lexicografía y metalexigrafía del español actual. *Lexicographica Series Mayor*, Tubinga, Max Niemeyer Verlag , vol. 47, p. 19-167
- JONES, KEVIN P. (1976). Towards a Theory of Indexing. *Journal of Documentation*, vol. 32, p. 118
- KINTSCH, W. (2001). Predication. *Cognitive Science*, vol. 25, p. 173-202.
- KOIKE, K. (2001). Colocaciones léxicas en el español actual: estudio formal y léxico-semántico. Madrid: Universidad de Alcalá de Henares.
- KORFHAGE, R.R. (1997). Information storage and retrieval. Nueva York: John Wiley and Sons.
- KORNAL, ANDRÁS (1999). Zipf's law outside the middle range. *Proc. Sixth Meeting on Mathematics of Language*. University of Central Florida, p.347-356.
- KROVETZ R. (1993). Viewing morphology as inference process, *Proceedings of the 16th ACM/SIGIR Conference*. Nueva York: Association for Computing Machinery, p. 191-202
- LANCASTER, F.W.; FAYEN, E. G. (1973). Information Retrieval on-line. Los Angeles, CA: Melville Publishing Co.
- LANDAUER, TH.; FOLTZ, P. Y LAHAM, D. (1998) An Introduction to Latente Semantic Analysis. *Discourse Processes*. vol. 25, n.2-3, p. 259-284.
- LE QUAN HA, E. I. SICILIA-GARCÍA, JI MING, F. J. SMITH. (2002). Extension of Zipf's law to words and phrases. *Proceedings of International Conference Computational Linguistics (COLING'2002)* p.315-320
- LOTKA, ALFRED J. (1926) The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*. vol.16, n.12, p. 317-323.
- LOVINS, J.B. (1968). Development of a stemming algorithm. *Mechanical translations and Computacional Linguistics*, vol. 11, n. 1-2, p. 22-31
- LUHN, H.P. (1957). A Statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research Development*, vol. 1, n. 4, p.309-317
- LUHN, H.P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, vol. 2 n. 2, p.159-165
- LYRA, M.L. Et. al. (2003). Generalized Zipf's Law in proportional voting processes. *Europhys. Lett.*, vol. 62 p.131
- MALACARNE, LC. Y MENDES, R.S. (2000). Regularities in football goal distributions. *Physica A*, vol. 286 n.1-2, p. 391-395

- MALACARNE, LC., MENDES, R.S., LENZI, E.K. (2002). q-Exponential distribution in urban agglomeration. *Physical Review E*, vol. 65 n.1
- MANDELBROT, B. (1953). An Information Theory of the Statistical Structure of Language. *Communication Theory*, ed. By Willis Jackson, New York: Academic Press, p. 486-502
- MANDELBROT, B. (1975). Les objets fractals: forme, hasard et dimensions, Paris: Flammarion
- MANDELBROT, B. (1982). The fractal Geometry of Nature, New York: W. H. Freeman & Company.
- MARTÍNEZ MÉNDEZ, FRANCISCO JAVIER; RODRÍGUEZ MUÑOZ, JOSE VICENTE. (2004). Reflexiones sobre la evaluación de los Sistemas de Recuperación de Información: necesidad, utilidad y viabilidad. *Anales de Documentación*, nº 7, p. 153-170
- MARTÍNEZ TAMAYO, A.M., Et. al. (2011). Interoperabilidad de sistemas de organización del conocimiento: el estado del arte. *Información, cultura y sociedad*, Instituto de Investigaciones Bibliotecológicas, INIBI. Facultad de filosofía y letras, Universidad de Buenos Aires, ISSN: 1851-1740, enero-junio, n. 24, p. 15-37
- MARON, M., & KUHNS, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, vol. 7 n. 3, p. 216-244
- MARQUET I FERIGLE, LLUÍS (1995). El llenguatge científic i tècnic. Barcelona: Associació d'Enginyers Industrials de Catalunya.
- MATTEO MARSILI Y YI-CHENG ZHANG (1998). Interaction individuals leading to Zipf's Law. *Physical Review Letters*, vol. 80 n. 12
- MÉNDEZ, E. Y GREENBERG, J. (2012). Datos enlazados para vocabularios abiertos y marco general HIVE. *El profesional de la información*, mayo-junio, vol. 21, n. 3, p. 236-244
- MÉNDEZ RODRÍGUEZ, E. (2002). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea.
- MONTEMURRO, MARCELO A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A* vol. 300, n. 3-4, p. 567-578
- MOOERS, CALVIN (1950). Coding, Information Retrieval, and Rapid Selector. *American Documentation*, vol. 1, n. 4, p. 225-229
- MOOERS, CALVIN (1951). Zetocoding Applied to Mechanical organization of Knowledge. *American Documentation*, vol. 2, n. 1

MOREIRO GONZÁLEZ, J.A.; MÉNDEZ RODRÍGUEZ, E. (1999). Lenguaje natural e indización automatizada. *Ciencias de la información*, vol. 30, n. 3, p. 11-24

MOREIRO GONZÁLEZ, J.A.; GARCÍA MARTUL, D. (2005). La visualización de la información en revistas electrónicas mediante la concurrencia de herramientas hipertextuales, mapas conceptuales, topic maps y ontologías. *Proceedings CINFORM - Encontro Nacional de Ciência da Informação VI*, Salvador - Bahia.

MOREIRO GONZÁLEZ, J.A., Et. al. (2008). Los lenguajes documentales en la gestión de la información ¿Un futuro prometedor o recursos del pasado?. *Actas I Encuentro internacional de expertos en teoría de la información. Un enfoque interdisciplinar (CD)*. 6-8 de Noviembre. León: Universidad de León. ISBN-13: 978-84-9773-451-6

MOREIRO GONZÁLEZ, J.A., Et. al. (2009). Actualización del concurso simultáneo en el uso del lenguaje libre y del controlado: folksonomías y taxonomías. *Memoria del I Simposio Internacional sobre Organización del Conocimiento: bibliotecología y terminología, del 27 al 29 de agosto de 2007*. México: UNAM, Centro Universitario de Investigaciones Bibliotecológicas, p. 359-386

MOREIRO GONZÁLEZ, J.A.; SÁNCHEZ CUADRADO, S. Y MORATO LARA, J. (2012). Mejora de la interoperabilidad semántica para la reutilización de contenidos mediante sistemas de organización del conocimiento. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, jan-abr., vol. 17, n. 33, p. 46-58

ORLOV, J.K. (1982). Linguostatistik: Aufstellung von Sprachnormen order Analyse des Redeprozesses? In J.K. Orlov, M.G. Boroda, & I.S. Nadarejsvili (Eds.), *Sprache, Text, Kunst. Quantitative Analysen*, Studienverlag Dr. N. Brockmeyer, Bochum. p. 1-55

PAICE, C. D. (1990). Another stemmer. *ACM SIGIR Forum*, vol. 24, n. 3, p. 56-61

PARRONDO, JUAN M.R. (2003). Números y palabras. *Investigación y Ciencia*, febrero, p. 86-87

PASTOR-SÁNCHEZ, J.A.; MARTÍNEZ-MÉNDEZ, F.J. Y RODRÍGUEZ-MUÑOZ, J.V. (2012). Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. *El profesional de la información*, mayo-junio, vol. 21, n. 3, p. 245-253

PEDRAZA-JIMÉNEZ, R.; CODINA, L. Y ROVIRA, C. (2007). Web semántica y ontologías en el procesamiento de la información documental. *El profesional de la información*, noviembre-diciembre, vol. 16, n. 6, p. 569-578

PORTER, M. F. (1980). An algorithm for Suffix Stripping. *Program*, vol. 14, n. 3, p.130-137

POPESCU I.-IOVITZ, GANCIU M., PENACHE M. C., PENACHE D (1997) On the Lavalette Ranking Law, *Romanian Reports in Physics*, vol. 49, p. 3-27

POPESCU IOAN-IOVITZ (2003). On a Zipf's Law Extension to Impact Factors. *Glottometrics*, vol. 6, p. 83-93

POWERS, DAVID M. W. (1998). Applications and Explanations of Zipf's Law. *New Methods in Language Processing and Computational Natural Language Learning*, ACL, p. 151-160

RÍOS-HILARIO, A.; MARTÍN-CAMPO, D. Y FERRERAS-FERNÁNDEZ, T. (2012). Linked data y linked open data: su implantación en una biblioteca digital. El caso de Europeana. *El profesional de la información*, mayo-junio, vol. 21, n. 3, p. 292-297

ROBERTSON, A., WILLET, P. (1999). Applications of n-grams in textual information systems. *Journal of Documentation*, vol. 54, n. 1, p. 28-47

ROBERTSON, S.E., SPARK JONES, K. (1976). "Relevance weighting on search term" *Journal of the American Society for Information Science*, vol. 27 n. 3, p. 129-146

RODRÍGUEZ LUNA, M. (2002). Stemming Process in Spanish Words with the Successor Variety Method. Methodology and Result. *Fourth International Conference on Enterprise Information Systems*. ICEIS-2002, p. 838-842

ROSENBERG, VICTOR. (1971). A study of statistical measures for predicting terms used to index documents. *Journal of the American Society for Information Science*, vol. 1, n. 22, p. 41-50

RUIZ-BAÑOS, R. y CONTRERAS CORTES, F. (1998). Cómo consultar eficazmente una base de datos bibliográfica. El método de las palabras asociadas, *XIII International Conference of the Association for History & Computing: "History in a new frontier"*, Toledo, 20-23 julio.

SALTON, G. (1968). Automatic information organization and retrieval, Nueva York: McGraw-Hill.

SALTON, G. Ed. (1971). The SMART retrieval system. Experiments in automatic document retrieval, Nueva Jersey: Prentice Hall Inc., Englewood Cliffs.

SALTON, G. (1988). On the relationship between theoretical retrieval models. *Infometrics*, vol. 87-88. Amsterdam: Elsevier, p. 263-270

SALTON, G. (1989). Automatic text Processing: the transformation, analysis and retrieval of information by computer, Massachussets: Addison-Wesley, p.318

SALTON, G., ALLAN, J., AND BUCKLEY, C. (1993). "Approaches to passage retrieval in full text information systems". *Proc. 16th Annual Intl. ACM SIGIR Conf. on R&D in Information Retrieval*.

SALTON, G., MCGILL, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

- SALTON, G., WONG, A., YANG, C.S. (1975). A vector space model for automatic indexing, *Communications of the ACM*, vol. 18, n. 11, p. 613-620
- SÁNCHEZ CUADRADO, S. Et. al. (2007). De repente, ¿todos hablamos de ontologías?, *El profesional de la información*, vol. 16, n. 6, p. 562-567
- SÁNCHEZ CUADRADO, S.; COLMENERO RUIZ, M.J Y MOREIRO, J.A (2012). Tesoros: estándares y recomendaciones. *El profesional de la información*, mayo-junio, vol. 21, n. 3, p. 229-235
- SÁNCHEZ-JIMÉNEZ, R. Y GIL-URDICIÁIN, B. (2007). Lenguajes documentales y ontologías. *El profesional de la información*, noviembre-diciembre, vol. 16, n. 6, p. 551-560
- SANGKON LEE, S. Et al. (2002). Extraction of field-coherent passages, *Information Processing and Management*, vol. 38, p. 173-207
- SAORÍN, T. (2012). Cómo linked open data impactará en las bibliotecas a través de la innovación abierta. *Anuario ThinkEPI*, vol. 6, p. 288-292
- SENSO, J.A. (2003). Herramientas para trabajar con RDF. *El profesional de la información*, marzo-abril, vol. 12, n. 2, p. 132-139
- SOLER MONREAL, M. C. (2009). Evaluación de vocabularios controlados en la indización de documentos mediante índices de consistencia entre indizadores. [Tesis doctoral]. Valencia: Universidad Politécnica de Valencia. Departamento de Comunicación Audiovisual, Documentación e Historia del Arte, DCADHA
- SPARK JONES, KAREN. (1972). "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 28 n.1, p. 11-20
- SPARK JONES, KAREN. (1972). "Experiment in relevance weighting of search term", *Information Processing and Management*, vol. 15 n. 3, p. 133-144
- SPARK JONES, KAREN Ed. (1981). *Information Retrieval Experiment*. London: Butterworths.
- TRAVALIA, CAROLINA (2006). Las colocaciones implícitas. *ELUA. Estudios de Lingüística*, vol. 20, p. 317-332
- TRAVALIA, CAROLINA (2006). Las colocaciones gramaticales en Español. *Anuario de Estudios Filológicos*, vol. XXIX, p. 279-293
- TSALLIS, CONSTANTINO. (2002). Nonextensive statistical mechanics: a brief review of its present status. *Annals of the Brazilian Academy of Sciences*, vol. 74 n. 3, p. 393-414
- TAUBE, MORTIMER. (1955). *The Uniterms System of Indexing Operating Manual*. Washington, *Documentation Inc.*, p. 47

UGARTE, M.D, MILITINO, A.F. (2002). Estadística Aplicada con S-PLUS. 2ª ed. revisada, Pamplona: Universidad Pública de Navarra.

URBIZAGASTEGUI-ALVARADO, R. (1999). La ley de Lotka y la literatura de bibliometría. *Investigación bibliotecológica*, vol. 13, n. 27, p. 125-141

VALERY I. FRANTS, SHAPIRO, J., VOISKUNSKII, VLADIMIR G. (1997). Automated Information Retrieval: Theory and Methods, New York, Academic Press.

VAN DIJK, M., VAN SLYPE, G. (1969). Le service de documentation face à l'explosion de l'information. París: Les Editions d'organisation, p. 53

VAN DIJK, M., VAN SLYPE, G. (1991). Los lenguajes de indización : Concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez: Pirámide.

VAN RIJSBERGEN, C.J. (1979). Information Retrieval. London: Butterworths.

VENEGAS, R. (2003). Análisis Semántico Latente: una panorámica de su desarrollo. *Revista Signos*, vol. 36, n. 53, p. 121-138

WENTIAN LI. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, vol. 38, n. 6, p.1841-1845

ZIPF, G. K. (1949). Human Behaviour and the Principle of Least Effort. Cambridge: Addison-Wesley.

APÉNDICE I: TEXTOS

<i>TIPOS DE TEXTOS</i>	<i>Autor</i>	<i>P</i> <i>(tamaño</i> <i>palabras)</i>	<i>V</i> <i>(vocabulario)</i>
LITERARIOS			
episo1.txt	Benito Pérez Galdós (1843-1920)	81.398	17.217
episogrande.txt	Benito Pérez Galdós (1843-1920)	204.757	29.554
Larra.txt	Mariano José de Larra (1809-1837)	78.719	16.449
larra1.txt	Mariano José de Larra (1809-1837)	76.550	16.521
erudi1.txt	Marcelino Menéndez y Pelayo (1856-1912)	53.940	13.470
calderon.txt	Calderón de la Barca (1600-1681)	172	142
fpoe.txt	Juan del Encina (1468-1529); Jorge Manrique(1440-1479); Marqués de Santillana (1398-1458); Anónimos	609	412
gg1.txt	Gabriel Galán (1870-1905)	6.821	3.171
ksoti.txt	José M ^a de Pereda (1833-1906)	56.854	14.045
ksoti1.txt	José M ^a de Pereda (1833-1906)	28.582	9.370
ksoti2.txt	José M ^a de Pereda (1833-1906)	28.279	8.555
seis.txt	San Francisco de Borja (1510-1572)	25.696	9.319
vari.txt	Varios	95.045	20.132
vari1.txt	Varios	380.054	48.636
castea.txt	Emilio Castelar (1832-1899)	74.939	17.979
casteb.txt	Emilio Castelar (1832-1899)	93.270	20.385
castec.txt	Emilio Castelar (1832-1899)	89.556	16.794
casted.txt	Emilio Castelar (1832-1899)	86.666	19.315
castee.txt	Emilio Castelar (1832-1899)	106.079	16.488
castela.txt	Emilio Castelar (1832-1899)	450.574	43.816

castela1.txt	Emilio Castelar (1832-1899)	112.466	17.097
castela2.txt	Emilio Castelar (1832-1899)	244.324	29.062
castela3.txt	Emilio Castelar (1832-1899)	367.570	38.094
castela4.txt	Emilio Castelar (1832-1899)	450.574	43.816
dos.txt	Emilio Castelar (1832-1899)	73.790	17.171
costa1.txt	Joaquín Costa (1846-1911)	279.141	42.169
costaa.txt	Joaquín Costa (1846-1911)	93.140	20.798
costab.txt	Joaquín Costa (1846-1911)	94.283	20.147
costac.txt	Joaquín Costa (1846-1911)	91.714	24.533
cervantes.txt	Miguel de Cervantes (1547-1616)	385.213	28.317

TIPOS DE TEXTOS	Temática	P (tamaño palabras)	V (vocabulario)
CIENTÍFICOS			
sonia.txt	Artículos temas científicos	32.852	6.348
aralpi.txt	Artículos temas legales	12.862	2.758
bachi1.txt	Artículos temas legales	73.173	9.161
legal1.txt	Artículos temas legales	97.670	11.080
cali1.txt	Artículos ISO	22.133	4.309
cali2.txt	Artículos ISO	11.745	3.054
cali3.txt	Artículos ISO	7.621	2.152
cali4.txt	Artículos ISO	2.155	908
cienti1.txt	Artículos ISO	101.164	13.668
icyt3.txt	Artículos ISO	8.148	2.294
icyt4.txt	Artículos ISO	66.260	10.515
imaa2.txt	Artículos ISO	86.913	13.476
imab2.txt	Artículos ISO	86.349	14.437
imac2.txt	Artículos ISO	46.362	9.545
ime1.txt	Artículos ISO	12.345	2.280
ime2.txt	Artículos ISO	55.696	6.324
imea2.txt	Artículos ISO	95.374	16.448
imeb2.txt	Artículos ISO	81.463	14.721
imec2.txt	Artículos ISO	79.744	14.765
isoc2.txt	Artículos ISO	70.158	13.999
upala.txt	Artículos ISO	76.990	13.273
upalb.txt	Artículos ISO	73.560	13.970
upalc.txt	Artículos ISO	70.523	11.598
medico1.txt	Artículos médicos	67.818	7.170
natalia.txt	Patentes	104.536	16.994
paten1.txt	Patentes	78.514	11.386

Referencias Bibliográficas de la edición manejada en los textos:

Benito Pérez Galdós (1843-1920)

Autor Principal: Pérez Galdós, Benito 1843-1920

Título: Zaragoza

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Notas de la Reproducción Original: Ilustraciones de los Sres. Mérida y Lizcano a partir de la edición del T. III, Madrid, Administración de La Guirnalda y Episodios Nacionales, 1882.

Portal: Biblioteca de Benito Pérez Galdós | Biblioteca Virtual Miguel de Cervantes | Literatura Española

CDU: 821.134.2-311.6"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela histórica española Siglo 19

Autor Principal: Pérez Galdós, Benito 1843-1920

Título: Trafalgar

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Notas de la Reproducción Original: Edición digital a partir de *Episodios Nacionales. T. I*, Madrid, Administración de La Guirnalda y Episodios Nacionales, 1882, pp.5-157.

Portal: Biblioteca de Benito Pérez Galdós | Biblioteca Virtual Miguel de Cervantes | Literatura Española

CDU: 821.134.2-311.6"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela histórica española Siglo 19

Mariano José de Larra (1809-1837)

Autor Principal: Larra, Mariano José de 1809-1837

Título: Ideario español

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2003

Notas de la Reproducción Original: Edición digital basada en la de Madrid, Biblioteca Nueva, 1910.

Portal: Biblioteca de Mariano José de Larra | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-92"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Literatura periodística española Siglo 19

Autor Secundario: González Blanco, Andrés 1888-1924

Autor Secundario: Alomar, Gabriel 1873-1941

Autor Principal: Larra, Mariano José de 1809-1837

Título: El doncel de Don Enrique el Doliente

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Notas de la Reproducción Original: Edición digital a partir de *Obras Completas*, Barcelona, Montaner y Simón, 1886, pp. 77-255.

Portal: Biblioteca de Mariano José de Larra | Novela Histórica Española | Literatura Española | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-311.6"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela histórica española Siglo 19

Marcelino Menéndez y Pelayo (1856-1912)

Autor Principal: Menéndez y Pelayo, Marcelino 1856-1912

Título: La ciencia española : polémicas, indicaciones y proyectos

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Notas de la Reproducción Original: Edición digital basada en la 2ª edición de Madrid, Imprenta Central a Cargo de Victor Saiz, 1879.

Portal: Servicio de Información Bibliográfica y Documental de la Universidad de Alicante | Biblioteca Virtual Miguel de Cervantes | Biblioteca de Marcelino Menéndez Pelayo

CDU: 001:1(460) - Ciencia y conocimiento en general. Organización del trabajo intelectual.

Encabezamiento de materia: Ciencias Filosofía

Calderón de la Barca (1600-1681)

Autor Principal: Calderón de la Barca, Pedro 1600-1681

Título: La vida es sueño

Publicación: [S.l.] : [s.n], 2009

Notas de la Reproducción Original: Edición digital a partir de la edición de Evangelina Rodríguez Cuadros, Madrid, Espasa-Calpe, 1997, 18ª ed.

Portal: Biblioteca de Calderón de la Barca | Biblioteca Virtual Miguel de Cervantes | Literatura Española

Materias:

CDU: 821.134.2-2"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Teatro español Siglo 17

Autor Secundario: Rodríguez Cuadros, Evangelina

Anónimos

Autor principal: Anónimo

Título: Tres morillas me enamoran en Jaén

Gabriel Galán (1870-1905)

Autor Principal: Gabriel y Galán, José María 1870-1905

Título: Epistolario de Gabriel y Galán

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Notas de la Reproducción Original: Edición digital basada en la edición de Madrid, Fernando Fe, 1918. -- Trabajo premiado en el certamen literario de Plasencia.

Portal: Servicio de Información Bibliográfica y Documental de la Universidad de Alicante | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-6"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Cartas Siglo 19. Literatura española Siglo 19

Autor Secundario: Santiago Cividanes, Mariano de

José M^a de Pereda (1833-1906)

Autor Principal: Pereda, José María de 1833-1906

Título: Sotileza

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Notas de la Reproducción Original: Edición digital basada en la de Madrid, Imprenta y Fundición de Manuel Tello, 1885.

Portal: Biblioteca de José María de Pereda | Biblioteca Virtual Miguel de Cervantes | Literatura Española

CDU: 821.134.2-31"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 19

San Francisco de Borja (1510-1572)

Autor Principal: San Francisco de Borja (1510-1572)

Título: Seis tratados muy devotos y útiles para cualquier fiel cristiano por San Francisco de Borja. Sermón sobre San Lucas, 19, 41-42

Emilio Castelar (1832-1899)

Autor Principal: Castelar, Emilio 1832-1899

Título: Crónica Internacional

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Notas de la Reproducción Original: Edición digital a partir de la edición de Dámaso Lario, Madrid, Editora Nacional, 1982.

Portal: Biblioteca Virtual Miguel de Cervantes

CDU: 327"11890/1898"(046) - Política internacional. Relaciones internacionales. Política exterior.

Encabezamiento de materia: Política Internacional 1890-1898 Artículos periodísticos

Autor Principal: Castelar, Emilio 1832-1899

Título: La Hermana de la Caridad

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Notas de la Reproducción Original: Edición digital basada en la de Madrid, Antonio de San Martín, [s.a.] . Localización: Biblioteca de Magisterio de la Universidad de Alicante, sig. FL FA/159 vol. 1 y vol. 2.

Portal: Servicio de Información Bibliográfica y Documental de la Universidad de Alicante | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"18" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 19

Joaquín Costa (1846-1911)

Autor Principal: Costa, Joaquín 1846-1911

Título: La fórmula de la agricultura española

Publicación: [S.l.] : [s.n], [19--]

Notas de la Reproducción Original: Edición digital a partir de la edición de *Agricultura armónica, expectante, popular*, Madrid, Imprenta de Fortanet, 1912, pp.1-203.

Portal: Servicio de Información Bibliográfica y Documental de la Universidad de Alicante | Biblioteca Virtual Miguel de Cervantes

CDU:63(460) - Agricultura. Silvicultura. Zootecnia. Caza. Pesca. 338:63 - Política económica. Organización, planificación y producción.

Encabezamiento de materia: Agricultura Aspecto económico. Agricultura España

Miguel de Cervantes (1547-1616)

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El ingenioso hidalgo Don Quijote de la Mancha

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 1999

Nota General: Contiene texto y voz

Notas de la Reproducción Original: Edición digital basada en la edición de Madrid, Ediciones de La Lectura, 1911-1913.

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes | IV Centenario Don Quijote de la Mancha

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: Novela de Rinconete y Cortadillo

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2002

Notas de la Reproducción Original: Edición digital a partir de *Obras completas de Miguel de Cervantes Saavedra. Novelas ejemplares. Tomo I*.Madrid, [s.n.], 1922 (Gráficas Reunidas), pp. 208 - 328.

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Schevill, Rodolfo

Autor Secundario: Bonilla y San Martín, Adolfo 1875-1926

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: Novela de la tía fingida [versión del Manuscrito 56-4-34 de la Biblioteca Colombina]

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2005

Notas de la Reproducción Original: Edición digital a partir del Ms. 56-4-34 de la Biblioteca Colombina.

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El licenciado Vidriera

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: La ilustre fregona

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001.

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: La gitanilla

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: La fuerza de la sangre

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: La española inglesa

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: Las dos doncellas

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El coloquio de los perros

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: La señora Cornelia

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes | Literatura Española

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El celoso extremeño

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El casamiento engañoso

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001.

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

Autor Principal: Cervantes Saavedra, Miguel de 1547-1616

Título: El amante liberal

Publicación: Alicante : Biblioteca Virtual Miguel de Cervantes, 2001

Portal: Biblioteca de Miguel de Cervantes | Biblioteca Virtual Miguel de Cervantes

CDU: 821.134.2-31"16" - Literaturas y obras literarias en los distintos idiomas

Encabezamiento de materia: Novela española Siglo 17

Autor Secundario: Sevilla Arroyo, Florencio

APÉNDICE II: BASES DE DATOS

BASE DE DATOS	NOMBRE DETALLADO	FUNCIONES
TOPOS	Recuentos de Palabras	<ul style="list-style-type: none"> -Cálculo de la ley de Zipf -Cálculo de la distribución transformada de Zipf -Cálculo de la ley reducida de Zipf -Ajustar fórmula de Zipf y Mandelbrot -Obtener predicción de la fórmula de Zipf y Mandelbrot -Genera textos sintéticos -Modelo Log-% en 10 tramos logarítmicos
RENOS	Representación Numérica	<ul style="list-style-type: none"> -Ajustes del exponente y coeficiente de Zipf -Cálculo de frecuencias de las palabras -Genera representación numérica de las palabras -Cálculo de vocabulario, la varianza y derivadas -Divide en partes el vocabulario -Ajusta la función potencial -Triple fórmula de Heaps
ARENA-1: PIEDRAS	Aprendiendo Recuperación de Información	<ul style="list-style-type: none"> -Convierte ficheros de tipo secuencial a ficheros de acceso aleatorio -Analiza léxico -Divide texto en documentos - Extrae palabras significativas y sus lemas o raíces -Extracción de raíces por Método de Variedad de Sucesores - Extracción de raíces por Método de Sufijos - Extracción de raíces por Método de TCP-% - Búsquedas para la recuperación de información mediante palabras, nº del documento o raíces. - Cálculos de Zipf
ARENA-2	Aprendiendo Recuperación de Información	<ul style="list-style-type: none"> - Recuperación de información: implementación del método vectorial - cálculo IDF

COPAS	Contando Palabras Asociadas	<ul style="list-style-type: none"> - Calcula las similitudes relativas de Dice de las palabras asociadas y almacena en tabla -Analiza un texto dividido en documentos -Cuenta y obtiene las palabras asociadas de un texto -Obtiene palabras asociadas de cada uno de los documentos en los que se divide un texto -Localiza las colocaciones (<i>colocations</i>) -Estudio de la distribución de valores de las similitudes de las palabras asociadas -Cálculo de umbrales para un número de palabras asociadas -Ajuste de los coeficientes de las fórmulas a los datos de cada texto -Distribución de frecuencias, Zipf y análogas
MALLOV	Sistema de Indización y Segmentación Automática	<ul style="list-style-type: none"> -Proceso previo del fichero -Analizador léxico -Construye vocabulario -Forma colección de raíces -Representación numérica -Triple fórmula de Heaps -Analiza variaciones de novedad en palabras -Divide el texto temáticamente en niveles -Localiza colocaciones -Selecciona términos de indización -Obtiene matriz de similitud entre términos -Calcula los términos en cada nivel siguiendo la representación Log-% -Procedimiento de clusterización dentro del grupo de términos seleccionado para definir la temática de cada nivel -Sistema de indización y Segmentación Automática: genera estructura arbórea de temas con términos en cada nivel.

APÉNDICE III: FORMULARIOS DESTACADOS DE LAS BASES DE DATOS:

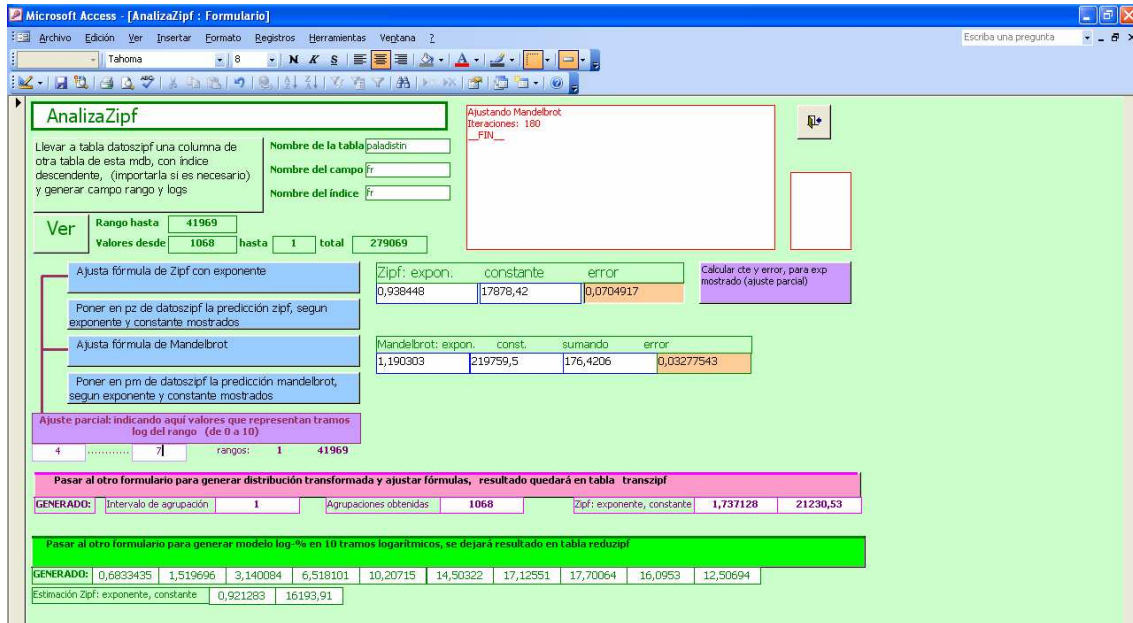


Figura 27. Aplicación TOPOS, formulario AnalizaZipf

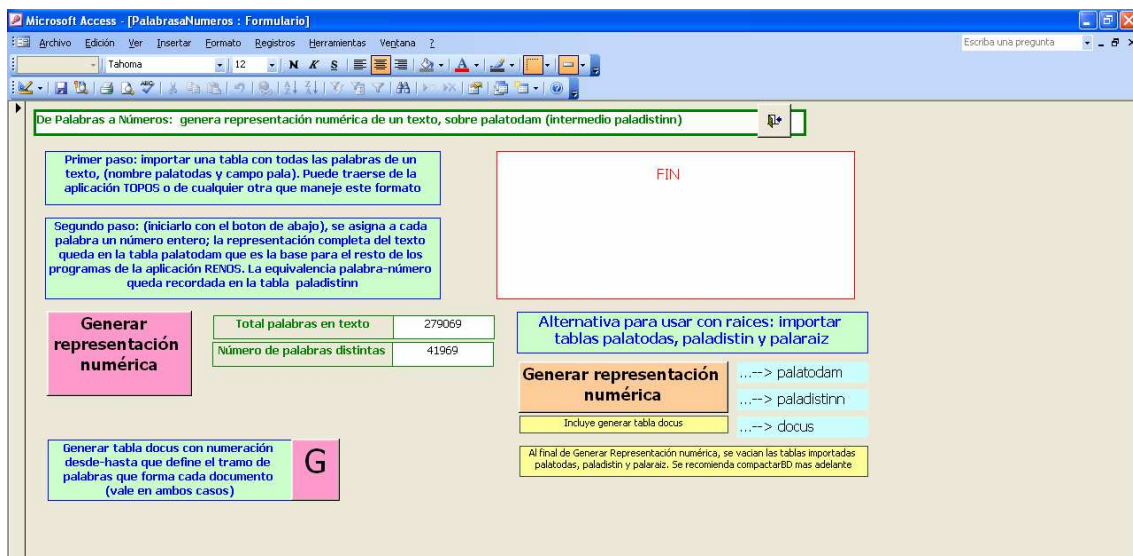


Figura 28. Aplicación RENOS, formulario Generar Representación Numérica

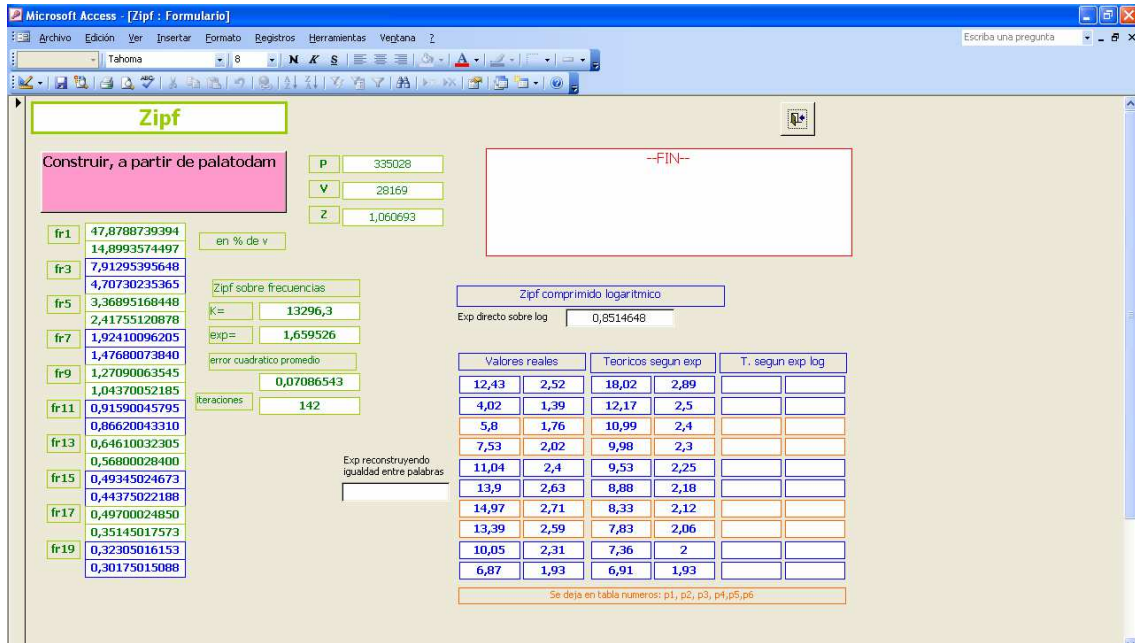


Figura 29. Aplicación RENOS, formulario Zipf

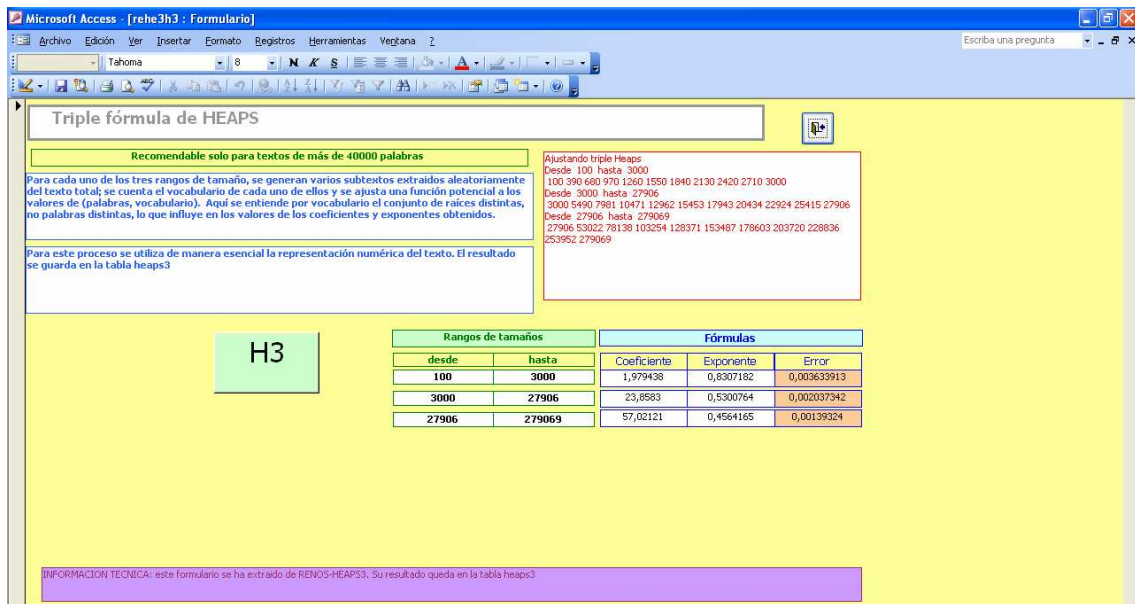


Figura 30. Aplicación RENOS, formulario Triple fórmula de Heaps

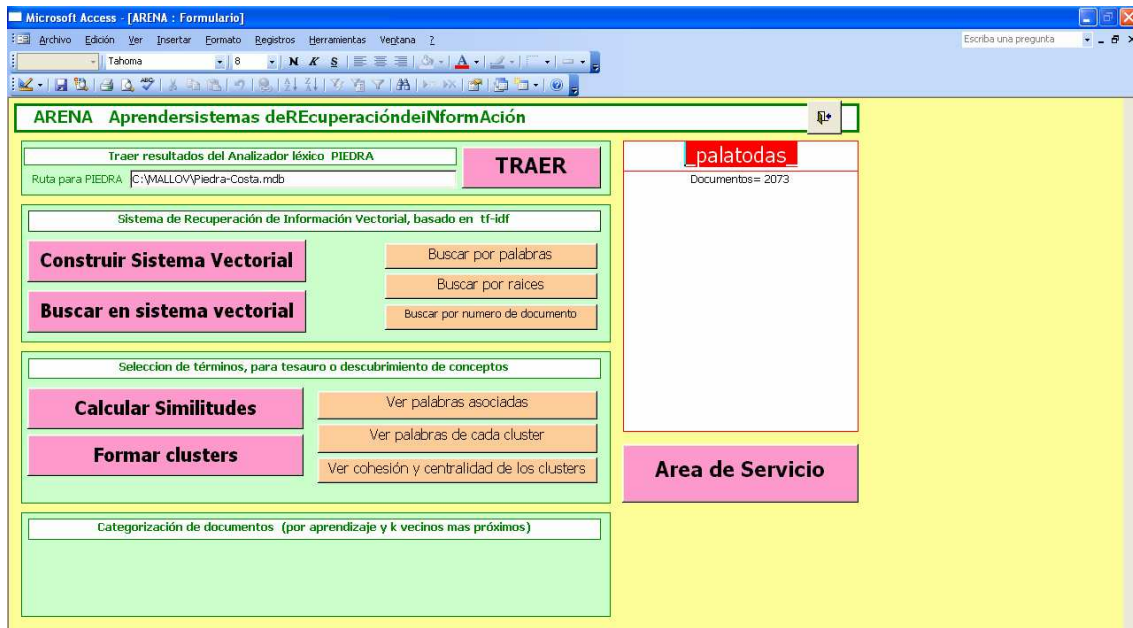


Figura 31. Aplicación ARENA, formulario Arena

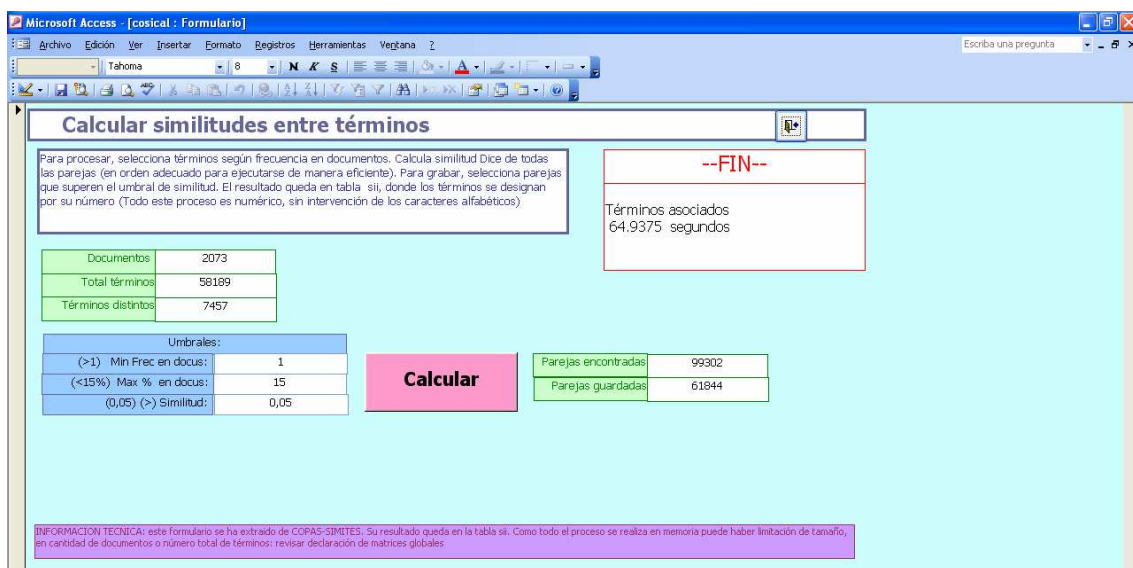


Figura 32. Aplicación COPAS, formulario Calcular Similitudes entre Términos

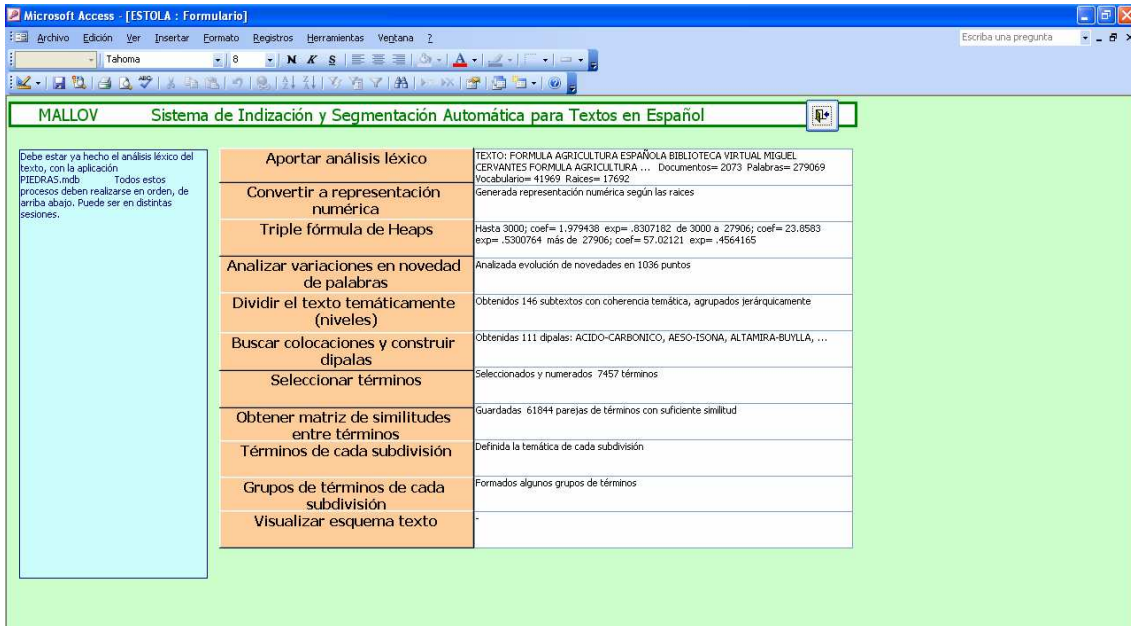


Figura 33. Aplicación MALLOV, formulario MALLOV

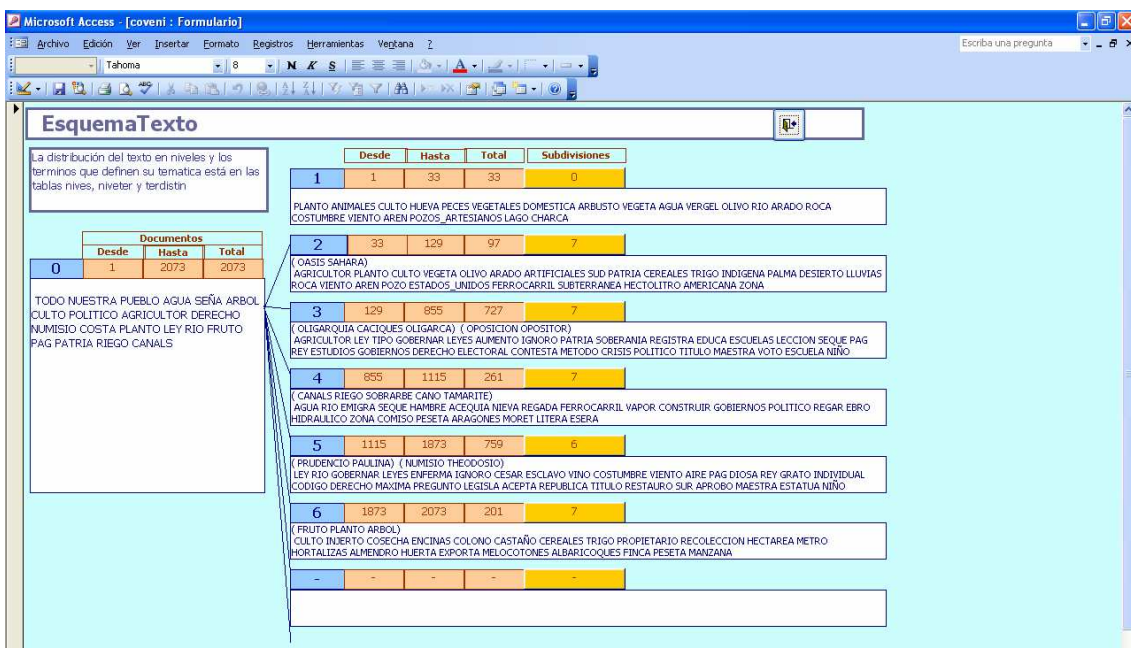


Figura 34. Aplicación MALLOV, formulario Sistema de Indización y Segmentación Automática