Mejoras en la usabilidad de la Web a través de una estructura complementaria

AUTOR: María Daniela López De Luise

DIRECTORES: Prof. Mela Bosch, Dr. Juan Ale

Tesis presentada para obtener el grado de Doctor en Ciencias Informáticas

Facultad de Informática – Universidad Nacional de La Plata

Noviembre de 2007

Índice

Índice de Tablas

Indice de Figuras

Agradecimientos

Primer deseo agradecer a todos los que colaboraron directa e indirectamente en la elaboración de esta tesis, desde el punto de vista técnico y afectivo. También agradezco a quienes permanentemente me impulsan a ir más allá de mis propios límites. Principalmente a mis directores de tesis, quienes depositaron su confianza en mí desde un comienzo.

Un especial agradecimiento a mis dos princesitas Sofía Ivana y Milena Oriana, y a mi esposo Luis Daniel, quienes debieron acomodarse a horarios extraños y soportar mis largas horas frente a la computadora. Sin su tolerancia y amor, no hubiera tenido la fuerza suficiente para avanzar hasta donde he llegado.

Finalmente agradezco a mis padres, quienes me enseñaron que todo paso en la vida es un comienzo, una lucha y un destino alcanzado...

Resumen

La Web ha motivado la generación de herramientas que permiten, con distintos grados de sofisticación y precisión, manipular sus contenidos. Para ello, tratan una serie de problemas, relacionados con la naturaleza imperfecta y cambiante de todas las actividades humanas. Ésta se refleja en fenómenos como las ambigüedades, contradicciones y errores de los textos almacenados.

Esta tesis presenta una propuesta para complementar la administración de contenidos en la Web y de esta manera facilitar el proceso de recuperación de información. Se presenta un prototipo, denominado Web Intelligent Handler (WIH), que implementa una serie de algoritmos básicos para manipular algunas características morfosintácticas de textos en castellano y, en base a ellas, obtener una representación resumida y alternativa de su contenido. En este contexto, se define una nueva métrica de ponderación para reflejar parte de la esencia morfosintáctica de los sintagmas. Además se define un esquema de interacción entre los módulos para regular la explotación de los textos. También se explora la capacidad de los algoritmos propuestos en el tratamiento de los textos, considerándolos como una colección de sintagmas, sujeta a factores tales como contradicciones, ambigüedades y errores.

Otro aporte de esta tesis es la posibilidad de evaluar matemáticamente y de manera automática tipos de estilos de texto y perfiles de escritura. Se proponen los estilos literario, técnico y mensajes. También se proponen los perfiles documento, foro de intercambio, índice Web y texto de sitio blog. Se evalúan los tres estilos y los cuatro perfiles mencionados, los que se comportan como distintos grados de una escala de estilos y perfiles, respectivamente, cuando se los evalúa con la métrica morfosintáctica aquí definida. Adicionalmente, utilizando la misma métrica, es posible realizar una valoración aproximada y automática de la calidad de cualquier tipo de texto. Esta calificación resulta ser invariante a la cantidad de palabras, temática y perfil, pero relacionada con el estilo del escrito en cuestión.

Abstract

The Web motivated a set of tools for content handling with several levels of sophistication and precision. To do so, they deal with many unsolved problems in saved texts. All of them are related to the mutable and imperfect essence of human beings such as ambiguities, contradictions and misspellings.

This theses presents a proposal to complement the Web content management and therefore to provide support to the information retrieval activity. A prototype named Web Intelligent Handler (WIH) is introduced to implement a set of algorithms that manage some morpho-syntactical features in Spanish texts. These features are also used to get a brief and alternate representation of its content. Within this framework, a new weighting metric is designed to reflect part of the syntagm morpho-syntactical essence. A module interaction approach is also outlined to rule the text processing output. Besides, this thesis analyzes the algorithms ability to handle texts considering them as a collection of syntagms affected by certain factors such as contradictions, ambiguities and misspellings.

Perhaps, the main contribution of this thesis is the possibility to automatically mathematical evaluation of text styles and profiles. Three initial three styles are proposed here: literary, technical and message. Furthermore, the following writer profiles are proposed also: document, foro, Web-index and blog. All the three styles and four profiles were evaluated. They behave respectively as a part of a graduated scale of styles and profiles when the morpho-syntactical metric defined here is used. It is also possible to perform a kind of automatic rough text quality valuation. This is invariant to the text word quantity, topic and profile, but it is related to its style.

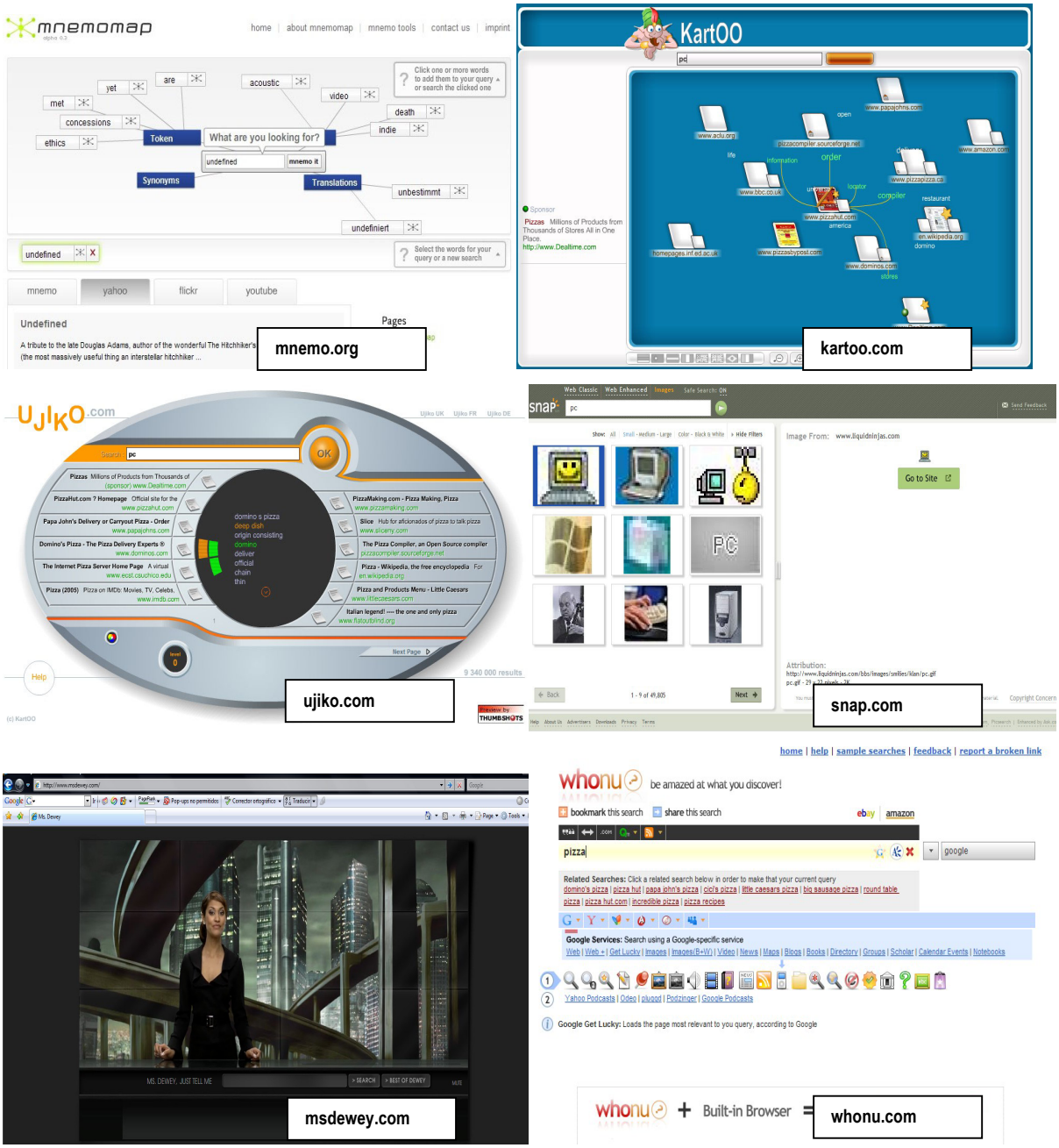
Capítulo 1. Introducción

La información almacenada en Internet ha dado lugar a la generación de una gama de herramientas que permiten, de manera más o menos sofisticada, manipular los contenidos remotos de la Web. Ya en el año 1993 [106] apareció el primer aplicativo conocido como Wandex, desarrollado por Mattew Gray en el MIT. Simplemente buscaba páginas organizadas en un índice temático. Compitió con otro buscador, poco conocido, que aún funciona: Aliweb. Pronto comenzó a sofisticarse el mecanismo de búsqueda y aparecieron motores de búsqueda por palabras claves o keywords (1994), criterios de ordenamiento o page-ranks (2001), combinaciones de palabras (tomadas como claves de búsqueda) y jerarquías de directorios (2004), etc.

Paralelamente se produjeron progresos en el diseño de las interfaces con el usuario, aunque actualmente, podría decirse que los buscadores más populares básicamente constan de un cuadro de texto para ingresar la consulta, una serie de hipervínculos y cierta información categorizada. Algunas interfaces también integran el acceso a un correo electrónico.

Sin embargo, existen una serie de buscadores que han incursionado en alternativas gráficas con diverso grado de sofisticación, recurriendo a presentaciones más amenas y equipados con funciones como auto completar, historiales de búsqueda, etc. Algunos ejemplos de estas alternativas pueden verse en la Fig. 1: organización espacial de contenidos tipo cluster (mnemo.org), forma de mapa (kartoo.com), disposición alternativa de contenidos interesantes (ujiko.com), mosaicos de imágenes (snap.com), listado sin paginar, con sonido y asistente de navegación (msdewey.com), búsqueda por imágenes o video o una combinación de búsquedas multimedia con íconos gráficos (whonu.com).

Fig. 1. Buscadores con interfaz gráfica.



De todas formas, un profundo estudio comparativo realizado en [106] muestra que los buscadores textuales suelen ser mucho más rápidos, más conocidos y aún preferidos respecto a sus alternativas gráficas, por los usuarios.

Otro aspecto importante que se ha incorporado es el tratamiento específico del contenido multimedia. En los últimos años la reducción del costo de obtención y manipulación de este tipo de contenidos ha marcado un crecimiento sostenido de sus almacenamientos y consultas en la WWW [19]. Lamentablemente las herramientas tradicionales no son totalmente adecuadas para su tratamiento, por lo que se hizo necesario plantear alternativas de procesamiento. Algunos buscadores (Google, IceRocket, Altavista, etc.) extienden el uso de keywords, simplemente asociándole un texto manualmente ingresado, que servirá para transformar la consulta en una tradicional recuperación de información textual. Otros trabajan con los denominados metadatos (información textual que describe de diversas maneras preestablecidas, el contenido del multimedia). También se han construido interfaces que permiten la navegación de imágenes organizadas por algún criterio de catalogación (ej. Banco de Imágenes), permitiendo hallar estructuras o relaciones entre los contenidos y seleccionar ítems centrando la atención visual en ellos. La navegación se implementa normalmente como hipertextos de vinculación en una misma página Web o entre distintas páginas. Al respecto se pueden distinguir tres tipos de navegación:

-Directo o específico: es sistemático y enfocado a un tópico. El usuario debe conocer exactamente lo que está buscando.

-Semi directo o predictivo: el usuario tiene un conocimiento vago de lo que busca. Suele requerir una búsqueda previa del tema a estudiar, en la que se visualizan las imágenes relacionadas al tópico en estudio. Una vez detectado el tema, la búsqueda procede como en el caso anterior. Estas búsquedas suelen requerir más tiempo.

-Indirecto o general: el usuario no tiene un objetivo definido. En estos casos suele navegarse azarosamente por mera recreación.

Es peculiar de este último tipo de búsqueda, la importancia de la información adicional asociada al contenido original para maximizar el éxito de la misma.

Asimismo, la complejidad y variabilidad de los archivos, unido a la inexistencia de una normalización universalizada de su almacenamiento y descripción, aumenta la complejidad de las herramientas y arquitecturas necesarias para su procesamiento y consulta. Entre estos sistemas, algunos indexan directamente su contenido. En estos casos se los denomina "sistema de recuperación por contenido". Se describirán brevemente los dos tipos de tratamiento multimedia predominantes: imagen y sonido.

a) Tratamiento de imágenes [19]: Esta categoría suele incluir a los videos, que son reducidos a una o más imágenes representativas por varios criterios. Los sistemas de recuperación por contenidos de imágenes suelen diferir entre sí básicamente en la forma en que extraen las características visuales para indexar y la forma en que se efectúan las consultas. Dichas consultas pueden realizarse a partir de una imagen modelo o bien de la especificación de las características que debieran contener las imágenes recuperadas en la consulta. Variantes exploradas para esto son:

-Búsqueda por invarianza de iluminación: se separa matemáticamente la información cromática de la imagen previamente normalizada según ciertos parámetros y se comprime antes de almacenarla. La comparación se realiza entre vectores análogamente derivados de las imágenes.

-Búsqueda por color: se ordenan las imágenes de acuerdo a un histograma de colores de cada una. Se extraen matemáticamente los colores predominantes.

-Búsqueda por textura: presentan un patrón repetitivo dispuesto de manera regular. Puede trabajarse con la fineza o granularidad, y la direccionalidad de la misma.

-Búsqueda por figuras: se detectan figuras dentro de la imagen y se guardan como descripción de la misma.

b) Tratamiento de sonidos [140]: generalmente se realiza una conversión de audio a texto (ej. SpeechBot y NewsTuner). A veces se complementa con otro tipo de

análisis como el estudio gramatical de la consulta (ej. ThisL), posibilidad de consultar usando micrófonos en vez de archivos de sonido (ej. ITESM), procesar los sonidos aún cuando están fuera del diccionario de palabras reconocibles (ej. OVSDTR), etc.
Para la extracción automática de las características del sonido, al igual que en el caso de las imágenes, existen variantes. Entre ellas cabe destacar:

-Búsqueda por distancia de Hamming: se calcula utilizando el algoritmo descrito por Hamming entre harmónicos de la muestra y de archivos de la base de datos.

-Búsqueda por ejemplos: suele denominarse QBE (Query By Example). Requiere que se disponga de una porción del archivo que se desea consultar.

-Segmentación de sonido: se descompone el sonido en fragmentos significativos y filtrados.

-Búsqueda probabilística por índice semántico latente: suele denominarse PLSI (Probabilistic Latent Semantic Indexing). Luego de extraer las palabras contenidas, realiza cálculos probabilísticos para hallar relaciones entre los grupos de palabras, generando asociaciones que pueden verse como subtemas de un documento.

Las dos primeras técnicas son las más usadas para consultas de sonidos musicales. La segmentación suele requerirse para el reconocimiento y extracción de palabras y PLSI permite una búsqueda temática.

Finalmente, un tópico de gran interés en la actualidad es el de la seguridad en la Web. La seguridad puede verse violada de diversas formas y en distintos momentos [121]: alteraciones de contenidos, escaneo de WSDL (descriptores de operaciones realizables con un Software), escaneo de parámetros (se intenta acceder a información o romper cierta aplicación), cargas recurrentes (reenvío de un mensaje indiscriminadamente para consumir recursos), cargas de gran tamaño (se incrementa

el contenido de cierto documento para consumir recursos), creación de mensajes SQL automáticamente (a veces robando claves de acceso a cierta Base de Datos), robo de información confidencial, intrusión en una sesión y hasta la infección o manipulación remota de hardware. Las técnicas tradicionales de validación de usuario y contraseña pueden no resultar suficientes para garantizar la seguridad. En este aspecto surgen alternativas como la organización de estándares (ej. OASIS, W3C), protocolos (como SSL) y herramientas de software (como firmas digitales, y técnicas de encriptado). Una generación de sistemas de seguridad con base en identificación biométrica está aún pendiente de ser investigada como alternativa posible para la Web. Conceptos como reconocimiento facial, son aún poco pensables en este entorno.

A pesar de la descripta plétora de ingredientes que intervienen en el estudio de la administración de contenidos, el grueso de la investigación se mantiene focalizado en el procesamiento de textos planos, ya sea por considerarse un punto álgido de la problemática, por moda o por tratarse de información muy diversamente consultada. Por tal motivo, en el capítulo siguiente se presentan los antecedentes específicos de este tratamiento.

1.1. Motivación y contribuciones

Existen varias estrategias seguidas por las diversas líneas de investigación para administrar los contenidos de Internet y proveer un acercamiento conceptual al usuario y sus necesidades, las que serán presentadas con mayor detalle en el próximo capítulo. Los tópicos normalmente estudiados abarcan aspectos tales como: mejoras en la capacidad y velocidad de indexación de nuevas páginas, mejoras en la capacidad y velocidad de recuperación de información, alternativas en la visualización e interrelación de contenidos, etc., siempre considerando visiones diversas de la información almacenada o información obtenida del comportamiento del usuario. A pesar de todos los avances presentados, es posible asegurar que el problema de la manipulación de contenidos no está totalmente resuelto, debido a que aún existen dificultades en la precisión de las respuestas ante una consulta, el tiempo insumido en responder, la complejidad en la preparación de los datos (para volverlos aptos para su manipulación según cierto criterio), el número de respuestas presentadas al usuario,

las alternativas de navegación, etc. Es probable que la solución no pase sólo por aumentar la complejidad de los buscadores, sino tal vez por simplificar y complementar los contenidos almacenados en la Web.

Este trabajo desarrolla y evalúa una herramienta para el estudio de contenidos de páginas en Internet pretendiendo simplificar y complementar la manipulación de los mismos por parte de otras técnicas. En ese sentido se enmarca en el área de recuperación textual y se orienta en el enfoque de trabajos como los desarrollados por el "Grupo de Estructuras de Datos y Lingüística" de la Universidad de Las Palmas de Gran Canaria. Este grupo ha desarrollado herramientas de reconocimiento y gestión morfológica [127] [128], con el objeto de proveer herramientas que aprovechen el enorme caudal de información lingüística que supone Internet.

Parte de la estrategia que se presenta, trabaja con stemming, algorítmica presentada originalmente por Porter[116] para extraer la raíz de las palabras, como elemento de modelización del lenguaje. Algo similar al enfoque de Allan y Kumaran en [5], puesto que ellos también consideran que más que una técnica de preprocesamiento, puede enfocarse con una nueva perspectiva, como una manera de simplificar y mejorar estimaciones estadísticas.

La solución presentada aquí consiste en una reorganización virtual de la información contenida en la Web, con un criterio alternativo. Por ejemplo, como parte de ese criterio, desaparece el concepto de "sitio Web" y de "documento" de manera explícita, preponderando la organización de palabras relacionadas por su co-pertenencia a una sentencia[1]. Esto requiere de una estructura complementaria de capas, aquí denominada **WIH**, que crezca progresivamente y con la que se sienten las bases para interactuar con criterios flexibles.

1.1.1. Principales contribuciones

Las principales contribuciones de este trabajo son:

-El diseño de una métrica que pondera sintagmas de textos en castellano sobre la base de algunas de sus características morfosintácticas sencillas, derivadas automáticamente, tales como su raíz (derivada con el mencionado algoritmo de Porter), su categoría léxica (sustantivo, verbo, etc.), cantidad de vocales, etc.

[1] Cabe aclarar al respecto, que la referencia a la página, sitio y documento originales nunca se pierde. Se almacena de manera especial dentro de la estructura de datos.

-La posibilidad de realizar una jerarquización de algunos sintagmas componentes del texto (gracias a la ponderación obtenida con la métrica mencionada en el ítem anterior). Es posible procesar incluso textos con contradicciones y ambigüedades sin que ello constituya un obstáculo al proceso global que se lleva a cabo.

-Con un procesamiento matemático sencillo, es posible utilizar la métrica para distinguir automáticamente el estilo general del escritor de un texto. Aquí se estudian tres estilos que se consideran básicos y distantes entre sí, tales como el estilo literario (escritos depurados y con gran dominio del lenguaje), técnico (donde el escritor tiene una finalidad más práctica y el contenido tiene un nivel de depuración muchas veces menor al literario, y su lenguaje suele ser del dominio corriente o de un área de especialidad), y mensajes (escritos normalmente muy cortos e informales). Estos grados diferentes de formalidad y dominio del idioma que intuitivamente son aparentes, se logran medir usando la métrica de ponderación aquí presentada.

-Algo similar sucede con los perfiles de conformación de los escritos, donde se logra establecer también un grado de conformación del perfil. Se trabaja aquí con la hipótesis de cuatro perfiles importantes para quienes realizan producción de textos: documentos (escritos con una cierta estructura, y que pretenden explicar un tópico o tema, o realizar una presentación escrita específica), foro de intercambio (comprende textos que buscan responder o preguntar acerca de cierto tópico o tema), índice Web (compilaciones enumerativas con acceso a otros sitios o páginas), y blog (escritos con contenido muy variable, normalmente centrados en algún interés común).

-Adicionalmente, con una extrapolación adecuada de las ponderaciones, es posible obtener un grado de calidad general y relativa entre textos, que resulta invariante a factores tales como el tipo o perfil del documento y la cantidad de sintagmas. El valor obtenido descansa especialmente en los estilos propuestos y en la ponderación inducida a nivel de sentencias del escrito.

-Un diseño de arquitectura general planeado para realizar tareas de manera muy flexible, al que responden todos los módulos, y donde la funcionalidad es desmembrada en una **categoría básica** y otra **categoría efectora de actividades**. De este modo, el flujo de interacción entre módulos es regulado de manera genérica por el sistema sobre la base de criterios de mínima y de máxima representados en un conjunto de **funciones de métrica**, administradas por un **sistema controlador**. La

actividad específica del instante dependerá, en cambio, de las denominadas **funciones efectoras** que implementan los aspectos funcionales correspondientes a la categoría efectora. La dinámica introducida por este diseño, es tal que permite la explotación de contenidos de manera autorregulada.

1.2. Objetivo y alcances de este trabajo

El objetivo de este trabajo es determinar los algoritmos y métricas básicas para el desarrollo adecuado de dicha estructura complementaria y complementar la manipulación de la Web, con las siguientes hipótesis de trabajo:

- Máxima autonomía de crecimiento de las capas constitutivas.
- Adaptación automática de nuevos contenidos.
- Flexibilidad para abarcar contradicciones, ambigüedades e información incompleta.
- Administración de contenidos (no de documentos).
- Navegación que haga transparente al usuario, la administración de aspectos técnicos.
- Purgado automático de contenidos que refieran a contenidos eliminados.
- Conexión con la Web con sus características actuales (sin necesidad de alterarla).
- Flexibilidad para interactuar con distintas interfaces visuales.

El presente estudio no realiza consideraciones de consumo y optimización de recursos; se realiza considerando aspectos técnicos teóricos y la evaluación estadística de cada propuesta realizada Para ello, se han implementado los módulos de Traducción al Lenguaje Interno (TLI), Motor de Composición (MC), Motor de Asimilación (MA) y Sistema Controlador (SC), correspondientes a la implementación completa de los módulos correspondientes a la capa más comprometida con esta propuesta.

1.3. Estructura de la tesis

El resto de la tesis está organizada de la siguiente manera:

- El capítulo 2: presentan los principales antecedentes de estrategias de procesamiento de textos en la Web para sustentar la actividad de recuperación de información, la influencia de otros factores no técnicos en el resultado efectivamente obtenible durante una consulta, la ingerencia de otras áreas como Text Mining y Lingüística Computacional. A continuación se describen propuestas categorizadas en dos criterios básicos: semántico y morfosintáctico. Finalmente se introduce el concepto de posicionamiento de los resultados frente al usuario y la problemática relacionada con las diferencias de los resultados obtenibles por las distintas estrategias.
- En el capítulo 3 se describe y fundamenta la estrategia global seguida para la manipulación de los contenidos de la Web y se la divide en tareas. Luego se presenta la asignación de dichas tareas a los distintos componentes del prototipo.
- El capítulo 4 presenta algunos aspectos de la implementación del prototipo WIH (Web Intelligent Handler), realizada sobre la base de los algoritmos y estrategias descriptos en el capítulo 3. Fundamentalmente se describe la arquitectura global del sistema.
- En el capítulo 5 se evalúan estadísticamente los resultados obtenidos al procesar datos reales empleando los módulos del prototipo desarrollado. Se hace hincapié en casos tradicionalmente requieren complejos análisis. Para facilitar la presentación y desarrollo de los casos, se introduce con mayor nivel de detalle el tratamiento de los datos desde su captura hasta su reorganización dentro de la base de datos.
- En el capítulo 6 se realiza un análisis de las características básicas del modelo presentado. Esto comprende un estudio comparativo del tratamiento y resultados obtenidos respecto a alternativas posibles, y un estudio focalizado en el propio WIH.

- En el capítulo 7 se detallan las principales conclusiones sobre la base del trabajo realizado.
- En el capítulo 8 se describen los posibles trabajos a futuro que serían convenientes para continuar con el desarrollo de la propuesta.

Capítulo 2. Antecedentes

La propuesta WIH puede interpretarse como una reorganización virtual de la información contenida en la Web. Esto se hace con una estructura complementaria de capas que crece progresivamente, y con la que se sientan las bases para interactuar con criterios alternativos y flexibles. El principal objetivo del trabajo es determinar los algoritmos y métricas básicas para el desarrollo adecuado de dicha estructura complementaria, considerando aspectos teóricos y estadísticos.

2.1. Procesamiento textual en la Web

¿Es dable pensar que la comunicación escrita en la Web tiene características especiales y distintivas? Si se tomaran todos los contenidos como expresión de un área de especialidad, según [10] debiera existir una Lengua de Especialidad (LE), ya que la existencia de las distintas especialidades sería un fundamento seguro para su definición. Según Schifko, la especialidad determina la temática, es decir, la base semántica y pragmática de los textos especiales correspondientes. De este modo, es dable pensar que en términos generales no puede decirse que toda comunicación en la Web deba tratarse como LE, salvo para los casos de los tratados técnicos que toman como área de estudio la Web en sí misma. Lo que sí es evidente, que muchos contenidos son expresiones válidas de distintas áreas de especialidad. Esto lleva a pensar que los tratamientos probados para textos en general están habilitados también para los contenidos en Internet.

A su vez, una lengua natural, tal como el español o castellano, por ejemplo, no representa un sistema homogéneo y unitario [10], sino un diasistema, un conjunto total de muchas variedades (dialectos), que se pueden clasificar de acuerdo con una serie de parámetros de variación: el tiempo (cronolectos), el espacio (topolectos o regiolectos), los distintos grupos de hablantes (sociolectos), el nivel estilístico (estratolectos), la especialidad (tecnolectos o funciolectos). Además, en este contexto también existe un problema: Todos los especialistas, que han trabajado en el dominio de la variación lingüística, así como todos los interlocutores que tienen contactos

comunicativos con personas que hablan distintas variedades y que dominan, a su vez, algunas de las mismas, saben que no existen límites claros ni entre las distintas clases de variedades (por ejemplo entre socio- y regiolectos), ni entre las variedades concretas de una misma clase (por ejemplo entre distintos regiolectos). Es decir, un idioma, como totalidad de sus variantes, es un universo de continuidades, graduaciones y matizaciones a veces casi imperceptibles, de manera que es difícil establecer con exactitud, dónde termina un dialecto y dónde empieza el otro. Además, dentro de las distintas LE existen variaciones sociolectales (la llamada variación o estratificación vertical): según su grado y tipo de especialización, los científicos teóricos de una disciplina hablan y escriben de otra manera que los profesionales prácticos o los obreros en los talleres. Todo ésto contradice el principio clásico de "definición", palabra derivada del latín "finis", ya que debe indicar claramente los límites entre fenómenos distintos.

En este marco, lo más inquietante es la relación entre lengua común (denominada léxico en términos lingüísticos) y lengua de especialidad (en lingüística denominada terminología[2]). La primera, representa al núcleo del diasistema de una lengua natural y es un instrumento que sirve para la comunicación general sobre asuntos corrientes entre todas las personas de una comunidad lingüística; la segunda, es un instrumento para la comunicación sobre asuntos especiales entre expertos que poseen conocimientos especiales de ciertos sectores del mundo. Deducir de esta constatación que aquellos instrumentos tienen que ser claramente distintos y distinguibles sería un gran error: la realidad de la comunicación especializada y de los textos especiales muestra que los signos y las construcciones utilizadas en ambos casos son, en gran parte, idénticos aún en los textos de las ciencias exactas y de la técnica. Es considerable la intersección entre lengua común y lengua de especialidad (en grados distintos), dado que involucra toda la gramática (morfología y sintaxis) y gran parte del vocabulario, aunque se utilicen con frecuencias e intensidades divergentes.

La herramienta exclusiva de las LE se reduce a los términos técnicos, frecuentemente con algunas particularidades morfológicas en la formación de las

[2] se diferencia en cuanto a métodos y objetivos de la Lexicografía.

palabras y, en parte, a ciertas estructuras textuales y tipos de textos. Una ilustración llamativa de la situación presentada es el hecho de que existen algunos lingüistas que niegan la existencia de lenguas de especialidad y admiten únicamente usos particulares de la lengua común, mientras que otros niegan la existencia de la lengua común y hablan de lenguas más o menos especializadas.

En el contexto de lo expresado anteriormente, puede entonces inferirse con razonable grado de acierto, que el tratamiento de textos en la Web conlleva implícitamente un grado de complejidad alto debido a la heterogeneidad de lenguas, especialidades, contextos y formatos. En virtud de ello, es sensato pensar que su tratamiento debe ser complementado por herramientas conceptuales diversas que permitan la simplificación directa o indirecta del manejo ontológico[3] formal de los conceptos que subyacen a los datos almacenados, como medio de administración práctica de la información en ellos contenida. En este criterio se basa la propuesta de este trabajo.

2.2. Factores subjetivos en la búsqueda de información

Pero, el éxito de una búsqueda en la Web no depende sólo de la potencia de un buscador ni de la organización de la información que éste ponga a disposición del usuario. Existen factores socio-culturales y psicológicos tales como la buena disposición subjetiva del individuo que influyen, además de los objetos tecnológicos específicos que puedan desarrollarse. Como parte del estudio preliminar para esta tesis, se realizó un estudio estadístico de factores sociales, culturales y emocionales para analizar sus influencias en el proceso de recuperación de información desde la Web [88]. Los resultados obtenidos indicaron una fuerte influencia de estos factores en el éxito final de la búsqueda (para más detalles ver el ***Apéndice B:*** estudio de aspectos no tecnológicos para el proceso de IR). La incidencia de estos aspectos fue comprobada también en un trabajo estadístico llevado a cabo con 30 voluntarios en [27], a quienes se les requirió el uso de un navegador con interfaz gráfica denominado

[3] Ontológico: referente a la ontología. disciplina que se suele identificar con la Metafísica general o bien indica una de las ramas de ésta que estudia lo que es en tanto que es y existe. [164]

WEBSOM que coloca los documentos en un mapa bidimensional agrupados por temas. Llamativamente los voluntarios (todos especialistas en Informática) no hallaron sencilla a esta herramienta, a pesar de las múltiples facilidades instaladas en ella (representación espacial de los documentos, roseta de navegación espacial, posibilidad de consultas, niveles de profundidad, links, etc.).

2.3. Text mining, lingüística computacional y recuperación de información

Es muy complejo abarcar los problemas relacionados con la manipulación de información textual, la cual requiere de una perspectiva multidisciplinaria. Los textos [18], al provenir de un repositorio con información no estructurada como Internet, son pasibles de estudio por parte de la minería textual, emparentada con la minería de datos (salvo que en esos casos se trata de repositorios estructurados). En ella confluyen distintas técnicas y principios teóricos de otras disciplinas como recuperación de textos (text retrieval) y la lingüística computacional.
Existen varias definiciones de minería de textos, una de ellas, la formulada por Dan Sullivan [138], quien la define como "cualquier operación realizada para extraer y analizar textos procedentes de distintas fuentes externas con el objetivo de obtener inteligencia". También la definió como el descubrimiento de información y conocimiento que anteriormente no se conocía, a partir de corpus textuales.
La definición más popular, de M. Hearst [51], señala que tiene cono objetivo descubrir información y conocimiento que previamente se desconocía, y que no aparecía en ninguno de los documentos analizados.

Si bien desde un punto de vista técnico existe una composición de técnicas de áreas como la recuperación textual y la lingüística computacional. Pero en el caso de la recuperación de información, el objetivo es la representación formal de los documentos sobre los que realizará una búsqueda y la formulación de las necesidades de información del usuario mediante un sistema de representación adecuado (no pretende facilitar el análisis ni extraer nuevos conocimientos). Por su parte, en cambio, la lingüística computacional agrupa una serie de técnicas para procesar textos

y tratar de hacerlos comprensibles para una computadora. Permite el análisis sintáctico y gramatical de textos en formato electrónico, la alineación e identificación de correspondencias entre textos escritos en diferentes idiomas, etc. Con frecuencia toma técnicas de la minería textual pero sus objetivos son distintos. El presente trabajo se halla en un punto más cercano a la recuperación textual, pero también usa conceptos de lingüística computacional.

2.4. Criterios semánticos

La generación de texto empieza con la conceptualización del mensaje que se transmitirá [39] y con la definición del nivel de generalización o de detalle en que se realizará. Existen criterios que intervienen en la construcción de la estructura, que no se consideran en el nivel de oración sino en el nivel del discurso completo tales como la coherencia, expuesta mediante enlaces entre oraciones. A su vez, la comprensión de texto es un proceso basado en conocimiento lingüístico, siendo más compleja que la generación. Surge de la representación de la información textual, es decir, de la cadena de palabras, y la traduce a diversas estructuras lingüísticas en varias etapas. Las transformaciones que se requieren en el análisis y la síntesis son tan complejas que se dividen, tanto en la teoría como en la aplicación, en etapas generales. Para que la computadora realice estas etapas, se requieren métodos adecuados para la descripción y construcción de las estructuras correspondientes, es decir, se requieren formalismos lingüísticos de representación y formalismos computacionales.
En la lingüística general se considera que son tres los niveles generales los que componen el proceso lingüístico: la morfología, la sintaxis y la semántica. En esta sección se presentan algunos de los principales desarrollos semánticos para el procesamiento de textos. En la próxima sección se presentarán algunas estrategias morfológicas[4], sintácticas y morfosintácticas.

La finalidad de las estrategias semánticas es reflejar en cierto grado la "semántica lingüística" de los documentos de manera más o menos expresa. Esto motivó a una serie de investigadores a incursionar en la fusión de la inteligencia artificial y la

[4] Ver en Diccionario y Nomenclatura, las definiciones correspondientes a morfología, y sintaxis para mayor detalle.

semántica de textos, generando una notación híbrida (lingüística y ontológica). Los especialistas, finalmente van hallando maneras de modelizar los recursos Web y su vocabulario para explicitar y manipular de alguna forma el significado subyacente en los términos de las páginas. A ésto se lo denomina manejo ontológico [46] [140]. Los conceptos en la ontología normalmente se hallan organizados en ciertas taxonomías. A veces la noción de ontología se encuentra algo diluida, en el sentido que las taxonomías son consideradas en sí mismas como una ontología también [136]. Los avances en esta área incluyen desarrollo de sistemas y proyectos, así como herramientas de anotación semántica (para expresar relaciones entre elementos del texto y para características de las palabras). A continuación se presentan algunos de estos avances clasificados según sus características preponderantes.

(i) Buscadores por concepto

Entre las técnicas para capturar el concepto subyacente en los textos se halla el uso de los SOM (Self Organizing Map) [22]. Esta algorítmica propuesta por T. Coñeen [74], se basa en una red neuronal con aprendizaje no supervisado, capaz de definir agrupamientos automáticamente. Esos agrupamientos son dables de ser utilizados como categorías de navegación.

En esta línea se halla el prototipo WEBSOM [77], diseñado para grandes colecciones de documentos. Crea un mapa bidimensional que representa gráficamente la cercanía conceptual de los documentos (representados como puntos). También tiene una jerarquía de mapas que permiten navegar de mayor a menor nivel de abstracción. Se aplicó en ciertas colecciones (ej. Artículos de grupos de noticias, también llamados “newsgroups”), pero su uso no es especialmente apto para consultas, sino más bien para navegación.

Basándose en los mismos principios, X. Lin propone la construcción de un mapa con un poco más de sencillez en su interfase [81]. En [142] se extiende el concepto anterior introduciendo el QFDN (Quasi Four Dimensional Neuroncube), una dimensión más al mapa bidimensional de Kohonen. En [107] se propone una modificación del SOM original: el HSOM (Hyperbolic SOM) donde se calcula de manera distinta la distancia de los puntos

(documentos) dentro del espacio del mapa (es decir, del SOM). Aquí las distancias ya no son euclidianas sino que crecen exponencialmente (es un espacio hiperbólico). Esto, según el autor, permitiría representar los detalles de relaciones complejas entre documentos con mayor claridad.

Conzilla [104] [105] es otro prototipo de buscador por concepto, con mapas conceptuales y referencias a documentos relacionados conceptualmente. Surgido alrededor del año 2000. Contiene una Base de Datos relacional, y se compone de varios módulos: "neurona", "tipos de neurona", "mapas de concepto" y "descripciones de contenidos". Se basa en MOF (Meta Object Facility) como estándar para meta-modelos, que interrelaciona objetos interconectados por CORBA, y usa XML.

Pero, el principal énfasis de la corriente semántica se halla en la denominada Web Semántica [66]. Basada originalmente en la conjunción de una serie de tecnologías propuestas por W3C, tiene como objetivo explicitar la semántica de manera organizada y propicia para el manejo informático de la misma y la mejora en la interacción con el hombre. Toma como base el texto escrito, como manifestación de la lengua natural. Su estructura es analizable en lo que Leech [79] describe cono niveles de anotación lingüística. Esos niveles son [2]:

-los lemas (asociación de cada token[5] léxico con una palabra específica)

-morfosintáctica (anotación[6] de la clase gramatical y opcionalmente su análisis morfológico)

-sintáctica (anotación de relaciones sintácticas entre categorías morfosintácticas)

[5] Token: en este contexto se refiere a un sintagma lingüístico (conjunto de letras que se agrupan en una unidad, por ejemplo, para conformar una palabra).

[6] Anotación refiere aquí a la explicitación estandarizada mediante tags.

-semántica (las anotaciones de relaciones semánticas entre elementos del texto o anotaciones de características semánticas de las palabras del texto)

-de discurso (con el enfoque Stenström de etiquetas con tipos de discursos, o bien la anotación anafórica con referencias pronominales y consideraciones de cohesión[7] del discurso).

Entre los instrumentos empleados para implementar estos niveles se pueden mencionar el RDF (Resource Description Framework, base para la interoperabilidad entre aplicaciones que intercambian información entendible por las máquinas en el contexto de la Web), RDF/S [73] (RDF Schema, que provee el vocabulario esencial para describir documentos), XML (que restringe la estructura del documento), DAML+OIL (que operan sobre RDF Schema y XML, para definir un lenguaje con semántica bien definida), OWL (Ontology Web Language, que extiende DAML+OIL con tres niveles de expresividad creciente: OWL Lite, DL y FULL), F-Logic [1], SHOQ(D), GOL (General Ontological Language), SWRL (Semantic Web Rule Language) [65], etc. Sin embargo, estos lenguajes, desde el punto de vista de su expresividad pueden necesitar de un conjunto de otros [65] como Alloy, UML/OCL, VDM, Z, y Object-Z. Se espera que éstos sigan evolucionando hasta comprender aspectos como el comportamiento de los Web Services semánticos, agentes, autómatas temporales, etc.

Las primeras iniciativas concretas fueron aplicadas a dominios muy específicos. Entre ellas UNSPSC [161] y RosettaNet [158].

Basándose en estas tecnologías se desarrollaron herramientas relacionadas con el razonamiento ontológico como RACER (que permite crear, mantener, borrar ontologías, conceptos, roles, etc.), FaCT (automatiza razonamientos de consistencia e hiponimia[8]), OntoEdit (editor de ontologías con capacidad de

[7] La cohesión [47] es el vehículo por el que se interconectan de una forma precisa los elementos de un texto o discurso, mediante el uso anafórico de pronombres, la repetición, etc.

[8] Relación de inclusión de significado entre términos. Dada una lengua, el vocabulario está organizado en relaciones jerárquicas. Un término es hipónimo de otro si el significado del segundo está incluído

inferencias), DOGMA (editor con escalabilidad y reusabilidad), etc. Aunque, no pueden incorporar ciertos razonamientos ontológicos complejos y deben complementar manualmente su actividad para completar el objetivo de las herramientas que tiene implementadas.

(ii) Frameworks[9]

En cuanto a los entornos de trabajo, son varias las propuestas y bastante amplias las alternativas. Algunas de ellas son mencionadas aquí.

El framework ODESeW (Semantic Web Portal Based on WebODE[10]) [26], es un entorno de trabajo basado en WebODE, que permite el desarrollo de portales con manejo automático de conocimientos.

El proyecto ContentWeb, es una plataforma basada en ontologías e integrada en WebODE [163], la cual permite a los usuarios interactuar con lenguaje natural dentro de un cierto ámbito temático. Esta tiene como objetivos integrarse con OntoTag [2] [3] (Ontological Semantics, implementada con RDF/S y XML, es un modelo de anotación lingüística y ontológica), OntoConsult (interfaz en lenguaje natural basada en ontologías), OntoAdvice (sistema de recuperación de información basado en ontologías). Como base para el desarrollo de ContentWeb se usaron lenguajes XML y RDF. Cada palabra recibe un URI (Uniform Resource Identifier). También asigna nuevos URI para relacionar elementos morfosintácticos.

En [1] se describe también WebSifter, un meta-motor[11] de búsquedas que usa ontologías para refinar las consultas realizadas por el usuario, lo que permite aumentar la precisión en las búsquedas Web.

En [137] se presenta una red denominada FrameNet Español, que estudia la organización conceptual de la red de clases semánticas que configura el léxico de predicados del español. Para ello realiza tres actividades: detecta las clases

---

(implicado) dentro del primero. Ej. el significado de animal está incluído en el de conejo. Animal es hiperónimo de conejo, Conejo es hipónimo de animal.

[9] Framework es el término inglés para "entorno de trabajo". En el Diccionario y Nomenclatura se halla la definición y mayor información sobre la palabra framework.

[10] WebODE es un entorno de trabajo para hacer ingeniería ontológica.

[11] Meta motor denota a un sistema de búsqueda que depende de un motor para para realizar las búsquedas. Su actividad es más bien complementaria y a la del motor que se le incorpore.

semánticas (a los que denomina marcos semánticos o frames), detecta los argumentos que las caracterizan (denominados frame elements), y realiza anotaciones semánticas y sintácticas cuando aparecen predicados pertenecientes a dichas clases. Una interfaz Web permite consultar sobre los marcos, argumentos y anotaciones almacenados.

En [75] se propone el diseño de un framework que integra los conceptos de Web Semántica con funcionalidades apropiadas para personas discapacitadas. La propuesta incluye consideraciones para distintos tipos de minusvalías y estratos socio-culturales.

(iii) Buscadores en dominios acotados

Si bien existen buscadores generales de información, también se ha explorado en prototipos y sistemas funcionales que se limitan a colecciones acotadas, algunas de ellas se mencionan a continuación.

En [141] se presenta una metodología para la identificación y selección de textos en el ámbito de una lengua de especialidad (específicamente lenguaje económico empresarial), con empleo de conceptología[12] en un contexto concreto y no genérico. Se opone a la conceptología fuera de todo contexto y ofrece una definición objetiva y científica de cada término como una unidad lingüística natural que activa un contenido semántico especializado en un contexto especializado.

En [149] se presenta una propuesta para la manipulación de porciones de código, de manera que una consulta en lenguaje natural pueda retornar componentes de software candidatos a ser usados para un problema específico. Con éste tipo de manejos se pretende fomentar el reuso de porciones de código. Aplica semántica para la descripción y procesamiento de la consulta.

Dentro del área de la medicina, MachineProse [139] se presenta como un entorno de trabajo que se desentiende de los vocabularios controlados e indexa

[12] Conceptología: teoría de los conceptos. Disciplina que analiza, describe y estudia los aspectos teóricos y prácticos de la creación, la estructura, el desarrollo y las aplicaciones interdisciplinarias (en Lingüística, en Ciencia y Tecnología, en Organización del Conocimiento, en Terminología, en Lógica, etc.) de los conceptos relacionados con cualquier rama del conocimiento.

artículos semánticamente. Tiene la capacidad de recuperar los artículos relevantes con precisión y puede contestar a preguntas acerca de los contenidos.

(iv) Buscadores sensibles a actividad del usuario

Existen buscadores que prestan atención a la solicitud de información y a la vez intentan perfilar el tipo de usuario para adaptarse a las preferencias del mismo. Para ello suelen rastrear y procesar la secuencia de pasos del usuario cuando está interactuando con el sistema. Algunos de estos buscadores se mencionan aquí.

En [50] se propone una interfaz de búsqueda que soporta tipos de actividad de usuario. Desarrolla un sitio que se adapta dinámicamente a la actividad de búsqueda del usuario, organizando los metadatos. También provee una facilidad intermedia entre la búsqueda y el manejo de hipertextos, que sugiere por dónde continuar la búsqueda.

En el proyecto FLAMENCO [33] también se propone el uso de metadatos para desarrollar soluciones orientadas a la actividad (task oriented) y no a colecciones de documentos (collection oriented). El objetivo es mejorar la búsqueda entre dominios distintos a partir de consultas. Según esta propuesta, una cantidad suficiente de metadatos manejada adecuadamente permitiría sugerir al usuario cómo continuar el rastreo de la información buscada, una vez iniciada. En [31] se presenta el mismo prototipo como una interfaz de búsqueda dentro de colecciones de imágenes.

SenseMaker [144] es un sistema de exploración de información de tipo interactivo para información heterogénea. Su modelo de conceptos tiene la característica de organizar interactivamente los documentos en una estructura conceptual según ciertas dimensiones contextuales. La estructura también permitiría la búsqueda y filtrado basándose en conceptos.

(v) Buscadores combinados con otras estrategias

En un paso más allá de los metadatos, English et. al. [33] propone una disposición jerárquica y su manipulación dinámica de acuerdo a la búsqueda en cuestión, para obtener mejores resultados que con las búsquedas estándar sobre los metadatos.

También los TM (Topic's Maps) [30] proponen un uso asociativo de los metadatos para expresar semántica similar. Se supone que estos mapas orientan a quienes buscan algo puntual y les ayuda a hallar información relacionada. No son recomendados para quienes navegan en términos generales. Entre estos se puede mencionar TM4L, específicamente diseñado para temas educativos que, además, provee vistas según cierta relevancia inferida.

(vi) Diccionarios y Sistemas de manipulación de relaciones semánticas

En un paso para la búsqueda semántica implícita en las palabras, existe un valioso desarrollo como el Thesauro [43], un diccionario que permite organizar los términos en tópicos relacionados para poder favorecer una mejor clasificación automática de documentos. Golub ha comparado la eficiencia de su uso conjunto con distintas técnicas de clasificación.

En [122] [123] se presenta un método de detección de relaciones semánticas en palabras compuestas del inglés dentro del contexto médico. Utiliza simplemente la conjunción de sustantivos en palabras compuestas. Con base en los sustantivos, utiliza una clasificación con redes neuronales para inferir si una nueva palabra tiene relación semántica con otras ya previamente conocidas, clasificadas y organizadas jerárquicamente. Esto serviría como base para deducir la relación semántica entre dichas palabras.

En [126] se presenta un estudio de creación automática de relaciones semánticas entre conceptos de un documento, dentro del dominio de la zoología. Sobre la base de un estudio fraseológico y partiendo del reconocimiento sintagmático, detecta tipos de relaciones predefinidas. La

metodología mostró en los resultados ser mucho más eficiente en dominios acotados que en los genéricos.

En [133] se trabaja con el manejo semántico de abreviaturas en inglés para textos en el área biomédica. Dada la aparente variabilidad y crecimiento en éstas, obliga a un tratamiento más flexible que la asociación de anotaciones semánticas estáticas. Por, ello luego de un sencillo pre-procesamiento, construye automáticamente pares del tipo <forma-corta, forma-larga> correspondientes a la abreviatura y una serie de palabras candidatas a ser la descripción larga. Luego extrae la descripción larga final comparando la primera letra de cada palabra candidata contra las siglas. Los resultados de este trabajo muestran que es un algoritmo sencillo y eficiente para text mining.

En [71] se presenta un sistema de adquisición de semántica semi-automático para el área de compañías de seguros. Es capaz de procesar información semi estructurada (en diccionarios) y documentos en lenguaje natural.

(vii) Otras aplicaciones

La semántica también se ha propuesto para la calificación automática del ranking de servicios Web (WS, o Web Services) en combinación con componentes genéricos de los WS (en general estos componentes constituyen atributos propios del servicio y los tipos de atributos que parametrizan la invocación del mismo). Como ejemplo pueden mencionarse WSAP (Web Service Agent Proxy) [93] y WSMF (Web Service Modeling Framework) [1]. También se ha estudiado la aplicación de semántica en el contexto de WS compuestos por dos o más WS. [60].

Más allá de todos estos avances y de la empatía que la sociedad científica pueda tener hacia la Web Semántica, pocos tratados se dedican exclusivamente a un análisis profundo sobre las fortalezas y debilidades de la misma [36]. Por caso, si se considera RDF/S, desde una perspectiva de semántica formal, existen roles y limitaciones en la Web. Para entender ésto se debe tener presente que hay dos concepciones de la palabra “semántica” en el contexto Web:

a) como los significados percibidos por los humanos acerca de ciertos recursos de información

b) como la semántica modelizada teóricamente.

En el caso a), un documento tiene etiquetas (normalmente denominada "tag"), como <TITULO> o <AUTOR>. En general se trata de un markup-language[13] o ML (como XML), donde el significado concomitante es el que resulta práctico para sus autores y usuarios. De esta manera un conjunto de tags con sentido para ciertas personas es útil para realizar transacciones en la Web.

En el caso b), se usan estructuras abstractas para asignar interpretaciones a contrucciones legítimas del lenguaje. Los defensores de esta concepción tienen como objetivo que la Web Semántica interprete estos tags y realice inferencias válidas acerca de los recursos de información. Esta última interpretación es la usada por los creadores de los ML.

Es curioso cómo la literatura técnica para la Web Semántica suele confundir ambas visiones dejando un objetivo que a veces mezcla en diversos grados de ambas[36]. Entre los mismos desarrolladores de los lenguajes para Web Semántica, estas dos visiones son punto de discusión y suele referirse como el "significado social" (social meaning). Algunos autores, ya en el año 2003, comenzaron a tomar posiciones al respecto.

Por otra parte, la modelización teórica con estructuras abstractas permitiría actividades tales como desambiguar en función de la estructura, interpretar una sentencia y su valor de verdad dentro de ese contexto, realizar ciertos razonamientos automáticos y controlar las contradicciones. Pero aún con esta modelización de estructuras abstractas, no se alcanzan a resolver completamente problemas como la determinación del significado real subyacente y la conexión entre la estructura abstracta (significado) y el mundo real:

[13] Markup language refiere a un lenguaje caracterizado por el uso de etiquetas o tags. Ver el diccionario.

a) El hecho de que la estructura abstracta no alcanza para expresar el significado real, se ve reflejado en que:

> -La estructura abstracta no permite a las máquinas entender conceptos en el sentido que los humanos lo hacen. De hecho los avances en ML van en sentido contrario. Al respecto en RDF/S se declara[14] en [49] que: "la principal utilidad del modelo semántico teórico no es proveer un análisis profundo de la naturaleza de las cosas que se describen con el lenguaje... sino proveer un mecanismo técnico para determinar cuándo el proceso de inferencia es válido".
>
> -Respecto a la lógica de primer orden (de la cual se extiende el modelo semántico teórico aplicado en los ML), Hodges afirma que las sentencias de primer orden nunca pretendieron significar nada sino condiciones que pueden ser satisfechas o no. [57]
>
> -Etchemendy y Barwise aseveran que la veracidad de las sentencias no es parte del alcance de la semántica. [34]
>
> -Hodges también declaró que el modelo semántico teórico de los lenguajes naturales es una forma de describir los significados de sentencias en lenguaje natural y no una manera de asignarles significado. [58]

b) Respecto a la desconexión con el mundo puede decirse que:

> -Cuando se manifiesta una expresión en lenguaje oral o escrito, ésta tiene múltiples interpretaciones, representativas de distintas esferas del mundo real. Por ejemplo, considerando el caso de la relación "es-madre-de", tiene muchas interpretaciones: legal, biológica, social, metafórica, etc. El modelo semántico teórico no provee tampoco ayuda para el discernimiento entre las mismas. [36]

[14] "[t]he chief utility of a formal semantic theory is not to provide any deep analysis of the nature of the things being described by the language...but to provide a technical way to determine when inference processes are valid"

-El modelo teórico no es una teoría semántica que relaciona los lenguajes naturales con la realidad física y social. Es una teoría matemática que relaciona estructuras matemáticas con otras estructuras matemáticas. [101]

-Los sistemas de modelización no ofrecen intrínsecamente una interfaz particularmente sencilla para interpretar el mecanismo de manejo de significados. [110]

2.5. Criterios morfosintácticos

Reznikov [120] asegura que "el signo funciona como vehículo de un significado, como soporte de una información con respecto a un objeto determinado". A su vez, el pensamiento humano se formaliza en signos, y es a través de un número relativamente limitado de signos como puede expresarse una infinita cantidad de objetos, ideas, propiedades, características, situaciones y relaciones. Es difícil decidir si la comprensión semántica de textos es posible para un sistema computacional, debido a que no se pudo acordar en la comunidad científica si el razonamiento humano, del cual sería fruto, es computable o no [154]. Algunas fuentes afirman que es contradictorio o bien sus reglas no pueden ser abarcadas en un sistema formal [99]. El probar cada una de estas propiedades ha dado origen a diversos trabajos de investigación.

Asimismo, los teoremas de Gödel imponen limitantes para que una teoría lógica sea completa (formalmente consistente). En el plano de la computación hay quienes traducen estas limitantes a nociones de no-computabilidad del razonamiento, aunque un análisis profundo como el de Hofstadter [59] parece indicar que la no-computabilidad del razonamiento está lejos de ser determinada. Yendo a niveles aún más subjetivos, existen también posturas como las de Joseph Weizenbaum [146] y John Searle [135] que postulan la existencia de una especie de entendimiento sólo propio de los seres vivientes, que impediría a las computadoras simularlo de manera alguna.

Si se limita el análisis al aspecto meramente lingüístico, el problema no deja de ser complejo. Tal como se expresa en [47], el significado de los textos no está solo en la semántica, sino entrecruzado en todos los niveles lingüísticos. No se trata sólo del significado de las palabras, ni el orden de su combinación (las reglas de sintaxis). El texto escrito hereda la complejidad del razonamiento que lo genera. Se establece la necesidad de reflejar en niveles superiores el significado aportado desde niveles inferiores del análisis. Esto hace pensar que una buena solución a niveles morfológico y sintáctico es crítica para el manejo semántico. Es sobre la base de ésto que se justifican los desarrollos de las alternativas presentadas en esta sección y en especial, la propuesta WIH.

A lo anterior se suma el hecho de que la manipulación de la semántica subyacente en un texto requiere no sólo la explicitación de los rasgos morfológicos y sintáctico, sino también aquellos de tipo fonématico y fonético[15] [111] [112], y sobre todo de contexto. Debe destacarse la importancia del contexto, pues sin duda la cadena sintagmática es un factor importante de desambiguación.

Existen varios estudios para manipulación de este mencionado significado morfológico y sintáctico, como medio de apoyar el procesamiento semántico en sí mismo o como herramientas y propuestas acoplables a otras. Algunas de ellas se describen a continuación, clasificadas por sus características fundamentales.

(i) Desambiguación

En ciertos casos, una palabra puede tener un uso ambiguo desde algún punto de vista (léxico, sintáctico, semántico, de discurso, etc.). Su tratamiento puede realizarse con técnicas alternativas. A continuación se presentan algunos tratamientos morfosintácticos.

En [125] se estudió la desambiguación de nivel léxico y se comprobó empíricamente que es muy importante considerar herramientas específicas según el área de aplicación (es decir, el tema en cuestión). La potencia de desambiguación adquirida se constató en textos médicos y textos genéricos

[15] Paul y Cant presentan como un aspecto fluctuante dentro de la realidad del uso de la lengua el hecho de que una partícula puede o no tener distintos fonemas y paralelamente esto puede o no tener significados distintos. Una distinción entre fonema y fonética, complica aún más el tratamiento completo de la semántica.

con la misma herramienta arrojando 0.4% contra un 0.7% de tags incorrectos respectivamente.

El grupo CLIC desarrolló el corpus CLIC-TALP [25] [92], con anotaciones morfológicas y sintácticas para el español (conocido como corpus sintáctico Cas3LB). Pero aunque el anotado morfológico es automático, el desambiguado y el anotado sintáctico se realizan manualmente. Se basa en el principio de constituyentes [23] (oracionales y no oracionales) y funciones sintácticas.

En [97] se presenta un desambiguador sintáctico para el italiano llamado SenSOR (the SEmaNtic Subject-Object desambiguator). Usa una base de conocimientos que se alimenta automáticamente de diccionarios en línea (on-line).

En [21] se presenta una técnica de desambiguación para el Coreano, con parsing incremental de un grafo de morfemas. El sistema empleado se denomina KCCG (Korean Combinatorial Categorical Grammar).

En [72] se desarrolla el concepto de hypertags. Asigna a cada palabra un árbol elemental a diferencia de los tradicionales procesamientos POS (Part Of Speech), donde se asignan a una sentencia. Este árbol representa una estructura sintáctica elemental) como técnica para mejorar la asignación de tags[16] al procesamiento de POS (Part Of Speech). Extiende así la información morfosintáctica con información sintáctica para el tratamiento de ambigüedades.

(ii) Traducción automática

Para realizar la traducción automática de textos de un idioma a otro, debe plantearse una cierta correspondencia entre partes del texto origen con el destino. A este proceso se lo denomina “alineación” entre idiomas. Requiere de tratamientos morfosintácticos complejos. Aquí se presentan sólo unas pocas de las muchas propuestas existentes.

En [102] se muestra que estos sistemas (denominados SMT, por Statistical Machine Translation) los resultados mejoran cuando aumenta la cantidad de

[16] TAG son las siglas de un entorno denominado Tree Adjoining Grammars. Refiere en este contexto a etiquetas para un ML.

información de entrenamiento, para inferir la estructura del modelo subyacente a partir de datos lingüísticos. La experiencia indica que es recomendable incorporar explícitamente conocimientos acerca de los lenguajes en juego [103]. En [115] se presenta una extensión del tratamiento, que incorpora información de las interdependencias morfosintácticas (de las formas base de los sintagmas léxicos y sus distintas derivaciones), para realizar el entrenamiento en los SMT.

El proyecto AMETRA [152] implementa una herramienta de traducción asistida por ordenador, combinando técnicas de análisis lingüístico y métodos estadísticos. Se procesan castellano y euskera. Su ámbito está acotado al administrativo (boletines oficiales, documentos legales, etc.). El análisis lingüístico se limita a etiquetaciones gramaticales de palabras y segmentado de frases. Luego, aplica lógica difusa para recuperar frases con cierto nivel de semejanza al buscado. Puede decirse que su nivel máximo es el morfosintáctico y luego continúa su tratamiento con modelos estadísticos.

(iii) Detección y manejo de patrones sintácticos.

Así como existen herramientas para manipulación semántica (ver Diccionarios y Sistemas de manipulación de relaciones semánticas) existen sistemas análogos para tratamiento sintáctico. Estos sistemas son muy variados, algunos de ellos son mencionados en esta sección.

HELAS [29] realiza una identificación de patrones sintácticos relevantes de manera automática. Da ciertas estadísticas de palabras y con la información lingüística adecuada es capaz de detectar secuencias de palabras relevantes dentro de un texto.

En [70] se describe un algoritmo implementado para resolver ciertos tipos de anáforas. En vez usar un parser[17] para el procesamiento, utiliza una combinación de análisis de constituyentes[18] e inferencias acerca de las relaciones funcionales entre palabras. Los resultados obtenidos solo presentan

[17] Parser: Parseador. Programa o módulo encargado del parseo de un texto. Ver en el diccionario Parseo.

[18] Análisis de constituyentes: [162] Este análisis dice si una unidad está constituida por otras menores, que serían sus constituyentes. Muestra la estructura de las lenguas.

un mínimo compromiso de la calidad del resultado a cambio de un procesamiento genérico con sólo anotaciones léxicas en el texto de entrada.

En [52] se presenta un estudio de las bases comunes del procesamiento sintáctico y su implementación usando un lenguaje específico denominado DRL (Dependency Representation Language).

En [16] se propone un método no supervisado para el aprendizaje automático de información sintáctica. Se basa en dos principios: detección de pistas morfosintácticas en vez de parsing de sentencias enteras[19] y el tratamiento probabilístico de estas pistas halladas.

En [118] se presentan técnicas para el parsing morfosintáctico para lenguas altamente inflexionales (lengua inflexional en gramática es la alteración de la forma de una palabra por el agregado de un afijo, o por cambio de la base, que reflejan características gramaticales tales como persona, número, etc. Puede indicar un patrón de paradigmas de conformación de palabras [160]).

En [80] se presenta una técnica para extraer un corpus morfosintáctico probabilístico a partir de un corpus[20] sin tags como estrategia para evitar el costo de asignar tags cuando el corpus es grande. El uso facilita la desambiguación morfológica en Hebreo y posiblemente en otras lenguas también.

En [113] se estudia la detección automática de colocaciones[21]. Se realizan estudios estadísticos que muestran claramente mejoras cuando aumenta el corpus (trabajando con estadísticas de co-ocurrencia combinadas con información lingüística sencilla), trabaja especialmente con combinaciones tipo verbo-sustantivo y sustantivo-adjetivo.

En [61] se detectan las variaciones sintácticas especiales en títulos y abstracts de documentos técnicos. Las tres variaciones detectadas son: permutación (cambio en la secuencia de palabras), expansión (un término es reemplazado por una cadena de palabras) y sustitución (un término reemplazado por otro). Estas variaciones luego son estudiadas en relación con la función gramatical

[19] parsing se sentencias: aplicación de parseo a sentencias. Ver en el diccionario parseo.
[20] Corpus: ver en el diccionario corpus.
[21] Según Manning son asociaciones de palabras estadísticamente significativas que representan una forma convencional de decir las cosas.

de los términos que involucra, y luego son clasificados en grupos. Con ésto logra detectar automáticamente los tópicos de investigación declarados en el texto.

(iv) Reconocimiento y corrección de errores en textos

En el área de corrección automática de textos escritos, existen tres problemas principales: [76] detección de cadenas que no corresponden a palabras válidas, corrección de palabras con errores, y corrección de palabras de acuerdo con el contexto. En [76] se presentan las principales estrategias hasta el año 1992. Algunas otras propuestas se muestran a continuación.

En [143] se presenta un sistema útil para procesar texto desde OCR, con análisis de errores tipográficos (errores de escritura o de OCR), ortográficos (transliteración errónea de fonemas en grafemas) y morfosintácticos (mala aplicación de inflexiones morfológicas y reglas de sintaxis) aplicados al alemán.

En [41] se presenta parte de un corrector de errores en textos escritos, donde se aplica parsing sintáctico basado en una jerarquía léxica en vez de un conjunto de categorías léxicas. Toma al texto como un conjunto de frases. Cada frase a su vez es una palabra modificada por el resto. De este modo todo termina organizado en un árbol. La información compleja (ej. Funciones sintácticas, relaciones semánticas) son asimilables a esta estructura arbórea utilizando principios análogos.

En [78] se proponen técnicas para normalización sintáctica de lenguaje hablado, proponiendo eliminaciones heurísticas y generalizaciones de ciertas características, para la corrección de sentencias mal formadas en idioma alemán.

En [20] se presenta un proyecto de corrector orto-tipográfico con morfosintaxis. Plantea distintas categorías para los diversos tipos de problemas a corregir: tipografía, ortografía, parentización, gramatical (patrones predeterminados), gramatical general. Se basa en procesamiento de tipo morfológico y usa lematización con etiquetación por tagger markoviano.

También define patrones y un experto con reglas de gramática sintagmática[22] para la gestión de errores gramaticales. Existen aplicaciones para diversos idiomas (Ej. en [32] se presenta un corrector morfosintáctico y sintáctico de textos escritos en francés) y en ámbitos de especialidad (Ej. En [42] se presenta un prototipo que realiza un procesamiento en tres niveles, de órdenes con indicaciones médicas. Dado lo acotado del área de aplicación, la verificación de consistencia sintáctica es más sencilla. En un primer nivel, se reconocen los caracteres escritos. En segundo nivel, se entrena al sistema para reconocer la escritura de un médico específico y se reconocen palabras. Luego se realiza el procesamiento sintáctico).

(v) Navegación Web

En el proceso de recuperación de documentos en la Web es imprescindible la óptima indización automática[23]. Para ello se debe incrementar la correspondencia entre términos de consulta y términos almacenados. Muchos sistemas agrupan las variantes de términos por medio de métodos de unificación (o conflación)[24]. Hay dos estrategias principales que manejan estas variantes de términos [40]:

a) stemmer: agrupa las variantes a un stem [116][148] [95] [68][94], definido como base o radical de la palabra. Existen varias técnicas para el inglés y otros idiomas [132], con distintos grados de fortaleza[25] , similitud[26] [38] y eficiencia [108] [48]. Específicamente el de Porter [116] ha sido implementado en inglés, francés, español, italiano, portugués, alemán, etc. y está disponible en la Web [159]. Es el algoritmo utilizado dentro de WIH para el procesamiento en la Estructura Virtual.

---

[22] Gramática sintagmática es la que estudia la vinculación entre las palabras en una estructura proposicional. Sintagma es una constitución de una o más palabras o elementos semánticos.

[23] Stevens en 1965 definió indización automática como el uso de máquinas para extraer o asignar términos de indización sin intervención humana.

[24] Los métodos de unificación son definidos como métodos computacionales encargados de la agrupación de variantes de términos, semánticamente equivalentes, a una forma normalizada.

[25] Fortaleza (strength of) mide el grado en que el algoritmo stemmer cambia las palabras y tiene relación con las métricas de precisión (precision) y recuperación (recall), así como con el grado de compresión del correspondiente índice.

[26] Similitud (similarity among) es una métrica complementaria a la fortaleza propuesta por Frakes y Fox. Consiste en una magnitud numérica que indica el grado de similitud entre las n-uplas generadas por $n \geq 2$ estrategias de stemming distintas.

b) lematizador: asocia el término a un lema, definido como el conjunto de palabras con igual raíz, y categoría léxico-gramatical o etiqueta POS (part of speech). Desde la lematización, la ayuda es especialmente necesaria para lenguas muy flexivas como el español o el alemán. Uno de los problemas del tratamiento de este nivel es su alta dependencia del idioma. Esto puede provocar que idiomas con morfologías relativamente sencillas como el español o el inglés [4] no permitan el descubrimiento y desarrollo de información relacionada con el lema de la palabra. Por caso EUSLEM es un proyecto para encarar la lematización y etiquetado automático de textos en euskara, idioma de tratamiento muy complejo. También pueden mencionarse los sistemas DAWeb y NAWeb [129], de soporte a la navegación y manejo fluido de Internet. Entre otras cosas, realizan detección de neologismos, definen métricas cuantitativas y cualitativas del uso de palabras y realizan el estudio masivo de los sitios con técnicas morfológicas.

Finalmente, cabe mencionar una arquitectura presentada en [114], que se basa en la extracción y uso de sintagmas[27] tanto para enriquecer y traducir las consultas de un usuario a Web como para acceder a la información desde los sintagmas sugeridos por el sistema. La extracción de sintagmas se realiza por medio de patrones morfosintácticos. Esta arquitectura afecta al proceso de indexación, al de recuperación y al modelo de interacción con el usuario. Para tratar la variación terminológica y multilingüismo en el acceso a la información recurre a la inferencia lingüística.

(vi) Indexación de textos

Como parte de un sistema de búsqueda existen estrategias para indexar los documentos accesibles, de manera que el proceso de evaluación y recuperación de textos sea más eficiente. Algunas de las soluciones implementadas se presentan a continuación.

En [13] una herramienta lingüística realiza la indexación automática de texto en lenguaje natural. Extrae las frases nominales (noun phrases, consideradas parte del concepto principal) pero sin llegar a realizar un análisis sintáctico

[27] Sintagma es una constitución de una o más palabras o elementos semánticos.

profundo. Se basa en el concepto de que los grupos de sustantivos son buenos para indexar cuando el manejo es por keywords.

En [145] se presenta una comparación entre los léxicos construidos sobre la base del procesamiento de palabras y los construidos basándose en morfemas[28], para el uso en sistemas de procesamiento de lenguaje natural. Muestra empírica y teóricamente la superioridad de los primeros.

En [8] se presenta un mecanismo de adquisición de conocimiento a partir de textos técnicos. Primero realiza un análisis morfosintáctico para extraer términos candidatos (con un aplicativo llamado LEXTER). Luego, aplica agrupamiento (clustering) e ingeniería del conocimiento para definir semi-automáticamente los campos conceptuales de dominio. En este trabajo, el concepto de semántica o significado se realiza por contraste con otras unidades morfosintácticas y no por asignación del mismo. Desde el punto de vista de Zernik [151], se podría decir que el tratamiento es "por uso de la palabra más que por significado" ("word usage rather than word meaning").

(vii) Expansión de términos

Cuando se trabaja con una palabra en cierto idioma para buscar información, es conveniente poder hallar textos que se relacionen conceptualmente aunque no contengan exactamente ese término. Al proceso involucrado en obtener otras palabras relacionadas semánticamente se le denomina expansión. Debajo se presentan unos pocos de estos sistemas para mostrar algunas de sus características.

Jacquemin [64] describe un sistema de expansión automática de términos basado en análisis morfosintáctico y sintáctico. Puede partir de un texto o bien de una lista de términos, como dispositivo de indexación automática para mejorar la actividad de recuperación de información.

En [17] se propone un método de normalización morfosintáctica automática de textos dentro del ámbito de un corpus relativo a sustancias tóxicas. También puede parafrasear (generar sentencias de igual información con

[28] Morfemas mínima unidad lingüística que tiene significado semántico [164].

distintas palabras). Para ello utiliza métodos simbólicos[29] aplicados a información sintáctica y morfológica y un corpus de conocimiento específico.

En [44] el proyecto TROLL encara el problema de la restricción de aplicabilidad en el tamaño y dominio de uso de los léxicos. Manipula las entradas léxicas al corpus como marcos (frames) conceptuales/fonológicos en vez de tomar como entradas a las palabras. También se complementa ésto con una serie de reglas de expansión que generan nuevas entradas a partir de los frames actuales. De esta manera un frame no sólo comprende las variantes de una palabra sino también palabras de distintas categorías que se derivan de la misma base semántica.

(viii) Frameworks

Se han desarrollado algunos entornos de trabajos o frameworks para facilitar el desarrollo y manipulación de un corpus morfosintáctico. Se refieren aquí algunos de ellos.

En [109] se presenta un entorno de trabajo gráfico denominado XTAG para el desarrollo de gramáticas asimilables a árboles (denominadas tree-adjoining grammar o TAG, contrarias a las gramáticas de reescritura de cadenas o string rewriting grammar puesto que la profundidad del árbol puede ser mayor a un nivel) y sus parsers.

En [62] se presenta un entorno de trabajo para anotaciones de nivel sintáctico flexible, capaz de procesar distintos niveles lingüísticos según los objetivos establecidos. Se basa en un modelo abstracto e implementa instancias con esquemas XML y RDF.

El entorno Morphologic Recognition Assistant [119] presenta un análisis morfosintáctico y sintáctico aplicado al proceso de texto generado automáticamente desde un sonido. El sistema realiza tres servicios: segmentación apropiada (separación de fragmentos respetando la coherencia del contenido), desambiguación de símbolos inespecíficos (derivados de sonidos no interpretados) y corrección de símbolos mal reconocidos.

[29] Método simbólico o por representación con símbolos.

(ix) Otros

Muchas son las variantes de aplicación de procesamiento morfosintáctico, algunas de las cuales se mencionan a continuación.

En [35] se presenta un sistema de reconocimiento de parafraseos, independiente de dominio y de gran escala. Se basa en reglas morfosintácticas y cuatro características sintáctico-semánticas asociadas a un conjunto de palabras predefinidas. Éste sería el soporte a un sistema de recuperación de información para mejorar las métricas de precisión (en inglés precision[30]) y recuperación (en inglés recall[31]).

En [130] se presenta un Web Service que genera etiquetas (tags) morfoléxicas de palabras en español. Es capaz de obtener distintas formas: canónicas o flexiones cuando se solicitan. También permite la manipulación de relaciones morfoléxicas.

En [131] se presenta una herramienta de recuperación morfoléxica aplicada a documentos escritos en formato Microsoft Word. Propone extender los menús tradicionales de este editor para considerar este tipo de tratamientos en tareas de búsqueda y reemplazos de texto. Específicamente el diseño trabaja con aspectos flexivos y derivativos de la lengua. Las facilidades extendidas son: buscar palabras con forma canónica acorde a alguna de las formas canónicas de la palabra en búsqueda, elegir grados de derivación y flexión de formas verbales y no verbales y manipulación del patrón de búsqueda (elecciones de flexión).

2.6. Information Retrieval (IR) y el posicionamiento

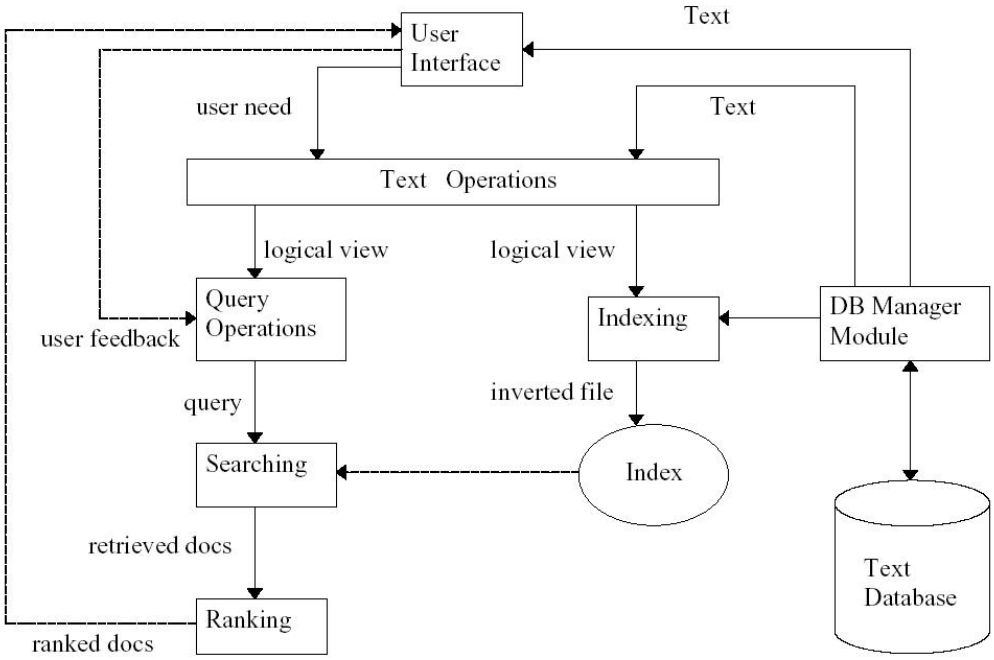
El lugar que una página de resultado ocupa luego de una búsqueda en la Web, tiene que ver en principio con la relevancia que, se estima, le otorgará el usuario. Esto impone un ordenamiento determinado y una modelización indirecta de las expectativas del mismo. Pero modelizar las expectativas requiere en cierta forma de

[30] Precision es la razón entre la cantidad de documentos relevantes recuperados por la cantidad de documentos recuperados.
[31] Recall razón entre la cantidad de documentos relevantes recuperados por la cantidad de documentos relevantes dentro de una colección.

modelizar al usuario mismo. Las secciones anteriores describen estrategias diversas para implementar la manera en que éste interpretaría la información escrita. Luego de esta interpretación, se requiere la priorización relativa de esos documentos que, presumiblemente, responden a su necesidad de información.

En términos generales el proceso puede esquematizarse como en la siguiente figura del autor Baeza-Yates [9].

Fig. 2. Proceso de recuperación de información.



Existen muchas alternativas de modelos para determinar la relevancia en el proceso de recuperación de información [117]. Se las puede clasificar en dos grupos:

a) Estrategias para recuperación de información: Modelos clásicos (Booleano, Vectorial o Probabilístico) o modelos estructurados (listas No sobrepuestas, nodos proximales) y a su vez responden a los distintos conjuntos teóricos (difusos/booleano extendido), algebraicos (vector generalizado, índice de semántica latente, redes neuronales), o probabilísticos (inferencia de red, redes de creencia).

b) Estrategias para navegación: plana, guiada estructurada o hipertexto.

Por ejemplo en [37] propone una categorización automática de los documentos como proceso de aprendizaje, donde el programa capta las

características que distinguen cada categoría o clase de las demás (las que deben poseer los documentos para pertenecer a esa categoría). La pertenencia es una escala graduada o coeficiente de pertenencia a cada clase existente. Las características se basan en la ocurrencia de ciertas palabras y en la realimentación del sistema basándose en consultas anteriores. Lo llamativo es que este tratamiento se realizó sin lematización previa de los documentos puesto que los autores la consideran riesgosa y una complicación innecesaria.

En este punto, es de rigor diferenciar relevancia y posicionamiento ya que no son lo mismo [124]. Mientras la relevancia corresponde a la posición relativa numéricamente calculada por cierto algoritmo interno de un buscador, el posicionamiento es la ubicación final real de cierto resultado en función de la relevancia y otros factores que son propios del navegador.

Dado que un usuario promedio sólo accede a los diez o veinte primeros resultados entregados por el buscador [134], el posicionamiento incidirá esencialmente en el éxito de una búsqueda. Por lo tanto, toda la tarea técnica que se realice para obtener los resultados conceptualmente más correctos para una búsqueda queda degradada si el posicionamiento no se relaciona con estos tratamientos.
Como parte de un estudio preliminar para contextualizar el presente trabajo, se estudiaron los factores de posicionamiento. Es interesante lo hallado en [124] con Rottestein puntualmente para Google, que es actualmente un buscador muy utilizado. Pudo mostrar estadísticamente que:

-El page rank (coeficiente técnico, numérico de priorización) no se correlaciona con el posicionamiento. (Ver Fig. 3).

-Los factores de posicionamiento: cantidad de enlaces, edad del dominio, cantidad de anchor texts, cantidad de páginas indexadas por Google en el dominio, dominios que enlazan la página, los meta-keywords (las dos etiquetas que se colocan en el encabezado de las páginas con la descripción de

las páginas) que suele considerarse correlacionados con el posicionamiento no tienen correlación con él.

-Los mismos factores no muestran tampoco correlación con el page rank.

-Si se estudian como grupos separados los diez primeros resultados (sean el grupo 1), los diez que le siguen (grupo 2) y los diez que siguen (grupo 3), se puede apreciar que el promedio de page rank decrece entre los grupos. Lo mismo sucede con el promedio de dominios que enlazan la Web, y la densidad promedio de keywords. Algunos de los resultados se reproducen en la Fig. 4.

Fig. 3. Posición y Page Rank en Google.

Los puntos marcan la relación entre la posición en los resultados de Google, y el Page Rank, la línea roja muestra como debería ser el gráfico, si existiera una relación lineal entre estos dos.

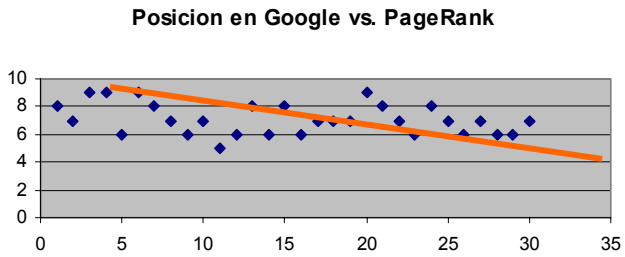


Fig. 4. Calidad de factores y posicionamiento.

Promedios de Page Rank, cantidad de enlaces (Factor 5), densidad de keywords en el texto de la página (Factor 6), páginas indexadas en Google de ese dominio (Factor 7).

Resultados en Google Clusters	Media del PageRank
Cluster A - 1 a 10 resultados	7,6
Cluster B - 11 a 20 resultados	6,9
Cluster C - 21 a 30 resultados	6,8

Resultados en Google Clusters	Media Factor 5
Cluster A - 1 a 10 resultados	725,4
Cluster B - 11 a 20 resultados	640,2
Cluster C - 21 a 30 resultados	517,0

Resultados en Google Clusters	Media Factor 6
Cluster A - 1 a 10 resultados	1,29%
Cluster B - 11 a 20 resultados	0,62%
Cluster C - 21 a 30 resultados	0,38%

Resultados en Google Clusters	Media Factor 7
Cluster A - 1 a 10 resultados	1.071.946
Cluster B - 11 a 20 resultados	769.090
Cluster C - 21 a 30 resultados	411.995

El trabajo logra establecer que ninguno de estos factores mencionados por los treinta especialistas internacionales convocados, hace que los resultados estén mejor posicionados en Google, sino que existirían otros criterios no considerados, por los que se determina la inclusión de un resultado en el grupo 1, 2 ó 3. Una vez dentro del grupo, habría un ordenamiento concreto decreciente por estos factores.

Es importante que, en el futuro, se trabaje sobre estos aspectos. Los diseñadores de buscadores deben tomar conciencia de que (más allá de todas las consideraciones eventuales respecto a la necesidad de posicionar un sitio) el dramático crecimiento documental de la Web exige cierta responsabilidad y el respeto de los factores que racionalmente se debieran vincular con el ordenamiento de los resultados.

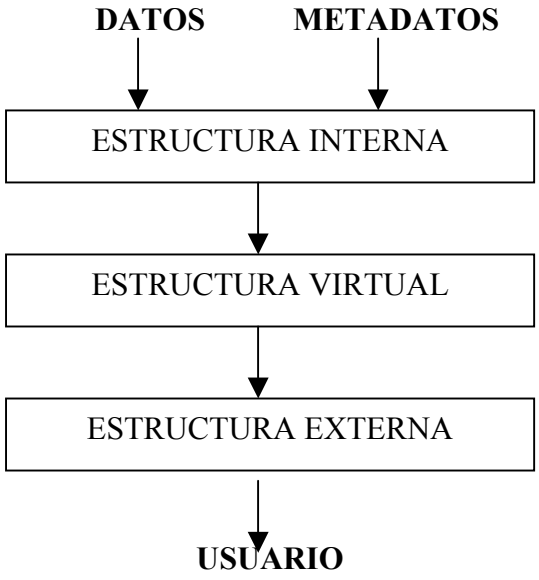
Capítulo 3. Descripción de la estrategia global

"If we spoke a different language, we would perceive a somewhat different world"

Ludwig Wittgenstein
(1889 – 1951)

La arquitectura general de esta propuesta ya presentada en [85], consta de tres niveles con distintos grados de abstracción (Fig. 5).

Fig. 5. Arquitectura global de la propuesta WIH



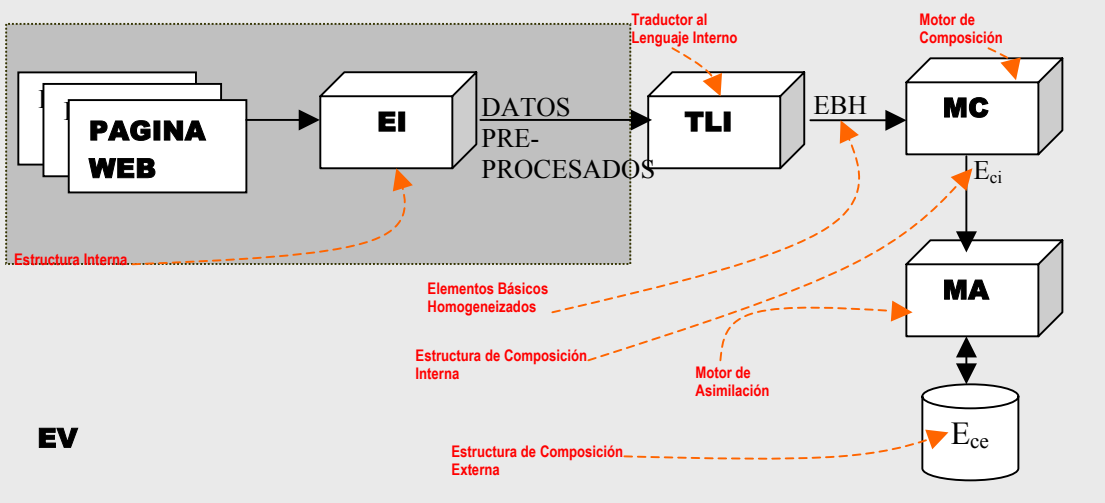
Esta arquitectura pretende ser una segmentación de la complejidad del problema según la visibilidad de datos. El trabajo consiste en el desarrollo de una propuesta de Estructura Virtual como punto de partida para el establecimiento de las restantes estructuras.

La Estructura Virtual es el corazón de la propuesta y su objetivo es almacenar de manera alternativa la información, como forma de reorganizar los datos de una manera más apropiada para la Estructura Externa. Desde este punto de vista podría decirse que es un middleware para transformar documentos desde una organización propia hacia una organización típica de los mecanismos de búsqueda.

3.0. Estructura Virtual

La **Estructura Virtual (EV)** tiene como objetivo realizar las transformaciones específicas necesarias para reorganizar los datos y metadatos capturados desde la Web y previamente preparados por la **Estructura Interna (EI)**. Deja, como resultado de su actividad, una red estructurada de componentes denominados $\mathbf{E_{ce}}$ (**Estructura de Composición Externa**), entendible por parte de la **Estructura Externa (EE)**, para que ésta pueda manipularlos al interactuar con un usuario. La arquitectura de la EV se detalla en la Fig. 6.

Fig. 6. Arquitectura de la Estructura Virtual (EV)

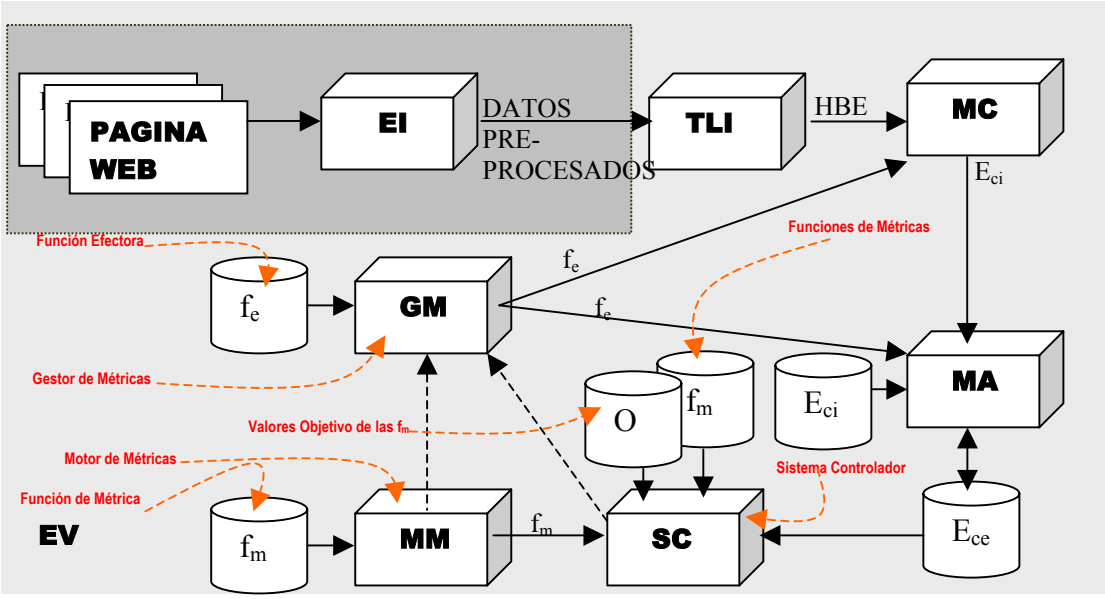


Básicamente, las páginas son obtenidas y preprocesadas por la Estructura Interna (indicada como EI) quien luego deja a disposición del Módulo de Traducción al Lenguaje Interno (en adelante TLI) los datos en formato plano, junto con la información necesaria para su procesamiento como página Web unívocamente determinada.

El TLI transforma los datos en ladrillos fundamentales denominados EBH (Elementos Básicos Homogenizados), que constituyen la base de codificación lógica de la información contenida en los datos. El Motor de Composición (MC), realizará un análisis rudimentario que le permitirá entrelazar los EBH y convertirlos en una estructura representativa de las frases. A esta estructura se la denomina E_{ci} (Estructura de Composición Interna). Las E_{ci} serán organizadas a su vez por el Motor de Asimilación (MA) en otra estructura representativa de todo el documento denominado E_{ce}, y finalmente insertadas en determinado locus dentro de la base de datos interna del sistema (denominada Red Virtual o RV).

Este mecanismo genérico vale para cualquier ingreso de información al WIH, pero debe ser extendido para los casos de manipulaciones de los contenidos por parte de sus dueños (ya sea por incorporación, modificación o eliminación de la misma). Para estos casos debe considerarse el esquema provisto en la Fig. 7.

Fig. 7. Arquitectura de la EV completa



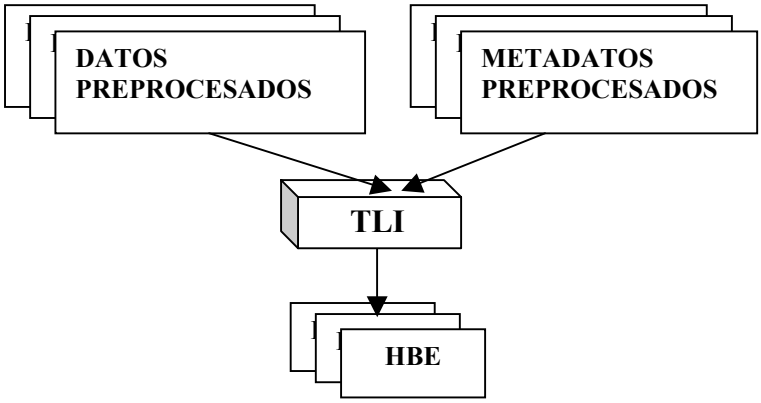
Se puede apreciar en el diagrama que el Motor de Composición en realidad actúa bajo la influencia de un Gestor de Métricas (GM) que dictamina las actividades del mismo, representadas en un conjunto de funciones denominadas funciones de efector (f_e). Lo mismo sucede con el Motor de Asimilación. Existe un Motor de Métricas (MM) que administra el conjunto de funciones de métrica que miden las actividades del sistema con una evaluación numérica específica para cada actividad objetivo a medir (luego este valor será comparado contra el objetivo representado con un valor umbral, que determina si se cumplen o no ese objetivo). En todo momento la actividad sobre las E_{ce} es monitorizada por el Sistema Controlador (SC), que detecta automáticamente los casos que consisten en modificaciones de páginas ya incorporadas. En estas circunstancias adapta convenientemente las f_e. Es interesante notar que este esquema de funcionamiento permite dar mucha flexibilidad al tipo y forma de procesamiento de la información no sólo para casos de modificaciones sino también para ajustar la mecánica de construcción de estructuras E_{ci} y E_{ce}, y en general para adaptar el funcionamiento según las necesidades globales del sistema a cada instante. Las

necesidades son reflejadas en un conjunto jerárquico de objetivos denotados en la figura como un tambor etiquetado con O, que alimenta al SC.

3.1. Módulo de Traducción a Lenguaje Interno

El Módulo de Traducción al Lenguaje Interno (TLI) tiene como objetivo la captura de datos ya preprocesados por la Estructura Interna, y la generación de un conjunto de elementos estandarizados (denominados Elementos Básicos Homogeneizados o EBH) listos para ser incorporados (ver la Fig. 8).

Fig. 8. Flujo del módulo TLI



Esta actividad es fundamental para la capacidad representativa de la **WIH** respecto al contenido original. El sistema genera ciertos campos que describen atributos morfosintácticos de las palabras y sentencias, y con ellos construye los EBH. Estos campos se extraen tomando como base el análisis realizado en [87] sobre los siguientes campos llamados "descriptores", como se detalla a continuación:

Descriptor 1-**tema (tema)**

El tema es un descriptor que representa en una sola palabra el sentido genérico del contenido de la página Web a la que pertenece la palabra.

Opciones consideradas: tema= {nuevo, orquídeas, topacio, natación}

Su finalidad es comparar estadísticamente (en el marco del trabajo [87]), si la temática incide en el comportamiento del resto de los descriptores. Este campo no es usado ni procesado en WIH.

Descriptor 2-**tipo de palabra (tipo-pal)**

La palabra es el principal medio formativo del concepto [156] y por ello es la base de la mayor parte de los descriptores que siguen. Este descriptor intenta considerar el criterio de Ortiz acerca de los enunciados sencillos: éstos sólo requieren de combinaciones nombre+verbo o nombre+adjetivo. En este trabajo sólo se considera la primera de estas dos posibilidades como un primer estadío de investigación en este descriptor. Cuando se realiza el procesamiento se combina este campo para cada palabra actual y su predecesora.

Opciones consideradas: tipo-pal= {sustantivo, verbo, otro}

El objetivo de este campo en el contexto del trabajo [87], es verificar estadísticamente la potencia del resto de los descriptores para inferir el tipo de sintagma léxico que corresponde a la palabra, por lo que se completó manualmente y luego se comparó contra las inferencias realizadas con árboles de inducción. Quedará como trabajo a futuro verificar la necesidad de incorporar otros tipos de sintagmas tales como adjetivos, preposiciones, conjunciones, etc.

En el contexto del prototipo WIH, basándose en los hallazgos del mencionado trabajo, el "tipo-pal" es inferido por un árbol de inducción.

Descriptor 3-**tipo de palabra anterior (pal-ant-tipo)**

Corresponde al tipo de palabra ubicada a izquierda de la palabra en proceso. Con este campo descriptor se pretende determinar alguna característica relevante acerca de la palabra anterior que sirva para inferir el tipo de sintagma de la palabra actual.

Ej. pal-ant-tipo={ninguna, otro}

Las opciones demarcadas en el ejemplo son las que se trabajaron, y permiten indicar si es primera palabra o no de una sentencia. Es un campo que se usa en el prototipo WIH. Algunas alternativas para trabajos a futuro, que podrían ser interesantes son: Artículo, preposición, pronombre-personal, pronombre-posesivo.

Descriptor 4-**tipo de página html (tipo-pag)**

Opción trabajada: tipo-pag={índice, contenido}

Corresponde la opción índice si en el URL de la página figura la palabra especial "index". Con este campo se estudió si el tipo de página es importante para determinar el contenido. Es uno de los descriptores incorporados por el prototipo de TLI.

Descriptor 5-**longitud de la palabra (long-palabra)**

Denota la cantidad de caracteres de la palabra en cuestión. Su objetivo es describir al sintagma en proceso y estudiar si tiene influencia en el proceso inferencial para pal-ant-tipo. Es un campo que se usa en el prototipo WIH.

Descriptor 6-**cantidad de vocales fuertes (cant-vocales-fuertes)**

Es la cantidad de vocales fuertes (a, e, o) en la palabra. Tiene el mismo objetivo que long-palabra. Es un campo que se usa en el prototipo WIH

Descriptor 7-**subfijo (terminacion)**

Opciones consideradas: terminación={ar, er, ir, or, ur, ra, re, ri, ro, ru, s, m, sa, se, si, so, su, an , en, in, on, un, cion, ciones}

Corresponde a la terminación de la palabra cuando es alguna de las opciones consideradas. Es null cuando la terminación no es una reconocida en la lista. Tiene el mismo objetivo que long-palabra. No es un descriptor usado ni procesado por WIH.

Descriptor 8-**cantidad de vocales débiles (cant-vocales-debiles)**

Es la cantidad de vocales débiles (i, u) de la palabra. Tiene el mismo objetivo que long-palabra. Es un campo que se usa en el prototipo WIH

Descriptor 9-**empieza con Mayúscula (empieza-mayuscula)**

Opciones posibles: empieza-mayuscula={si, no}

Denota si una palabra fue escrita con mayúscula o no. Su valor es "si" cuando la palabra empieza con mayúscula, de lo contrario es “no”. Tiene el mismo objetivo que long-palabra. Es un campo que se usa en el prototipo WIH

Descriptor 10-**resaltada (resaltada)**

Opciones posibles: resaltada={si, no}

Denota las veces en que la frase en la que figura la palabra que ha sido resaltada con comillas dobles o simples o bien cuando toda la palabra ha sido escrita en caracteres en mayúscula. Su valor es "si" cuando la frase está entre " o ' o en mayúsculas toda la palabra. Tiene el mismo objetivo que long-palabra. Es un campo que se usa en el prototipo WIH

Descriptor 11-**es título (es-titulo)**

Opciones posibles: es-titulo={si, no}

Denota las situaciones en las que una palabra pueda ser candidata a conformar el título de la página a la que pertenece. Su especial característica es que no se basa en la

existencia ni procesamiento de tags. Simplemente se considera título cuando pertenece a la primera oración de contenido dentro de la página.

Su valor es "si" cuando la palabra es título de la página. Al igual que en los casos anteriores, tiene igual objetivo que el campo long-palabra. Es un campo que se usa en el prototipo WIH.

Descriptor 12-**longitud de la oración (long-oracion)**

Es la longitud de la oración en cantidad de palabras válidas en castellano (es decir, sin contar los números ni signos especiales como palabras). Una oración es considerada como la colección de sintagmas hasta el próximo punto o señal de fin de línea. Tiene igual objetivo que long-palabra. Es un campo que se usa en el prototipo WIH.

Descriptor 13-**STEM (stem)**

Es una partícula reducida que representa la raíz de la palabra. La algorítmica de su extracción fue diseñada por Porter [116] y tal como se explica en [117], junto con la lematización (eliminación de palabras comunes en el lenguaje como artículos, conjunciones, etc.) consiste en una de las técnicas tradicionales de indexación de términos (aunque la lematización ha caído en desuso por haberse demostrado en la práctica que no es apropiada para todos los casos). Este descriptor es usado por su mencionada capacidad de representar las palabras con sólo un simple procesamiento local. Es un campo que se usa en el prototipo WIH

Descriptor 14-**palabra(id-palabra)**

Es la palabra en cuestión, origen de todos los descriptores mencionados. No es un campo usado en el prototipo WIH luego de la extracción de los descriptores.

Los estudios realizados en el trabajo [87] sobre 47820 palabras de 340 páginas, permiten afirmar que los descriptores son un representante bastante razonable del tipo de palabra originalmente hallada en el texto (94% de las veces se puede inferir correctamente el tipo de palabra en cuestión).

Descriptor 15-identificador de página (id-caso)

Denota unívocamente la página Web a la que pertenece la palabra. Si bien este descriptor se definió desde el principio junto al resto, no se procesa y está destinado a proveer acceso al documento original desde dentro de E_{ci} pertenecientes a la RV. Es un campo implementado en el prototipo WIH.

Descriptor 16-**profundidad página Web (Web-profundidad-pagina)**

Es el conteo de la cantidad de subdirectorios dentro del servidor, que se refleja en el texto del URL correspondiente a la página de donde se extrae la palabra en proceso. No es un campo usado en el prototipo WIH. Se estudió si tiene influencia en el tipo de palabras del texto.

Descriptor 17-**cantidad ocurrencias (cant-ocurrencias)**

Denota la cantidad de veces que la palabra actual figura en el documento. Se estudió si tiene influencia en el tipo de palabras del texto y sentencias. No es un campo usado en el prototipo WIH.

Descriptor 18-**cantidad de palabras (cant-pal-pagina)**

Denota la cantidad de palabras en el documento. Se estudió si tiene influencia en el tipo de palabras del texto y sentencias. No es un campo usado en el prototipo WIH.

Descriptor 19-**cantidad de números(cant-numeros)**

Denota la cantidad de números en el documento. Se estudió si tiene influencia en el tipo de palabras del texto y sentencias. No es un campo usado en el prototipo WIH.

Descriptor 20-**cantidad de signos especiales (cant-signos-especiales)**

Denota la cantidad de signos especiales en el documento que no califican como palabras ni sintagmas válidos del lenguaje. Se estudió si tiene influencia en el tipo de palabras del texto y sentencias. No es un campo usado en el prototipo WIH.

Descriptor 21-**palabra anterior (pal-anterior)**

Es la palabra que figura antes en el documento. Se usó sólo para verificaciones en los resultados y seguimientos de casos. No es un campo usado en el prototipo WIH.

Descriptor 22-**país de radicación (pais-radicación)**

Es el código de país codificado en el dominio. Ej: us, ar, es, cu, mx, etc. del URL donde figura la palabra. Con este campo se estudió si el país origen es importante para determinar el contenido. No es un campo usado en el prototipo WIH.

Descriptor 23-**sigue un signo de puntuación (sigue-puntuacion)**

Opciones posibles: sigue-puntuacion={si, no}

Define si hay un signo de puntuación válido luego de la palabra en proceso. Los signos considerados son: “.”, “,”, “;” y “:”.Su valor es "si" cuando se verifica la existencia de alguno de éstos. Tiene el mismo objetivo que long-palabra. No es un campo usado en el prototipo WIH.

Descriptor 24-**tipo de dominio de la Web (clase-pag)**

Opciones posibles: sigue-puntuacion={org, com, net, otro}

Define el tipo de organización declarada en el dominio del URL donde figura la palabra. Tiene el mismo objetivo que pais-radicacion. No es un campo usado en el prototipo WIH.

Descriptor 25-**es una frase especial (frase-especial)**

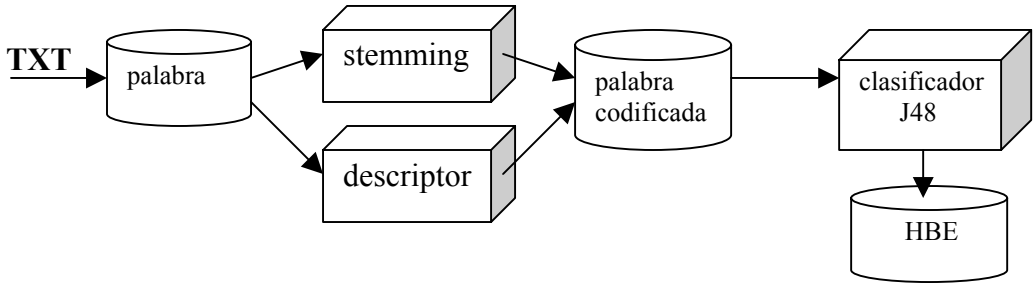
Opciones posibles: sigue-puntuacion={si, no}

Define si la palabra es parte de una frase especial porque ha sido destacada con alguno de los recursos válidos de la gramática. Concretamente se verifica si la frase que engloba la palabra está encerrada entre los signos: “¡!”, “¿?”, “< >”, “( )”, “[ ]”, “{ }” o “«»”.Su valor es "si" cuando se verifica la existencia de alguno de éstos. Tiene el mismo objetivo que long-palabra. No es un campo usado en el prototipo WIH.

Los descriptores fueron estudiados y, finalmente, sólo unos pocos demostraron ser útiles para el procesamiento del TLI. En general fueron descartados aquellos que exigen un costo computacional al nivel de páginas o sentencias, probado que la información que aportan no cambia el nivel de precisión de la actividad. Otros fueron descartados porque no aportan información significativa al proceso. El detalle de la selección se halla en una experiencia previa [87].

Para la obtención de los descriptores, se procedió a la implementación de dos módulos java de procesamiento independiente: Un módulo de stemming y un módulo descriptor. La salida de ambos módulos se ha coordinado de manera que entre ambos generen una nueva base de datos con todas las palabras reemplazadas por sus correspondientes descriptores. En esa base cada palabra ocupa un registro de datos y está ordenada según su orden de aparición en el texto original. Los campos corresponden a los descriptores enumerados. De acuerdo con ésto, el esquema de funcionamiento es el que se plantea en la Fig. 9.

Fig. 9. Esquema de procesamiento de descriptores.



Como puede apreciarse, el proceso incluye la obtención de un código denominado stemming. Este es un método para obtener el descriptor 13, que reduce una palabra a un radical común denominado stem. Fue diseñado para tratar búsquedas textuales. Una palabra es asociada con una familia o stem para poder ampliar la base de búsqueda, de esta forma no sólo se obtiene una descripción generalizada de la palabra, sino que también puede ser usado para agilizar el tiempo de búsqueda y ajustar la precisión asociada a los resultados [98]. En [14] se muestra que la expansión morfológica usada para Information Retrieval tiene un mejor recall que otras alternativas. Esta técnica se puede implementar según el algoritmo propuesto por Porter en [116]. A este código se le agregan otros descriptores morfosintácticos que mejor describen cada palabra. Luego, se procesa todo con un árbol clasificador basado en el algoritmo J48 [53] (implementación de C4.5 dentro del aplicativo WEKA ©) y se obtiene por inferencia el tipo de palabra (tipo-pal). Con toda esa información se construye un **EBH** por cada palabra. Los **EBH** se organizan en la misma secuencia que se hallan en el texto y se van sometiendo al módulo **MC** a medida que se completa la conversión de una oración completa.

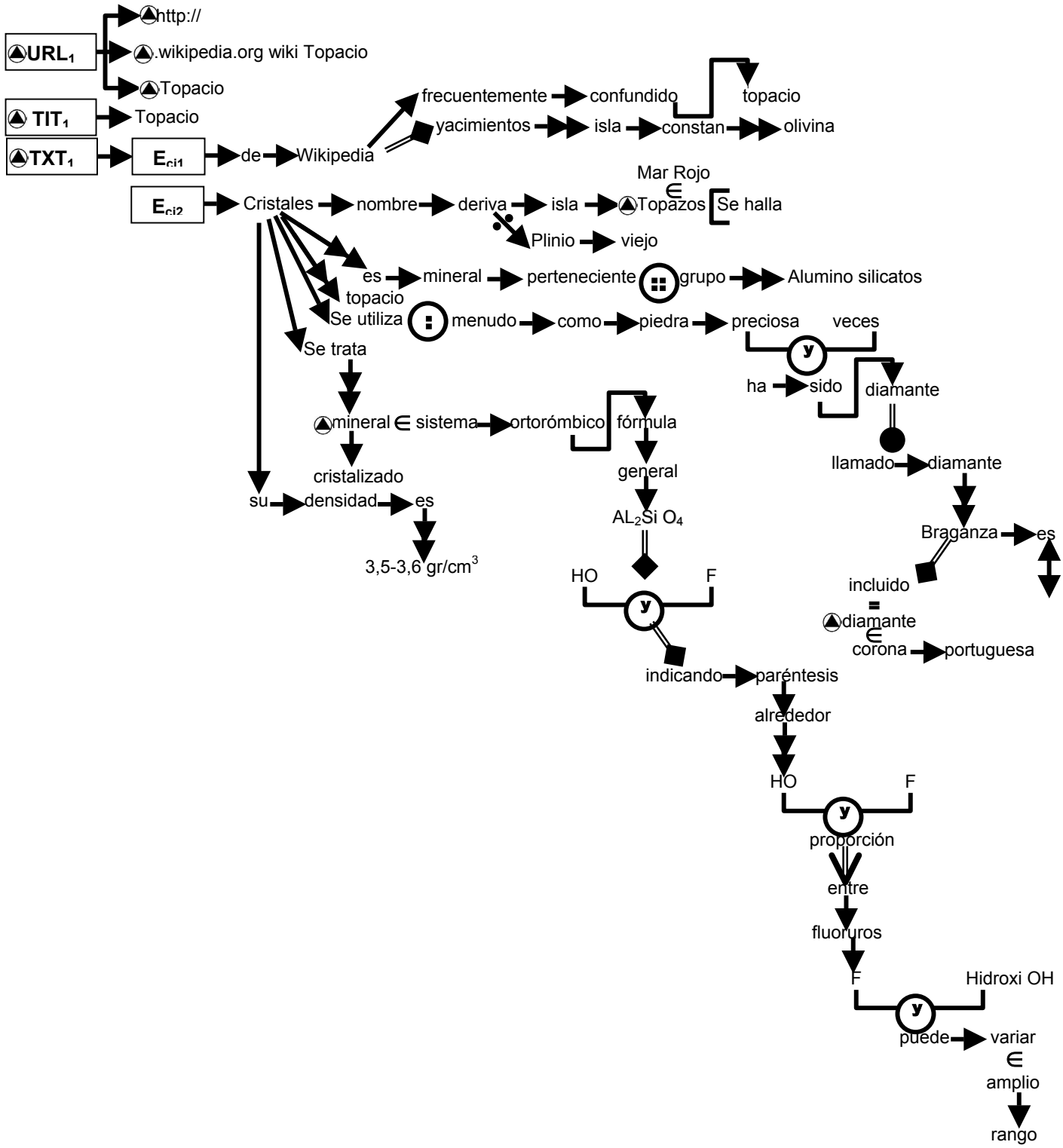
3.2. Módulo Motor de Composición (MC)

Este módulo toma los **EBH** preparados por el **TLI** y comienza el "enhebrado" de los mismos de manera lógica para representar ciertos aspectos de la información

contenida por los datos en el texto original. Para realizar su trabajo, se emplea un criterio de elaboración de las partículas **EBH** similar al desarrollado en [86]. En la versión allí presentada se realiza una transformación del texto a un conjunto de símbolos que se insertan en un grafo orientado. En el **MC** en cambio no hay necesidad de tal tipo de representación y por lo tanto se utiliza directamente el EBH, y al grafo se lo plasma en una codificación equivalente[32]. En la Fig. 10 se puede apreciar el resultado obtenido en [86] de aplicar la transformación simbólica descrita allí, usando como base la Tabla I, Tabla II y Tabla III (luego de cierto preprocesamiento) sobre el texto perteneciente a una página Web (el mismo se transcribe en Apéndice E: Texto base para codificación en símbolos).

[32] Esta codificación, actualmente propietaria, será pasada a un equivalente XML en versión futura.

Fig. 10. Ejemplo de E_{ci} con representación simbólica.



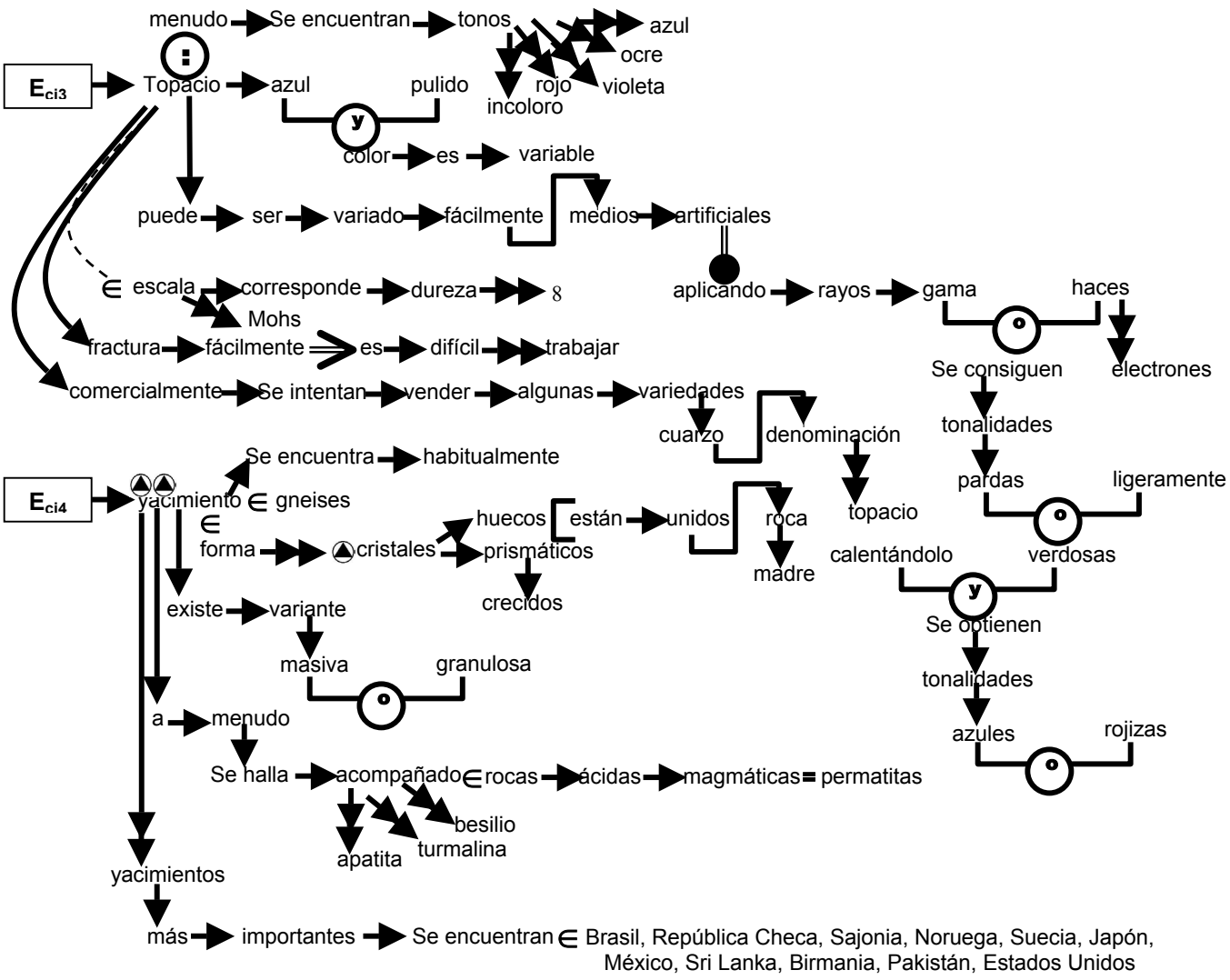
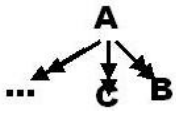
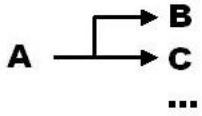
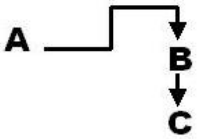
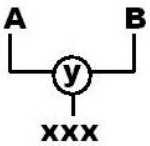
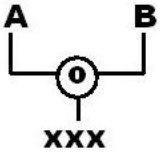
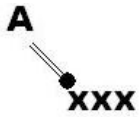
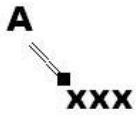
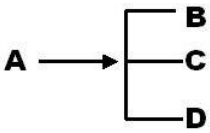
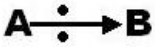

menudo
Se encuentran
tonos
azul
ocre
violeta
rojo
incoloro
Eci3
Topacio
azul
pulido
y
color
es
variable
puede
ser
variado
fácilmente
medios
artificiales
∈ escala
corresponde
dureza
8
Mohs
fractura
fácilmente
es
difícil
trabajar
comercialmente
Se intentan
vender
algunas
variedades
aplicando
rayos
gama
haces
Se consiguen
electrones
cuarzo
denominación
tonalidades
Se encuentra
habitualmente
Eci4
yacimiento
∈ gneises
pardas
ligeramente
∈
forma
cristales
huecos
están
unidos
roca
topacio
prismáticos
madre
calentándolo
verdosas
crecidos
y
existe
variante
Se obtienen
masiva
granulosa
tonalidades
a
menudo
azules
rojizas
Se halla
acompañado
∈ rocas
ácidas
magmáticas
permatitas
besilio
turmalina
apatita
yacimientos
más
importantes
Se encuentran
∈ Brasil, República Checa, Sajonia, Noruega, Suecia, Japón,
México, Sri Lanka, Birmania, Pakistán, Estados Unidos

Tabla I.Conversión de sintagmas a símbolos y conectadores orientados

ID	caso	símbolo
1	A de B [y | o C[y |o ...]] A de B [, C[,...y | o...]] Disgrega A en B, C,...Luego, si sólo existe B continúa la inserción en B, si existen C, D, etc. Continúa en A.	
2	A para B [, C...] Compone los "para" como selección	
3	A con B Vincula conceptos / sustantivos	
4	A y|e B [XXX] Junta una enumeración. Coordina el último sustantivo a izquierda con los sustantivos a derecha de "y" procesando lo que sigue (representado como "XXX"). Si C contiene más de una palabra, entonces saltea desde el primer sustantivo que halla hasta el próximo ".", ":", ",", ";", o fin de texto. Si en B aparecen los casos 1-3, 7-9, 11-15, 18,21, 24-31, entonces se tratan estos casos primero y se continúa en XXX	
5	A, B Coordina B al mismo nivel que el último sustantivo generando un caso 4. Si al retroceder en el texto halla un caso 3, deberá tomar el primer sustantivo que se presente antes de la ocurrencia del caso 3. Si queda como primera palabra después de punto, entonces ignora la coma y continúa insertando donde estaba. -Aplica el marcador ▲ a B -aplica a los casos en que aparece ";:." antes de otra coma	
6	A por B Subordina A al B. -Invierte la secuencia de inserción hasta el próximo sustantivo	
7	A o B [XXX] Si en B aparecen los casos 1-3, 7-9, 11-15, 18,21, 24-31, entonces se tratan estos casos primero y se continúa en XXX	
8	**A en B[, C,...[y |o ...]]**	**A ∈ B[,C...]**

ID	caso	símbolo
	Busca el próximo sustantivo anterior de la estructura E_{ci} y le marca con ▲ también le asocia con el símbolo especial a B (∈). Continúa en B la inserción. Si no existe un sustantivo o hay antes un caso 4 o 7, entonces elimina "B" y "en" y continúa la inserción donde estaba. Si en la búsqueda encuentra casos 1 o 3, los saltea y continúa la búsqueda hacia atrás.	
9	▲ marca para futura entrada de rastreo o búsqueda	
10	**A: XXX** Delimita lo que sigue hasta el final de la frase (. ó ;) y subordina XXX. Luego sigue insertando en A	
11	**A Contra B**	
12	**A de su B** Es similar al caso 1 pero invierte el proceso de inserción de los sucesivos componentes.	
13	**"XXX" (XXX) {XXX} ,XXX,** Similar al caso 10, pero el texto delimitado se lo subordina al sustantivo anterior. En caso de no existir se marca el primer sustantivo del contenido con ▲ y se genera una nueva E_{ci}. Luego la inserción continúa desde el punto A.	
14	**A a B** Enlaza A con B	
15	**A al B** Enlaza A con B	
16	**A entre las cuales B [, C ...]y D** Este caso se aplica sólo si A es sustantivo	
17	**,y** Elimina la "y" y actúa como si fuera un caso 5	
18	**A quien(es) B** Vincula B al anterior sustantivo. Similar al caso 14	
19	**A, para quien(es) B** Vincula B al anterior sustantivo. Similar al caso 18 pero invierte A con B	
20	**A para quien(es) B** Similar al caso 2 pero se elimina B.	

ID	caso	símbolo
21	**A que B**	
22	**;** Busca el primer sustantivo desde el principio de la E_{ci} y retoma el proceso de inserción de lo que siga a partir de allí. Si no hay un sustantivo, toma la primera palabra que no sea un caso especial. Si no tiene éxito, genera una nueva E_{ci}.	
23	**.** Si luego sigue un sustantivo genera una nueva E_{ci}, de lo contrario: Busca el primer sustantivo desde el principio de la E_{ci} y retoma el proceso de inserción de lo que siga a partir de allí. Si no hay un sustantivo, toma la primera palabra que no sea un caso especial. Si no tiene éxito, genera una nueva E_{ci}. Los subesquemas que se estén desarrollando se suspenden.	
24	**A ante B**	
25	**A según B**	
26	**A sin B**	
27	**A so B**	
28	**A sobre B** Conecta A con B	
29	**A tras B**	
30	**A como B**	
31	**A se B** No genera flecha direccional entre se y B	
32	**A. Por esta razón B** **A. Por esto B** **A. Por eso B** **A Por lo tanto B** **A, Por esta razón B** **A, Por esto B** **A, Por eso B** **A, Por lo tanto B** **A, y Por esta razón B**	

ID	caso	símbolo
	A, y Por esto B	
	A, y Por eso B	
	A, y Por lo tanto B	

Tabla II.Conectadores de eliminación

La	Ella
El	Ellas
Le	Les
Los	Ello
Las	Ellos
Los cuales	Un
Las cuales	Una
La cual	Esta
El cual	Este
Estas	Estos
Eso	Esos
Esa	Esas
Tan	Su
Sus	según*
Sin*	so*
sobre*	tras***
que	Además***
También***	Incluso***
Sin embargo	No obstante

* **Estos casos se aplican cuando inmediatamente antes hay un caso 13.**

** **Cuando están precedidos de caso 4 o 7.**

*** **Cuando son primer palabra después de un punto.**

* Estos casos se aplican cuando inmediatamente antes hay un caso 13.
** Cuando están precedidos de caso 4 o 7.
*** Cuando son primer palabra después de un punto.

Tabla III.Conectadores especiales de textos Web

Objeto	Representación
URL	URLj▲. Genera un suevo E_{ci}
Link	$LINK_j$
Objetos wav	WAV_j
Objetos avi	AVI_j
Objetos jpg, jpeg	JPG_j
Objetos tiff	$TIFF_j$
Objetos doc	DOC_j
Objetos pdf	PDF_j
Objetos bmp	BMP_j
Tag <TITLE>	Sustantivo TIT, genera un nuevo E_{ci}.

Como se expuso, en la Tabla I se detallan los sintagmas detectables y el correspondiente símbolo o conectador en el que derivan. Sólo se describen ciertos casos a detectar dentro del texto en proceso. Cuando el caso se verifica, se codifica como indica la columna "Símbolo", ubicada a la derecha.

La Tabla II lista los sintagmas que deben ser eliminados y no se codifican de ninguna manera especial dentro del grafo. En la Tabla III se describe cómo se codifican y cómo se insertan en el grafo los datos identificatorios de las páginas Web con el texto. De los nombres de las tablas puede deducirse que, en el contexto de este tratamiento todo se reduce a símbolos o conectadores, aún en el caso en que el sintagma deba ser eliminado del grafo. Obsérvese también que algunas reglas de conversión tienen precondiciones relacionadas con otras.

Los casos se presentan usando la siguiente convención:

Tabla IV.Convención para describir casos

Símbolo	Significado	Ejemplo
Letras mayúsculas	Sintagma cualquiera	A
[ ]	opcionalidad	A [B] B puede o no suceder después de A
...	Posible repetición	A ... Puede o no haber otros sintagmas después de A
|	alternativa	y|e puede suceder y o bien e
XXX	cadena de sintagmas	"XXX" entre comillas habrá una cadena de sintagmas
( )	Caracteres que pueden aparecer opcionalmente	quien(es) puede aparecer quien o quienes

Resumiendo, entre las actividades desarrolladas por el **MC** se pueden enumerar:

-distinción de ciertos sintagmas

-distinción de algunos sintagmas

-eliminación de los **EBH** que no son considerados esenciales al proceso

-detección de subordinaciones

-detección de enumeraciones

-elaboración de **EBH** opuestos

-elaboración de **EBH** ambiguos

Cabe aclarar que la concepción de **EBH** opuestos y ambiguos no son los que corresponden a palabras opuestas ni ambiguas aunque están relacionados. A continuación se describen estos tratamientos.

Más allá de la utilidad del grafo orientado para el propio sistema WIH, se analizó la legibilidad de este lenguaje como expresión representativa de cierto texto. Para ello se realizó una encuesta con 44 voluntarios (publicado en [85]). La encuesta consta de seis preguntas agrupadas en dos temas:

-Cuatro preguntas relacionadas con una enfermedad denominada linfedema.

-Dos preguntas relacionadas con leyes.

Los voluntarios eran todos argentinos nativos de tres extracciones distintas:

-pacientes de una clínica privada dedicada al tratamiento de linfedemas

-médicos especialistas de la misma clínica

-otros, principalmente estudiantes de informática

La muestra se definió con estas características y merecen destacarse detalladamente las implicancias de la selección:

-Los pacientes suelen tener un rango de edad y nivel cultural bastante variado.

-Por tratarse de una enfermedad crónica con tratamiento de por vida, los pacientes suelen adquirir muchos conocimientos acerca de la misma.

-Los estudiantes de informática seleccionados fueron los que desconocían temas legales y la enfermedad en cuestión.

El formulario utilizado (replicado en el Apéndice F: Formulario de encuesta para representación simbólica) contiene la representación simbólica simplificada de un texto médico denominado "Normas de prevención del linfedema". La simplificación se realizó cambiando los símbolos de la Tabla I por rectángulos similares, y sólo se preservó el criterio de conexión y direccionalidad de los arcos. En el caso del símbolo ⏶, el rectángulo fue resaltado con un doble rayado[33]. Los EBH se reemplazaron por las palabras que los originan a fin de aumentar la legibilidad del gráfico. El texto original nunca fue presentado. En base a esta representación se realizaron las cuatro primeras preguntas. Obsérvese que esto es casi equivalente a realizar las preguntas en base a las E_{ci} del texto.

[33] De esta manera se destacan sólo las EBH indicadoras.

Las preguntas restantes apuntan a evaluar el poder de inferencia de los encuestados, en base a muy poca información. Sólo las palabras correspondientes a los EBH marcados con ▲ fueron extraídas y listadas. Estas palabras, corresponden a los denominados términos indicadores (en este caso serán palabras indicadoras, referidas en el trabajo como "pointer words"). Obsérvese que esto es equivalente a realizar las preguntas sobre parte de la E_{ce} del documento original, puesto que sólo se promueven a ese nivel las indicadoras y/o las ponderadas favorablemente con la métrica de ponderación p_o [83] (ver 3.2.3. Justificación de la estrategia con po) según cuál sea la función efectora f_e activa.

En el trabajo se obtienen números interesantes sobre los siguientes estudios:

-eficiencia de la selección de palabras para la E_{ci} en general.

-eficiencia de la selección de palabras para la E_{ci} de acuerdo al nivel de conocimientos del tema.

-eficiencia de la selección de palabras para la E_{ci} de acuerdo al nivel de conocimientos de búsqueda en la Web.

-representatividad de la información con la E_{ci}

-representatividad de las palabras indicadoras

En base a los resultados obtenidos, se puede afirmar que las E_{ci} representan eficiente mente el texto aún cuando no existe conocimiento especial del área temática o de informática. Tampoco incide la experiencia en el manejo de la Web. También se mostró que las indicadoras pueden resultar una versión reducida de las E_{ci} y éstas del texto original. Lo que sigue es un análisis presentado en [82] del tratamiento de ciertos casos lingüísticos detectados y procesados especialmente por WIH.

3.2.1. Tratamiento de EBH opuestos y contradicciones

La generación de texto escrito implica una comunicación [10], la que ostentará una coherencia determinada. Los textos tienen esquemas estructurales [7] y elementos cohesivos particulares a cada tipo de texto. Hay una unidad de sentido en la totalidad del texto, cuando éste es coherente (coherencia intratextual). Pero existe otro nivel de coherencia que trasciende al texto y se manifiesta dentro de una cultura y en una situación dada. El tratamiento de este tipo trasciende ampliamente los objetivos de la

estrategia aquí propuesta, que sólo puede trabajar con un nivel muy rudimentario de incoherencia en los textos, y prepararlos para la interpretación por parte del usuario. Por este motivo se presentará aquí brevemente el concepto de contradicción como falta de coherencia intratextual y luego se definirá con mayor precisión el tipo de tratamiento que puede realizar WIH. En las subsecciones correspondientes al Capítulo 5. Estudio de casos y resultados, se presentarán ejemplos concretos para ilustrar ambos casos.

(i) La contradicción: un marco teórico

Existen dos modos de ver la contradicción, desde el lenguaje ordinario y en los escritos filosóficos. Desde un punto de vista formal puede afirmarse que desde la época de Aristóteles [67] existe un campo de estudio dedicado a las denominadas falacias y sofismas. Se refiere al error lingüístico asociado a un error de razonamiento o viceversa, con intenciones manipuladoras (sofisma) o por descuido o ignorancia (falacia).

Una clasificación tradicional[34] divide las falacias en formales e informales; éstas, a su vez, en falacias de ambigüedad y falacias materiales; y éstas, a su vez, en falacias de pertinencia y falacias de datos insuficientes.

Por otro lado, y desde un lenguaje más ordinario [45], se puede considerar una contradicción de dos modos diferentes:

1) concepto lógico o analítico: es un vicio de razonamiento. Por ejemplo, dentro de la filosofía clásica se puede mencionar el principio de no-contradicción de Aristóteles según el cual una cosa no puede ser y no ser al mismo tiempo. En el mismo sentido, en la lógica contemporánea existe el principio de congruencia, según el cual la presencia de una contradicción (conjunción de una proposición y su negación) dentro de un sistema lo invalida para todo efecto, pues una falsedad es capaz de producir todas las otras mediante mediante reglas de inferencia.

2) concepto dialéctico o sintético: la contradicción se concibe en un contexto filosófico, como cifra o signo de contenido esencial. Se toma como síntoma de

[34] Dentro de los estudios no clásicos, hay distintas versiones e, igual que ocurre con las paradojas, de vez en cuando se formulan nuevas falacias o esquemas de error, como la falacia categorial (category-mistake), formulada por Gilbert Ryle. Se habla también de falacias genética, idealista, naturalista, etc.

densidad ontológica (Unamuno, Kiekegaard), como forma de expresar lo inefable (filosofía existencialista), como método de entender el devenir (Hegel), etc.
Cabe destacar que, desde la perspectiva de la tecnología computacional actual sólo el primero es encarable de algún modo.

(ii) Definición de opuesto en el contexto WIH

Sobre la base de lo expuesto anteriormente, el principio de congruencia puede detectarse en algunos casos de manera sencilla, ya que ciertas palabras por su construcción morfológica reflejan una ***oposición o falta de*** en virtud de la presencia de ciertos prefijos o palabras. La presencia de éstos se reflejará en un EBH y lo convertirá en ***EBH opuesto***. WIH contempla las sentencias con EBH opuesto y las califica con una magnitud denominada ponderación de fortaleza p_o. Dicha ponderación pretende caracterizar el contenido de la sentencia como un todo y no reflejar la semántica subyacente.

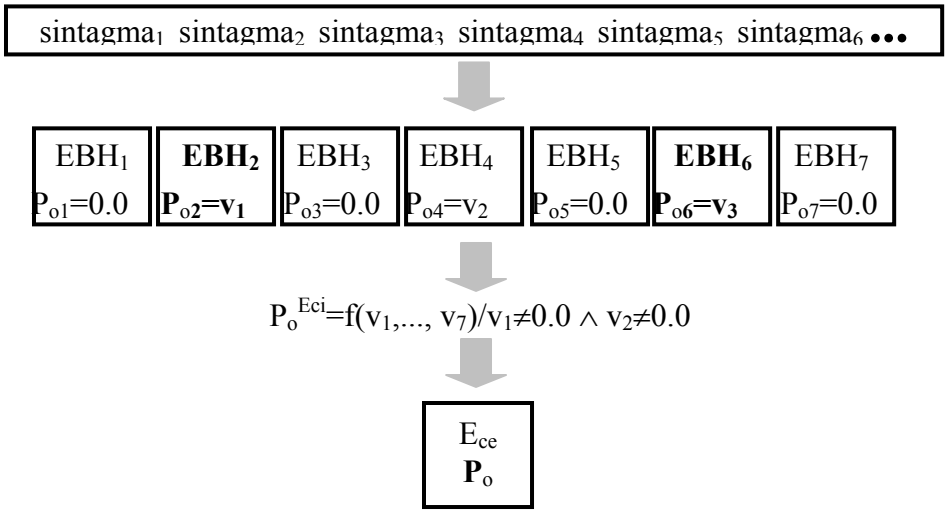
Lo que interesa a WIH como sistema, es la fortaleza p_o de los términos vertidos en cierta frase y luego la de la frase como un todo. Esta fortaleza la detecta reconociendo la existencia de ciertos patrones en los EBH, es decir, en las palabras que cierto autor ha empleado para redactar su texto. La idea detrás de ésto es la categorización de una frase respecto a otras. Una vez determinada la categoría de la frase, se toma como característica para ciertos EBH destacados como ***términos indicadores*** representantes de la misma. Estos términos, seleccionados automáticamente de acuerdo a las reglas consideradas en Tabla I y Tabla III, corresponden a los EBH que resulten denotados por un "▲".

En la Fig. 11 se presenta un proceso genérico para caracterización de sentencias: Sea una sentencia que deriva en la secuencia de EBH numerados del 1 al 7. Todos los EBH son ponderados como tales con un valor $v_i \in [-1,+1]/ v_i \neq 0.0$. Los valores de v_i son administrados de manera automática (en general, de acuerdo a ciertos prefijos originales de la palabra que se detallan en la próxima sección), con una asignación

por defecto de v_i =0.0. De todos los EBH hay tres, EBH_2, EBH_4 y EBH_6, que contienen prefijos especiales y sólo dos de ellos, además, son ***términos indicadores*** (EBH_2 y EBH_6, marcados con negrita). En términos generales, el valor de su fortaleza será menor en valor absoluto cuando se considera que el EBH casi no es afectado por el significado del prefijo. En contraste, los valores de $|v_i|\approx 1$, corresponderán a prefijos que podrían afectar radicalmente al EBH.

Por ejemplo en el EBH correspondiente a la palabra: ***ANTI****tetánica*, el prefijo estaría afectando fuertemente al EBH, significando un opuesto a *tetánica*, por lo que su valor sería v_i=-1.

Fig. 11. Ejemplo de EBH opuesto y derivación desde el texto hasta el E_{ce}.



Como se observa en la figura, el valor de p_o obtenido será usado por WIH para determinar la fortaleza de las E_{ci} y E_{ce}. En el caso de las E_{ci}, dependerá de cierta función efectora del momento que las combine apropiadamente. En el caso de la E_{ce}, será función de la efectora activa en ese momento saber si corresponde o no promover a E_{ce} desde la E_{ci}. De esta manera se pretende inferir aproximadamente las

características del texto que representan. En las secciones que siguen se bosqueja la implementación de este mecanismo en WIH y se analizan datos reales procesados con este criterio.

(iii) Implementación de la propuesta en WIH

Sea X un sintagma, se dice que un **EBH** es opuesto a X si se identifica la palabra origen del mismo con alguno de los patrones descritos en la Tabla V.

Tabla V.Patrón de EBH opuesto

EBH opuesto
no X
sin X
desX
inhX
antiX
disX

La implementación tiene las siguientes características:

1. Es un listado indicativo y perfectible. No es completo. La falta de completitud de la misma hará que el sistema genere valores de p_o cercanos al 0.0 en muchos casos, y de este modo el sistema manifieste una caracterización muy similar en casos heterogéneos. Esto generará un promedio de caracterizaciones (denominados internamente PDC_ECE) muy similares al nivel de los E_{ce}. El sistema controlador detectará una necesidad de ajuste al contrastar PDC_ECE contra un umbral de mínima ponderación (denominado internamente PDC_ECE_UMBRAL). Como consecuencia si PDC_ECE< PDC_ECE_UMBRAL, se generará una alarma administrativa requiriendo mayor información. Nótese que a pesar de ser perfectible, continuará el funcionamiento primario.

Será opción del administrador ajustar el PDC_ECE_UMBRAL para reducir la sensibilidad o bien alimentar con más patrones al sistema.

2.Los patrones opuestos están en relación con las $\mathbf{f_e}$ activas, que a su vez dependen del objetivo y sensibilidad del sistema en ese momento. Un cambio de objetivo puede

hacer que el sistema controlador cambie el conjunto de $\mathbf{f_e}$ activas haciendo que la elaboración de p_o para las E_{ce} cambie. En el caso de la Fig. 11, cambiaría la ecuación de $P_o=f(v_1,..., v_7)$.

3. El tratamiento es orientativo pero no determinista. La adecuación es variable según lo que indique el Motor de Métricas (MM). Si bien se supone la eventual interacción con un controlador humano, mejorando las listas según la información recabada por el mismo sistema durante su funcionamiento, no se descarta que el sistema no evolucione en tal sentido si la métrica de evaluación contra el PDC_ECE_UMBRAL no es lineal o eventualmente es cambiada. Nótese que con este tipo de tratamiento es posible la convivencia de información procesada durante la prevalencia de distintas métricas[35].

El Motor de Asimilación (MA) trabaja con un conjunto de funciones efectoras $\mathbf{f_e}$ activadas por el Gestor de Métricas (GM) en algún momento. Este conjunto { $\mathbf{f_e}$ } implementará alguna versión de la ecuación de $P_o=f(v_1,..., v_7)$ y son referidas dentro del sistema como las FuncionEfectorReglaAsimilacionPesoECE, tomadas de un diccionario (llamado FuncionEfectorReglaAsimilacionPesoECE_dicc) a disposición del Gestor de métricas (GM) según indicaciones del Sistema controlador (SC).

A su vez, la métrica de evaluación contra el PDC_ECE_UMBRAL será implementada por alguna función de métrica $\mathbf{f_m}$, también activada por el Motor de Métricas (MM) en algún momento.

En Capítulo 5. Estudio de casos y resultados se muestran algunos resultados obtenidos con este tipo de procesamiento.

3.2.2. Tratamiento de EBH ambiguos y ambigüedades

Las ambigüedades en los textos son raras debido a que normalmente se tiene en claro cuál es el tema general del mismo [12]. Lo que ya se conoce del entorno hace que se seleccione de manera automática e inconsciente el buen punto de vista, el compatible

[35] Probablemente eso suceda debido a distintas evoluciones del lenguaje o cambios de contexto correspondientes al contenido procesado.

con el contexto. El único caso en el que la ambigüedad se hace consciente es ante el fracaso de la comprensión.

La ambigüedad es intrínseca en las lenguas naturales [6] tanto en el ámbito morfológico como sintáctico y semántico. Esta característica hace que los analizadores sintácticos diseñados para tratar el lenguaje natural sean más complejos que los algoritmos dedicados al análisis de los lenguajes de programación.

Existen varios tipos de ambigüedades:

-ambigüedad léxica

-ambigüedad sintáctica

-ambigüedad semántica

-anáforas

Existen dos variantes tradicionales de enfoque para su estudio: racionalista (de N. Chomsky) y empiricista (de Z. Harris). Este último sostiene que el conocimiento lingüístico se puede inferir a partir de la experiencia, que se puede recoger a través de corpus textuales, por mecanismos sencillos como generalización o asociación.

A continuación se presentarán brevemente distintos tipos de ambigüedades y luego se definirá con mayor precisión el tipo de tratamiento que puede realizar WIH. En las subsecciones correspondientes al Capítulo 5. Estudio de casos y resultados, se presentarán casos concretos para ilustrar las alternativas.

(i) La ambigüedad: un marco teórico

La concepción de ambigüedad responde a un modelo racionalista [23]. Existe un bagaje de conocimientos previos iniciales a su consideración. Chomsky considera el lenguaje, como un sistema bien definido. Pero cuando un mensaje pasa de su fuente en el cerebro, en su viaje hacia la persona a la que va dedicado, el mensaje se deforma. El receptor debe dar un sentido al mensaje, librarlo del ruido, reconstruirlo en forma tal que permita interpretar adecuadamente el mensaje original. A menos que haga ésto, la comunicación será imposible.

Chomsky establece una distinción crítica entre lo que llama “competencia” y “actuación”. La competencia es el conocimiento táctico que un parlante posee de todas las frases bien formadas de su propia lengua, y sólo de ellas. La actuación, es el uso del lenguaje en los encuentros cotidianos entre la gente. Chomsky, centrará su interés principalmente en la competencia.

Según su teoría, la ambigüedad, que en realidad es una especie de ruido que oscurece el sentido del mensaje, puede resolverse remitiéndose a la estructura de la cual el mensaje es una transformación.

En la lingüística de Chomsky, constantemente sujeta a revisión y cambio, un conjunto de reglas básicas genera una “estructura profunda” que es el plan abstracto de una frase. La estructura profunda es lo que está más cerca del sentido intentando por el parlante, y lo menos afectado por distorsiones y ambigüedades.

Chomsky considera el lenguaje como sistema de gran eficiencia para procesar información porque, aunque el habla misma puede ser desordenada y corrupta, existen regularidad y orden debajo de ella. La gramática, que es parte de la competencia, actúa como filtro, eliminando errores; es un recurso contra el azar que mantiene las frases regulares y apegadas a una ley. Es un código sistemático aplicado a la fuente del mensaje.

La gramática universal, como la describe Chomsky, es una teoría de las gramáticas en general. No sólo explica el número limitado de hechos ya conocidos, sino que predice la existencia de hechos adicionales que aún se desconocen.

En 1983 propone su Teoría “Government and Binding” en la que da mayor importancia al léxico, reduciendo el papel de la gramática a una serie de principios de buena formación. En esta línea aparecen gramáticas como:

-Gramáticas de Estructura de Frase Generalizadas (GPSG)

-Gramáticas Léxico Funcionales (LFG)

-Gramáticas de Unificación Funcionales (FUG)

A partir del trabajo de Colmerauer surgen las gramáticas lógicas:

-Gramáticas de Cláusulas Definidas

(ii) La ambigüedad en los textos: marco teórico

Este estudio se centra en tales casos ya que WIH tratará muy sencillamente sólo textos escritos. Son menos frecuentes las ambigüedades en los textos porque sus entornos de singularidades indican cuál es el tema general de los mismos [11]. Como se ha dicho, el conocimiento del entorno colabora en la selección de un punto de vista compatible con el contexto. Por caso, el término "manzana" puede referir a un fruto o a un grupo de casas, pero sólo el sentido más compatible con el contexto viene a la conciencia.

Como destacó Bréal, en sus Ensayos de semántica a finales del siglo pasado:

« Las palabras están empleadas cada vez en un entorno que predetermina su valor. [...] No es siquiera necesario suprimir los otros sentidos de la palabra: estos sentidos no existen para el lector, no pasan al umbral de conciencia. »

El único caso en el cual una ambigüedad es posible, es cuando fracasa la comprensión: porque el punto de vista elegido no permite entender las relaciones de una singularidad con otras, o bien no parecen compatibles con lo que se conoce del mundo descrito. Si estas relaciones no son importantes, se continúa escuchando o leyendo; si lo son, se busca un punto de vista que permita entender esas relaciones y el enunciado se lee desde esta nueva perspectiva. Muchos chistes funcionan así, obligando al coenunciante a interpretar dos veces el último enunciado.

A continuación se estudia el enfoque que se implementa en WIH para el tratamiento de situaciones ambiguas y la definición que se hace en este marco de un "EBH ambiguo".

(iii) Fundamentos de la estrategia WIH

Si bien el concepto de ambigüedad afecta al proceso como un todo, como en el caso del PLN (Procesamiento del Lenguaje Natural), existen ciertas diferencias sustanciales que simplifican notablemente el problema:

-En PLN es necesario interpretar todo el texto. WIH no pretende ser una interpretación de los textos originales sino una referencia a los mismos con suficiente información como para que permita conocer su existencia y contenido aproximado.

-En PLN se incluye el plano semántico como parte del problema. WIH sólo realiza un sencillo procesamiento que no excede los aspectos morfosintáticos correspondientes a cada palabra. Dicho de otra forma, realiza un tratamiento local de ambigüedades.

-En PLN se trabaja con el mensaje y su interpretación conforme los siguientes niveles de información [79]:
1-Nivel léxico: vocabulario de la lengua.
2-Nivel morfológico: morfemas de género, número y persona.
3-Nivel sintáctico: estructuras de secuencias de unidades léxicas.
4-Nivel semántico: significado o sentido de los elementos y estructuras oracionales.
5-Nivel pragmático: relación de las unidades lingüísticas con el contexto extralingüístico.
En el caso de WIH, sólo interesan algunos aspectos de los tres primeros niveles. El resto descansa en la actividad intelectual del usuario.

-En PLN el estudio sintáctico pretende ser lo suficientemente robusto como para sustentar los niveles de abstracción superior. En el caso de WIH esto no es así, dado que el estudio es mucho más simple.

En lo sucesivo se presentarán ejemplos de distintos tipos de ambigüedad y los resultados correspondientes por el procesamiento en WIH.

(iv) Situaciones ambiguas que procesa WIH

Por lo visto anteriormente, el tratamiento aquí es sencillo: si la frase ambigua es considerada por WIH como parte representativa del contexto, entonces se promueve como E_{ci}, y sólo trascienden a nivel E_{ce} las palabras de las frases que, por su posicionamiento sintáctico y conformación, sean declaradas representativas (o indicadora en terminología WIH).

Si bien se consideran de esta manera todas las sentencias, aún las que son linguísticamente ambiguas (delegando en el usuario la interpretación correcta), existe un tratamiento para algo que WIH considera una "situación de ambigüedad": ciertos términos dentro del léxico vistos por sí mismos reflejan imprecisión en cuanto a la magnitud o intensidad de algo. Si bien podría decirse que en rigor estos términos no son ambigüedades, en esta propuesta son tratados especialmente debido a que el lector no puede establecer con real precisión su significado completo utilizando los elementos contextuales.

Desde este punto de vista se considera como situación de ambigüedad a la ocurrencia de uno o más patrones descritos en la Tabla VI.

Tabla VI.Patrón de EBH ambiguo

EBH ambiguo
muy X
tan X
algo X
poco/a(s) X
mucho/a(s) X
bastante(s) X
escaso/a (s) X
escasamente X
excesivamente X
excesiva/o X
abundantemente X
abundante(s) X
demasiado/a(s) X
exageradamente X

EBH ambiguo
exagerada/o(s) X

Nótese que la lista nuevamente no es exhaustiva, por las mismas razones expresadas para las EBH opuestas. En este caso el sistema adiciona un descriptor adicional ya presentado como $p_o(c)$, con el cual pretende establecer una base de precisión para su tratamiento. Cuando una EBH ambigua es promovida a E_{ci}, se considera su $p_o(c)$ como factor de peso de la sentencia E_{ci} en proceso[36]. Eventualmente si a alguno de los términos de esa E_{ci} corresponde un patrón, es promovido a E_{ce}, y también se verá afectado por cierto peso $p_o(c)$. Si se detecta más de una situación ambigua, entonces la ponderación resultante se hará según lo describa la f_e erncargada de la promoción.

3.2.3. Justificación de la estrategia con p_o

El uso de la ponderación p_o [84] para las EBH, fue presentada en [83] y tiene su origen en la necesidad de establecer una caracterización en la forma de redactar sentencias por parte de los distintos autores y en la necesidad de reflejar parte del contenido de manera concisa y representativa. Básicamente, una lengua se expresa muy heterogéneamente conforme el nivel de cultura, temática, situación geográfica, etc. [10]. El valor p_o derivado por WIH, actuará no sólo como apoyo al tratamiento sino también como criterio de elaboración de estructuras dentro de la RV. También podrá destinarse opcionalmente para soportar actividades de navegación o búsqueda. A continuación se detalla la propuesta. En Capítulo 5. Estudio de casos y resultados, se muestran algunos resultados obtenidos sobre la base de lo aquí presentado.

(i) La f_e y el procesamiento por Lógica Difusa

La f_e encargada de la promoción de términos, trabaja según unos principios básicos para obtener los $p_o(c)$ correspondientes a las E_{ci} y E_{ce}. En general, sin importar la implementación activa las f_e esos principios son:

[36] Recuérdese que la misma sentencia E_{ci} es un conjunto de palabras que se procesan en conjunto, y que han sido eventualmente extraídas de una misma sentencia del documento original.

-Deben caracterizar situaciones morfosintácticas sencillas de detectar a nivel palabra.

-Los valores deben tener dominio [-2.0 ; +2.0].

-La mayoría de los valores serán 0.0

-Las ponderaciones de términos deben ser cercanas a 0.0 cuando la situación no modifica radicalmente la palabra/sentencia involucrada.

-Las ponderaciones de términos deben ser lejanas a 0.0 cuando la situación modifica radicalmente la palabra/sentencia involucrada.

-La ponderación de una sentencia se traduce en una ponderación p_o^{Eci} al nivel de E_{ci}, donde se combinan los p_o dentro de la misma. La forma específica es esta combinación variará según el comportamiento del sistema controlador SC, definido por el estado del sistema. Siguiendo lo presentado por [84] se usa la fórmula:

(ec. 1) $p_o = (p_i + p_{i-1})/2$ **,** $\forall$ pal_i

-La ponderación de un párrafo se traduce en una ponderación p_o^{Ece} a nivel E_{ce}, resultante de la elaboración de todas las p_o^{Eci} dentro de la misma E_{ce}. Ganará la p_o^{Eci} más optimista[37].

En el Capítulo 5. Estudio de casos y resultados, se observan algunos resultados obtenidos siguiendo estos criterios.

Este esquema interpreta que, durante la redacción, existen perfiles de narración y tipos de sentencias (más o menos representativas, de menor o mayor calidad). Una

[37] El concepto de optimista variará según el comportamiento del sistema controlador SC, definido por el estado del sistema.

eventual búsqueda o consulta, podrá usar lógica difusa sobre las ponderaciones y determinar si la frase merece ser posicionada mejor en un listado de respuestas candidatas debido a la calidad probable de su redacción, obtenida en función del perfil y el tipo de representatividad del texto en cuestión. A continuación se realizará el estudio estadístico que justificaría el uso de p_o.

(ii) Análisis estadístico de datos

En esta sección se realiza un estudio preliminar de las características de los datos que permita justificar el tratamiento de ponderaciones propuesto. A continuación se presenta el análisis sobre dos clasificaciones fundamentales de los textos a trabajar:
-Tipos de texto.
-Perfiles de narración.
Cada uno de estos aspectos se modelizará con lógica difusa conforme a lo presentado en [84].

A continuación, siguiendo los pasos de rigor para un modelado con lógica difusa [100], se realizan los siguientes pasos:

-Análisis de frecuencias.

-Estudio de patrones e interrelaciones.

-Descripción del modelo matemático.

La validación del modelo en la práctica se realiza en 5.4. Práctica de la estrategia con po. Los detalles numéricos de donde se extraen tablas y figuras están descritos en el Apéndice D: Análisis de datos po.

1. Análisis de frecuencias

a) tipos de texto

Los textos escritos tienen una estructura típica que todo autor debiera seguir para cubrir las expectativas de su interlocutor. En este sentido, se propone diferenciar al menos los siguientes estilos en la Web: literario, técnico y mensajes.

Se muestra en la Fig. 12, el resultado de graficar cada uno de estos estilos, tomando las clases como indica la Tabla VII. La muestra incluye 50 casos de cada uno de estos tipos mencionados, un total de 150 casos.

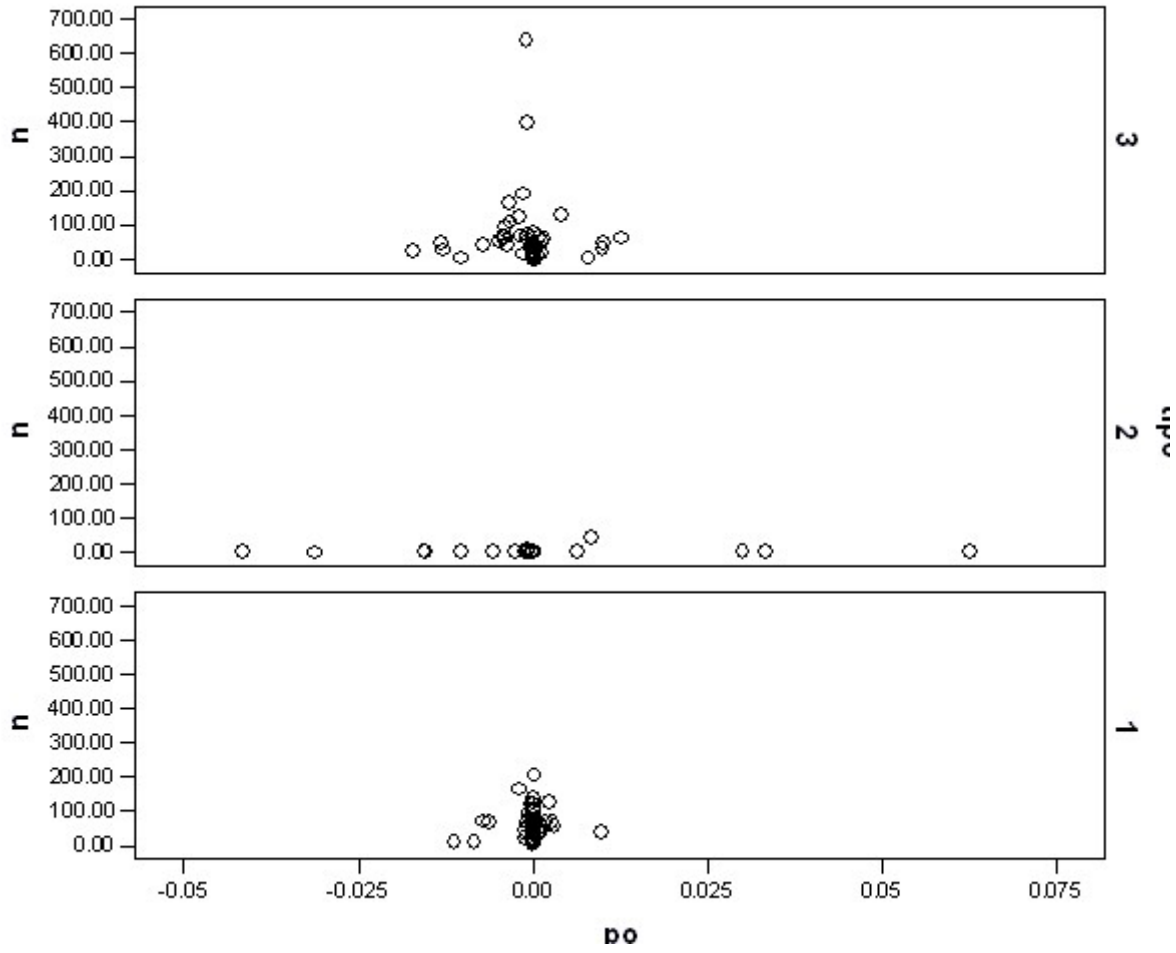
En la gráfica, "n" es la frecuencia de elementos especiales (EBH opuestas y contradictorias) dentro del texto y, p_o es el valor promedio hallado de aplicar las ponderaciones respectivas según la fórmula dada en (ec. 1).

Si bien se aprecian ciertos valores que podrían ser outliers, no se tratarán como tales dado que la información procesada no permite considerarlos como tales con justificación razonable y que se considera podría corresponderse con el tipo de comportamiento poblacional.

Tabla VII.Codificación de tipos

tipo	descripción
1	literario
2	mensajes
3	técnico

Fig. 12. Diagrama de puntos de cada tipo de texto.



b) perfiles de narración

Los textos escritos estimulan al autor de los mismos en cierto sentido según el objetivo perseguido. En este sentido, se propone diferenciar al menos los siguientes

estilos en la Web: foro, índice de Web (página que sólo sirve a los efectos de indexar una serie de otras páginas), documento y blog.

Se muestra en la

Fig. 13, el resultado de graficar cada uno de estos subconjuntos, tomando las clases como indica la Tabla VIII. La muestra incluye 50 casos de cada uno de estos tipos mencionados, un total de 200 casos.

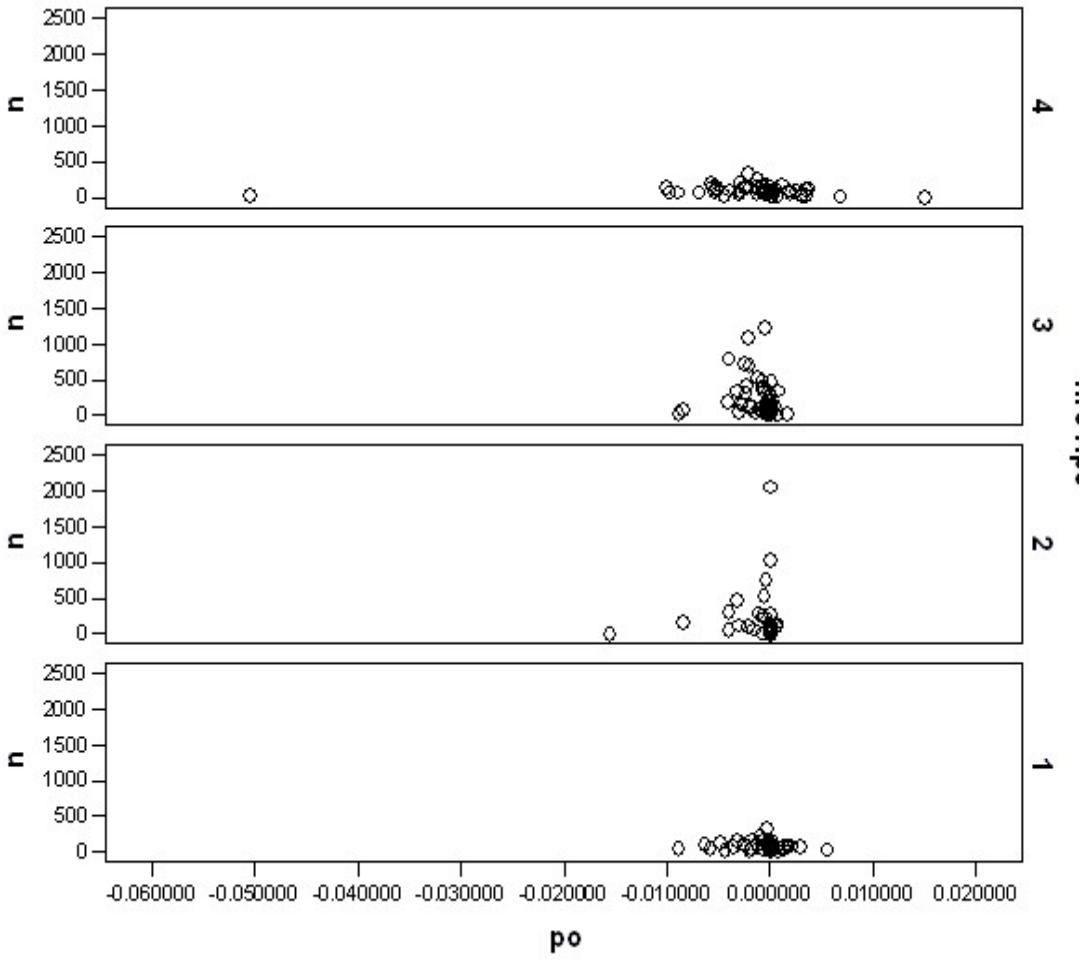
En la gráfica "n" es la frecuencia de elementos especiales (EBH opuestas y contradictorias) dentro del texto y, p_o es el valor promedio de hallado de aplicar las ponderaciones respectivas según la fórmula dada en (ec. 1).

Si bien se aprecian ciertos valores que podrían ser outliers, no se tratarán como tales dado que la información procesada no permite considerarlos como tales con justificación razonable y que se considera podría corresponderse con el tipo de comportamiento poblacional.

Tabla VIII.Codificación de perfiles

tipo	descripción
1	foro
2	webindex
3	doc
4	blog

Fig. 13. Diagrama de puntos de cada perfil de narración.



2. Estudio de patrones e interrelaciones

A continuación se evalúan las interdependencias y se estudia la posibilidad de extraer patrones de comportamiento para cada clasificación propuesta en las secciones anteriores.

2.a) Estudio de normalidad para n y p_o

Se estudió el grado de similitud a una curva normal (de Gauss), mediante el test de normalidad de Shapiro-Wilks. Dada la hipótesis nula H_0: "la distribución es normal", y la hipótesis alternativa H_a: "no se puede asegurar normalidad". Se obtuvieron los resultados que se detallan en la Tabla IX:

Tabla IX.Prueba de normalidad (Shapiro-Wilks modificado) para tipos de texto.

class	Variable	n	Media	D.E.	W*	p (una cola)
literario	po	50	-3.5E-04	2.9E-03	0.73	<0.0001
literario	n	50	61.74	41.38	0.89	0.0005
mensajes	po	50	1.1E-03	0.02	0.66	<0.0001
mensajes	n	50	4.14	5.94	0.34	<0.0001
tecnico	po	50	-1.0E-03	0.01	0.86	<0.0001
tecnico	n	50	65.46	105.24	0.56	<0.0001

Tabla X.Prueba de normalidad (Shapiro-Wilks modificado) para perfiles de narración.

class	Variable	n	Media	D.E.	W*	p (una cola)
blog	po	50	-2.0E-03	0.01	0.64	<0.0001
blog	n	50	97.74	66.09	0.92	0.0108
doc	po	50	-1.2E-03	2.0E-03	0.77	<0.0001
doc	n	50	246.08	273.52	0.78	<0.0001
foro	po	50	-6.8E-04	2.3E-03	0.85	<0.0001
foro	n	50	80.00	63.03	0.86	<0.0001
webindex	po	50	-9.1E-04	2.6E-03	0.47	<0.0001
webindex	n	50	167.54	336.36	0.51	<0.0001

En todos los casos p<<0.05 por lo que *se puede rechazar la presunción de normalidad*. Dado que las poblaciones no siguen una distribución normal, estos valores sólo se toman como orientación para el comportamiento genérico de cada muestra.

2.b) Estudio de correlación

El análisis de correlación de Pearson (ver Tabla XI) entre n y p_o para los tipos de texto, indica que existe una progresiva correlación lineal entre n y p_o conforme se pasa de la clase "literario", a "mensajes" y "técnico". En este último caso es altamente significativo (0.94).

Tabla XI.Correlación lineal entre n y p_o para los tipos

tipo	correlación
literario	.43
mensajes	.69
técnico	**.94**

Lo mismo se comprueba con los perfiles de narración (Tabla XII).

Tabla XII.Correlación lineal entre n y p_o para los perfiles

tipo	correlación
doc	.44
foro	.49
webindex	.84
blog	.97

Se puede observar que *los grupos tienen un comportamiento progresivamente lineal*, siendo significativo el coeficiente para los blog (0.97).

2.c) Estudio de variabilidad Kruskal Wallis para p_o

Se estudió la variabilidad de p_o con este ANOVA no paramétrico que toma como hipótesis nula H_0: "las muestras son de la misma población", y como hipótesis

alternativa H_a: “no son de la misma población”. En la Tabla XIII se detallan los resultados obtenidos.

Tabla XIII.Significación Chi-Cuadrado para los subgrupos.

Tipo de archivo	subgrupo	parámetro	valor
	literario	Chi-Cuadrado	1.014
	mensajes	grados libertad	2
	técnico	Significación	0.602
perfil de narración	doc	Chi-Cuadrado	7.555
	foro	grados libertad	3
	webindex	Significación	0.056
	blog		

Como puede verse, la significación para los perfiles es 0.056 > 0.05, y para los tipos de narración es 0.602 >0.05, no puede rechazarse la hipótesis nula, por lo que *puede afirmarse que las poblaciones tienen variabilidad similar*.

2.d) Estudio de medianas para p_o

Además de verificar que los subgrupos dentro de los tipos y perfiles no tienen igual variabilidad, se estudió si las medianas se comportan igual en los subgrupos. En la Tabla XIV se muestran los valores de significación obtenidos.

Tabla XIV.Significación Chi-Cuadrado para los subgrupos.

Tipo de archivo	subgrupo	parámetro	valor
	literario	Chi-Cuadrado	5.303
	mensajes	grados libertad	2
	técnico	Significancia	0.071
perfil de narración	doc	Chi-Cuadrado	18.720
	foro	grados libertad	3
	webindex	Significancia	0.00
	blog		

Este estudio toma como hipótesis nula H_0: las medianas poblacionales son iguales, y como hipótesis alternativa H_a: las medianas poblacionales no son iguales.
El valor de significación p=0.071 > 0.05, *para los tipos de archivo indica que no puede rechazarse la hipótesis original y por lo tanto p_o no tiene una mediana poblacional distinta en cada subgrupo*. En cambio, para *los perfiles da una significación p=0.0 < 0.05, con lo que puede rechazarse la hipótesis original y por lo tanto p_o no tiene una mediana poblacional igual en cada subgrupo*.

2.e) Estudio de medianas y variabilidad para n

De manera similar a lo estudiado con p_o, se realizó un estudio de medianas y variabilidad de n para los subgrupos. La Tabla XV, muestra los valores obtenidos.

Tabla XV.Significancia de medianas y variabilidad para n

grupo	parámetro	estudio medianas	estudio variabilidad
Tipo de archivo	Chi-Cuadrado	74.880	90.848
	grados libertad	2	2
	Significancia	0.000	0.000
perfil de narración	Chi-Cuadrado	15.607	11.680
	grados libertad	3	3
	Significancia	0.001	0.009

Donde p <0.05 indica que realmente se trata de poblaciones distintas según el subgrupo. Esto indicaría que *n es buen discriminante de tipos de archivo* ya que tiene comportamiento significativamente distinto para cada subgrupo.

Por lo tanto las poblaciones tienen un comportamiento que permitiría afirmar que son subgrupos distintos.

2.f) Estudio de p_o a nivel documento

Se estudió el valor promedio de p_o calculado para cada frase (tal como se lo usa en WIH para ponderar los EBH) y se contó la cantidad de frases menores, mayores e iguales a cero. Estas frecuencias se identificaron como las variables MENOR_CERO, MAYOR_CERO e IGUAL_CERO respectivamente. Es posible observar nuevamente un comportamiento perfectamente distinguible por cada uno de los subgrupos. A continuación se muestran los resultados de estimar la curva de comportamiento como Binomial y la bondad de ajuste resultante en cada caso. En la Tabla XVI se muestran los resultados obtenidos.

Tabla XVI.Bondad de ajuste a Binomial para n=50

Tipo de archivo	subgrupo	MAYOR CERO	MENOR CERO	IGUAL CERO
		parámetro p (Chi-Cuadrado)	parámetro p (Chi-Cuadrado)	parámetro p (Chi-Cuadrado)
perfil de narración	literario	0.00720 (0.9434)	0.01000 (0.6795)	0.00280 (0.9998)
	mensajes	0.00320 (0.9996)	0.00640 (0.9718)	0.01040 (0.6196)
	técnico	0.00600 (0.9809)	0.01000 (0.6795)	0.00400 (0.9984)
	doc	0.00200 (>0.9999)	0.01520 (0.0577)	0.00280 (0.9998)
	foro	0.00400 (0.9984)	0.00960 (0.7347)	0.00640 (0.9718)
	webindex	0.00280 (0.9998)	0.00640 (0.9718)	0.01080 (0.5563)
	blog	0.00720 (0.9434)	0.01200 (0.3608)	0.00080 (>0.9999)

Un nivel de significación menor al valor de significación nominal de la prueba conduce a un rechazo del modelo distribucional propuesto. En casi todos los casos se presentan valores que permiten indicar que las muestras son modelizables como distribuciones Binomiales con los parámetros especificados en el encabezado de cada tabla.

Además, es posible afirmar que las distribuciones tienen significación respectivamente superior para Binomiales que para Poisson, en cada subgrupo. Es posible que en el caso de Poisson, la curva en estos rangos y parámetros sea similar a la Binomial, pero no represente el comportamiento real poblacional.
En consecuencia podría afirmarse con razonable grado de precisión que *las distribuciones son Binomiales*.

2.g) Estudio de p_o al nivel de significación

De los diversos tipos de archivos, se puede decir, basándose en los resultados previos, que cada uno tiene características propias en cuanto a cantidad de sentencias típicas, y distribución de p_o. Por ese motivo se seleccionaron dos muestras para estudiar el nivel de significación de frases en relación con p_o. Una muestra es el subgrupo de mensajes y la otra es el perfil "documento".
Inicialmente se observaron todos los documentos. Hay un porcentaje que presenta el 100% de los valores p_o en cero. La Tabla XXIV muestra los porcentajes respectivos para documentos y mensajes.

Tabla XVII.Muestras con 100% de valores p_o en cero

	mensajes	documentos
100%	52	13
<100%	48	87

De los documentos con valores de p_o distintos de cero, se tomaron los valores de p_o mínimo y máximo para cada documento. No se consideraron otros valores cercanos (en la práctica es necesario el uso de los mismos como se muestra en 5.4. Práctica de la estrategia con po). Se extrajeron las frases correspondientes a estos p_o seleccionados y se estudió si la misma se relaciona al tema principal o a un tema secundario. Los criterios seguidos para considerar cada caso se detallan en Apéndice D: Análisis de datos po.
En la Tabla XVIII se muestran los porcentajes cuando se consideran tan sólo el mínimo p_o y el máximo.

Tabla XVIII.Significado de las sentencias con mínimo y máximo p_o

	%mensajes	%documentos
Mínimo p_o		
-tema principal	92	77
-tema secundario	8	15
-otra frase	0	8
Máximo p_o		
-tema principal	83	38
-tema secundario	17	54
-otra frase	0	8

Puede apreciarse que en ambos casos el tema principal se ve reflejado y en parte el/los tema(s) secundario(s) también. Se detectaron muy pocas frases de otro tipo.

Un fenómeno especial es la aparición del tema principal en la primera frase del texto. En el caso de los mensajes tiene una frecuencia muy alta que no se observa, en los documentos estudiados en la sección de perfiles de narración. En Tabla XIX se aprecian los porcentajes correspondientes.

Tabla XIX.Primera sentencia

	%mensajes	%documentos
máximo p_o	29	0
mínimo p_o	33	0
otro	38	26

2.h) Consideraciones finales

Se ha mostrado estadísticamente hasta aquí que, dado p_o (valores de ponderación de frases) y ***n*** (cantidad de valores p_o de un documento):

-El comportamiento de n y p_o no es asimilable al de una curva normal.

-Existe una correlación progresivamente lineal entre p_o medio (promedio de p_o para todo el documento) y ***n***. Este comportamiento se mantiene para los tipos de texto y perfiles de narración. La correlación es significativa sólo para el tipo "técnico" y perfil "blog".

-El comportamiento de variabilidad de p_o es esencialmente igual para los tipos y perfiles.

-El comportamiento de medianas de p_o es esencialmente igual para los tipos, pero no es así para perfiles.

-El comportamiento poblacional de ***n*** en medianas y variabilidad es esencialmente distinto para los tipos y perfiles.

-Cada subgrupo de tipo de documento y perfil de narración tiene un comportamiento Binomial distinguible cuando se contabiliza el valor p_o como: MENOR_CERO, IGUAL_CERO, MAYOR_CERO.

-Existe una relación entre el nivel de importancia de una sentencia y la cercanía de p_o a cero.

De lo anterior, no puede afirmarse que los tipos de documentos se traten de poblaciones distintas, pero sucede lo contrario con los perfiles, dado que las pruebas de variabilidad dan diferencias significativas. Por lo tanto se podría afirmar que *esos*

subgrupos se comportan como poblaciones distintas sólo para los perfiles de narración.

A partir de ésto puede decirse que p_o es una nueva métrica invariante al tamaño de un documento y tipo de texto, que permitiría diferenciar el perfil del narrador por sus medianas, y que mediría razonablemente bien la relevancia relativa de las frases dentro de un documento.

3. El modelo matemático

Los niveles de significación al nivel de sentencias son útiles para manejar los documentos almacenados y administrados por WIH. Tratamientos con manejo de niveles de representatividad no son nuevos y fueron propuestos ya por varios autores [28], [90], [89]. Los modelos propuestos suelen partir de un conjunto de documentos $D= \{d_1, d_2,..., d_m\}$, que constituyen la base de datos sobre la cual se desean realizar consultas. Se suelen asociar ciertas palabras (denominados normalmente "labels"), seleccionadas de un conjunto predeterminado. A cada una de estas palabras se les establece niveles de representatividad en referencia al texto asociado. Cuando se realiza una consulta (denominada normalmente "query"), se utilizan comparaciones contra estas etiquetas y sus niveles de representatividad a fin de filtrar los documentos que más probablemente sean pertinentes. A tal fin se definen una serie de modelos matemáticos que indican manejos difusos alternativos [54].
En el caso de WIH, hay una marcada diferencia con el procedimiento tradicional en algunos aspectos:

-No se predefiniría un conjunto específico de palabras sino que cualquiera de las que figuran en un texto son candidatas a derivar EBH relevantes. Las mismas palabras del texto (y no etiquetas asociadas) son detectadas y derivadas en elementos homogeneizados ponderados automáticamente.

-La ponderación de un EBH se proyecta de cierta manera en toda la estructura E_{ci} que representa una sentencia, y se combina con la ponderación del resto de las palabras de la frase.

-Las ponderaciones se derivan de la posición y conformación de las palabras (morfosintaxis).

-Las ponderaciones no se restringen al rango [0...1]. Pueden abarcar [-1...+1]

Las propuestas con operadores difusos, han probado ser tan competitivas y eficientes en la etapa de recuperación como las performantes algorítmicas clásicas[91]. Si bien los primeros trabajos comenzaron con la manipulación difusa (con lógica difusa) de los términos de una consulta [90] (para evaluar grados de necesidad de requerimientos durante la formulación de las queries), se ha mostrado que el modelado con lingüística difusa ayuda a mejorar los resultados de las búsquedas [55] Algunas propuestas, incluso, van más allá: en [91] se propone una combinación de palabras cuantificadas con sentencias ponderadas usando operadores de cuantificación semi difusos. Algunas de estas ideas fueron aplicadas incluso a la fase de IR para mejorar el funcionamiento de un CBR (Case Based Reasoner) [63]. En el caso de WIH, también se cuantifican palabras y sentencias.

El estudio del modelo difuso y su implementación corresponde a la Estructura Externa (por ser parte del manejo y no de la reestructuración de los datos, según lo descripto en el

Capítulo 3. Descripción de la estrategia global), que no es parte del alcance de este trabajo. Sin embargo, se plantearán en (iii) El modelo difuso como **fe** en WIH algunos lineamientos básicos que permitan visualizar los condicionamientos mínimos que se imponen a la Estructura Virtual, para habilitar a la Estructura Externa al uso de este tipo de algorítmica.

(iii) El modelo difuso como f_e en WIH

Manejo difuso en la base de documentos:

Según lo presentado en la sección anterior, es posible realizar un tratamiento difuso de los documentos, según la relevancia de los contenidos.

A tal fin, los siguientes pasos son realizados por la Estructura Virtual como conjuntos de f_e:

-Los documentos son almacenados.

-Se definen las estructuras E_{ci} y E_{ce}.

-Se deriva el grado de significación de las sentencias de cada documento, asignando el valor de p_o correspondiente a palabras y sentencias.

La Estructura Externa determina el tratamiento de las consultas, la definición de operadores adecuados para filtrar, recuperar y ordenar los documentos de respuesta. Los problemas más comunes en esta actividad suelen relacionarse con [90]:

a) La descripción matemática del cuantificador.

b) La adecuación de cierto valor numérico como resultado.

El primero está resuelto principalmente por el Motor de Composición y su definición de p_o. El tratamiento deberá completarse implementando el uso adecuado de esta ponderación para filtrar los resultados ya sea en función de una consulta o de una navegación.

La segunda problemática será controlada por una serie de métricas de calidad, evaluadas por el Sistema Controlador con apoyo del Motor de Métricas (ver [84]).

Manejo difuso en las consultas:

El uso de lógica difusa introduce el manejo de la ambigüedad propia de la lengua natural cuando se usa como parte de la "query" o consulta. Si bien es responsabilidad de la Estructura Externa la definición y tratamiento al respecto, se realizará una apreciación inicial acerca de los elementos y conveniencia de este tipo de tratamiento. En estos casos se requiere al usuario una apreciación numérica de sus prioridades dentro de la búsqueda. Pero muchos usuarios no están habilitados para dar una apreciación numérica de sus necesidades de información. Por eso se categoriza lingüísticamente (con términos como muy importante, importante, etc.) [54]. El estudio y manejo de términos sopesados en la consulta, manifestó la necesidad de definir un nuevo operador de agregación mucho más eficiente llamado LOWA (Linguistic Ordered Weighted Averaging), con un costo limitado, pero un tanto más restrictivo en cuanto a su implementación.
Los pasos a seguir por parte de la Estructura Externa serán:

-Predefinir una secuencia corta de etiquetas (7 ó 9) que califican lingüísticamente la importancia para expresar {null, very_low, low, medium, high, very_high, total}

-Definir las funciones de pertenencia de cada etiqueta.

-Definir los operadores: operador negación, de comparación (MAX (s_i, s_j), MIN (S_i, s_j)), de agregación (LOWA).

-Reducir las queries a un conjunto de términos (con peso) combinados por AND, OR, NOT, permitiendo que el usuario disponga de la prioridad de los términos según sus necesidades de información.

Muchos procesos reales tienen manejos ambiguos similares a los de toma de decisión [56]. En estos casos se usa fuzzy-sets. El problema surge cuando coexisten varias opiniones en forma de relaciones lingüísticas de preferencias (relaciones porque son apreciaciones relativas a una escala preexistente, lingüísticas porque se establecen etiquetas lingüísticas a cada categoría ordenada, que es a su vez un fuzzy set). Por

ello debe proveerse de un adecuado operador de agregación, (además de los tradicionales para manipular fuzzy sets), para proveer un resultado racional, tal como LOWA[56]. Este tratamiento de agregación puede mirarse como un caso especial de otro: agregación de información en el campo de teoría de decisión con múltiples personas y múltiples criterios. El operador LOWA se basa en un operador OWA (Ordered Weighted Averaging) definido por Delgado, y la combinación convexa de etiquetas lingüísticas. Las etiquetas $\{a_1,..., a_m\}$ ordenadas y con preferencias $\{w_1,..., w_m\}$.

Propiedades del operador: monotonicidad, conmutatividad, es un operador ORAND.

Axiomas que sigue: dominio sin restricción, unanimidad o idempotencia, asociación positiva de valores sociales e individuales, independencia de alternativas irrelevantes, soberanía cívica, neutralidad, neutralidad respecto a la escala de intensidades, proceso de votación descomponible.

Una mención especial merece el tratamiento de los perfiles de narración, como medio para discriminar modalidades de uso del lenguaje. Desde este punto de vista es un hallazgo interesante que podría profundizarse para determinar la calidad presunta del documento (yendo más allá del análisis de las sentencias comprendidas).

3.3. Módulo Motor de Asimilación (MA)

Este módulo comprende todas las tareas necesarias para la inserción de un conjunto de $\mathbf{E_{ci}}$ en una $\mathbf{E_{ce}}$ representativa de todo el contenido de esa página. Realiza la búsqueda de un locus dentro de la red de $\mathbf{E_{ce}}$ preexistente y la inserción en la misma (esa inserción puede no ser definitiva ya que diferentes circunstancias pueden movilizar su locus, como bajas y modificaciones de las páginas que los originan).

Las principales actividades del módulo comprenden:

-La extracción de palabras indicadoras.

-Ponderación de las indicadoras.

-Inserción en los índices de la RV.

La estructura de funcionamiento del **MA** se compone de los módulos indicados en la Fig. 14.

Se puede apreciar en la figura, que el subsistema Extractor se encarga asincrónicamente de procesar las $\mathbf{E_{ci}}$. Toma de las mismas la mínima información posible (en la actual implementación el idPalabra, las marcas de palabras indicadoras, las raíces STEM y la ponderación p_o).

Para completar la inserción en los índices de la Red Virtual, el módulo Asimilador consta de un GestorIV y un GestorIEV. Los mencionados índices actualmente tienen la estructura de la Fig. 15.

Fig. 14. Estructura del MA

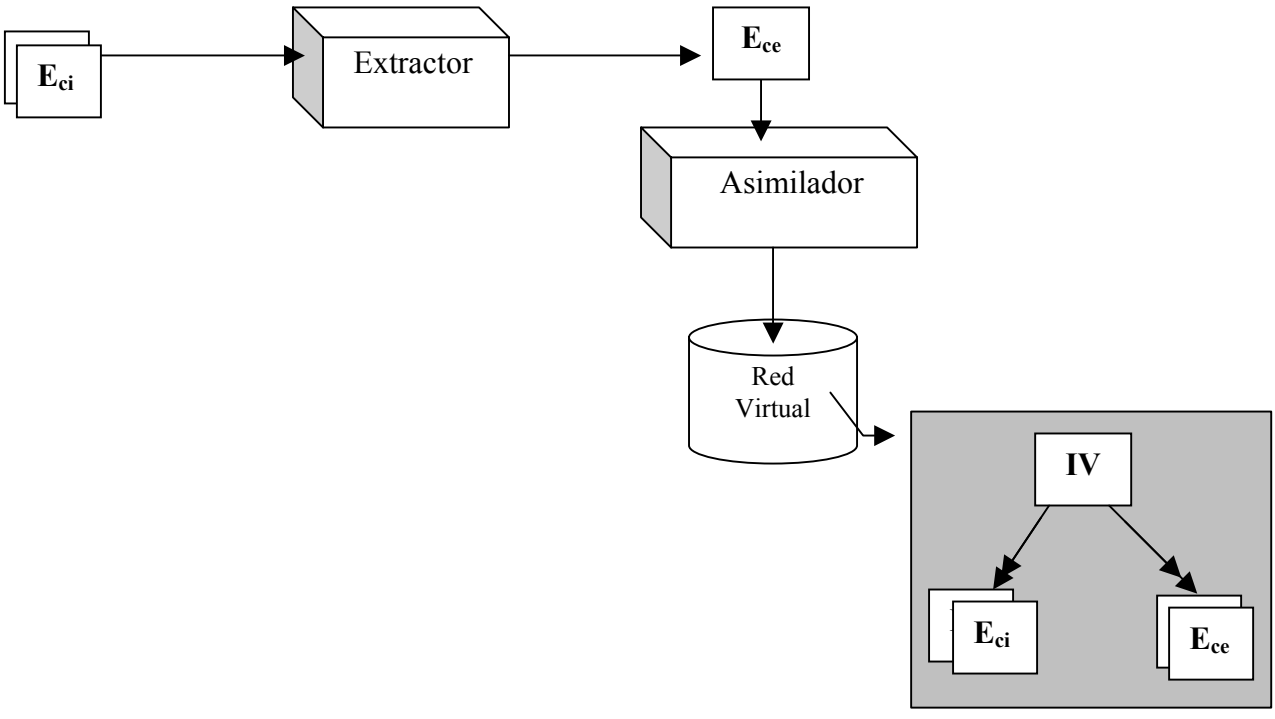
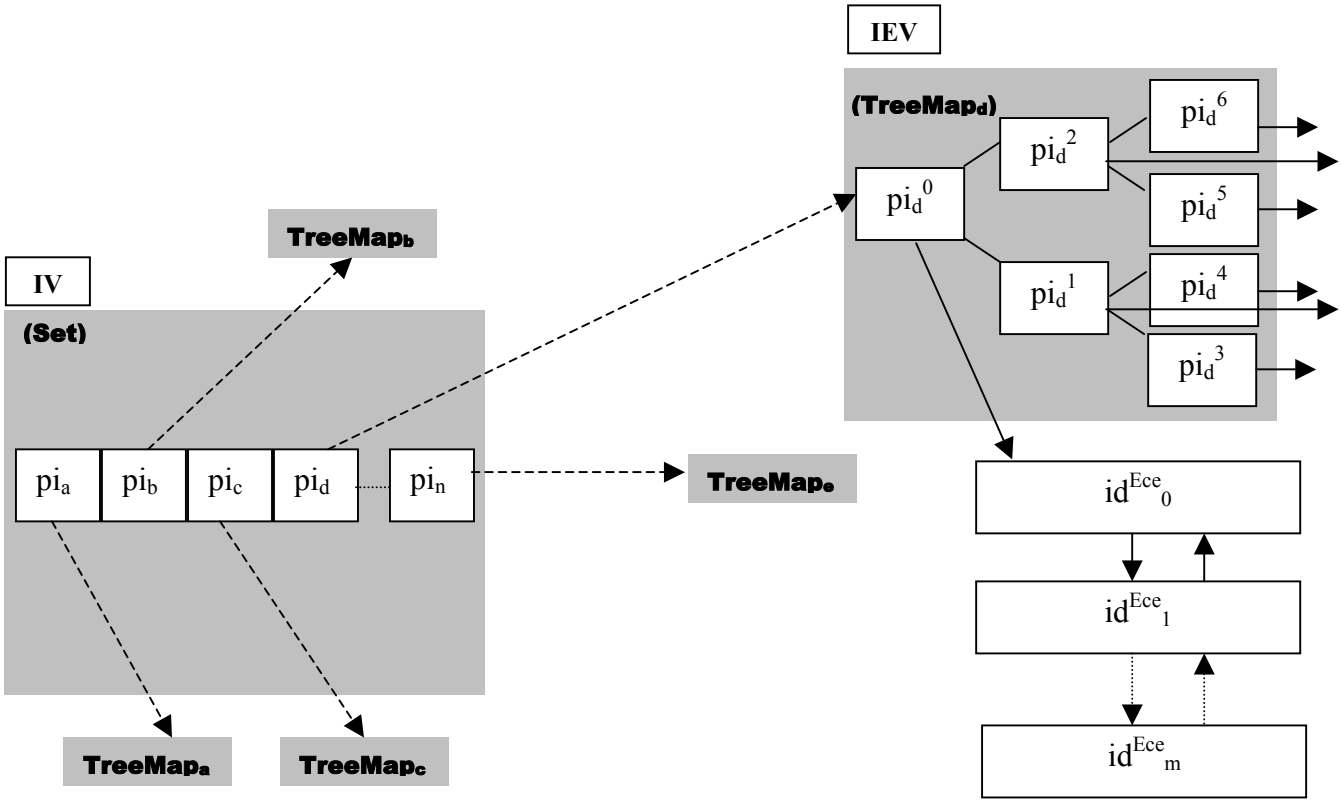


Fig. 15. Estructura de índices de la Red Virtual



El Indice Virtual se compone de dos subíndices: IV (índice virtual compacto, administrado por el GestorIV) e IEV (índice virtual extendido, administrado por el GestorIV).

El subíndice IV: tiene una estructura de datos correspondiente a una lista ordenada sin repeticiones de elementos (un Set de datos, en Java). Sus nodos se componen por STEM (raíz de la palabra), idE_{ci} (puntero lógico al E_{ci} que contiene la palabra) y p_o (ponderación de la palabra).

El subíndice IEV: tiene una estructura de árbol ordenado. Sus nodos se componen por idE_{ce} (puntero al E_{ce} que se relaciona con el E_{ci} y la palabra), pero se administra como un mapa de datos (un TreeMap de Java), donde se asocia cada idE_{ce} a un idPalabra. Esto hace que todas las palabras con el mismo STEM estén ordenadas y asociadas a uno o más E_{ce} (dependiendo en qué E_{ce} figuran).

El subsistema Asimilador toma cada nuevo E_{ce} y le extrae los indicativos. Le pide al GestorIV que asimile en el IV las raíces de las palabras nuevas (los primeros caracteres, la cantidad de caracteres es denominada LONGITUD_INDICATIVO), luego invoca al GestorIEV. Este toma el conjunto de palabras cuyas raíces coinciden y las ubica en un subíndice IEV propio de esa raíz.

Con estos subíndices es posible administrar las E_{ci} con las palabras buscadas y considerar su p_o (en IV), o bien directamente administrar los E_{ce} con esas palabras (en IEV). Además, este esquema permitiría manejar algún tipo de manipulación indirecta del ordenamiento lógico y dinámico de los documentos involucrados. En el Capítulo 6. Sensibilidad y capacidad de adaptación se continuará el análisis de este tema.

3.4. Sistema de adaptación: GM, MM y SC

Las estructuras descriptas hasta ahora no contemplan un hecho claramente distintivo de los contenidos Web: los cambios permanentes en tipo y cantidad. Muchas veces, las algorítmicas se ven complicadas por la necesidad de adaptación. Estos cambios pueden fluctuar en temporalidad y extensión.
Para alcanzar la flexibilidad necesaria y administrar los cambios de manera similar a la administración de la información tradicional, se agregan a la arquitectura de WIH tres componentes adicionales: el Gestor de Métricas (GM), el Motor de Métricas (MM) y el Sistema Controlador (SC). Todos éstos están íntimamente ligados y su objetivo primario es sensar la actividad actual y cambiar el comportamiento del MC y MA sin intervención exterior. A continuación se describe brevemente cada uno.

(i) El Sistema Controlador (SC)

Es el administrador de las actividades de sensado (realizada por el MM) y cambio de comportamiento (realizado por el GM).

Entre sus responsabilidades están:

-Evaluar el estado actual de la actividad utilizando las funciones de métrica (f_m) que activa el MM, y los umbrales p^* que éste le provee.

-Solicitar al GM el cambio de las funciones efectoras (f_e) cuando lo considera necesario.

-Solicitar al MM el cambio de las f_m si el contexto del funcionamiento lo requiere.

-Cambiar los umbrales de sistema, y de las estructuras E_{ce} y E_{ci} (denominados internamente O, O_{ece}, O_{eci}).

En forma permanente el SC evalúa el funcionamiento del sistema como un todo, y al nivel de las estructuras E_{ce} y E_{ci}. Para ello dispone de una serie de umbrales para cada uno de los parámetros que debe evaluar y una función de métrica (f_m) activada para cada caso. Realiza siempre la misma sencilla comparación contra el umbral. Por ejemplo para el caso de las E_{ci}:

$$\textbf{(ec. 2)} \quad p^{*eci} \leq \left| f_m^{eci}(O_{eci,} E_{ci}) \right|$$

Para las E_{ce}:

$$\textbf{(ec. 3)} \quad p^{*ece} \leq \left| f_m^{ece}(O_{ece,} E_{ce}) \right|$$

Y para el sistema en su totalidad:

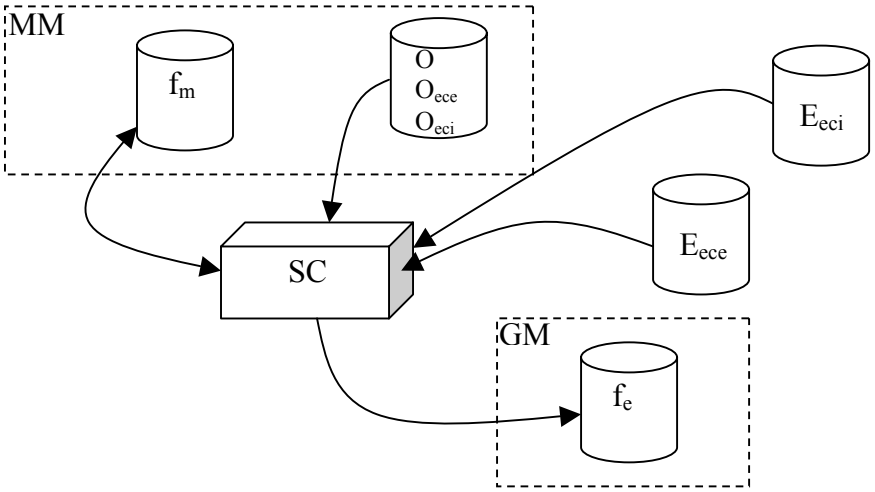
$$\textbf{(ec. 4)} \quad p^{*S} \leq \left| f_m^{s}(O_{,} S) \right|$$

En los casos en que no se verifica la relación, el SC estará detectando una necesidad de cambio de las f_e, y procederá a solicitar un cambio al GM. En caso de que la

relación sea superada ampliamente requerirá el cambio del método de evaluación (reflejado en el conjunto actual de f_m) al MM. En este último caso, según el estado del sistema también podrá requerir el campo de umbrales O, O_{ece}, O_{eci}.

En la Fig. 16 se observa un esquema general del intercambio de controles e información que realiza el SC con el resto de los módulos.

Fig. 16. Flujos del Sistema Controlador



La figura muestra al SC alimentándose de las f_m y los umbrales para saber cómo sensar la actividad, recurriendo a las estructuras del sistema a sensar y eventualmente indicando al MM y/o GM un cambio.

Los cambios realizados en las f_m impactará en la frecuencia y modalidad del cambio en las f_e. Dado que tanto el MC como el MA realizan su actividad enteramente a través de las f_e, el SC está cambiando su actividad indirectamente.

(ii) El Motor de Métricas (MM)

Es el responsable de la actividad de sensado del estado y actividades del sistema. Provee las facilidades necesarias para que un operador realice la administración completa de las f_m.

Entre las responsabilidades del MM están:

-Realizar el alta, baja y modificación de las f_m.

-Informar el resultado de aplicar una métrica al GM, cuando éste lo requiere.

-Informar el resultado de aplicar una métrica al SC, cuando éste lo requiere.

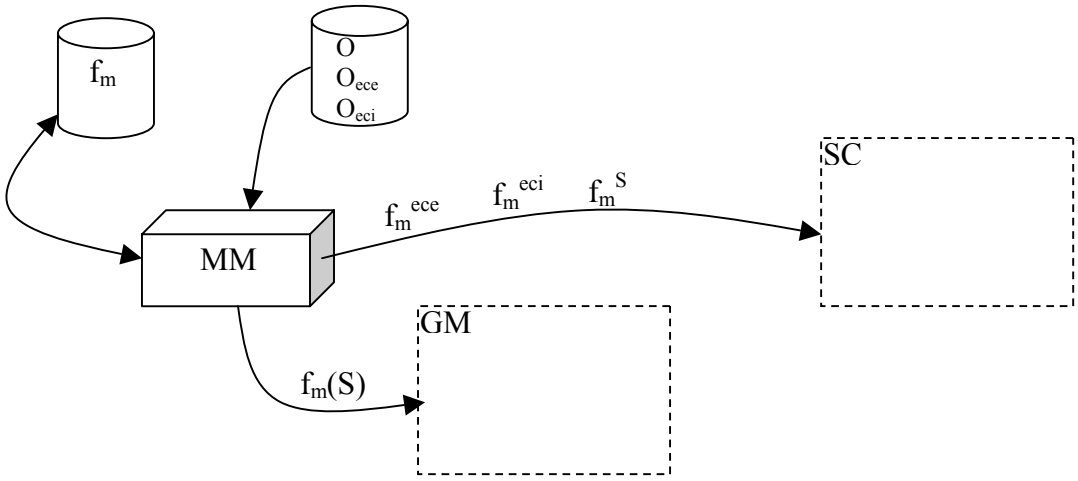
-Informar acerca de las métricas activas al SC, cuando éste lo requiere.

-Activar /desactivar las métricas que el SC le indique.

Podría decirse que este módulo es el dueño de las métricas y responsable de su manipulación directa bajo las órdenes del SC.

En la Fig. 17 se puede apreciar una representación esquemática el flujo de control e información correspondiente: su capacidad de realizar actividades de generación, modificación, activación y desactivación de métricas. Es capaz de informar al GM el resultado de aplicar las funciones de métricas activas y luego contrastar ésto contra los umbrales y objetivos del sistema. También puede informar al SC acerca de las funciones de métrica activas.

Fig. 17. Flujos del Motor de Métricas



Algunas funciones definidas e implementadas como métricas dentro de WIH son:

-Definir numéricamente la pertenencia de los perfiles (blog, doc, foro, Web index) que describen los modos en que una persona escribe texto.

-Definir numéricamente el funcionamiento del sistema para evaluar un eventual cambio del conjunto actuante de funciones de métrica para los E_{ci} y/o E_{ce}.

-Definir numéricamente la pertenencia de los tipos de texto (literario, mensajes, técnico) que describen los tipos de texto en la Web.

(iii) El Gestor de Métricas (GM)

Es el responsable del comportamiento del sistema. Provee las facilidades necesarias para la administración de las f_e, que son la implementación de las funciones de cada módulo.

Este módulo es el encargado de activar y desactivar las distintas funciones f_e, a pedido del SC. Dado que las funciones a cumplir dentro de WIH son múltiples, cada una de éstas es implementada por un conjunto de f_e. A su vez, la actividad realizada determina cuáles son los parámetros del caso.

Algunas funciones definidas e implementadas como efectoras dentro de WIH son:

-Reglas de Composición (para el MC): Realizan de maneras alternativas la composición de una E_{ci} a partir de una página Web.

-Reglas de Identificación (para el MC): Permiten extraer distintos tipos de información acerca de la página o documento original al que refieren las estructuras y palabras en proceso.

-Reglas de Asimilación (para el MC): Realizan de maneras alternativas la composición de una E_{ce} a partir de una E_{ci}.

-Reglas de extracción (para el MA): Extraen la información necesaria para interrelacionar las estructuras nuevas.

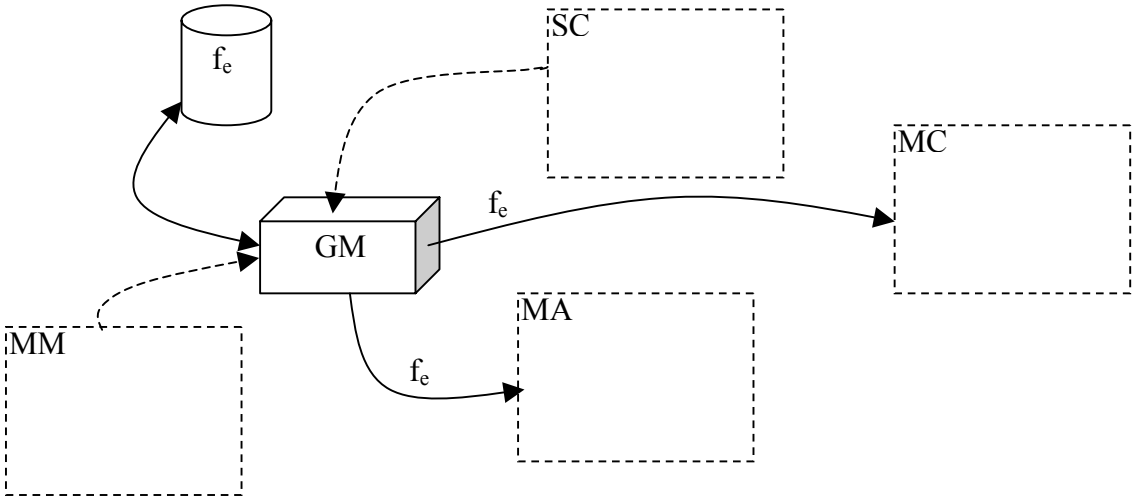
-Reglas de inserción (para el MA): Usan la información extraída para interrelacionar las estructuras E_{ci} y E_{ce} nuevas a las preexistentes.

Todas estas y otras reglas del sistema son generadas, mantenidas y utilizadas directamente sólo por el GM. En la Fig. 18 se puede ver el flujo de control e información de este módulo.

En la figura se refleja que es capaz de generar, modificar, activar o desactivar las funciones correspondientes. A su vez, el sistema SC puede ordenar un cambio en las activaciones. El MM puede informarle el resultado de aplicar cierta métrica cuando

sea necesario. Por último, el conjunto activo de funciones efectoras es un servicio que le brinda al MC y MA.

Fig. 18. Flujos del Gestor de Métricas



Capítulo 4. Descripción del prototipo implementado

4.1. Arquitectura de software y hardware

El modelo descrito se ha implementado en Java S.E. v1.6, puesto que, como se menciona en [117] es uno de los más populares para sistemas de recuperación de información. Se trabajó en un procesador Intel Pentium III con disco de 16GB con 6 GB de espacio libre, y 256MB de memoria RAM. El sistema operativo fue Windows XP Home Edition, versión 2002 Service Pack 2. La conexión a Internet se realizó a través de un módem externo (con capacidad máxima de 56Kbps) conectado a línea telefónica.

El pre-procesamiento se implementó como una clase separada en un paquete independiente. El resto del sistema se implementó en proyectos organizados en paquetes según lo indicado en la Tabla XX:

Tabla XX.Módulos de WIH

Módulo de arquitectura	Proyecto	Paquete	Clases importantes[38]
TLI	WIH00	motor.tli	Estadistica GestorPalabra GestorArchivo PreProcesadorArchivo MotorTLI
MC	WIH01	motor.composicion	MotorComposicion Eci EBH Regla ReglaComposicion ReglaEliminacion

[38] A fin de hacer más clara la presente descripción sólo se presentan aquí algunas de las clases que componen al sistema real.

Módulo de arquitectura	Proyecto	Paquete	Clases importantes[38]
MA	WIH02	motor.asimilacion	ReglaIdentificacionPagina MotorAsimilacion ECE IV IEV Asimilador Extractor GestorIV GestorIEV
SC	WIH03	sistema.controlador	SistemaControlador
MM	WIH04	motor.metricas	MotorMetricas FuncionMetrica SelectorPalabras
GM	WIH05	gestor.metricas	GestorMetricas FuncionEfectora

4.2. Diagrama de clases reducido

En esta sección se describen los diagramas de clases para los módulos TLI, MC y MA, que constituyen el corazón de WIH.

(i) motor.tli

Estadística: genera estadísticas al nivel de palabras y archivos para el sistema controlador y para los descriptores que luego se transformarán en parte del E_{ci}.

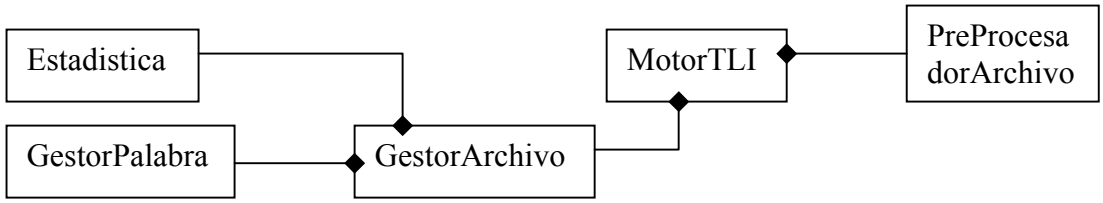
GestorPalabra: administra al archivo actual como un conjunto de palabras que se pueden navegar.

GestorArchivo: administra la palabra actual como un conjunto de caracteres que se pueden navegar.

PreProcesadorArchivo: realiza las conversiones preliminares necesarias para que un texto puro en html se convierta en palabras válidas (en texto).

MotorTLI: coordina la actividad de procesamiento relacionado con la traducción de un html a un archivo EBH.

Fig. 19. Diagrama de clases para TLI



(ii) motor.composicion

MotorComposicion: coordina la actividad de composición de un EBH en un E_{ci}.

E_{ci}: Estructura de Composición Interna.

EBH: cadena de caracteres básico preprocesado (Elemento Básico Homogenizado)

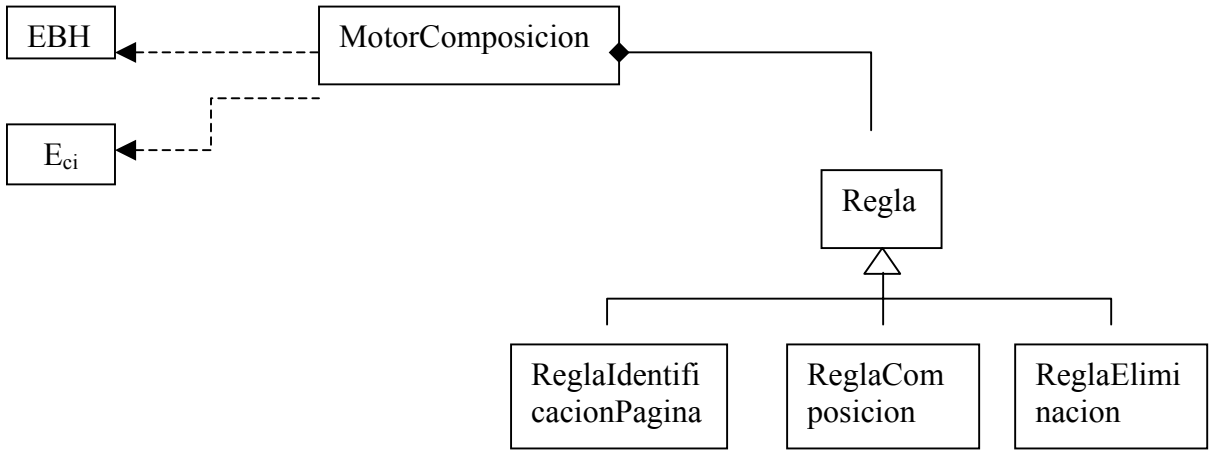
Regla: regla a aplicar a un determinado EBH

ReglaComposicion: realiza la inserción de un EBH dentro de una determinada E_{ci} si correspondiera.

ReglaEliminacion: realiza la eliminación de un EBH si corresponde.

ReglaIdentificacionPagina: genera la identificación de la página html con los EBH adecuados.

Fig. 20. Diagrama de clases para MC



(iii) motor.asimilacion

MotorAsimilacion: coordina la actividad de transformar una E_{ci} en una E_{ce} y su incorporación a la Red Virtual.

ECE: Estructura de Composición Externa.

IV: (Indice Virtual) que concentra información reducida sobre las E_{ci} y E_{ce} que se están procesando.

IEV: (Indice Extendido Virtual) concentra información complementaria para localizar y entender una E_{ce}.

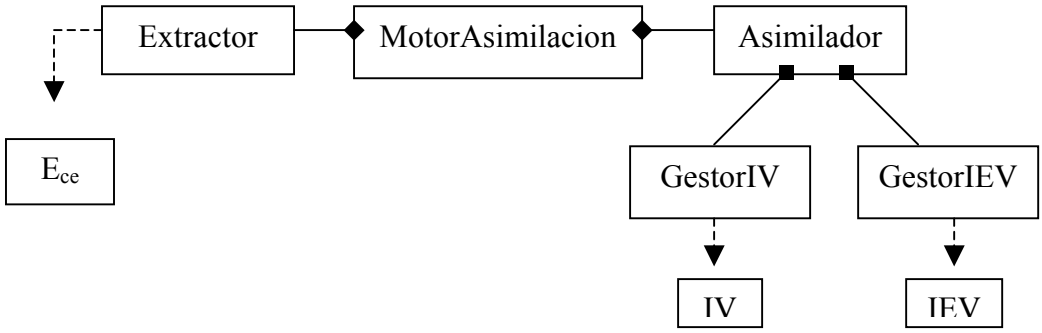
Asimilador: realiza la asimilación efectiva de una E_{ce} a la Red Virtual.

Extractor: realiza la construcción efectiva de la E_{ce} a partir de la E_{ci}.

GestorIV: actualiza el IV según las órdenes del Asimilador. Hay uno por cada hilo de ejecución (thread) despertado por el Asimilador.

GestorIEV: actualiza el IEV según las órdenes del Asimilador. Hay uno por cada thread despertado por el Asimilador.

Fig. 21. Diagrama de clases para MA



4.3. Metodología desarrollo

A continuación se detallan algunas consideraciones específicas al proceso de desarrollo.

1) Editor: La herramienta de desarrollo utilizada fue el JCreator Pro ©2000 - 2005 Xinox Software, versión 3.50.013.

2) Desarrollo: Los módulos se desarrollaron como prototipos evolutivos y de manera tal que cada uno puede ser invocado independientemente del resto. En todos los casos los parámetros de funcionamiento fueron levantados desde archivos de configuración con extensión ".cfg".

3) Funcionamiento: La actividad de los módulos fue diseñada para facilitar la rápida adaptación de las partes críticas a threads paralelos.

4) Tests: Cada módulo fue depurado individualmente, siguiendo una lista de casos prediseñados y luego se realizó una verificación integral del funcionamiento conjunto de a pares consecutivos de módulos.

5) Estadísticas: En algunos casos necesarios se extendieron las estadísticas definidas originalmente para poder realizar verificaciones puntuales (por ejemplo para estudiar detalladamente el comportamiento de perfiles y tipos de archivos, o para estudiar más profundamente el comportamiento de p_o). Estas extensiones actualmente son parte de la implementación, pero son obtenibles opcionalmente según los argumentos de arranque del módulo correspondiente.

6)GUI: En el estado actual de la implementación no se cuenta con interfaces gráficas, dado que este tipo de desarrollo trasciende el objetivo inicial del presente trabajo y no hace a la esencia del mismo.

Capítulo 5. Estudio de casos y resultados

5.1. Conceptos preliminares

Para comprender los casos en su totalidad es necesario describir algunos conceptos que se usarán para presentar los ejemplos.

(i) Descriptores EBH.

Son campos que describen cada palabra y son procesados para cada EBH. Estos descriptores fueron introducidos en 3.1. Módulo de Traducción a Lenguaje Interno. En la Tabla XXI se listan y explican los que sirven para clarificar el comportamiento y el tratamiento de WIH.

Tabla XXI.Descriptores HBE

Columna	Descripción
ID	Corresponde al descriptor idPalabra. Denota unívocamente a la palabra del EBH en proceso. A los fines mostrativos se determinó que sea la palabra original, base de los descriptores.
tema	Denota el tema al que corresponde el texto dentro de la página html.*
TP	Corresponde al descriptor tipoPal. Denota si la palabra actual es verbo, sustantivo, u otro.
PAT	Corresponde al descriptor palAntTipo. Denota sólo si existe una palabra que le antecede.
TPG	Corresponde al descriptor tipoPag. Denota si en el URL fue indicado que el contenido corresponde a un índice.
EM	Corresponde al descriptor empiezaMayuscula. Denota si la palabra comienza con mayúscula o no.
LP	Corresponde al descriptor longPalabra. Denota la cantidad de caracteres que forman la palabra actualmente en proceso.
CVF	Corresponde al descriptor cantVocalesFuertes. Denota la cantidad de veces que figuran las vocales a, e, o en la palabra actual.

Columna	Descripción
TER	Corresponde al descriptor terminacion. Denota el tipo de terminación según ciertas categorías prefijadas. *
CVD	Corresponde al descriptor cantVocalesDebiles. Denota la cantidad de veces que figuran las vocales i, u en la palabra actual.
RE	Corresponde al descriptor resaltada. Denota si la palabra pertenece a un fragmento de texto resaltado con comillas, admiración, etc.
ET	Corresponde al descriptor esTitulo. Denota si la palabra fue empleada en una posición asimilable a un título.
STEM	Denota la raíz de la palabra extraída según el algoritmo de stemming.

***Este campo es incorporado temporariamente en el sistema sólo a los fines mostrativos pero no forma parte de la implementación definitiva de WIH.**

(ii) Estructura de una E_{ci}.

Una E_{ci} se presenta aquí como una colección de EBH, cada una de las cuales es representada por una línea con el siguiente formato:

<IR>pAnt pal [pal/p_o (tpal."url")] S D

En la Tabla XXII se explica el significado de cada sigla del formato.

Tabla XXII.Campos E_{ci}

Campo	descripción
<IR>	Es el número de identificación de la regla de composición aplicada a la palabra. Cuando es 0 significa que ninguna regla fue aplicada. Las reglas de composición definen cómo se inserta dentro de la E_{ci} y el formato de su representación interna.
pAnt	Es la palabra que precede a la que se está representando con esta línea. En los casos en que una regla se aplica, se incorpora en una línea adicional donde pAnt vale null.
Pal	Es la palabra que se está procesando. En los casos en que una regla se aplicó, vale null y va acompañada de una pAnt con el valor de la última

Campo	descripción
	palabra procesada antes de la aplicación de la regla.
p_o	Es la ponderación característica de la palabra.
tpal	Es el tipo de palabra detectado.
url	Es el URL de la página de donde se extrajo la palabra.
S	Es "+" cuando la palabra sea marcada como indicativa del contexto, de lo contrario es "-".
D	Es ">" cuando el sentido de lectura de las palabras es hacia delante, y "<" cuando debe leerse hacia atrás.

***En algunos casos se omitirá el campo p_o cuando su presentación resulte innecesaria.**

(iii) Estructura de una E_{ce}.

Una E_{ce} se presenta aquí como una colección de EBH, cada una de las cuales es representada por una línea con el siguiente formato:

pal/p_o

En la Tabla XXIII se explica el significado de cada sigla del formato.

Tabla XXIII.Campos E_{ce}

Campo	descripción
Pal	Es la palabra que se está procesando. En los casos en que una regla se aplicó, vale null y va acompañada de una pAnt con el valor de la última palabra procesada antes de la aplicación de la regla.
p_o	Es la ponderación característica de la E_{ce}, calculada como combinación de las p_o de palabras por obra de las f_e.

5.2. Tratamiento de EBH opuestos y contradicciones

A continuación se presentan los resultados de procesar casos de contradicción con el prototipo WIH y casos de EBH opuestos según lo presentado en el apartado 3.2.1. Tratamiento de EBH opuestos y contradicciones.

(i) Estudio de contradicciones usando WIH

Se procesaron tres casos de frases contradictorias, extrayéndose las correspondientes E_{ci} y E_{ce}. Por cada caso se transcribe la sentencia del texto original, los EBH, el E_{ci} y E_{ce} que genera y una conclusión del caso.

Caso 1 de contradicción

La sentencia procesada es: "Las orquídeas vegetativamente son muy diferentes pero iguales.". En la Tabla XXIV se transcriben los EBH generados. La Fig. 22 y la Fig. 23 muestran las correspondientes E_{ci} y E_{ce} respectivamente.

Tabla XXIV. EBH del caso 1 de contradicción

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Las	orquideas	otro	ninguna	contenido	3	1	s	0	si	no	no	Las
orquídeas	orquideas	sustantivo	otro	contenido	9	3	s	2	no	no	si	orquid
vegetativamente	orquideas	otro	otro	contenido	15	6	null	1	no	no	no	veget
son	orquideas	verbo	otro	contenido	3	1	on	0	no	no	si	son
muy	orquideas	otro	otro	contenido	3	0	null	1	no	no	no	muy
diferentes	orquideas	otro	otro	contenido	10	3	s	1	no	no	no	diferent
pero	orquideas	otro	otro	contenido	2	1	en	0	no	no	no	en
iguales	orquideas	otro	otro	contenido	1	0	null	0	no	no	no	igual

Fig. 22. E_{ci} del caso 1 de contradicción

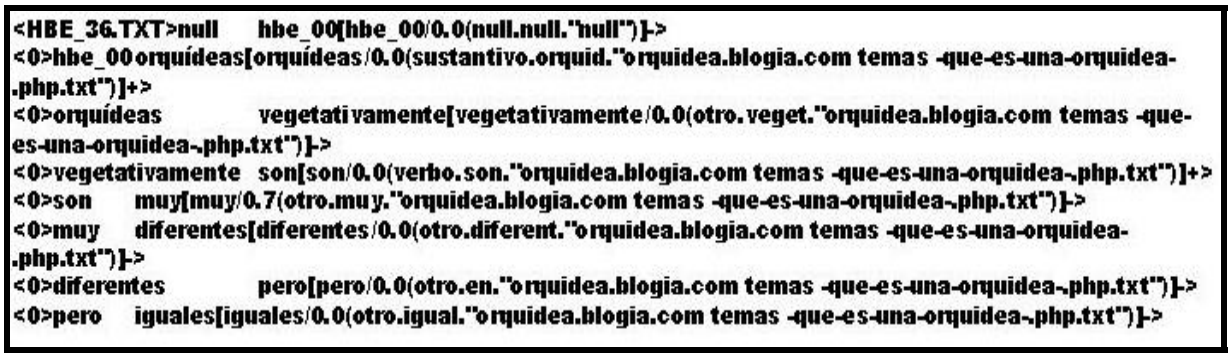


Fig. 23. E_{ce} del caso 1 de contradicción

Puede observarse que la contradicción semántica no sólo no es detectada por el sistema sino que, además, queda oculta al nivel del texto original. La E_{ci} procesada alcanza a detectar que existe una definición de orquídeas promoviendo al nivel de E_{ce} los dos términos EBH más representativos de la sentencia (se pueden ver resaltados dentro de la E_{ci} transcripta): orquídeas y son. El lector que acceda a los contenidos a través de la E_{ce} sólo estará en conocimiento de que hay una definición pero no chocará con la contradicción a menos que acceda al texto o a la E_{ci} correspondiente.

Caso 2 de contradicción

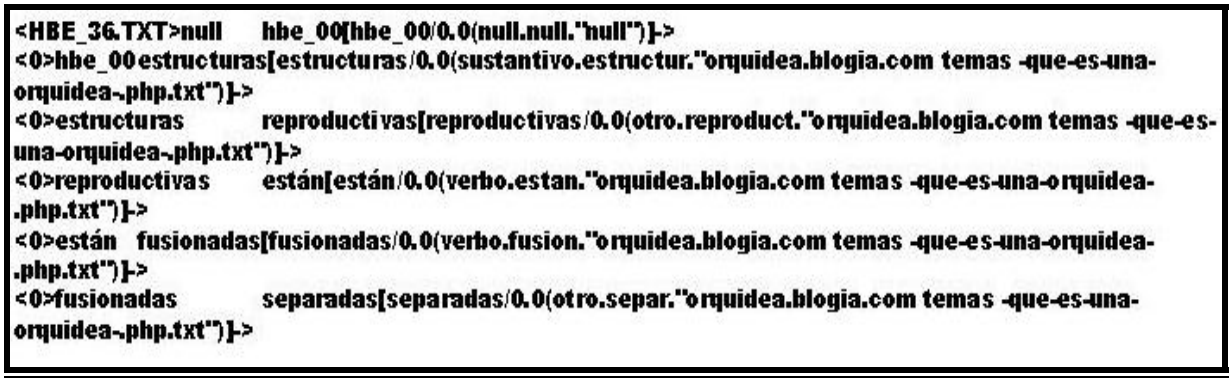
La sentencia procesada es: "Las estructuras reproductivas están fusionadas separadas.".

En la Tabla XXV se transcriben los EBH generados. La Fig. 24 muestra la correspondiente E_{ci}, pero no se promueven términos a nivel E_{ce}.

Tabla XXV. HBE del caso 2 de contradicción

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Las	orquideas	otro	ninguna	contenido	3	1	s	0	no	no	si	las
estructuras	orquideas	sustantivo	otro	contenido	11	2	s	2	no	no	no	estructur
reproductivas	orquideas	otro	otro	contenido	13	3	s	2	no	no	no	reproduct
están	orquideas	verbo	otro	contenido	5	2	null	0	no	no	no	estan
fusionadas	orquideas	verbo	otro	contenido	10	3	s	2	no	no	no	fusion
separadas	orquideas	otro	otro	contenido	4	2	null	0	no	no	no	separ

Fig. 24. E_{ci} del caso 2 de contradicción



Puede observarse que la contradicción semántica queda nuevamente oculta a nivel del texto original. La frase es clasificada como superflua por el sistema debido a que no alcanza a promover términos al nivel de E_{ce}.

Caso 3 de contradicción

La sentencia procesada es: "Tienen y no tienen dos tipos básicos de crecimiento simpodial."

En la Tabla XXVI se transcriben los EBH generados. La Fig. 25 muestra la correspondiente E_{ci}, pero no se promueven términos a nivel E_{ce}.

Tabla XXVI. HBE del caso 3 de contradicción

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Tienen	orquideas	verbo	ninguna	contenido	6	2	en	1	si	no	no	Tien
y	orquideas	otro	otro	contenido	1	0	null	0	no	no	no	y
no	orquideas	otro	otro	contenido	1	0	null	0	no	no	no	no
tienen	orquideas	verbo	ninguna	contenido	6	2	en	1	si	no	no	tien
dos	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	dos
tipos	orquideas	otro	otro	contenido	5	1	s	1	no	no	no	tip
básicos	orquideas	otro	otro	contenido	7	2	s	1	no	no	no	basic
de	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	de
crecimiento	orquideas	sustantivo	otro	contenido	11	3	null	2	no	no	no	crecimient
Simpodial	orquideas	otro	ninguna	contenido	9	2	null	2	si	no	no	Simpodial

Fig. 25. E_{ci} del caso 3 de contradicción

```
<HBE_36.TXT>null      hbe_00[hbe_00/0.0(null.null."null")]->
<0>hbe_00 Tienen[Tienen/0.0(verbo.Tien."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<4>Tienen null[null/0.0(null.null."null")]->
<0>null      no[no/-1.0(otro.no."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>no        tienen[tienen/0.0(verbo.tien."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>tienen  dos[dos/0.0(otro.dos."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>dos     tipos[tipos/0.0(otro.tip."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>tipos   básicos[básicos/0.0(otro.basic."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<11>básicos          null[null/0.0(null.null."null")]+>
<0>null      crecimiento[crecimiento/0.0(sustantivo.crecimient."orquidea.blogia.com temas -que-es-una-
orquidea-.php.txt")]->
<0>crecimiento       Simpodial[Simpodial/0.0(otro.Simpodial."orquidea.blogia.com temas -que-es-una-
orquidea-.php.txt")]->
```

Puede observarse que la contradicción semántica queda nuevamente oculta a nivel del texto original. La frase es clasificada como superflua por el sistema y nuevamente no alcanza el nivel de una E_{ce}.

(ii) Estudio de caso con la estrategia de opuestos

Se procesó la siguiente frase:
"Si algun día desapareciera de la faz de la tierra, el hongo Rhizostoma y el cultivo in vitro salvaría a estas preciadas plantas que superan las 1000 especies."

Como puede apreciarse, la palabra **desapareciera** tiene un prefijo **des**, por lo que WIH caracterizará la palabra de manera especial. Por ejemplo, se le podría ponderar con -1. Considerando ésto, se construirá la E_{ci} como en la Fig. 26.

Fig. 26. E_{ci} del caso de opuestos

```
<HBE_36.TXT>null      hbe_00[hbe_00/0.0(null.null."null")]->
<0>hbe_00 Tienen[Tienen/0.0(verbo.Tien."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<4>Tienen null[null/0.0(null.null."null")]->
<0>null      no[no/-1.0(otro.no."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>no        tienen[tienen/0.0(verbo.tien."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>tienen  dos[dos/0.0(otro.dos."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>dos       tipos[tipos/0.0(otro.tip."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<0>tipos     básicos[básicos/0.0(otro.basic."orquidea.blogia.com temas -que-es-una-orquidea-.php.txt")]->
<11>básicos            null[null/0.0(null.null."null")]+>
<0>null      crecimiento[crecimiento/0.0(sustantivo.crecimient."orquidea.blogia.com temas -que-es-una-
orquidea-.php.txt")]->
<0>crecimiento         Simpodial[Simpodial/0.0(otro.Simpodial."orquidea.blogia.com temas -que-es-una-
orquidea-.php.txt")]->
```

Se puede observar que la característica ($\mathbf{p_o}$) sería 0.0 en todos los casos, a no ser por la palabra **desapareciera**. Esta sentencia tendrá una redacción perfilada por algún valor $\mathbf{p_o}^{\mathbf{Ece}}$, derivado de la combinación específica de valores $\mathbf{p_o}$ (más adelante se verá que también pueden aparecer otros tipos de partículas que inciden en esta caracterización). Las $\mathbf{f_e}$ activas serán las encargadas de definir la manera en que se realizará dicha derivación. Para este caso específico, el sistema consideró que la única palabra significativa a promover es plantas, y se calculó:

$$p_o^{Ece} = \sum_{i=1}^{i=n} p_o^i / n = \text{-}0.0625 \qquad (1)$$

Siendo **n** la cantidad de instancias **pal**abras distintas de **null**. Cuando el E_{ce} resulta del proceso de varias sentencias, se obtiene un valor que resulte ser representativo de todas. Una vez más, las f_e activas serán responsables de la actividad. Para asimilar la E_{ce} al resto de la Red Virtual, el sistema se limitará a insertar una referencia a cada E_{ce} con su mejor valor de p_o, y una referencia a las E_{ci} y URL de origen.

5.3. Tratamiento de EBH ambiguos y ambigüedades

A continuación se presentan los resultados de procesar casos de ambigüedad con el prototipo WIH y casos de EBH ambiguos según lo presentado en el apartado 3.2.2. Tratamiento de EBH ambiguos y ambigüedades.

(i) Estudio de casos de ambigüedad con la propuesta WIH

Por cada caso se describirán los tópicos que se enumeran a continuación:

-un caso tipo.

-una explicación concisa no formal.

-estudio de casos. En cada caso se describirá a su vez:

a) un texto ambiguo capturado de un dato real.

b) las preguntas orientativas de las ambigüedades generadas.

c) URL de donde se extrajo el texto.

d) identificador del HBE que la contiene.

e) descriptores del HBE.

f) E_{CI} donde se contuvo esa porción de texto.

g) E_{CE} que representa directa o indirectamente a la E_{CI}. Se presenta como una columna de radicales de palabras, con las palabras resaltadas extraídas del contexto de ese párrafo.

1. Ambigüedad léxica

Caso tipo:

Luis dejó el periódico en el banco. Cuando estaba en el banco, abrió el libro / fue a la ventanilla.

Definición:

Ambigüedad dada por el significado de una palabra. Debe diferenciarse entre lo que es léxico, (lenguaje común) de lo que es terminológico (lenguaje de especialidad de una disciplina). Para la parte léxica hay diccionarios lexicológicos desarrollados sobre la base de propiedades que permiten deducir a cuál ocurrencia se refiere.

Estudio de casos:

Caso 1:"Ha despertado las más inimaginables pasiones en los hombres."

Pregunta: ¿en los hombres como género o en los hombres como especie?

URL: es.wikipedia.org wiki Orchidaceae

HBE: HBE32

Descriptores:

Tabla XXVII.Descriptores ambigüedad léxica 1

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
ha	orquideas	verbo	otro	contenido	2	1	null	0	no	no	no	ha
despertado	orquideas	verbo	otro	contenido	10	4	null	0	no	no	no	despert
las	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	las
más	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	mas
inimaginables	orquideas	otro	otro	contenido	13	3	s	3	no	no	no	inimagin
pasiones	orquideas	sustantivo	otro	contenido	8	3	s	1	no	no	no	pasion
en	orquideas	otro	otro	contenido	2	1	en	0	no	no	no	en
los	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	los
hombres	orquideas	sustantivo	otro	contenido	7	2	s	0	no	no	no	hombr

Fig. 27. E_{ci} ambigüedad léxica 1

```
<0>ha       despertado[despertado(verbo.despert.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>despertado       más[más(otro.mas.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>más      inimaginables[inimaginables(otro.inimagin.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>inimaginables      pasiones[pasiones(sustantivo.pasion.'es.wikipedia.org wiki Orchidaceae.txt")]+>
<8>pasiones       null[null(null.null."null")]->
<0>null     hombres[hombres(sustantivo.hombr.'es.wikipedia.org wiki Orchidaceae.txt")]->
```

Fig. 28. E_{ce} ambigüedad léxica 1

Se obtiene un EBH que, de toda la frase, detecta dos verbos y dos sustantivos. Al generarse la E_{ci} con el algoritmo interno de WIH lo único que interesa es que ***pasiones*** resulta ser una palabra indicativa y por lo tanto deberá pasar a ser parte de la E_{ce} una vez que se la reduzca a su raíz o STEM. De esta manera, toda la frase queda reducida al stem ***pasion***, denotando que se habla en algún sentido de la pasión en relación al tema orquídeas (que es el tema en proceso). Puede apreciarse que la ambigüedad queda oculta dentro del texto original y no se propaga hacia los niveles de mayor abstracción. Quienes lean la E_{ci} o el texto original chocarán con la ambigüedad y deberán resolverla por el contexto.

Caso 2: "se conocen plantas recolectadas a mediados del siglo pasado que todavía están creciendo y floreciendo saludables en muchas colecciones"

Pregunta: ¿colección de plantas por parte de coleccionistas o colección como grupo de plantas en general?

URL: es.wikipedia.org wiki Orchidaceae

HBE: HBE32

Descriptores:

Tabla XXVIII. Descriptores ambigüedad léxica 2

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
se	orquideas	verbo	otro	contenido	2	1	se	0	no	no	no	se
conocen	orquideas	verbo	otro	contenido	7	3	en	0	no	no	no	conoc
plantas	orquideas	sustantivo	otro	contenido	7	2	s	0	no	no	no	plant
recolectadas	orquideas	verbo	otro	contenido	12	5	s	0	no	no	no	recolect
a	orquideas	otro	otro	contenido	1	1	null	0	no	no	si	a
mediados	orquideas	otro	otro	contenido	8	3	s	1	no	no	no	medi
del	orquideas	otro	otro	contenido	3	1	null	0	no	no	no	del

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
siglo	orquideas	sustantivo	otro	contenido	5	1	null	1	no	no	no	sigl
pasado	orquideas	verbo	otro	contenido	6	3	null	0	no	no	no	pas
que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
todavía	orquideas	otro	otro	contenido	7	3	null	1	no	no	no	todav
están	orquideas	verbo	otro	contenido	5	2	null	0	no	no	no	estan
creciendo	orquideas	verbo	otro	contenido	9	3	null	1	no	no	no	crec
y	orquideas	otro	otro	contenido	1	0	null	0	no	no	no	y
floreciendo	orquideas	verbo	otro	contenido	11	4	null	1	no	no	no	florec
saludables	orquideas	otro	otro	contenido	10	3	s	1	no	no	no	salud
en	orquideas	otro	otro	contenido	2	1	en	0	no	no	no	en
muchas	orquideas	otro	otro	contenido	6	1	s	1	no	no	no	much
colecciones	orquideas	sustantivo	otro	contenido	11	4	s	1	no	no	no	coleccion

Fig. 29. E_{ci} ambigüedad léxica 2

```
<0>null     conocen[conocen(verbo.conoc.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>conocen          plantas[plantas(sustantivo.plant.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>plantas          recolectadas[recolectadas(verbo.recolect.'es.wikipedia.org wiki Orchidaceae.txt")]-
>
<13>recolectadas    null[null(null.null."null")]+>
<0>null     mediados[mediados(otro.medi.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>mediados         del[del(otro.del.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>del      siglo[siglo(sustantivo.sigl."es.wikipedia.org wiki Orchidaceae.txt")]+>
<8>siglo    null[null(null.null."null")]->
<0>null     muchas[muchas(otro.much.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>muchas           colecciones[colecciones(sustantivo.coleccion.'es.wikipedia.org wiki
Orchidaceae.txt")]->
<0>siglo    pasado[pasado(verbo.pas.'es.wikipedia.org wiki Orchidaceae.txt")]->
<20>pasado          null[null(null.null."null")]->
<0>null     todavía[todavía(otro.todav.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>todavía          están[están(verbo.estan.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>están    creciendo[creciendo(verbo.crec.'es.wikipedia.org wiki Orchidaceae.txt")]->
<4>creciendo        null[null(null.null."null")]->
<0>null     floreciendo[floreciendo(verbo.florec.'es.wikipedia.org wiki Orchidaceae.txt")]->
<0>floreciendo      saludables[saludables(otro.salud.'es.wikipedia.org wiki Orchidaceae.txt")]->
```

Fig. 30. E_{ce} ambigüedad léxica 2

Se obtiene un HBE que tiene varias palabras. El E_{ci} que genera destaca dos cosas: *siglo* y *null*. Pero null es indicador de que lo relevante es el lugar de la frase y no una

palabra en especial. Por ello se promueve sólo la raíz de la palabra siglo al E_{ce} correspondiente, indicando que la frase refiere a algo relativo a los siglos, sin importarle qué. Nuevamente la ambigüedad queda oculta en el texto y no interfiere en su representación a niveles superiores.

Caso 3:"Las orquídeas son realmente las flores de lo superlativo "

Pregunta: ¿superlativo como algo grande?, ¿de suma calidad?

URL: es.wikipedia.org wiki Orchidaceae

HBE: HBE32

Descriptores:

Tabla XXIX. Descriptores ambigüedad léxica 3

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Las	orquideas	otro	ninguna	contenido	3	1	s	0	si	no	no	Las
Orquídeas	orquideas	sustantivo	otro	contenido	9	3	s	2	no	no	no	orquid
Son	orquideas	verbo	otro	contenido	3	1	on	0	no	no	no	son
Realmente	orquideas	otro	otro	contenido	9	4	null	0	no	no	no	realment
Las	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	las
Flores	orquideas	sustantivo	otro	contenido	6	2	s	0	no	no	no	flor
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Lo	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	lo
Superlativo	orquideas	otro	otro	contenido	11	3	null	2	no	no	no	superl

Fig. 31. E_{ci} ambigüedad léxica 3

```
<0>hbe_00 orquídeas[orquídeas(sustantivo.orquid.")es.wikipedia.org wiki Orchidaceae.txt")]->
<0>orquídeas son[son(verbo.son.")es.wikipedia.org wiki Orchidaceae.txt")]->
<0>son realmente[realmente(otro.realment.")es.wikipedia.org wiki Orchidaceae.txt")]->
<0>realmente flores[flores(sustantivo.flor.")es.wikipedia.org wiki Orchidaceae.txt")]->
<11>flores null[null(null.null."null")]->
<0>null lo[lo(otro.lo.")es.wikipedia.org wiki Orchidaceae.txt")]->
<0>lo superlativo[superlativo(otro.superl.")es.wikipedia.org wiki Orchidaceae.txt")]->
```

Fig. 32. E_{ce} ambigüedad léxica 3

```
«...
Orchidacea
organ
orquid
palabr
pasion
pedicel ...»
```

Se obtiene un HBE que deriva en una E_{ci} sin palabras relevantes por el contexto. A pesar de ello, el algoritmo interno de WIH resalta la palabra orquídea por ser el

primer sustantivo que figura en la frase. De esta manera se promueve en la E_{ce} la correspondiente raíz "orquid". La ambigüedad trasciende hasta el nivel de una E_{ci} pero toda la frase queda representada por la palabra orquídea. Se puede pensar que la frase habla en general de las orquídeas pero no vierte un concepto que valga para ser promovido a nivel E_{ce}.

2. Ambigüedad sintáctica

Caso tipo:

El vendedor de periódicos del barrio. Luis vio al niño con los prismáticos.

Definición:

Deriva de la estructura sintáctica y está dada por relación lógica entre proposiciones subordinadas por anidamiento (nesting) o encastramiento (embeding).

Estudio de casos:

Caso 1:" ¿Sabías que la orquídea Brassavola Digbyana es flor nacional de Honduras?"

Pregunta: ¿es de Honduras? ¿Es la flor representante de Honduras?

URL: 63.173.68.43 sites 7dias content.cfm id 276 PageName Insolit

HBE: HBE01

Descriptores:

Tabla XXX. Descriptores ambigüedad sintáctica 1

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Sabías	orquideas	otro	ninguna	contenido	6	2	s	1	si	no	no	Sab
Que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	la
Orquídea	orquideas	sustantivo	otro	contenido	8	3	null	2	no	no	no	orquide
Brassavola	orquideas	sustantivo	otro	contenido	10	4	null	0	si	no	no	Brassavol
Digbyana	orquideas	sustantivo	otro	contenido	8	2	null	1	si	no	no	Digbyan
Es	orquideas	verbo	otro	contenido	2	1	s	0	no	no	no	es
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	la
Flor	orquideas	sustantivo	otro	contenido	4	1	or	0	no	no	no	flor
Nacional	orquideas	otro	otro	contenido	8	3	null	1	no	no	no	nacional
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Honduras	orquideas	sustantivo	otro	contenido	8	2	s	1	si	no	no	Hondur

Fig. 33. E_{ci} ambigüedad sintáctica 1

Fig. 34. E_{ce} ambigüedad sintáctica 1

Se obtiene una E_{ci} donde el algoritmo interno no aprecia palabras destacables, pero termina promoviendo *Sabías* como representativa del párrafo. Puede considerarse que el párrafo no introduce ningún concepto que merezca representarse al nivel de E_{ce}, pero está representado como conjunto en esta palabra que indica que se presenta un comentario referencial.

Caso 2:" La parte de la flor que produce el polen."

Pregunta: ¿La flor produce el polen y esta es la parte de esa flor? ¿Dentro de la flor esta parte produce el polen?

URL: orquidea.blogia.com temas -que-es-una-orquidea-.php

HBE: HBE36

Descriptores:

Tabla XXXI. Descriptores ambigüedad sintáctica 2

ID	TEMA	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
La	orquideas	otro	ninguna	contenido	2	1	null	0	si	no	no	La
Parte	orquideas	sustantivo	otro	contenido	5	2	null	0	no	no	no	part
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	de

La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Flor	orquideas	sustantivo	otro	contenido	4	1	or	0	no	no	no	flor
Que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
Produce	orquideas	verbo	otro	contenido	7	2	null	1	no	no	no	produc
El	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	el
Polen	orquideas	sustantivo	otro	contenido	5	2	en	0	no	no	no	pol

Fig. 35. E_{ci} ambigüedad sintáctica 2

Si se contempla la E_{ci} correspondiente se puede ver que WIH considera que no hay ninguna palabra especial que merezca ser promovida a nivel E_{ce}. Sólo detecta un punto de la frase que es especial donde se describen partes de la flor. Desde el texto original, es bastante aparente por qué esta parte del texto no llega a promoverse. A continuación se transcribe parte de la página que contiene a la frase en cuestión:

«…

* El fruto es una cápsula que contiene grandes cantidades de semillas muy pequeñas.

* Las flores de muchas especies giran 180º antes de abrirse, para exponer el labelo a los polinizadores. Este evento se conoce como resupinación.

VOCABULARIO

ANTERA: La parte de la flor que produce el polen

COLUMNA: Estructura central de la flor de una orquídea constituída por la fusión de los órganos femeninos con los masculinos.

CORMO: …»

Puede apreciarse que, como parte de un texto mucho más amplio se sucede una sección de vocabulario que describe varias partes de una orquídea. El algoritmo interno de WIH ha considerado que esta sección no es representativa de lo que se transmite en el texto y por ello no ha promovido estas definiciones. En caso de que el usuario desee navegarlas, deberá recurrir al nivel de las E_{ci} donde se observa que están escuetamente indicadas. Para mayor detalle, se deberá recurrir al texto original.

De todo esto se deduce que nuevamente la ambigüedad queda oculta dentro del texto y no se promueve a niveles superiores.

Caso 3:" Para ello, tras la floración se corta la vara por encima de un nudo sobre la mitad de su longitud."

Pregunta: ¿Se corta la vara luego del proceso de floración o se corta la vara detrás de las flores producidas por la floración?

URL: orquidea.blogia.com

HBE: HBE35

Descriptores:

Tabla XXXII. Descriptores ambigüedad sintáctica 3

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Para	orquideas	otro	ninguna	contenido	4	2	ra	0	si	no	no	Par
Ello	orquideas	otro	otro	contenido	4	2	null	0	no	no	no	ello
Tras	orquideas	otro	otro	contenido	4	1	s	0	no	no	no	tras
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Floración	orquideas	sustantivo	otro	contenido	9	3	null	1	no	no	no	floracion
Se	orquideas	verbo	otro	contenido	2	1	se	0	no	no	si	se
Corta	orquideas	verbo	otro	contenido	5	2	null	0	no	no	no	cort
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Vara	orquideas	sustantivo	otro	contenido	4	2	ra	0	no	no	no	var
Por	orquideas	otro	otro	contenido	3	1	or	0	no	no	no	por
Encima	orquideas	otro	otro	contenido	6	2	null	1	no	no	no	encim
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Un	orquideas	otro	otro	contenido	2	0	un	1	no	no	no	un
Nudo	orquideas	sustantivo	otro	contenido	4	1	null	1	no	no	no	nud
Sobre	orquideas	otro	otro	contenido	5	2	re	0	no	no	no	sobr
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Mitad	orquideas	otro	otro	contenido	5	1	null	1	no	no	no	mit
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Su	orquideas	otro	otro	contenido	2	0	su	1	no	no	no	su
Longitud	orquideas	sustantivo	otro	contenido	8	1	null	2	no	no	no	longitud

Fig. 36. E_{ci} ambigüedad sintáctica 3

Si se contempla la E_{ci} correspondiente, se puede apreciar que no existen palabras que deban ser promovidas a nivel E_{ce}, lo que se explica mirando el texto original:

«… Un keiki es un hijuelo que la planta madre emite en la vara floral, tras la floración.

Por supuesto no siempre ocurre. Pero se puede estimular su emisión.

Para ello, tras la floración se corta la vara por encima de un nudo sobre la mitad de su longitud. Luego se retira con cuidado la pielecilla que cubre las yemas de los entrenudos, con mucho cuidado para no dañar éstos. Con ello conseguiremos que les llegue más luz.

También se puede añadir "pasta para keikis" que es una pasta especial con una hormona, benziladenina, que estimula su emisión. Una vez el keiki ha emitido raíces de unos 4 cm, se puede separar de la planta madre, aunque hay quien prefiere dejarlo más tiempo unido, a la vez que lo coloca sobre una macetita con substrato, para que desarrolle más las raíces.

Otro sistema es el que nos cuentan en la wikipedia, buscando keiki:….

Otro metodo muy utilizado en Phalaenopsis es mediante siembra de estacas.

La técnica de sembrado es la siguiente: …»

Sí se promueven a E_{ce} las palabras: madre (habla de las plantas madre y cómo se obtienen keikis), es (qué es un keiki y el proceso de obtención), sistema (diferentes métodos de obtención de nuevas plantas entre los cuales están los keikis), método (como sinónimo de sistema). Estas palabras se han subrayado en el texto original. Nuevamente la ambigüedad, queda encubierta en el texto original y se promueve al

nivel de E_{ci} donde se cuenta con mayor porción de texto como para desambiguar el sentido de la frase.

3. Ambigüedad semántica

Caso tipo:

Luis dio un pastel a los niños.

- ¿Uno para todos?
- ¿Uno para cada uno?

Definición:

Se relaciona con la estructura lógica entre proposiciones pero no subordinadas como en el caso de la ambigüedad sintáctica.

Estudio de casos:

Caso 1:" Se corta unos 3 cms por arriba y por abajo de la yema."

Pregunta: ¿se corta un tajo de 3 cm. pasando de arriba hacia debajo de la yema? ¿Se cortan 3 cm. encima y 3 debajo?

URL: orquidea.blogia.com

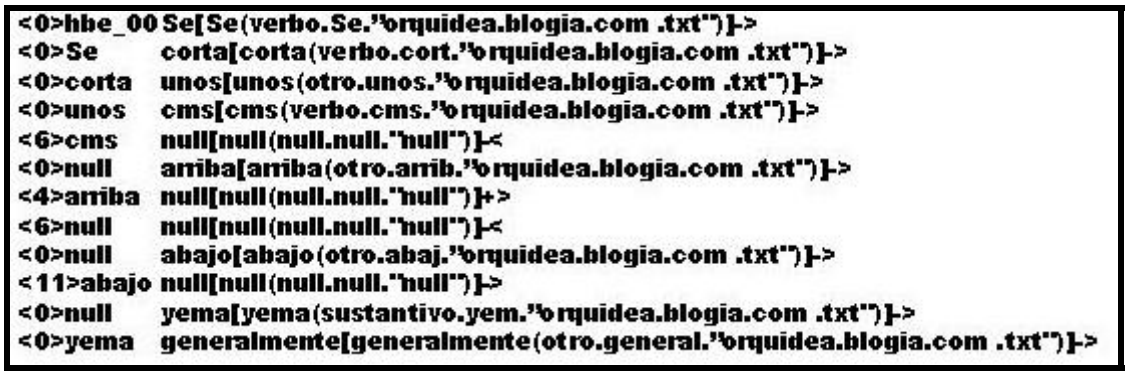
HBE: HBE35

Descriptores:

Tabla XXXIII. Descriptores ambigüedad semántica 1

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Se	orquideas	verbo	ninguna	contenido	2	1	null	0	si	no	no	Se
Corta	orquideas	verbo	otro	contenido	5	2	null	0	no	no	no	cort
Unos	orquideas	otro	otro	contenido	4	1	s	1	no	no	no	unos
Cms	orquideas	verbo	otro	contenido	3	0	s	0	no	no	no	cms
Por	orquideas	otro	otro	contenido	3	1	or	0	no	no	no	por
Arriba	orquideas	otro	otro	contenido	6	2	null	1	no	no	no	arrib
Y	orquideas	otro	otro	contenido	1	0	null	0	no	no	si	y
Por	orquideas	otro	otro	contenido	3	1	or	0	no	no	no	por
Abajo	orquideas	otro	otro	contenido	5	3	null	0	no	no	no	abaj
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Yema	orquideas	sustantivo	otro	contenido	4	2	null	0	no	no	no	yem

Fig. 37. E_{ci} ambigüedad semántica 1



WIH considera que no deben promoverse palabras indicadoras al nivel E_{ce}. La ambigüedad queda como en el caso anterior, al nivel de E_{ci}. El texto original:

«…La técnica de sembrado es la siguiente:

Se corta unos 3 cms por arriba y por abajo de la yema (generalmente la vara floral tiene 2 o 3 yemas debajo de la flor mas baja) con un cuchillo o navaja bien afilado para no dañar en exceso los tejidos. …»

Se promueve sólo la palabra sembrado. Se puede apreciar que el algoritmo interno decide que el texto puede representarse con esta palabra y deja la ambigüedad de lado.

Caso 2:" Según he leído en algunas Webs, se puede acelerar agregando al agua algunas hormonas, pero yo nunca las he usado, así que no opino al respecto."

Pregunta: ¿Las hormonas las usa para sí o para otra cosa? ¿Del contexto sale que es agua de riego?

URL: orquidea.blogia.com

HBE: HBE35

Descriptores:

Tabla XXXIV. Descriptores ambigüedad semántica 2

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Según	orquideas	otro	otro	contenido	5	1	null	1	si	no	no	Segun
He	orquideas	verbo	otro	contenido	2	1	null	0	no	no	no	he

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Leído	orquideas	verbo	otro	contenido	5	2	null	1	no	no	no	leid
En	orquideas	otro	otro	contenido	2	1	en	0	no	no	no	en
Algunas	orquideas	otro	otro	contenido	7	2	s	1	no	no	no	algun
Webs	orquideas	otro	otro	contenido	4	1	s	0	no	no	no	webs
Se	orquideas	verbo	otro	contenido	2	1	se	0	no	no	si	se
Puede	orquideas	verbo	otro	contenido	5	2	null	1	no	no	no	pued
Acelerar	orquideas	verbo	otro	contenido	8	4	ar	0	no	no	no	aceler
agregando	orquideas	verbo	otro	contenido	9	4	null	0	no	no	no	agreg
Al	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	al
Agua	orquideas	sustantivo	otro	contenido	4	2	null	1	no	no	no	agu
Algunas	orquideas	otro	otro	contenido	7	2	s	1	no	no	no	algun
hormonas	orquideas	sustantivo	otro	contenido	8	3	s	0	no	no	no	hormon
Pero	orquideas	otro	otro	contenido	4	2	ro	0	no	no	no	per
Yo	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	yo
Nunca	orquideas	otro	otro	contenido	5	1	null	1	no	no	no	nunc
Las	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	las
He	orquideas	verbo	otro	contenido	2	1	null	0	no	no	no	he
Usado	orquideas	verbo	otro	contenido	5	2	null	1	no	no	no	usad
Así	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	asi
Que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
No	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	no
Opino	orquideas	verbo	otro	contenido	5	2	null	1	no	no	no	opin
Al	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	al
Respecto	orquideas	otro	otro	contenido	8	3	null	0	no	no	no	respect

Fig. 38. E_{ci} ambigüedad semántica 2

Si bien no se promueven palabras se detecta contextualmente que hay una opinión y que se habla de otras Webs, lo que se denota al detectar con indicadores los null que siguen. Esto, como en casos anteriores, deja al nivel de E_{ci} la ambigüedad. Observando el texto original salta a la vista justamente este hecho:

«… Pasadas 2 semanas debemos de abrir el recipiente para ver si se ve algún crecimiento en la yema. Si lo hicimos bien, se debe de observar un incremento de tamaño de la "ramita". Si lo requiere, se vuelve a pulverizar y se tapa de nuevo para esperar hasta que pasen las 10 semanas donde debemos ver una plántula de unos 5 o 6 cms. Podemos ir abriendo cada 2 semanas para revisar que tenga suficiente humedad, pero sin exceder la dosis.

Este proceso es algo lento. Según he leído en algunas Webs, se puede acelerar agregando al agua algunas hormonas, pero yo nunca las he usado, así que no opino al respecto. Otro método un poco mas rápido, es hacer esto mismo pero In-Vitro, el detalle es la esterilización de los materiales (lo demás es casi lo mismo, y he estado planeando hacer algunos experimentos más adelante).

Pues esto es lo que he hecho. Le he cortado las dos varas florales a la phal y de ahí he sacado 6 yemas. A cada yema le he quitado la especie de piel que la cubre en forma de triangulo para dejar al descubierto la yema con unas pinzas. …»

El párrafo habla de pasos para el crecimiento de una nueva planta, y aporta una información colateralmente a la que el mismo autor no le da mayor importancia.

Caso 3:" A cada yema le he quitado la especie de piel que la cubre en forma de triangulo para dejar al descubierto la yema con unas pinzas."

Pregunta: ¿La piel tiene forma de triángulo o la sacó con esa forma?

URL: orquidea.blogia.com

HBE: HBE35

Descriptores:

Tabla XXXV. Descriptores ambigüedad semántica 3

ID	tema	TP	PAT	TPG	LP	CVF	EM	CVD	EM	RE	ET	STEM
A	orquideas	otro	otro	contenido	1	1	null	0	si	no	si	A
Cada	orquideas	otro	otro	contenido	4	2	null	0	no	no	no	cad
Yema	orquideas	sustantivo	otro	contenido	4	2	null	0	no	no	no	yem
Le	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	le
He	orquideas	verbo	otro	contenido	2	1	null	0	no	no	no	he
Quitado	orquideas	verbo	otro	contenido	7	2	null	2	no	no	no	quit
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Especie	orquideas	otro	otro	contenido	7	3	null	1	no	no	si	especi
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Piel	orquideas	sustantivo	otro	contenido	4	1	null	1	no	no	no	piel
Que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Cubre	orquideas	verbo	otro	contenido	5	1	re	1	no	no	no	cubr
En	orquideas	otro	otro	contenido	2	1	en	0	no	no	no	en
Forma	orquideas	otro	otro	contenido	5	2	null	0	no	no	si	form
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Triangulo	orquideas	otro	otro	contenido	9	2	null	2	no	no	no	triangul
Para	orquideas	otro	otro	contenido	4	2	ra	0	no	no	no	par
Dejar	orquideas	verbo	otro	contenido	5	2	ar	0	no	no	no	dej
Al	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	al
descubierto	orquideas	otro	otro	contenido	11	3	null	2	no	no	no	descubiert
La	orquideas	otro	otro	contenido	2	1	null	0	no	no	si	la
Yema	orquideas	sustantivo	otro	contenido	4	2	null	0	no	no	no	yem
Con	orquideas	otro	otro	contenido	3	1	on	0	no	no	no	con

ID	tema	TP	PAT	TPG	LP	CVF	EM	CVD	EM	RE	ET	STEM
Unas	orquideas	otro	otro	contenido	4	1	s	1	no	no	no	unas
Pinzas	orquideas	sustantivo	otro	contenido	6	1	s	1	no	no	no	pinz

Fig. 39. E_{ci} ambigüedad semántica 3

```
<0>hbe_00 A[A(otro.A.'orquidea.blogia.com .txt")]->
<0>A      cada[cada(otro.cad.'orquidea.blogia.com .txt")]->
<0>cada   yema[yema(sustantivo.yem.'orquidea.blogia.com .txt")]->
<0>yema   he[he(verbo.he.'orquidea.blogia.com .txt")]->
<0>he     quitado[quitado(verbo.quit.'orquidea.blogia.com .txt")]->
<0>quitado          especie[especie(otro.especi.'orquidea.blogia.com .txt")]+>
<11>especie         null[null(null.null."null")]->
<0>null   piel[piel(sustantivo.piel.'orquidea.blogia.com .txt")]+>
<8>piel   null[null(null.null."null")]->
<0>null   forma[forma(otro.form.'orquidea.blogia.com .txt")]+>
<11>forma null[null(null.null."null")]->
<0>null   triangulo[triangulo(otro.triangul.'orquidea.blogia.com .txt")]->
<18>triangulo       null[null(null.null."null")]->
<0>null   dejar[dejar(verbo.dej.'orquidea.blogia.com .txt")]->
<14>dejar null[null(null.null."null")]+>
<0>null   descubierto[descubierto(otro.descubiert.'orquidea.blogia.com .txt")]->
<0>descubierto      yema[yema(sustantivo.yem.'orquidea.blogia.com .txt")]->
<3>yema   null[null(null.null."null")]->
<0>null   unas[unas(otro.unas.'orquidea.blogia.com .txt")]->
<0>unas   pinzas[pinzas(sustantivo.pinz.'orquidea.blogia.com .txt")]->
<20>piel  null[null(null.null."null")]->
<0>null   cubre[cubre(verbo.cubr.'orquidea.blogia.com .txt")]->
```

Fig. 40. E_{ce} ambigüedad semántica 3

```
«... especi
estudi
experient
FAMILIA
flor
form
gener
GÉNEROS
histori
Laeli
madr
mar
metod
muert
navaj
ORDEN
orquid
Orquidacea
Orquidal
piel
plant....»
```

Se promovieron las palabras especie, piel y forma, que se redujeron a su stem. Esto indica que a nivel E_{ce} se habla de esto pero no interesa el detalle del trabajo sobre la

planta. A nivel E_{ci} se muestra la frase con la ambigüedad para ser leída sin intentar resolver la misma. Una vez más queda a criterio del lector su interpretación recurriendo o no al texto de pleno.

4. Anáforas

Caso tipo:

Ella le dijo que lo pusiera debajo.

-¿Quién habló?

-¿A quién?

-¿Que pusiera qué?

-¿Debajo de dónde?

Definición:

Ambigüedad introducida por elementos textuales sin carga semántica. Dado que sintácticamente son correctas no son realmente ambigüedades.

Estudio de casos:

Caso 1:" Unas serán fértiles, otras estériles."

Pregunta: ¿quiénes serán fértiles? ¿Quiénes serán estériles?

URL: orquidea.blogia.com

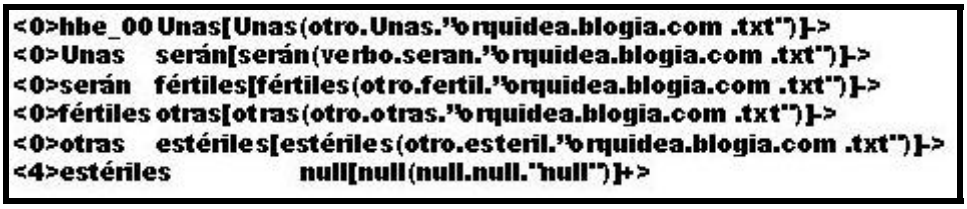
HBE: HBE35

Descriptores:

Tabla XXXVI. Descriptores anáfora 1

ÍD	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Unas	orquideas	otro	otro	contenido	4	1	s	1	si	no	no	Unas
Serán	orquideas	verbo	otro	contenido	5	2	null	0	no	no	no	seran
Fértiles	orquideas	otro	otro	contenido	8	2	s	1	no	no	no	fertil
Otras	orquideas	otro	otro	contenido	5	2	s	0	no	no	no	otras
estériles	orquideas	otro	otro	contenido	9	3	s	1	no	no	no	esteril

Fig. 41. E_{ci} anáfora 1



Si se observan los descriptores E_{ci}, se nota que, si bien se considera que se aporta algo de interés, no hay términos que merezcan promoverse a E_{ce}. El texto original dice:

«… Los insectos, atraídos por el olor o la forma de la flor son quienes propician el desarrollo de los tubos polínicos, luego de haber trasladado el polinio —masa de polen— al estigma. Solo así ocurre la fecundación y, más tarde, el paulatino crecimiento de una cápsula de miles y miles de diminutas semillas que posteriormente se diseminarán en el entorno por la acción del viento. Unas serán fértiles, otras estériles. …»

Se resaltó la palabra semilla, que es la que se promueve como indicadora del contexto. Esto es así dado que el párrafo habla de las mismas.

Caso 2:" Debajo de esta hay una pequeña "ramita"."

Pregunta: ¿debajo de quién?

URL: orquidea.blogia.com

HBE: HBE35

Descriptores:

Tabla XXXVII. Descriptores anáfora 2

ID	stem	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Debajo	orquideas	otro	otro	contenido	6	3	null	0	si	no	no	Debaj
De	orquideas	otro	otro	contenido	2	1	null	0	no	no	no	de
Esta	orquideas	otro	otro	contenido	4	2	null	0	no	no	no	esta
Hay	orquideas	verbo	otro	contenido	3	1	null	0	no	no	no	hay
Una	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	una
pequeña	orquideas	otro	otro	contenido	7	3	null	1	no	no	no	pequeñ
Ramita	orquideas	sustantivo	otro	contenido	6	2	null	1	no	si	no	ramit

Fig. 42. E_{ci} anáfora 2

Si se observa el texto original:

«… con mucho cuidado, se retira la membrana que cubre la yema, (tiene aspecto de una pequeña hoja de forma triangular). Debajo de esta hay una pequeña "ramita". Esta es la yema. …»

Se puede apreciar que la frase en cuestión es parte de una descripción en relación con un procedimiento para descubrir las yemas de la planta. El algoritmo de WIH decide entonces no promover palabras de la frase, pero sí del contexto donde se encuentra. Las palabras resaltadas son las que se definen como descriptoras del contexto: forma y yema. La ambigüedad queda a nivel de E_{ci} donde se la halla inmersa en el párrafo que le da sentido.

Caso 3:" Con ello conseguiremos que les llegue más luz."

Pregunta: ¿con qué? ¿A quién?

URL: orquidea.blogia.com

HBE: HBE35

Descriptores:

Tabla XXXVIII. Descriptores anáfora 3

ID	tema	TP	PAT	TPG	LP	CVF	TER	CVD	EM	RE	ET	STEM
Con	orquideas	otro	otro	contenido	3	1	on	0	si	no	no	Con
Ello	orquideas	otro	otro	contenido	4	2	null	0	no	no	no	ello
conseguiremos	orquideas	verbo	otro	contenido	13	4	s	2	no	no	no	consegu
Que	orquideas	otro	otro	contenido	3	1	null	1	no	no	no	que
Les	orquideas	otro	otro	contenido	3	1	s	0	no	no	si	les
Llegue	orquideas	verbo	otro	contenido	6	2	null	1	no	no	no	lleg
Más	orquideas	otro	otro	contenido	3	1	s	0	no	no	no	mas
Luz	orquideas	sustantivo	otro	contenido	3	0	null	1	no	no	no	luz

Fig. 43, E_{ci} anáfora 3

Si se observa el texto original:

«… Para ello, tras la floración, se corta la vara por encima de un nudo sobre la mitad de su longitud. Luego se retira con cuidado la pielecilla que cubre las yemas de los entrenudos, con mucho cuidado para no dañar éstos. Con ello conseguiremos que les llegue más luz. …»
En este caso no se destacaron del texto palabras para promoverse. Esto se debe a que el algoritmo no consideró que el párrafo sea destacable con relación al texto en que se está.

(ii) Estudio de caso con la estrategia de EBH ambiguo

La sentencia procesada es:
"No hay pregunta demasiado sencilla ni técnica, siempre se puede aprender algo."

Se ha seleccionado por contener una de las partículas de la tabla de ambiguos descrita en la sección Situaciones ambiguas que procesa WIH.

Fig. 44. E_{ci} del EBH ambiguo

```
<HBE_34.TXT>null     hbe_00[hbe_00/0.0(null.null."null")]->
<0>hbe_00 No[No/-1.0(otro.No."mx.groups.yahoo.com group orquidea .txt")]->
<0>No      hay[hay/0.0(otro.hay."mx.groups.yahoo.com group orquidea .txt")]->
<0>hay     pregunta[pregunta/0.0(sustantivo.pregunt."mx.groups.yahoo.com group orquidea .txt")]->
<0>pregunta          demasiado[demasiado/1.7(otro.demasi."mx.groups.yahoo.com group orquidea .txt")]-
>
<0>demasiado         sencilla[sencilla/0.0(otro.sencill."mx.groups.yahoo.com group orquidea .txt")]->
<7>sencilla          null[null/0.0(null.null."null")]->
<0>null     técnica[técnica/0.0(otro.tecnic."mx.groups.yahoo.com group orquidea .txt")]->
<0>técnica           siempre[siempre/0.0(otro.siempr."mx.groups.yahoo.com group orquidea .txt")]->
<30>siempre          null[null/0.0(null.null."null")]->
<0>null     puede[puede/0.0(verbo.pued."mx.groups.yahoo.com group orquidea .txt")]->
<0>puede  aprende[aprende/0.0(verbo.aprend."mx.groups.yahoo.com group orquidea .txt")]->
<0>aprende           algo[algo/0.3(otro.algo."mx.groups.yahoo.com group orquidea .txt")]->
```

La E_{ci} obtenida se muestra en Fig. 44. Puede apreciarse que la presencia de valores del p_o, distintos a 0.0 en las líneas asociadas a las palabras: No, demasiado y algo. Como consecuencia la ponderación característica que tendrá la sentencia no será 0.0 (frase regular) sino otro, por ejemplo $\mathbf{p_o}^{\mathbf{Ece}}$=0.1613, si la $\mathbf{f_e}$ activa usa la fórmula:

$$p_o = (p_i + p_{i-1})/2 \text{ , } \forall \text{ pal}_i \qquad \textbf{(2)}$$

Para cada palabra $\mathbf{pal_i}$ dentro de la sentencia actualmente en proceso. Este valor representará a toda palabra que sea promovida a nivel E_{ce}.

5.4. Práctica de la estrategia con p_o

En este apartado se mostrará la manipulación práctica simplificada que se hizo de la estrategia con p_o, una encuesta para validar resultados. La estrategia sirve para caracterizar sentencias y pretende ser una contribución para la determinación de palabras que representen un texto abultado. El texto completo del que se extrajo la anterior E_{ci} (ver Apéndice A: texto de EBH35) fue procesado con esta estrategia. Luego se filtraron los p_o que cumplen $|p_o|<0.11$ (más cercanos a 0.0) y se obtuvieron los resultados de la Tabla XXXIX.

Tabla XXXIX. p_o cercanos a 0.0

palabra base	p_o	palabras completas
estudi	0.108	estudio/ada
subespeci	0.108	subespecies
DIVISIÓN	0.000	division
especi	0.000	especie
FAMILIA	0.000	familia
gener	0.000	género/s
ORDEN	0.000	orden
Orquidacea	0.000	Orquidacea
Orquidal	0.000	orquidales
Planta	0.000	Planta
REINO	0.000	REINO
seccion	0.000	seccion
semill	0.000	semilla/s
distribu	0.000	distribucion
varied	0.000	vaqriedad/es

Por otra parte, un resumen muy escueto del contenido, utilizando las palabras originales del texto podría resultar el siguiente par de ítems:

1.-**estudio** de germinación de **planta Orquidácea** desde **semilla**.

2.-estudio de **especies**, **subespecies**, **distribución** de los **orquidales** en México, **variedades**, clasificación científica (**Reino**, **División**, **Género**, **Orden**, **Familia**, Subfamilia).

Obsérvese que las palabras resaltadas son prácticamente todas las de la tabla. Continuando con el caso en estudio, puede decirse que el texto tiene como datos menores:

3.-habla de una experiencia de una persona que pretende germinar en **envases** de aceitunas dentro de una **caja**, unas **orquídeas** para producir keikis

4.-se hace un seguimiento **semanal** del **crecimiento**.

5.-relata que en México hay varias especies.

6.-Describe la **distribución** geográfica, nombres y usos.

7.-También describe las especies. Describe sus amenazas y lista especies en riesgo de extinción

La Tabla XL muestra los valores p_o más lejanos a 0.0:

Tabla XL. p_o lejanos a 0.0

idPalabla	p_o	palabra
ser	0.998	ser
orquid	0.992	orquídeas
crecimient	0.973	crecimiento
tamañ	0.973	tamaño
distribu	0.938	distribuidas/distribución
potencial	0.938	potencial
seman	0.927	semana/semanas
caj	0.857	caja
envas	0.857	envases

Si bien es bastante obvia la relación conceptual con las palabras extraídas, se realizó una encuesta con 35 personas voluntarias a fin de validar este resumen contra la opinión de algunos lectores, constatar si existe un nivel mínimo de adecuación para el texto y que estas palabras son efectivamente representativas. Los voluntarios eran personas de ambos sexos, con edad entre 22 y 56 años, que usan Internet menos de 10 horas semanales.

A cada voluntario se les presentó el texto del Apéndice A y se les pidió que escribieran de 2 a 10 palabras pertenecientes al texto que usarían para representar al mismo en una consulta en un buscador. También se les pidió que listen los tópicos de los que habla el texto. En el Apéndice C: relevancia po se transcriben el formulario y resultados obtenidos, de los cuales es interesante observar de la Fig. 45 que:

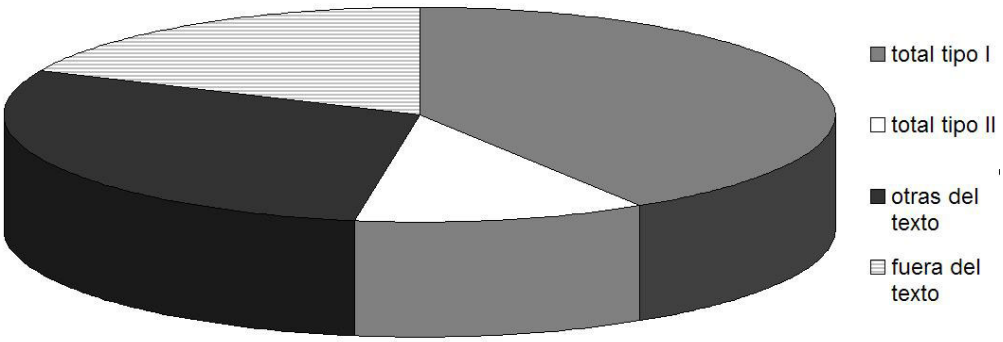
-Las palabras detectadas con p_o cercano a 0.0 son un superset de las palabras más comúnmente preferidas por las personas (ver ***total tipo I***).

-Existe una proporción de personas que, aún cuando se les requiere que sólo empleen palabras del texto, han recurrido a otras palabras que no figuran en él (ver ***fuera del texto***).

-Otras personas han seleccionado palabras describiendo detalles diversos que consideraron interesante (ver ***otras del texto***).

-las palabras con p_o lejano a 0.0 (ver ***total tipo II***) y las cercanas a 0.0 juntas superan ampliamente al uso de otras palabras del texto.

Fig. 45. Tortas por tipo de palabras



Respecto a la distribución temática, se puede apreciar desde los números originales (ver apéndice) que:

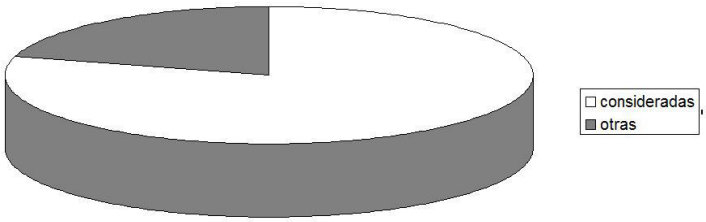
-Los temas más frecuentes son los representados por las frases armadas con las palabras cercanas al p_o =0.0.

-Existe una buena porción de los temas contestados que no resultaban acorde al texto (por ejemplo hubo encuestados que colocaron como tema el nombre de una persona mencionada en el texto, o bien algún elemento trivial).

-La mayoría de los tópicos respondidos fueron extraíbles desde las palabras con p_o =0.0 (ver la Fig. 46).

-Un pequeño porcentaje de los tópicos no fueron extraíbles de las palabras especialmente seleccionadas. En este grupo se consideraron los tópicos contestados erróneamente. (Ver ***área otros*** en la Fig. 46.

Fig. 46. Tortas por temas



Por último, obsérvese que los valores de p_o no están sesgados por otro factor que no sean las partículas opuestas o ambiguas propuestas en 3.2.2. Tratamiento de EBH ambiguos y ambigüedades y detectadas dentro de cada frase del E_{ci}. Para comprobar ésto puede observarse que en la gráfica con los promedios del valor de p_o (Fig. 48) para los distintos tipos de palabras, es prácticamente el mismo. Esta similaridad es notoria debido a que no parece sesgarse por la categoría léxica (la distribución en el gráfico con la distribución de cada tipo en el caso de estudio corresponde al de la Fig. 47).

Fig. 47. Distribución de palabras en el caso de estudio

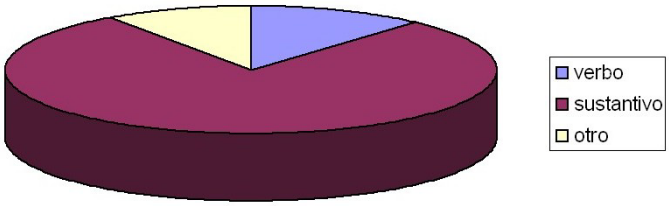
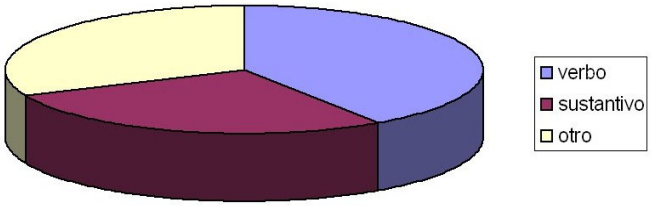


Fig. 48. Promedio de p_o según el tipo de palabra.



Capítulo 6. Sensibilidad y capacidad de adaptación

En este capítulo se estudian los aspectos de WIH relacionados con la sensibilidad a cambios de parámetros de funcionamiento y con la adaptación a distintos contextos de funcionamiento del propio sistema.

6.1. Manipulación de contenidos

Parte esencial del funcionamiento pasa por la manipulación de contenidos. Es interesante estudiar si la relación de los contenidos textuales originales respecto a las estructuras generadas permite una reducción de contenidos adecuada como para transmitir la esencia de los contenidos con la mínima información. Esto se describe en las siguientes secciones.

(i) Tasa de reducción de palabras en E_{ci}

Si se considera la proporción entre la cantidad de palabras E_{ci} respecto al texto original, puede apreciarse una reducción marcada. Se tomó una muestra de 50 páginas Web procesadas con WIH y se contaron las palabras contenidas luego del preprocesamiento y se comparó con la cantidad dentro de la E_{ci}. La proporción promedio es 1668 palabras aunque el desvío estándar es 1337 (ver Tabla XLI).

Tabla XLI. Reducción de palabras E_{ci}

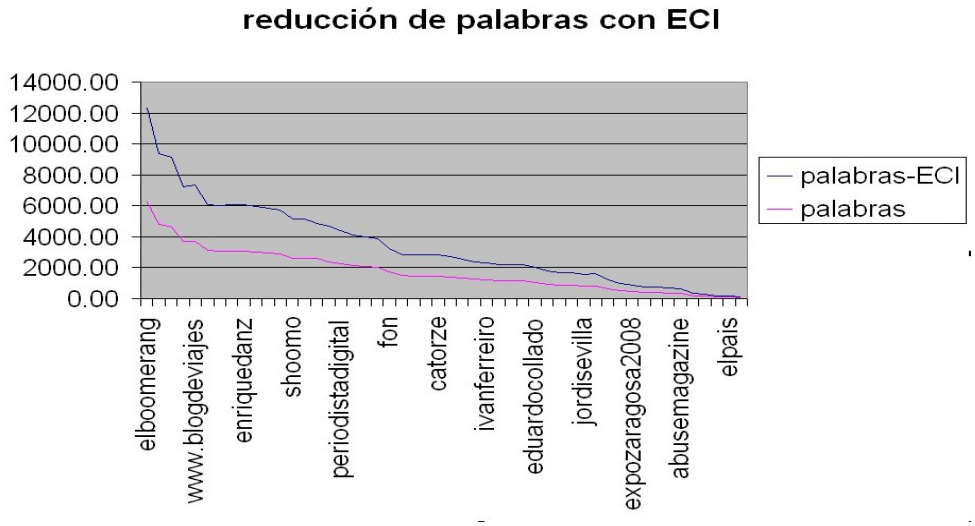
palabras ECI	WEB	palabras-ECI	proporcionECI
939.00	85antoniofumero	854.00	9.052183
333.00	37abusemagazine	296.00	11.11111
355.00	21dejaboo	334.00	5.915493
4628.00	142despegar	4486.00	3.06828
1712.00	185fon	1527.00	10.80607
837.00	101jordisevilla	736.00	12.06691
3096.00	107sedic	2989.00	3.456072
2615.00	97shoomo	2518.00	3.709369
59.00	5wsmaldone	54.00	8.474576
6236.00	125elboomerang	6111.00	2.00449

palabras	ECI	WEB	palabras-ECI	proporcionECI
87.00	18	elpais	69.00	20.68966
2269.00	142	periodistadigital vocacion	2127.00	6.258264
1163.00	149	periodistadigital alejovidal	1014.00	12.81169
1045.00	62	eduardocollado	983.00	5.933014
2962.00	68	elcuadernodepepeblanco	2894.00	2.295746
633.00	62	eventoblog	571.00	9.794629
2369.00	71	filmica	2298.00	2.997045
406.00	52	internetblog	354.00	12.80788
1228.00	165	ivanferreiro	1063.00	13.43648
2054.00	129	jamillan	1925.00	6.280428
495.00	11	jamillan	484.00	2.222222
1359.00	152	jorgecornell	1207.00	11.18469
170.00	66	macanudoliniers	104.00	38.82353
1281.00	182	octaviorojas	1099.00	14.20765
2028.00	191	share	1837.00	9.418146
3074.00	138	weblogs clarin	2936.00	4.489265
1467.00	97	www.alasbarricadas	1370.00	6.612134
3064.00	71	arcadi	2993.00	2.317232
3015.00	93	www.blogalaxia	2922.00	3.084577
3712.00	93	www.blogdeviajes	3619.00	2.505388
2583.00	329	briefblog	2254.00	12.73713
1425.00	44	catorze	1381.00	3.087719
1179.00	156	eblog	1023.00	13.23155
114.00	37	eduardoharotec	77.00	32.45614
1447.00	33	e-global	1414.00	2.280581
406.00	32	eblogsalmon	374.00	7.881773
2115.00	107	elmundo	2008.00	5.059102
1166.00	134	emezeta	1032.00	11.49228
3064.00	71	enriquedanz	2993.00	2.317232
439.00	31	expozaragosa2008	408.00	7.061503
1482.00	111	frenaelcambioclimatico	1371.00	7.489879
2902.00	95	javiermarias	2807.00	3.273604
4827.00	257	libertaddigital	4570.00	5.324218
1392.00	88	luchadealmohadas	1304.00	6.321839
867.00	36	lucia-etxebarria	831.00	4.152249
3726.00	212	mafius	3514.00	5.689748
2614.00	76	nachovigalondo	2538.00	2.907422

palabras	ECI	WEB	palabras-ECI	proporcionECI
869.00		36pablomanini	833.00	4.142693
836.00		70radiocable	766.00	8.373206
175.00		15x-flash	160.00	8.571429
1766.38	97.74		1668.64	8.23
1370.80	66.09		1337.85	7.05

Si se observa la relación oculta entre la cantidad de palabras procesadas y la cantidad configurada en las E_{ci} se refleja una proporcionalidad razonable. La Fig. 49 muestra un diagrama ordenado por cantidad de palabras procesadas, donde la curva inferior representa la cantidad de palabras procesadas y recibidas como EBH. La línea superior se construyó por encima con la cantidad de palabras representadas dentro de las E_{ci} de esas mismas páginas Web. Puede apreciarse que a medida que desciende la cantidad de palabras originales, también desciende la cantidad de palabras en las E_{ci}. El comportamiento es razonable si se considera que las E_{ci} funcionan como resumen extraído sobre las correspondientes E_{ce}, por lo tanto es de esperar que la reducción sea proporcional pero no dramática.

Fig. 49. Reducción del número de palabras al construir una E_{ci}.



(ii) Tasa de reducción de palabras en E_{ce}

Si se considera ahora la estructura E_{ce} como un índice por keywords extraídos de las correspondientes E_{ci}, la proporción de términos entre ambas debiera seguir algún criterio similar al hallado para las E_{ce}. La tabla Tabla XLII muestra los resultados obtenidos para la muestra: existe una nueva reducción en la cantidad, pero el promedio ahora es muy inferior (8.23 contra los 1668.64 para las E_{ci}).

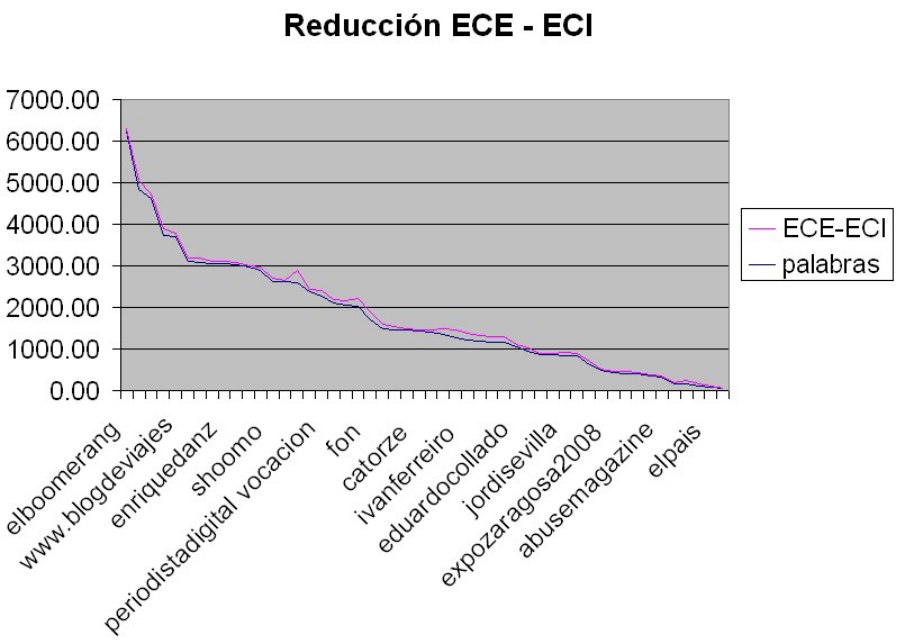
Tabla XLII. Reducción de elementos en E_{ce}

palabras ECE	WEB	ECI-ECE	palabras-ECE	proporcionECI	proporción ECE
939.00	4antoniofumero	81	935.00	9.052183	0.425985
333.00	12abusemagazine	25	321.00	11.11111	3.603604
355.00	4dejaboo	17	351.00	5.915493	1.126761
4628.00	33despegar	109	4595.00	3.06828	0.713051
1712.00	10fon	175	1702.00	10.80607	0.584112
837.00	9jordisevilla	92	828.00	12.06691	1.075269
3096.00	25sedic	82	3071.00	3.456072	0.807494
2615.00	22shoomo	75	2593.00	3.709369	0.8413
59.00	1wsmaldone	4	58.00	8.474576	1.694915
6236.00	51elboomerang	74	6185.00	2.00449	0.817832
87.00	1elpais	17	86.00	20.68966	1.149425
2269.00	1periodistadigital vocacion	141	2268.00	6.258264	0.044072
1163.00	5periodistadigital alejovidal	144	1158.00	12.81169	0.429923
1045.00	11eduardocollado	51	1034.00	5.933014	1.052632
2962.00	28elcuadernodepepeblanco	40	2934.00	2.295746	0.945307
633.00	2eventoblog	60	631.00	9.794629	0.315956
2369.00	2filmica	69	2367.00	2.997045	0.084424
406.00	3internetblog	49	403.00	12.80788	0.738916
1228.00	24ivanferreiro	141	1204.00	13.43648	1.954397
2054.00	24jamillan	105	2030.00	6.280428	1.168452
495.00	3jamillan	8	492.00	2.222222	0.606061
1359.00	16jorgecornell	136	1343.00	11.18469	1.177336
170.00	4macanudoliniers	62	166.00	38.82353	2.352941
1281.00	13octaviorojas	169	1268.00	14.20765	1.014832
2028.00	16share	175	2012.00	9.418146	0.788955
3074.00	24weblogs clarin	114	3050.00	4.489265	0.780742
1467.00	12www.alasbarricadas	85	1455.00	6.612134	0.817996
3064.00	27arcadi	44	3037.00	2.317232	0.881201
3015.00	15www.blogalaxia	78	3000.00	3.084577	0.497512

palabras ECE	WEB	ECI-ECE	palabras-ECE	proporcionECI	proporción ECE
3712.00	27www.blogdeviajes	66	3685.00	2.505388	0.727371
2583.00	23briefblog	306	2560.00	12.73713	0.890437
1425.00	17catorze	27	1408.00	3.087719	1.192982
1179.00	7eblog	149	1172.00	13.23155	0.593723
114.00	1eduardoharotec	36	113.00	32.45614	0.877193
1447.00	4e-global	29	1443.00	2.280581	0.276434
406.00	7eblogsalmon	25	399.00	7.881773	1.724138
2115.00	27elmundo	80	2088.00	5.059102	1.276596
1166.00	14emezeta	120	1152.00	11.49228	1.200686
3064.00	27enriquedanz	44	3037.00	2.317232	0.881201
439.00	2expozaragosa2008	29	437.00	7.061503	0.455581
1482.00	5frenaelcambioclimatico	106	1477.00	7.489879	0.337382
2902.00	14javiermarias	81	2888.00	3.273604	0.482426
4827.00	42libertaddigital	215	4785.00	5.324218	0.870106
1392.00	17luchadealmohadas	71	1375.00	6.321839	1.221264
867.00	9lucia-etxebarria	27	858.00	4.152249	1.038062
3726.00	41mafius	171	3685.00	5.689748	1.100376
2614.00	28nachovigalondo	48	2586.00	2.907422	1.071155
869.00	4pablomanini	32	865.00	4.142693	0.460299
836.00	5radiocable	65	831.00	8.373206	0.598086
175.00	4x-flash	11	171.00	8.571429	2.285714
1766.38	14.54	83.20	1751.84	8.23	0.96
1370.80	12.18	60.73	1360.23	7.05	0.62

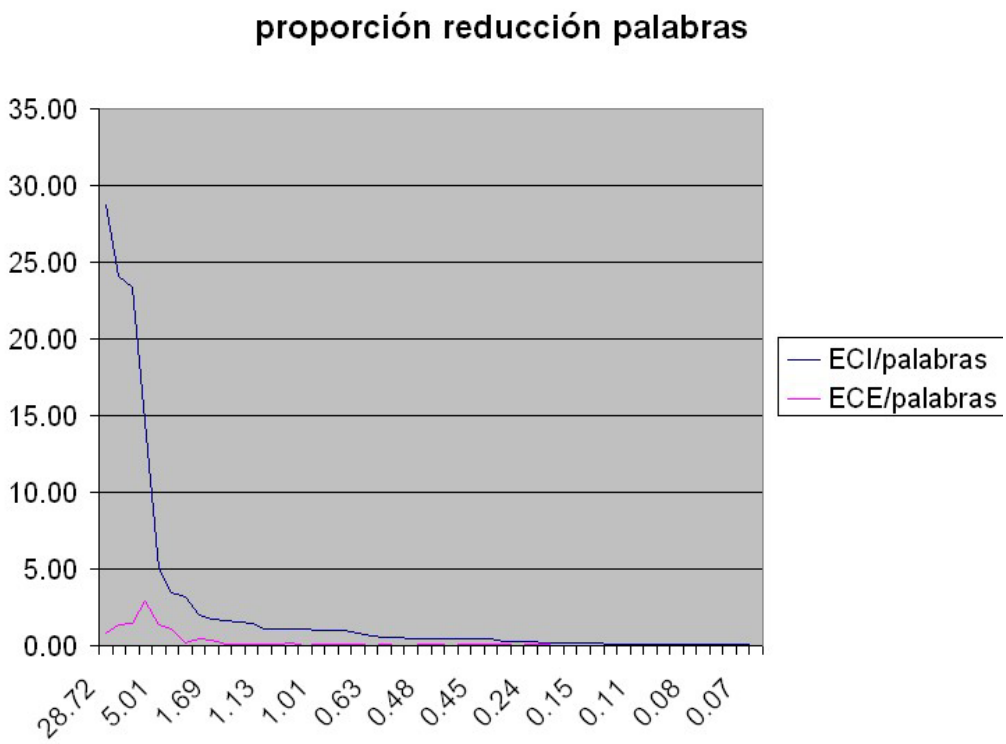
Se ordenó la tabla Tabla XLII respecto a la cantidad de palabras procesadas como EBH y se graficó la cantidad de palabras que se redujeron al extraer las E_{ce}. La curva oscura indica la cantidad de palabras procesadas como EBH en las E_{ce} y la curva por encima representa la cantidad de palabras reducida al pasar de E_{ci} a E_{ce}. Puede observarse que la cantidad de palabras reducida es ligeramente superior.

Fig. 50. Reducción del número de palabras al construir una E_{ce}.



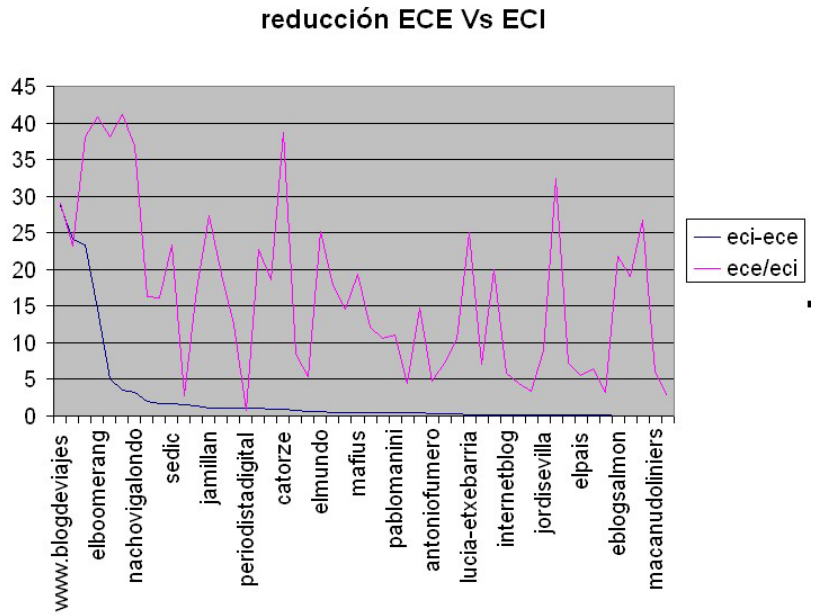
En la Fig. 51 se aprecia una comparación entre la proporción de palabras reducidas en las E_{ce} y las E_{ci}. Nuevamente se refleja la mayor reducción por parte de las E_{ci}. Es mucho más claro con esta gráfica, que el comportamiento reducidor de las E_{ce} es errático respecto al de las E_{ci}. Esto se debe a que la finalidad de las E_{ce} es transformar el contenido extenso y organizado (en un grafo direccional) en un conjunto de palabras significativas del contenido esencial (esto se desarrolló en las secciones correspondientes a 3.2. Módulo Motor de Composición (MC) y Capítulo 5. Estudio de casos y resultados).

Fig. 51. Proporción de reducción de E_{ce} y E_{ci}.



Para hacer más evidente el comportamiento de reducción del algoritmo para E_{ce}, se trazó en la Fig. 52 la proporción de reducción de E_{ce} respecto a E_{ci} pero ordenando la cantidad de palabras dentro de la respectiva E_{ci}.

Fig. 52. Proporción de reducción de E_{ce} respecto a las palabras en E_{ci}.



6.2. Estudio comparativo respecto a otras alternativas.

(i) Uso de una métrica propia.

Entre las técnicas tradicionales de indexación suelen hallarse: el etiquetado manual de contenidos con un conjunto predefinido de palabras (esto incluye una amplia gama de alternativas de ponderación y manipulación como los presentados en [28], [90], [91], [55], etc.), indexaciones automáticas con sistemas que buscan la representación estadística de ciertas palabras del texto (como en [139], [13], [8], etc.) En el caso de WIH se presenta una técnica propia basada en el uso de la métrica p_o. Tal como ya se ha mostrado, esta ponderación sirve de base para poder distinguir perfiles de narración. Esto podría ser útil a la hora de definir la Estructura Externa, puesto que podría constituir un ítem a considerar durante una búsqueda / navegación.

(ii) Indexación automática.

En el caso de WIH, la indexación consiste en la extracción automática de términos. En esto ya existen alternativas implementadas (ej. en [114], [64], etc.). Donde se distingue la actual propuesta es en la estructuración de un nivel E_{ce} y de un nivel E_{ci} para lograr una jerarquía donde cada uno de estos estratos tienen peculiaridades distintas: mientras las E_{ce} trabajan a nivel documento, las E_{ci} son un conjunto de palabras la representación de una sentencia. En muchos otros casos (Ej. SOM [69], estructuras jerárquicas [33], reducción de dimensión, etc.) es posible ver jerarquías pero típicamente consistentes de palabras. En esto WIH se podría acercar un poco a la propuesta C-Bird [150], en el sentido que se aplica a datos multimediales y textuales, en caso de que se implementen las funcionalidades correspondientes como funciones efectoras y de métrica.

(iii) Localización de información.

En buscadores tradicionales como Google, Yahoo y Hotbot, la forma de localizar la información consiste esencialmente en preguntas (denominadas queries) o en la navegación por una estructura jerárquica de directorios.

En el caso de la presente propuesta, no se ha definido la manera de recuperar información ya que esta tarea corresponderá a la Estructura Externa. Lo que sí puede adelantarse es que dicha estructura podrá implementar cualquiera de las dos alternativas mencionadas y aún una nueva: la navegación directa sobre las estructuras E_{ce} o E_{ci}. La navegación sobre las estructuras E_{ce} podrá actuar como un diccionario ordenado alfabéticamente con facilidad de acceder a las páginas correspondientes. La navegación sobre E_{ci} en cambio podrá trabajarse como un esquema de navegación de frases (sin importar la página que la contiene), con facilidad de acceder a las páginas correspondientes.

(iv) Extracción de resúmenes automáticos.

Gracias al uso de p_o, es posible generar un texto reducido significativamente representativo del contenido de la página correspondiente. Esto distingue esta propuesta de las actuales implementaciones de navegadores, donde se suele transcribir parte del contexto inmediato a cierta palabra buscada, como resumen orientativo de la búsqueda.
En el caso de WIH, el resumen una vez extraído conforme lo descrito en 5.4. Práctica de la estrategia con po, constará realmente de un conjunto de palabras cuya ponderación tenga determinado rango. Dado que las palabras son ponderadas desde el nivel de las E_{ci}, y se proyecta al nivel de E_{ce}, existen algunas alternativas posibles para la presentación de un "resumen" al usuario:

-Presentar al usuario el conjunto de palabras selectas promovidas a E_{ce}, en bruto. Este sería un compacto de las palabras mejor calificadas, ya que E_{ce} hace las veces de un índice por ciertos EBH.su vez promovidas.

-La representación simbólica a nivel E_{ci}, del tipo presentado en la sección 3.2. Módulo Motor de Composición (MC): Las E_{ci} con una especie de resumen no compacto de sentencias del texto original, organizado como un grafo dirigido.

-Una representación alternativa de las palabras promovidas a E_{ce}, considerando de alguna manera la representación de la ponderación relativa.

-Una representación alternativa de las E_{ci}, interpretando visual o textualmente las relaciones representadas en el grafo.

(v) Tratamiento de la información.

Algunas características del tratamiento de información de WIH fueron extraídas de la literatura, y se las combinó de manera diferenciadora:

-La estructura de las sentencias se refleja como una estructura igual que en [41], pero allí se presenta como un grafo jerárquico con cierta palabra especialmente seleccionada como centro o raíz. En WIH, la frase se refleja como una estructura bastante más libre. Algunas partes pueden ser jerárquicas parcialmente dependiendo de las EBH en proceso.

-De igual manera que en [41] cada frase merece una estructura. Pero en el caso de WIH, las estructuras pueden fusionarse o enlazarse.

-Otro aspecto similar a [41] es la capacidad de asimilar sentencias con estructuras complejas de manera análoga a las sencillas, y de generar de forma totalmente automática la estructura correspondiente.

-En [13] se trabaja con conjuntos de sustantivos para trabajarlos como índices de keywords en cambio WIH maneja conjuntos de palabras de cualquier tipo, siempre que sean calificadas en el rango de valores de p_o adecuado para ser tratada como indexadora.

-A diferencia de tratados como [70] donde se realiza análisis de constituyentes, en esta propuesta sólo se hace un filtrado y se procesa el tipo de relación del EBH actual respecto al anterior y siguiente. Esto podría tener

alguna analogía con los tratamientos contextuales con contexto de dimensión 3.

-En la propuesta WIH se realizan inferencias sólo para ciertos tipos de lexemas declarados de interés para la aplicación, a diferencia de otros tratados como [70], donde las inferencias se realizan sobre todas las relaciones funcionales.

-En tratamientos como en [8] se extraen palabras candidatas, se realiza análisis morfosintáctico y luego clustering de términos. Los términos están vinculados (linkeados) sintácticamente, en cambio en esta propuesta los términos se hallan todos simbólicamente. No existe un procedimiento de extracción sino un previo proceso de eliminación de ciertos EBH, que son los que se convierten en vinculantes simbólicos.

-En tratamientos como [8] se trabaja con frases nominales (noun phrases) como unidad de trabajo, en cambio aquí se trabaja con EBH como unidad.

-En algunas propuestas se define una semántica por uso (ej.[16]) o por asignación (ej. [151]), en cambio en esta propuesta no se diseña ningún tipo de manejo semántico.

-Existen propuestas que extraen descripciones de temas de manera automática, con resultados muy buenos (ej. En [61], se logra usando variaciones sintácticas), en el caso del prototipo de este trabajo, la extracción de la descripción se hace sobre la base de la compaginación de los EBH cuyo valor de p_o está en cierto rango correspondiente al de una descripción secundaria.

-En [114] se asume que el único que puede otorgar significados a la información procesada es el usuario y que a los sistemas automáticos sólo les cabe ofrecerle un procesamiento parcial de la información para ayudarle a recorrer el camino de manera más eficiente. En ésto es compatible con la

propuesta de WIH, aunque no se cierra la alternativa de incorporar progresivamente niveles semánticos ya existentes en la comunidad.

-En propuestas como [114] se justifica el uso de sintagmas directamente y considera que los stems no son apropiados dado que no tiene variaciones. En el caso de WIH, subsana los problemas que allí se mencionan apoyando el uso de stems con un conjunto de descriptores.

(vi) Flexibilidad

En WIH, existen varios mecanismos para proveer flexibilidad al funcionamiento del sistema:

-Al nivel de funcionamiento global: la estructura jerárquica de objetivos **O**, $\mathbf{O_{ece}}$, $\mathbf{O_{eci}}$ determinan indirectamente el comportamiento generado por el **MM** a través del **GM**, ya que el **SC** los toma como umbrales disparadores de cambios en el funcionamiento. Las f_e implementan las distintas alternativas de comportamiento y todas las tareas esenciales del sistema. De modo que existen tipos de funciones efectoras dependiendo del módulo al cual responden.

-En la extracción automática de resúmenes: La ponderación p_o [84] se presentó en 3.2.3. Justificación de la estrategia con po, y se ejemplificó en con una manipulación numérica directa simplificada en 5.4. Práctica de la estrategia con po. Allí se estipuló un umbral a partir del cual se determina si una EBH tiene ponderación *cercana* o *lejana* al valor $|p_o|$=0.0. Este valor establece indirectamente el grado de compactación de los resúmenes automáticos. No obstante ésto, y ya presentados los resultados de este tratamiento, se prevé una mayor flexibilización de la implementación con una modelización difusa de los conceptos p_o *cercano* y *lejano* a 0.0. Se pretende definir un punto de corte α de acuerdo a opción del usuario partiendo de un valor propuesto por el Sistema Controlador. Este valor, se deduciría del estado detectado con las funciones de métricas $\mathbf{f_m}$ activas. Dichas funciones se limitan a evaluar métricas categóricas según los criterios estudiados en [84], donde se lista una serie de criterios para su definición.

-Funcionamiento del SC sobre los componentes difusos: A continuación se presenta brevemente la propuesta desarrollada en [84] para regular el funcionamiento del SC

de WIH. Considerando las métricas sobre los componentes difusos como mediciones efectivas de calidad de funcionamiento de acuerdo a cierta necesidad de información definida, en un contexto dado, es posible categorizarlas:

Como aparato de requisitos: métricas directas subjetivas, se compadece con el grado de satisfacción subjetiva por cada requisito. Las métricas e indicadores de este grupo se alimentan de la interacción con un usuario final, que eventualmente pueda reflejar sus necesidades de información para que el sistema controlador dispare el proceso de adaptación conveniente.

Como aparato prescriptivo: métricas directas y objetivas, son métricas indirectas. Se relaciona con el resultado obtenido con relación al esperado. Las métricas e indicadores de este grupo estudian los resultados obtenidos por el módulo.

Como aparato descriptivo: métricas indirectas. Relacionada con la caracterización del resultado obtenido. Las métricas e indicadores de este grupo sirven para formalizar y organizar la información para el usuario.

Las métricas propuestas se presentan en la Tabla XLIII según las categorías descritas. A partir de estas métricas, se derivan los indicadores de la Tabla XLIV, con los cuales se obtiene valores apropiados para el funcionamiento del SC.

Tabla XLIII.Métricas propuestas para SC

tipo	nombre	descripción	fórmula
Métricas para el Aparato de Requisitos	Lista de Requisitos, LR	lista de conceptos calculables	$LR = \{r_i\}$
	Lista de Requisitos Ponderada	para LRP, ponderar entre 0 y 1 los requisitos según su importancia relativa	$LRP = \{\rho_i\} / \rho_i \in [0..1]$
Métricas para el Aparato prescriptivo	#CAT(Número de CATegorías):	determina la cantidad de categorías definidas para cada variable	$\#CAT = \{\#cat_i\}$

tipo	nombre	descripción	fórmula
		difusa i dentro del problema.	
	SCAT(Simetría de curva de la CATegoría)	determina la medida de simetría de la curva solución.	$SCAT = \int (y - \bar{y})^3 .dy / \sigma^3$ con: $\bar{y} = \int y.dy$
	KCAT(valor K de curva de la CATegoría):	determina la medida de achatamiento de la curva solución.	$KCAT = \int (y - \bar{y})^4 .dy / \sigma^4$ con: $\bar{y} = \int y.dy$
	RCAT(Reglas de la CATegoría):	calcula la cantidad de reglas en las que figura la categoría i como consecuente (N_{sol}) en relación con la cantidad de reglas totales (N_{tot}).	$RCAT = N_{sol} / N_{tot}$
Métricas para el Aparato Descriptivo	A%(porcentaje del Area total):	calcular la probabilidad de las categorías en la respuesta, mediante el cociente del área bajo la curva solución (A_{sol}) sobre la sumatoria de las áreas correspondientes a las funciones de pertenencia i involucradas en la solución (A^i_{tot}).	$A\% = A_{sol} / \sum_i A^i_{tot}$
	R%(porcentaje de Reglas):	calcular la proporción de	$R\% = \sum_i R^i_{sol} / R_{tot}$

tipo	nombre	descripción	fórmula
		categorías en la respuesta, mediante el cociente de la sumatoria de las categorías i involucradas en la solución (R^i_{sol}) sobre la cantidad total de categorías (R_{tot}).	
	P%(porcentaje de Puntos solución):	calcular la representatividad del punto solución respecto al total de puntos en el área respuesta, mediante la inversa del área solución (A_{sol}).	$P\% = 1/A_{sol}$

Tabla XLIV.Indicadores propuestos para SC

tipo	nombre	descripción	fórmula
elemental	i_{disp}(indicador de requisitos efectivamente dispuestos):	evalúa la heurística como aparato de requisitos.	$i_{disp} = \left(\sum_i r_i\right)/\#LR$
	i'_{disp}(indicador2 de requisitos efectivamente dispuestos)	Interpreta el grado de satisfacción del requisito original considerando las ponderaciones relativas de cada requisito. Estas	$i'_{disp} = \left(\sum_i r_i.\rho_i\right)/\#LRP$

tipo	nombre	descripción	fórmula
		ponderaciones pueden corresponder a distintos tipos de requisitos (de implementación, económicos, legales, restricciones externas, etc.)	
	i_{rep}(indicador de representatividad)	Interpreta el grado de representatividad de la solución en función del subconjunto de categorías involucradas en la respuesta a las cuales representa. Si este valor es cercano a 1 indica que la solución hallada es poco precisa en términos de categorías.	$i_{rep} = \#CAT_{sol} / \#CAT_{tot}$
	i'_{rep}(indicador2 de representatividad)	Se basa en la métrica SCAT. Cuando SCAT es cercana a 0 indica	$i'_{rep} = \begin{cases} SCAT \approx 0 \Rightarrow x_m \\ SCAT >> 0 \Rightarrow ? \end{cases}, x_m = (x_{\max} - x_{\min})/2 + x_{\min}$

tipo	nombre	descripción	fórmula
		que la solución es simétrica. En estos casos, será un mejor valor si es cercano al x_m (punto medio). En caso de que SCAT no sea cercano a 0, esto no es cierto y no se puede asegurar nada al respecto. Interpreta el grado de representatividad de la solución en función de la curva solución.	
	i''$_{rep}$(indicador3 de representatividad)	Interpreta en KCAT la representatividad de la solución en función de la curva solución. KCAT < 0 indica que la curva es achatada, con mayor dispersión de valores. Lo contrario sucede con un valor muy	$i''_{rep} = \begin{cases} KCAT \le 0 \rightarrow soluc.pobre \\ KCAT > 0 \rightarrow soluc.buena \end{cases}$

tipo	nombre	descripción	fórmula
		positivo. En consecuencia los valores positivos indican que la solución será mejor.	
	i_{reg}(indicador de reglas involucradas)	Interpreta en RCAT el grado de representatividad del subconjunto solución como conjunto de reglas. RCAT cercano a 0 indica que la solución es altamente ajustada. Lo contrario sucede con un valor positivo. Se puede determinar algún valor M_r de ajuste mínimo para considerar la respuesta como una respuesta de buena calidad.	$i_{reg} = \begin{cases} RCAT < M_r \rightarrow soluc.buena \\ RCAT \geq M_r \rightarrow soluc.pobre \end{cases}$
	i_{prop}(indicador de proporcionalidad)	Interpreta en A% el grado de representatividad del subconjunto	$i_{prop} = \begin{cases} A\% < M_r \rightarrow soluc.buena \\ A\% \geq M_r \rightarrow soluc.pobre \end{cases}$

tipo	nombre	descripción	fórmula
		solución en función del área normalizada. A% cercano a 0 indica que la solución es altamente ajustada. Lo contrario sucede con un valor positivo. Se puede determinar algún valor M_r de ajuste mínimo para considerar la respuesta como una respuesta de buena calidad.	
	i_{cat}(indicador de categorías)	Interpreta en R% el grado de representatividad del subconjunto solución en función del las categorías implicadas. R% cercano a 0 indica que la solución es altamente ajustada. Lo contrario sucede con un valor	$i_{cat} = \begin{cases} R\% < M_r \rightarrow soluc.buena \\ R\% \geq M_r \rightarrow soluc.pobre \end{cases}$

tipo	nombre	descripción	fórmula
		positivo. Se puede determinar algún valor M_r de ajuste mínimo para considerar la respuesta como una respuesta de buena calidad.	
	i_{peso}(indicador de peso relativo)	Interpreta en P% el grado de representatividad del subconjunto solución en función del las categorías implicadas. P% cercano a 0 indica que la solución es altamente ajustada. Lo contrario sucede con un valor positivo. Se puede determinar algún valor M_r de ajuste mínimo para considerar la respuesta como una respuesta de buena calidad.	$i_{cat} = \begin{cases} P\% < M_r \rightarrow soluc.buena \\ P\% \geq M_r \rightarrow soluc.pobre \end{cases}$

tipo	nombre	descripción	fórmula
	i_{prec}(indicador de precisión)	Interpreta el grado de representatividad del subconjunto solución en función del la precisión de la respuesta. Se calcula como el producto de los indicadores P% y R%; i_{prec} cercano a 0 indica que la solución es altamente ajustada. Lo contrario sucede con un valor positivo. Se puede determinar algún valor M_r de ajuste mínimo para considerar la respuesta como una respuesta de buena calidad.	$i_{prec} = \begin{cases} P\%.R\% < M_r \rightarrow soluc.buena \\ P\%.R\% \geq M_r \rightarrow soluc.pobre \end{cases}$
globales	I_{disp} (indicador global sobre indicadores i_{disp})	evalúa la completitud de la respuesta hallada. Es el cociente de los indicadores	$I_{disp} = i'_{disp} / i_{disp}$

tipo	nombre	descripción	fórmula
		i'_{disp} sobre i_{disp}. Si supera un parámetro M_r es relevante, para la resolución completa del problema.	
	I_{rel}(indicador global de relevancia)	evalúa la eficiencia global de la respuesta hallada. Es el producto de los indicadores i_{cat} con i_{peso}. Si supera un parámetro M_r es relevante.	$I_{rel} = i_{cat}.i_{peso}$
	I'_{rel}(indicador global2 de relevancia):	evalúa la bionomía global de la respuesta hallada conforme a restricciones de calidad. Es la conjunción lógica de los valores de verdad sobre las restricciones impuestas a los indicadores i_{rep}, i'_{rep} y i''_{rep}.	$I'_{rel} = (i_{rep} << 1) \wedge (i'_{rep} \cong 0) \wedge (i''_{rep} > 0)$

-Ponderaciones de las E_{ci} y E_{ce}: las valuaciones de los distintos p_o de cada palabra, el cálculo correspondiente al nivel de E_{ci} y de E_{ce}, puede regularse con los parámetros de funcionamiento de las efectoras pertenecientes al MA. La manera en que se combinan las ponderaciones, más allá de la propuesta inicial de este trabajo en (extraído de lo presentado en [84]), también es ajustable. Estas dos flexibilizaciones permiten ajustar el comportamiento ante eventuales mejoras del algoritmo, sin alterar esencialmente el sistema de base.

-Tratamiento de errores y omisiones: la sensibilidad al resumen y extracción de palabras indicadoras depende del listado de EBH ambiguos y contrarios, y el p_o que se les define. Las omisiones/errores incidirán en el proceso sólo cuando alteren la ponderación de la sentencia. En el peor de los casos, una sentencia puede ser infravalorada respecto a su peso real (ver casos presentados en 3.2. Módulo Motor de Composición (MC)).

-Presentación al usuario: si bien en la Estructura Virtual se describen alternativas y facilidades para la presentación de la información (ej. las estructuras ponderadas E_{ce} y E_{ci}, el manejo difuso de parte de las necesidades del usuario, la posibilidad de extraer resúmenes automáticos, etc.), es responsabilidad de la Estructura Externa la interacción de cualquier tipo con el usuario. De modo que cualquier alternativa (gráfica, textual o con multimedia), puede desarrollarse y conectarse a la Estructura Virtual a través de las estructuras de la Red Virtual.

6.3. Restricciones y funcionalidades mínimas

A fin de garantizar el correcto acoplamiento entre capas, la Estructura Virtual debe cumplir con ciertas condiciones mínimas. A continuación se presenta una enumeración rápida de los requisitos relevados hasta el momento para la EV:
-La estructura debe procesar texto plano, pero garantizando que toda otra información (multimedia) pueda ser extraída convenientemente (reemplazándola al menos por un EBH que describa mínimamente la existencia del objeto extraído).

-La estructura debe ser capaz de recibir meta datos aunque no sea capaz de procesarlos adecuadamente como tales. En principio es sólo cuestión de implementar (igual que en el caso anterior), las efectoras correspondientes con habilidad de transformarla en EBH, incorporándole al mismo la información semántica contenida.
-El funcionamiento de EV debe ser acoplable a otros sistemas por medio de nuevas funciones efectoras y parámetros de control para regular sus resultados.
-La EE debe proveer el tratamiento adecuado de las consultas de parte del usuario o de otra aplicación y de los operadores adecuados (ej. El valor α de corte para el filtro difuso de ponderaciones, recuperar y ordenar los documentos de respuesta). La EV no intercambiará de manera directa información, sino a través de la RV (Red Virtual) conformada.
-La EE será la única que realice actividades de rastreo de locus de información y tratamiento del nivel de profundidad que requiere el usuario. Al respecto las EI y EV no se acoplan a la tarea.
-La EE debe proveer métodos de expansión usando los descriptores de palabras introducidos en 3.1. Módulo de Traducción a Lenguaje Interno; probablemente con algorítmicas que exploten la raíz stem (como en [98], [14], etc.) y nuevas propuestas a partir del resto de los descriptores.

Capítulo 7. Conclusiones

De lo presentado en este trabajo y en los concomitantes, se puede apreciar que existe la necesidad de respetar cierta correlación entre page rank y posicionamiento para que las técnicas que se aplican sean visualizables por el usuario final.

A partir de allí, las conclusiones de este trabajo son:

- Es posible lograr autonomía de crecimiento en las capas constitutivas de WIH, dado que se han automatizado actividades que normalmente requieren la intervención humana, tales como la categorización léxica de palabras (es decir la determinación de si una palabra es sustantivo, verbo, etc.) y la determinación del nivel de relevancia de una palabra dentro de un texto.
- Se ha mostrado que es dable la adaptación automática del sistema a nuevos contenidos, a través de la implementación de un sistema controlador basado en ciertas métricas como p_o y técnicas como la lógica difusa.

Respecto al manejo WIH de la ambigüedad:

- Según Chomsky la ambigüedad, es una especie de ruido que oscurece el sentido del mensaje. Como se evaluó en los casos, salvo que este ruido sea parte del mensaje esencial, no trasciende a niveles de abstracción superiores (nivel E_{ce}) salvo cuando realmente tienen grandes posibilidades de representar al texto.

En cuanto a las contradicciones:

- Se ha probado aquí con casos concretos, que las contradicciones resultan irrelevantes para la manipulación de contenidos por parte de este sistema. Esto es así gracias al manejo de la métrica po, que no se ve afectada por las mismas y mantiene en su capacidad para extraer las palabras relevantes en forma aceptable.
- Se ha logrado independizar la administración de contenidos de la existencia previa de los documentos, desde el momento en que las estructuras generadas no toman en cuenta la cohesión lógica del documento como totalidad, sino sólo la cohesión a nivel de sentencias. A pesar de esto, el diseño propuesto permite conocer cuál fue el documento de origen en cada caso.

- Sobre la base de lo presentado hasta aquí y respetando los requisitos para la Estructura Externa, es posible dejar planteada la posibilidad de futuros desarrollos que permitan al usuario la navegación de manera transparente a aspectos técnicos.
- Queda planteado un esquema básico de purgado automático de contenidos en base al ajuste adecuado del Gestor de Métricas, el Motor de Métricas y el Sistema Controlador.
- Resulta evidente por su característica de estructura complementaria a la Web, que la misma no se verá alterada tal como se la conoce por el uso de WIH como herramienta.
- Queda planteada la flexibilidad para el desarrollo de interfaces visuales que permitan manejos alternativos de los datos generados por la estructura propuesta aquí.

Es pertinente aclarar con respecto al manejo de la gramática, que existe un recurso (según Chomsky) que de alguna manera mantiene una regularidad básica (un conjunto de reglas básicas que genera una "estructura profunda"). El procesamiento de WIH toma la gramática no como generadora sino como reflejo de algo generado por un emisor inteligente en plan de transmitir una estructura profunda. Por lo tanto intenta detectar sólo ciertas regularidades y en función de su existencia relaciona los componentes textuales en curso.

Respecto al manejo de contexto, la ambigüedad escrita se puede resolver mayormente por nuestra interpretación contextual (Bréal [11]). Dado que esta interpretación recae en el usuario, recaerá en el mismo la contextualización correcta. Este es el principio en el que WIH se basa para este tratamiento y que se ha intentado demostrar, puesto que al nivel de desarrollo en que se encuentra le es transparente la administración de estructuras semánticas como tales.

Respecto a la caracterización de frases, WIH hace una caracterización especial de las mismas. Considera ciertas EBH especialmente, y consecuentemente las pondera.

Luego proyecta la ponderación en las representaciones de las sentencias (E_{ci}s) y de los textos (E_{ce}s).

Es interesante notar que su esquema de funcionamiento le permite dar flexibilidad (sin modificar la Web actual) al tipo y forma de procesamiento de la información no sólo para casos de modificaciones sino también para ajustar la mecánica de construcción de estructuras (E_{ci} y E_{ce}) según las necesidades del conjunto a cada instante.

Capítulo 8. Trabajo futuro

Queda mucho por hacer sobre el prototipo implementado y en la hilación técnica teórica de su administración de contenidos.

Por el lado de los estudios teóricos:

Aún debe completarse el análisis de resumen automático:

-Se estudiará más profundamente la sensibilidad de la extracción automática de resúmenes respecto del punto de corte de p_o

-Se implementará y evaluará el manejo difuso de los conceptos p_o *cercano* y *lejano* a 0.0. Evaluar un punto de corte α de acuerdo a opción del usuario partiendo de un valor propuesto por el Sistema Controlador a partir del estado detectado con las funciones de métricas $\mathbf{f_m}$ activas.

En cuanto al estudio del motor MC, quedan por evaluar nuevas alternativas para **tipo de palabra anterior**, tales como artículo, preposición, pronombre personal, pronombre posesivo.

Algo similar cabe para **tipo de palabra**, donde se procesó sustantivo, verbo, y otro y deberá evaluarse la extensión a adjetivo para avanzar sobre lo expresado por Ortiz [156] y verificar el porcentaje de veces que suceden las combinaciones sencillas dentro de distintos tipos de textos.

En cuanto a la inducción por árbol de tipo de palabra debe estudiarse la posibilidad de reducción de descriptores a mínima expresión y de trabajar con inducción de otros tipos de palabras como artículo, preposición, pronombre personal, pronombre posesivo, preposición, conjunción, etc...

En la métrica p_o: aún deben profundizarse el estudio otros tipos de texto (científico y técnico, periodístico, literario, didáctico, histórico, informativo, de entretenimiento, argumentación) y perfiles.

También se estudiará p_o en relación con la cantidad de contradicciones y ambigüedades, y su distribución según el número de sentencia (en este trabajo se analizó sólo la primer sentencia).

En cuanto a los **perfiles de narración** detectados, se desarrollarán como medio para discriminar modalidades de uso del lenguaje. Desde este punto de vista es un hallazgo interesante que se profundizará para determinar la calidad presunta del documento y no de las sentencias comprendidas. También podría estudiarse el uso de estos resultados para trabajar complementariamente con un subsistema de retroalimentación (feedback) de parte del usuario y refinar el sistema de navegación en la EV y hasta eventualmente colaborar en la generación de sitios Web adaptables en su contenido.

Finalmente queda por estudiar la interacción con la Web semántica y sus herramientas (por ahora sólo se capturan meta datos en WIH, no se desarrollaron funciones efectoras apropiadas, ni se extendió el módulo TLI para su tratamiento).

Desde la implementación:

Cabe aún extender el procesamiento a meta datos, y la implementación de las efectoras de extracción de multimediales (y reemplazo por EBH descriptivos al caso).

Por supuesto falta la implementación de la Estructura Externa y con ella se elaborará alternativas de presentación visual/textual de resumen automático al usuario y el manejo de queries y browsing usando E_{ci} y E_{ce}.

Quedan pendientes también aspectos más pragmáticos del desarrollo tales como un estudio para paralelizar la implementación, pruebas de estrés, etc.

Diccionario y Nomenclatura

Corpus: Tradicionalmente, los lingüistas han definido corpus como "un conjunto de datos (auténticos) del lenguaje manifestados de forma natural y utilizables como base para la investigación lingüística" ([79]). En la actualidad, el término corpus se aplica a "un conjunto de material lingüístico que existe en forma electrónica y que puede ser procesado por una computadora con distintos fines como la investigación lingüística y la ingeniería del lenguaje". Según el Corpus Encoding Standard [153], es toda colección de datos lingüísticos, haya sido seleccionada o estructurada según un criterio de diseño o no. Según esta definición un corpus puede contener todo tipo de texto, desde la prosa pura, poemas, dramas, etc. Hasta listas de palabras, y diccionarios. El CES también pretende tratar con transcripciones de texto hablado.

EBH: Elementos Básicos Homogeneizados. Denota las estructuras básicas que recibe la Estructura Virtual desde la Estructura Interna.

E_{ce}: Estructura de Composición Externa. Estructura que contiene toda la información necesaria para vincular una E_{ci} con otras y con la página o documento Web original.

E_{ci}: Estructura de Composición Interna. Grafo orientado que contiene la representación de una frase o sentencia.

Framework: En el desarrollo de software, un framework es una estructura de soporte definida en la cual otro proyecto de software puede ser organizado y desarrollado. Típicamente, un framework puede incluir soporte de programas, bibliotecas y un lenguaje de scripting entre otros softwares para ayudar a desarrollar y unir los diferentes componentes de un proyecto.

f_e: función efectora. Funciones que contienen porciones de código que implementan alternativas de actividades dentro del sistema.

f_m: función de métrica. Funciones que realizan mediciones sobre la actividad de la Red Virtual para detectar cambios de contenidos y niveles de eficiencia en las distintas actividades (siempre respecto a ciertos objetivos).

GM: Gestor de Métricas. Módulo del prototipo WIH encargado de administrar las funciones efectoras (f_e).

IR: Information Retrieval. Término técnico en inglés para referir a la actividad de recuperación de información que se halla almacenada en Internet.

MA: Motor de Asimilación. Módulo del prototipo WIH encargado de componer estructuras E_{ce}.

MC: Motor de Composición. Módulo del prototipo WIH encargado de componer estructuras E_{ci}.

Metadato: (del griego μετα, meta, "después de "y latín datum, "lo que se da", "dato"). Literalmente "sobre datos", son datos que describen otros datos. En general, un grupo de metadatos se refiere a un grupo de datos, llamado recurso. [164].

ML: siglas para el término en inglés Markup-Language. Lenguaje de descripción que puede ser insumo en una programación. Se caracteriza por el uso de etiquetas [15].

MM: Motor de Métricas. Módulo del prototipo WIH encargado de administrar las funciones de métrica (f_m).

Morfología: (< griego *μορφ*-, *morph* ['forma'] + λογία *logía* ['tratado']) es la rama de la lingüística que estudia la estructura interna de las palabras para delimitar, definir y clasificar sus unidades, las clases de palabras a las que da lugar (morfología flexiva) y la formación de nuevas palabras (morfología léxica). La palabra 'morfología' fue introducida en el siglo XIX.

Ontología: En filosofía, la **ontología** (del griego οντος, genitivo del participio del verbo ειμι, ser, estar) y λογος, ciencia, estudio, teoría) es una disciplina que se suele identificar con la *Metafísica general* o bien indica una de las ramas de ésta que estudia lo que es en tanto que es y existe [164]. En informática hace referencia al intento de formular un exhaustivo y riguroso esquema conceptual dentro de un dominio dado, con la finalidad de facilitar la comunicación y el intercambio de la información entre diferentes sistemas.

Parsing morfosintáctico: también llamado Etiquetado sintáctico, expresa la estructura de constituyentes de los enunciados [157].

Parseo: (Parcing). Proceso de analizar una secuencia de símbolos a fin de determinar su estructura gramatical con respecto a una gramática formal dada. Formalmente es llamado análisis de sintaxis. Un parseador (parser) es un programa de computación que lleva a cabo esta tarea. El parseo transforma una entrada de texto en una estructura de datos (usualmente un árbol) que es apropiada para ser procesada.

Generalmente los parseadores primero identifican los símbolos de la entrada y luego construyen el árbol de parseo para esos símbolos.

P_o: ponderación. Valoración de ciertas EBH que se propagan a las E_{ci}.

SC: Sistema Controlador. Módulo del prototipo WIH encargado de controlar el funcionamiento local y global del sistema, usando un conjunto de objetivos y funciones f_m.

Semántica: Rama de la Lingüística que se ocupa del sentido o el significado de los signos, así como de la relación entre los mismos, tanto desde un punto de vista sincrónico como diacrónico.

Sintagma: En Lingüística, cualquier combinación seriada de elementos morfológicos, que adquieren determinada unidad, e incluso estabilidad, cuando la combinación se estereotipa por el uso (por ej., el sintagma "por favor").

Sintaxis: La sintaxis, una subdisciplina de la lingüística y parte importante del análisis gramatical, se encarga del estudio de las reglas que gobiernan la combinatoria de constituyentes y la formación de unidades superiores a estos, como los sintagmas y oraciones.

Tag: término para expresar las etiquetas generadas en forma estandarizada por algún lenguaje ML.

W3C: El Consorcio World Wide Web (W3C) es un consorcio internacional donde las organizaciones miembro, personal con dedicación exclusiva y el público en general, trabajan conjuntamente para desarrollar estándares Web [165].

Web semántica: (del inglés Semantic Web) es la idea de añadir metadatos semánticos a la World Wide Web. Esas informaciones adicionales (describiendo el contenido, el significado y la relación de los datos) deben ser dadas de manera formal, de forma que sea posible evaluarlas automáticamente por máquinas. El objetivo es ampliar la interoperabilidad entre sistemas informáticos y reducir la necesaria mediación de operadores humanos. [164].

WIH: Web Intelligent Handler. Nombre del prototipo generador de la Red Virtual, que reorganiza virtualmente los contenidos que le son visibles.

Referencias

[1] Aberer K.”Report on the First International Conference on Ontologies, Databases and Applications of Semantics (ODBASE)”. Part of the Federated Conference Ön the Move to Meaningful Internet Systems 2002. SIGMOD Record. vol. 32. No. 1. 2003.

[2] Aguado de Cea G., Álvarez de Mon y Rego I., Pareja Lora A.”Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: Onto Tag”. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No. 17. pp. 37 - 49. 2002.

[3] Aguado de Cea G., Álvarez de Mon y Rego I., Pareja Lora A., Plaza Arteche R.”RFD(S)/XML LINGUISTIC ANNOTATION OF SEMANTIC WEB PAGES”. International Conference on Computational Linguistics. Proceedings of the 2nd workshop on NLP and XML - Volume 17.pp 1 - 8. 2002.

[4] Aldezábal I., Alegria I., Ezeiza N., Urizar R., Aduriz I.”Del analizador morfológico al etiquetador/lematizador: unidades léxicas complejas y desambiguación”. Procesamiento del lenguaje natural, ISSN 1135-5948, Nº. 19. pp. 90-100.1996.

[5] Allan J., Kumaran G.”Stemming in the Language Modeling Framework”. Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canadá. Pp.455 - 456. 2003.

[6] Alonso Pardo M. A. “Interpretación tabular de autómatas para lenguajes de adjunción de árboles”. Tesis Doctor en Informática por la Universidad de La Coruña. España. 2000.

[7] Álvarez Muro A. “Análisis de oralidad: Una poética del habla cotidiana”. Univ. de los Andes. Grupo de Lingüística Hispánica. Mérida. Venezuela. 2001.

[8] Assadi H.”Knowledge Acquisition from Texts: Using an Automatic Clustering Method Based on Noun-Modifier RelationShip”. European

Chapter Meeting of the ACL. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Madrid, Spain. pp. 504 - 506. 1997.

[9] Baeza Yates R., Ribeiro Neto B.”Modern Information Retrieval”. Addison Wesley. ISBN 10:020139829X, ISBN 13:9780201398298. 1999.

[10] Bargalló M., Forgas E., Garriga C., Rubio A.”Las lenguas de especialidad y su didáctica”. J. Schnitzer Eds. Universitat Rovira i Virgili. Tarragona, cap. 1 (P. Schifko, Wirtschaftsuniversität Wien), pp. 21-29. 2001.

[11] Bernard G.”Compréhension de texte, opérations linguistiques, linguistique textuelle, reference”. 1990.

[12] Bernard G., Feat J., San Gines P., Sabido V.”Comprensión de textos: elementos de modelización”. Caminos del texto (Actes des colloques de Madrid et Grenade). Universidad de Granada. 1995.

[13] Berrut C., Palmer P.”Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing”. ACM Conference on Research and Development in Information Retrieval. 1986.

[14] Bilotti M. W., Katz B., Lin J. “What works better for Question Answering: Stemming or Morphological Query Expansion?”. Proc. of Information Retrieval for Question Answering (IR+QA), SIGIR 2004, Sheffield: England, 2004.

[15] Bosch M.”Documentos y lenguaje de marcado: conceptos, problemas y tendencias”. El profesional de la información. Barcelona. v. 10, n. 11, pp. 4-9. 2001.

[16] Brent M. R., Hopkins J.”From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax”. Computational Linguistic, Vol. 19, No. 2. pp. 243-262. 1993.

[17] Brun C., Hagège C.”Normalization and Paraphasing Using Symbolic Methods”. ACL: Second International workshop on Paraphrasing, Paraphrase Acquisition and Applications, Sapporo, Japan. 2003.

[18] Brun R.E., Senso J. A.”Minería textual”. El profesional de la información, vol. 13, nro 1. Enero-Febrero 2004.

[19] Bruschetti C.E.”Queries de Imágenes Basadas en Contenido”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Julio de 2006.

[20] Castellón Y., Collet A., Gonzalvo A., LLoré X., Ortega Á., Pérez G., Pérez J., Trotzig D. “El corrector de Planeta”. Procesamiento del lenguaje natural, ISSN 1135-5948, Nº 26, Págs. 143-146. 2000.

[21] Cha J., Lee G.”Structural disambiguation of morpho-syntactic categorical parsing for Korean”. International Conference on Computational Linguistics. Proceedings of the 18th conference on Computational linguistics - Volume 2. Saarbrücken, Germany. pp. 1002 - 1006. 2000.

[22] Chen H., Houston A. L., Sewell R. R., Schatz B. R.”Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques”. International Conference on Digital Libraries. Proceedings of the second ACM international conference on Digital libraries. Philadelphia, Pennsylvania, United States .pp 257. 1997.

[23] Chomski N.”Aspectos de la teoría de la sintaxis”. Ed. Gedisa. ISBN: 8474326729. ISBN-13: 9788474326727. 1965.

[24] Chow C. K., Liu C. N.”Approximating discrete probability distributions with dependency trees”. IEEE Trans. on Information Theory, IT-14. pp. 426 - 467. 1968.

[25] Civit M., Martí M.A.”Estándares de Anotación Morfosintáctica para el español”. Taller de Herramientas y Recursos Lingüísticos para el español y el portugués. Tonantzintla, México. 2004.

[26] Corcho O., López Cima A., Gómez Pérez A.”A Platform for the Development of Semantic Web Portals”. ICWE’06. Palo Alto. California. USA. ACM. 2006.

[27] Corti P. C., Guerisoli S. M.”Estudio Crítico de la Interfaz Gráfica WEBSOM”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Julio de 2007.

[28] Delgado M., Sánchez D., Serrano J.M., Vila M.A.”A Survey of methods to evaluate quantified sentences”. Mathware and Soft Computing. vol. 7 (2 - 3): pp. 149 - 158. 2000.

[29] Dias G.”Multiword Unit Hybrid Extraction”. Annual Meeting of the ACL .Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18. pp. 41 - 48. 2003.

[30] Dichev D. C.”View-Based Semantic Search and Browsing”. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI’06). Computer Society. 2006.

[31] Elliott A.”Flamenco Image Browser: Using Metadata to Improve Image Search During Architectural Design”. Conference on Human Factors in Computing Systems. Proceedings of CHI ‘01. 2001.

[32] Emirkanian L., Bouchard L. H.”Knowledge integration in a robust and efficient morpho-syntactic analyzer for French”. Proceedings of the 12th conference on Computational linguistics - Volume 1. Association for Computational Linguistics. 1988.

[33] English J., Hearst M., Sinha R., Swearingen K., Yee K.”Hierarchical Faceted Metadata in Site Search Interfaces”. Conference on Human Factors in Computing Systems. CHI '02 extended abstracts on Human factors in computing systems. Miniápolis, Minnesota, USA. pp. 628 - 639. 2002.

[34] Etchemendy J., Barwise J. “Model-theoretic Semantics”. Foundations of Cognitive Science. M. Posner, MIT Press, pp. 207-243, 1989.

[35] Fabre C., Jacquemin C.”Boosting Variant Recognition with Light Semantics”. International Conference on Computational Linguistics. Proceedings of the 18th conference on Computational linguistics - Volume 1. Saarbrücken, Germany. pp. 264 - 270. 2000.

[36] Farrugia J.”Model-Theoretic Semantics for the Web”. Journal of the XX World Wide Web Conference. 2003.

[37] Figuerola C. G., Zazo A.F., Berrocal J. L. A.”Categorización automática de documentos en español: algunos resultados experimentales”. ReLIS, Jornadas de Bibliotecas Digitales. vol. 14. http:// imhotep.unizar.es/jbidi/jbidi2000/14_2000. 2000.

[38] Frakes W. B., Fox C. J.”Strength and Similarity of Affix Removal Stemming Algorithms”. ACM SIGIR Forum. Volume 37, Issue 1.pp.26 - 30. 2003.

[39] Galicia Haro S.N., Gelbukh A.”Investigaciones en análisis sintáctico para el español”. Serie Ciencia de la Computación. Primera edición. Instituto Politécnico Nacional. ISBN 970-36-0265-7. México. 2007.

[40] Gálvez C.”El diccionario electrónico: un instrumento para la unificación de términos en la indización automática”. Linguax: Revista de Lenguas Aplicadas (Universidad Alfonso X El Sabio), ISSN 1695-632X. 2006.

[41] Genthial D., Courtin J., Kowarski I.”Contribution of a Category Hierarchy to the Robustness of Syntactic Parsing”. 13th CoLing, Helsinki, Finland, Vol. 2, pp 139-144. 1990.

[42] Godlewski G., Piasecki M., Sas J.”Application of Syntactic Properties to Three-level Recognition of Polish Hand-written Medical Texts”. DocEng’06. The Netherlands. Amsterdam. ACM. 2006.

[43] Golub K.”The Role of Different Thesauri Terms and Captions in Automated Subject Classification”. Web Intelligence. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Pp.961-965 .2006.

[44] Gulla A. A., N. Moshagen S.”A sign Expansion Approach to Dynamic, Multi-purpose Lexicons”. International Conference on Computational Linguistics. Proceedings of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark. pp. 478 - 483. 1996.

[45] Gutiérrez C. “La contradicción: ¿vicio formal o cifra de contenido?”. Rev. Crítica. Vol. 18. México. 1972.

[46] Gruber R.”A translation approach to portable ontology specification”. Knowledge Acquisition. Vol. 5. pp 199 - 220. (1993)

[47] Halliday M., Hassan R.”Cohesion in English”. Longman, Londres. 1976.

[48] Harman D.”A failure Analysis on the Limitations of Suffixing in an Online Environment”. ACM SIGIR Forum. Volume 23, Issue 1-2. pp. 5 - 11. 1988.

[49] Hayes P.”RDF Semantics”. W3C Working Draft 23. 2003. http://www.w3.org/TR/rdf-mt/.

[50] Hearst M. A.”Next Generation Web Search: Setting Our Sites”. IEEE Data Engineering Bulletin, Vol. 23(3): pages 38--48.2000.

[51] Hearst M. A.”Untangling text data mining”. Proceedings of ACL’99. The 37th annual meeting of the Association for Computational Linguistics. 1999.

[52] Hellwig P.”Dependency Unification Grammar”. E. Hajicova (ed.), Functional Description of Language, Charles University, Prague, pp. 67-84. 1983.

[53] Hernández O. J., Ferri Ramírez C.”Práctica de Minería de Datos. Introducción al WEKA”. Universitat Politécnica de Valencia. Valencia. 2006. [CXLIV]

[54] Herrera Viedma E., López Herrera A. G., Luque M., Porcel C. “A Fuzzy Linguistic IRS Model Based on a 2-Tuple Fuzzy Linguistic Approach. V Congreso ISKO. pp. 148 - 157. España. España. 2001.

[55] Herrera Viedma E., Pasi G.”Approaches to access information on the Web: recent developments and research trends”. Fuzzy. Proc. International Conference on Fuzzy Logic and Technology (EUSFLAT 2003), pp. 25–31, Zittau (Germany). 2003.

[56] Herrera F., Herrera Viedma E., Verdegay J. L.”Aggregating linguistic preferences: properties of lowa operator. Proc. of VI IFSA World Congress, Sao Paulo. Brazil. Vol. II. pp. 153 - 157. 1995.

[57] Hodges W. “Classical Logic I: First-/order Logic”. The Blackwell Guide to Philosophical Logic. Goble Lou (Editor). pp. 9-32. 2001.

[58] Hodges W. “Model Theory”. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/model-theory/.2001.

[59] Hofstadter D. R., Göedel E., Bach E. “An Eternal Golden Braid”. New York: Vintage Books, pp. 26. ISBN-10: 0465026567. ISBN-13: 978-0465026562. 1980.

[60] Hull R., Su J.”Tools for Design of Composite Web Services”. SIGMOD 2004. Paris. Francia. ACM. 2004.

[61] Ibekwe Sanjuan F.”Terminological variation, a means of identifying research topics from texts”. International Conference on Computational Linguistics. Proceedings of the 17th international conference on Computational linguistics - Volume 1. Montreal, Quebec, Canadá. pp. 564 - 570. 1998.

[62] Ide N., Romary L.”A Common Framework for Syntactic Annotation”. Annual Meeting of the ACL. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Toulouse, France. pp. 306 - 313. 2001.

[63] Jackzynski M., Trousse B.”Fuzzy Logic for the retrieval step of a Case-Based Reasoner”. Proc. of the Second European Conference on Case-Based Reasoning, pp. 313-322. 1994.

[64] Jacquemin C., Klavans J. L., Tzoukerman E.”Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax”. Conf: Meeting of the Association for Computational Linguistics. ACL: Annual Meeting of the ACL. 1997.

[65] Jin Song D.”From Semantic Web to Expressive Specifications: A Modeling Languages Spectrum”. ICSE’06. Shangai. China. ACM. 2006.

[66] Jin Song D. “Software Modeling Techniques and the Semantic Web”. Proc. of 26th International Conference on Software Engineering (ICSE’04). Computer Society. 2004.

[67] José Padrón G. “notas sobre análisis del lenguaje”. Cap. 1: Modelo General de los lenguajes. Maracaibo. La Universidad del Zulia -Doctorado en ciencias Humanas. 1997.

[68] Kantrowitz M., Mohit B., Mittal V.”Stemming and its effects on TFIDF Ranking”. Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Athens, Greece. Pp. 357 - 359. 2000.

[69] Kaski S.”Data Exploration using self organizing maps”. Acta Polytechnica Scandinava, mathematics, computing and management in engineering series. No 82. PHD. Tech. Thesis. Helsinki University of Tech. Finland. 1997.

[70] Kennedy C., Boguraev B.”Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser”. International Conference on Computational Linguistics. Proceedings of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark. pp. 113 - 118. 1996.

[71] Kietz J.U., Maedche A., Volz R.”A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet”. Proceedings of the EKAW’00 Workshop on Ontologies and Text. Juan- Les-Pin es, France. 2000.

[72] Kinyon A.”Hypertags”. International Conference on Computational Linguistics. Proceedings of the 18th conference on Computational linguistics - Volume 1. Saarbrücken, Germany. pp. 446 - 452. 2000.

[73] Klyne G., Carroll J.”Resource Description Framework (RDF): Concepts and Abstract Syntax”. Klyne G., Carroll J. (Editors). W3C Working Draft. 2003.

[74] Kohonen T.”The self-organizing map”. Neurocomputing nro. 21. 1998.

[75] Kouroupetroglou C., Salampasis M., Manitsaris A.”A Semantic-Web based Framework for Developing Applications to Improve Accessibility in the WWW”. ACM International Conference Proceeding Series; Vol. 134 .Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A): Building the mobile web: rediscovering accessibility? Edinburgh, U.K.SESSION: Understanding accessibility. pp 98 - 108. 2006.

[76] Kukich K.”Techniques for Automatically Correcting Words in Text”. ACM Computing Surveys, Vol. 24, No. 4. 1992.

[77] Lagus K., Honkela T., Kohonen T.”WEBSOM - A Status Report”. Proc. of STeP96. T. Honkela and M. Jakobsson (Eds.). Finnish Artificial Intelligence Society, pp. 73 - 78. 1996.

[78] Langer H.”Syntactic Normalization of Spontaneous Speech”. Proceedings of COLING. International Conference on Computational Linguistics. Proceedings of the 13th conference on Computational linguistics - Volume 3. Helsinki, Finland. pp 180 - 183. 1990

[79] Leech G.”Introducing corpus annotation”. Corpus Annotation: Linguistic Information from Computer Text Corpora. R. Garside, G. Leech, A.M. McEnery Eds. London: Longman. 1997.

[80] Levinger M., Ornan U.”Learning Morpho-lexical Probabilities from an Untagged Corpus with an Application to Hebrew”. Levinger M., Ornan U. Computational Linguistics. Volume 21, Issue 3. pp. 383 - 404. 1995.

[81] Lin X. “Visualization for the document Space”. Journal of the American Society for Information Science. 49 (1). PP 40 - 54. 1997.

[82] López De Luise M. D.”Ambiguity and Contradiction From a Morpho-Syntactic Prototype Perspective”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2007. Springer. Tarek Sobh & Khaled Elleithy Editors. Aceptado para publicación. 2007.

[83] López De Luise M. D.”A Metric for Automatical Word Categorization”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2007. Springer. Tarek Sobh & Khaled Elleithy Editors. Aceptado para publicación. 2007.

[84] López De Luise M. D., Agüero M. J.”Aplicabilidad de métricas categóricas en sistemas difusos”. IEEE Latin America Magazine. Vol. 5. Issue 1. Editor Jefe José Antonio Jardini. 2007.

[85] López De Luise M. D.”A Morphosyntactical Complementary Structure for Searching and Browsing”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2005. Springer. Tarek Sobh & Khaled Elleithy Editors. Pp. 283 – 290. 2005.

[86] López De Luise M. D.”Una representación alternativa para textos”. Ciencia yTecnología. Colección C&T. ISSN 1850 0870. 2007-04. Bs As. Argentina. Pp. 119-130. 2007.

[87] López De Luise M. D., Ale J. M.”Induction Trees for Automatic Word Classification”. Anales XIII Congreso Argentino de Ciencias de la Computación (CACIC07). Corrientes. Argentina. pp. 1702. 2007.

[88] López De Luise M. D., Ale, J.”Non-technological Aspects on Web Searching Success”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2007. Springer. Tarek Sobh & Khaled Elleithy Editors. Aceptado para publicación.

[89] Losada D. E., Barro Ameneiro S., Bugarín Diz A. J., Díaz Hermida F.”Experiments on using fuzzy quantified sentences in adhoc retrieval”. ACM Symposium on Applied Computing. vol. 0, pp. 1059 - 1066. 2004.

[90] Losada D. E., Díaz Hermida F., Bugarín A. “Semi-fuzzy quantifiers for information retrieval”. International Journal of approximate reasoning. vol. 34, pp. 49-88. 2003.

[91] Losada D. E., Díaz Hermida F., Bugarín A., Barro S. ”Experiments on using fuzzy quantified sentences in adhoc retrieval” Proc. SAC-04, the 19th ACM Symposium on Applied Computing - Special Track on Information Access and Retrieval, Nicosia, Cyprus. 2004.

[92] Martínez Fernández P., García Serrano A.M. ”Interacción persona-web empleando recursos lingüísticos”. Revista Iberoamericana de Inteligencia Artificial. Nro. 16. pp 55 - 65. 2002.

[93] Maximilien E.M., Singh M.P. ”Conceptual Model of Web Service Reputation“. SIGMOD Record. vol. 31. No. 4. 2003.

[94] Mayfield J., McNamee P.”Single N-gram Stemming”. Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada. POSTER SESSION: Posters. Pp.415 - 416. 2003.

[95] Melucci M., Orio N.”A Novel Method for Stemmer Generation Based on Hidden Markov Models”. Conference on Information and Knowledge Management. Proceedings of the twelfth international conference on Information and knowledge management. New Orleans, LA, USA. SESSION: Information retrieval session 3: cross language retrieval. pp. 131 - 138. 2003.

[96] Mitchell T. M.”Machine Learning”. Mc. Graw Hill. ISBN 0070428077. Boston. 1997.

[97] Montemagni S., Federici S., Pirelli V.”Resolving syntactic ambiguities with lexico-semantic patterns: an analogy-based approach”. International Conference on Computational Linguistics. Proceedings of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark. pp. 376 - 381. 1996.

[98] Monz C., Rijke M. “The University of Amsterdam at CLEF 2001”. Proc. of Cross Language Evaluation Forum (CLEF), pp.165 - 169, Amsterdam, 2001

[99] Morales Luna G. “El pensamiento natural y las limitantes formales”. Avance y Perspectiva vol. 21, pp. 355 - 360. Noviembre 2002.

[100] Morillas Raya A.”Introducción al análisis de datos difusos”. Depto. de Estadística y Econometría. Univ. de Málaga. España. www.eumed.net/libros/2006b/amr/. ISBN: 84-689-9208-2. 2006.

[101] Mosterín J.”Model Theory”. Preface to Manzano María. Oxford. 1999.

[102] Nießen S., Ney H.”Improving SMT quality with morpho-syntactic analysis”. International Conference on Computational Linguistics. Proceedings of the 18th conference on Computational linguistics - Volume 2. Saarbrücken, Germany. pp. 1081 - 1085. 2000

[103] Nießen S., Ney H.”Statistical Machine Translation with Scarce Resources Using Morpho-Syntactic Information”. Computational Linguistics. Volume 30, Issue 2. Pp. 181 - 204. 2004.

[104] Nilsson M.”The Conzilla Design. The definitive reference”. Technical Report CID/NADAKTH. Department of Numerical Analysis and Computing. http://conzilla.sourceforge.net/doc/conzilla-design/conzilla-design.html. 2000.

[105] Nilsson M., Palmer M.”Conzilla. Towards a Concept Browser”. Technical Report CID-53, TRITA-NA-D9911, Department of Numerical Analysis and Computing. 1999.

[106] Nys M. A.”Buscadores Web: Análisis Comparativo entre Interfaces Gráficas y Textuales”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Julio de 2007.

[107] Ontrup J., Ritter H.”Hyperbolic Self-Organizing Maps for Semantic Navigation”. Proceedings of NIPS. 2001.

[108] Paice C. D.”An evaluation method for Stemming Algorithms”. Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland. pp. 42 - 50. 1994.

[109] Paroubek P., Schabes Y., Joshi A. K.”XTAG - A Graphical Workbench for Developing Tree-Adjoining Grammars”. Third Conference on Applied Natural Language Processing, Trento (Italy). 1992

[110] Patel-Schneider P., Hayes P., I. Horrocks I.”Web Ontology Language (OWL) Abstract Syntax and Semantics”. Patel-Schneider P., Hayes P., I. Horrocks I. (Editors). W3C Working Draft. 2003.

[111] Paul I. J., Cant N. “Chomsky and the duality of the phonemic distinction”. En Beyond Chomsky. 2002.

[112] Paul I. J., Cant N.”Guidelines for post-Chomsky an linguistic theory”. Alternatives to Chomsky conference. New Jersey, USA. 2000.

[113] Pazos Bretaña J. M., Pamies Bertrán A.”Detección automatizada de colocaciones y otras unidades fraseológicas en un corpus electrónico”. .. Letras de Hoje. Porto Alegre. v. 41, nro. 2. pp. 23-36. 2006.

[114] Peñas Padilla A. “Website Term Browser. Un sistema interactivo y multilingüe de búsqueda textual basado en técnicas lingüísticas”. Tesis doctoral. Depto. de Lenguajes y Sistemas Informáticos. Universidad Nacional de Educación a Distancia. España. Cap. 4. pp. 85 - 118. (2002).

[115] Popović M., Ney H.”Improving Word Alignment Quality using Morpho-syntactic Information. International Conference on Computational Linguistics. Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland. Article No. 310. 2004.

[116] Porter M.F. “An Algorithm for suffix Stripping”. Program, vol. 14 (3) pp. 130 - 137. Jul 1980.

[117] Proal Aguilar C. “Innovación y Servicios de Información”. Tesis de Maestría en Ciencias con Especialidad en Ingeniería en Sistemas Computacionales. Univ. UDLA. Puebla. Cap. 4. pp. 43 - 73. 2003.

[118] Prószéky G., Kis B.”A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages”. Annual Meeting of the ACL. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland. pp. 261 - 268. 1999.

[119] Prószéky G., Naszódi M., Kis B.”Recognition Assistance: Treating Errors in Texts Acquired from Various Recognition Processes”. International Conference on Computational Linguistics. Proceedings of the 19th

international conference on Computational linguistics - Volume 2.pp 1 - 5. 2002.

[120] Reznikov L. O. “El rol semiótico de los diagramas en la resolución mental”. Alberto Corazón, Madrid. Cap. 4. 1970.

[121] Rodríguez A. H.”Validación de Sesión por Presencia Utilizando Reconocimiento Facial y Servicios Web”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Julio de 2006.

[122] Rosario B., Hearst M.”Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy”. Proceedings of Conference on Empirical Methods in Natural Language Processing. p. 82-90. 2001.

[123] Rosario B., Hearst M.A., Fillmore C.”The Descent of Hierarchy, and Selection in Relational Semantics”. Meeting of the Association for Computational Linguistics (ACL-02). 2002.

[124] Rottenstein V. E.”Análisis de los principales factores para el posicionamiento orgánico en Google”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Julio de 2006.

[125] Ruch P., Gaudinat A.”Comparing corpora and lexical ambiguity”. Workshop on Comparing Corpora, Proc. of ACL, Hong-Kong. 2000.

[126] Sánchez Cuadrado S., Llorens J., Morato J., Hurtado J. A.”Extracción Automática de Relaciones Semánticas”. 5ta Conferencia Iberoamericana en Sistemas, Cibernética e Informática. CISCI .2006.

[127] Santana Suárez O., Hernández Figueroa Z., Rodríguez Rodríguez G.”DAWeb: Un descargador y analizador morfológico de páginas Web”. Sociedad Española para el Procesamiento del Lenguaje Natural. Procesamiento del lenguaje natural, ISSN 1135-5948, Nº 30, 2003, Págs. 75-88. 2003.

[128] Santana Suárez O., Pérez J., Carreras F., Hernández Z., Rodríguez G.”FLANOM: Flexionador y lematizador automático de formas nominales”. Lingüística Española Actual XXI, 2. Ed. Arco/Libros, S.L. España. 1999.

[129] Santana Suárez O., Hernández Figueroa Z., Rodríguez Rodríguez G. ”Morphoanalysis of Spanish Texts: Two Applications for Web Pages”. Lecture Notes in Computer Science, (2722). pp. 511-514. ISSN: 03029743. 2003.

[130] Santana O., Pérez J., Carreras F., Hernández Z, Rodríguez G.”The Spanish Morphology in Internet”. Lecture notes in computer science (Lect. notes comput. sci.). ICWE 2003: international conference on web engineering, Oviedo, España. 2003.

[131] Santana Suárez O., Hernández Figueroa Z., Rodríguez Rodríguez G., Losada García L.”Una herramienta de Recuperación Morfoléxica Aplicada a Microsoft Word”. The Association for Computers and the Humanities, the Association for Literary and Linguistic Computing ACH/ALLC Conference 05.University of Victoria. 2005

[132] Savoy J.”Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages”. Symposium on Applied Computing. Proceedings of the 2006 ACM symposium on applied computing. Dijon, France. SESSION: Information access and retrieval (IAR). Pp.1031 - 1035 .2006.

[133] Schwartz A. S., Hearst M. A.”A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text”. Proceedings of the Pacific Symposium on Biocomputing (PSB 2003). Kuwait. 2003.

[134] Seamus D., Ciara H., Karen N.”Combining website search engine optimization with advanced web log analysis”. National University of Ireland. Cork, O’Rahilly Building, UCC, Ireland. 2006.

[135] Searle J. R. “El misterio de la conciencia”. Paidós. ISBN: 844930895X. ISBN-13: 9788449308956. 2007.

[136] Studer R., Benjamins R., Fensel D.”Knowledge Engineering: Principles and Methods”. Data and Knowledge Engineering (DKE), vol. 25 (1 - 2). pp 161 - 197. (1998).

[137] Subirats Rüggeberg C.”FrameNet Español. Una red semántica de marcos conceptuales”. E. Serra y G. Wotjak, eds. Cognición y percepción lingüísticas. Valencia: Universidad de Valencia y Universidad de Leipzig, pp. 182-196. 2005.

[138] Sullivan D.”Document warehousing and text mining”. New York. Wiley Computer Publishing. XVIII. 2001.

[139] Ton T., Lee Y.”Modeling Biomedical Assertions in the Semantic Web”. SAC’07. Seoul. Corea. ACM. 2007.

[140] Trujillo W. R.”Queries sobre archivos de voz”. Tesis de grado dirigida por López De Luise M. D. en el contexto de las investigaciones de esta tesis doctoral. Facultad de Ingeniería de la Universidad de Palermo. Diciembre de 2006. [140]

[141] Vangehuchten L.”La identificación y la selección de léxico a partir de un corpus de discurso económico empresarial en Español como lengua Extranjera con fines específicos”. Proc. TALC 06. España. 2004.

[142] Von Guericke O. “Growing Multi-Dimensional Self-Organizing Maps”. Von Guericke O. (Eds.): Self-Organizing Neural Networks - Recent Advances and Applications, pp. 95-120, Springer, Heidelber. 2001.

[143] Vosse T.”Detecting and Correcting Morpho-Syntactic Errors in Real Texts”. The Third Conference on Applied Natural Language Processing, pp. 111-118. 1992.

[144] Wang Baldonado M. Q.”An interactive, structure-mediated approach to exploring information in a heterogeneous, distributed environment”. Disertación doctoral. Depto. Computación científica. Sanford University. 1997.

[145] Wehrli E.”Design and Implementation of a Lexical Data Base”. European Chapter Meeting of the ACL. Proceedings of the second conference on European chapter of the Association for Computational Linguistics. Génova. Suiza. pp. 146 - 153. 1985.

[146] Weizenbaum J. “Computer Power and Human Reason: From Judgement to Calculation”. San Francisco: W.H. Freeman & Company. ISBN-10: 0716704633. ISBN-13: 978-0716704638. pp. 222. 1976.

[147] Witten I. H., Frank E.”Data Mining Practical Machine Learning Tools and Techniques”. Witten I. H., Frank E. Second Ed. Morgan Kaufman. San Francisco. ISBN 0-12-088407-0. USA. 2005.

[148] Xu J., Croft W. B.”Corpus Based Stemming Using Co occurrence of Word Variants”. ACM Transactions on Information Systems (TOIS). Volume 16, Issue 1. Pp. 61 - 81. 1998.

[149] Yao H., Etzkorn L.”Towards A Semantic-Based Approach for Software Reusable Component Classification and Retrieval”. ACMSE’04. Huntsville. Alabama. USA. 2004.

[150] Zaïane O. R.”Resource and Knowledge discovery from the internet and multimedia repositories”. PHD. Thesis. School of Computer Science, Simon Fraser University. 1999.

[151] Zernik U.”Corpus-Based Thematic Analysis”. Text-Based Intelligent Systems. P. S. Jacobs Eds. Lawrence Erlbaum. NJ. 1993

Referencias Web:

[152] AMETRA, http://www.ametra.fr/, 2004

[153] Corpus Encoding Standard. http://www.cs.vassar.edu/CES/

[154] Diálogo Universitario. www.dialogo universitario. Paden R.L, Wolfer J.

[155] Diccionario de informática. http://www.alegsa.com.ar/Dic/

[156] Discurso y pensamiento. http://apuntes.rincondelvago.com/discurso-y-pensamiento.html. Tema III: Discurso y Pensamiento de Luria. Ortiz A. México.

[157] La constitución de los corpus orales. http://liceu.uab.es /~joaquim/language_resources/ spoken_res/Const_corp_oral.html

[158] RosettaNet (Lingua Franca for eBusiness), http://www.rosettanet.org, 2002.

[159] Snowball Web Site. http://snowball.tartarus.org/. Algoritmo de Porter.

[160] The free dictionary. http://www.thefreedictionary.com/inflexional

[161] UNSPSC (Universal Standard Products and Services Classification), http://www.unspsc.org, 2002.

[162] Webé. http://www.webe.laencrucijada.com/2005/12/tema_11_la_lingueistica_nortea.html

[163] WebODE, http://delicias.dia.fi.upm.es/webODE, 2003.

[164] WIKIPEDIA. http://es.wikipedia.org/wiki

[165] W3C. http://www.w3c.es/Consorcio/

Apéndice A: texto de EBH35

Transcripción del texto correspondiente al hbe35, perteneciente al URL, orquidea.blogia.com, capturado en Febrero de 2006:

Estos son los envases que voy a utilizar:
Phalaenopsis yema

Así han quedado dispuestos dentro de la caja:
Phalaenopsis yema

Y así se van a quedar unas semanas.
Phalaenopsis yema

Los envases los puedes comprar en cualquier tienda, yo en mi caso son de aceitunas lavados, por supuesto. El sustrato que le he puesto es corteza de pino y lo he rociado bien de agua para que mantenga la humedad. He puesto los dos envases en un lugar luminoso pero sin sol directo. Tenia una caja mas grande pero he preferido dividirlos en dos grupos por si acaso alguno se malogra tener el otro grupo dispuesto.

Quizas los tenga que tirar dentro de una semana pero en mi afán de experimentar he decidido hacerlo solo por saber mas cosas de este maravilloso mundo. A ver que tan sale todo y si va bien os pondre fotos de su desarrollo para que vosotros también podais probar.

Si todo va bien, a las 4 semanas de sembrado, se debe ver en la yema una pequeña elevación de escasos 2 cm.

A las 8 semanas debe de medir de 1 a 3 cms y alas 10 sem de 4 a 5 cm, las raíces deben de aparecer alrededor de ese tiempo.

Las fotos y el texto de este metodo de siembra para producción de keikis son por gentileza de : Erick Damián de Tapachula (Chiapas) México.
Obtenido de "http://es.wikipedia.org/wiki/Keikis"

Categorías: Orchidaceae
Orquídeas de México

En México, todas las regiones situadas al sur del Trópico de Cáncer, desde las costas del Pací-fico y las del Golfo hasta las regiones que rebasan los 3 500 m sobre el nivel del mar en los estados de Michoacán, Guerrero, Oaxaca, Veracruz y Chiapas albergan la mayor riqueza de orquí-deas, aunque todos los estados cuentan por lo menos con una especie.

En estos estados son conocidas con nombres populares que aluden ya sea a la época en que florecen, a festividades religiosas o bien a la forma que asemeja la flor, por ejemplo torito, calaverita, flor de mayo, flor de Candelaria, flor de muerto, entre otros.

Sin duda las flores más admiradas son las del género Laelia muy conocidas por su uso tradicional en las ofrendas de muertos y en fiestas como el dí-a de las madres o el de la Virgen de Guadalupe.

Aunque el número de especies mexicanas es menor que el de otros paí-ses América tropical (Colombia, Ecuador, Perú, Brasil, etc.), México cuenta con un conocimiento taxonómico más avanzado de sus especies. En un estudio realizado en 1995, Miguel Àngel Soto, investigador dedicado al estudio de las orquídeas, habla de 1 106 especies y subespecies mexicanas descritas, distribuidas en 159 géneros. De éstas, señala: "Existen 444 especies o subespecies endémicas, las cuales corresponden a 40% del total registrado en el país. Esta característica convierte a la orquideoflora

mexicana, en una de las más ricas en endemismos entre los principales países de América tropical, quizás sólo superada por Brasil".

Aunque todavía no se cuenta con un inventario completo de orquídeas, muchas Aéreas de nuestro país cuya flora no había sido estudiada, comienzan a ser de gran interés para los botánicos. La región de Chimalapa, ubicada en el Istmo de Tehuantepec, Oaxaca, pese a que posee la vegetación natural mejor conservada del trópico de México, no cuenta con un inventario de su flora, por lo menos en lo que a orquídeas se refiere. Sin embargo, mediante el proyecto Diversidad y conservación de las orquídeas de la región de Chimalapa, Oaxaca, México.

Además de inventariar y evaluar los taxa, una de las acciones prioritarias es desarrollar planes para la conservación de la diversidad de especies de orquídeas, especialmente de aquellas que han sido declaradas amenazadas o en peligro de extinción. Los investigadores aseguran que la causa principal de que muchas especies hayan sido declaradas en algún estado de riesgo, es la pérdida de sus hábitats naturales causada por la destrucción de bosques para abrir paso principalmente a la agricultura. Por ello, "la estrategia más importante es la preservación del hábitat. Las prioridades para preservar estos ambientes naturales deben estar dictadas por la riqueza de especies y endemismos de un determinado hábitat. Como estas dos condiciones pueden variar en magnitud, en tipo de amenaza y en condiciones económicas y sociales, las estrategias de conservación tienen que ser manejadas a nivel nacional y regional".

Las orquídeas se concentran generalmente en Aéreas muy específicas, que son importantes por la riqueza y diversidad de sus poblaciones o por sus endemismos. Se estima que en México existen seis Áreas muy diversas, con menos de 100,000 ha cada una, localizadas en diferentes regiones florísticas del país, las cuales poseen 50% del total de orquídeas registrado y que representan tan solo 0.003% del territorio mexicano. Es muy importante identificar y conocer muy bien dichos centros y enfocar hacia ellos los planes de conservación".

Por otra parte, es necesario impulsar el cultivo y propagación, especialmente de las especies que por ser de gran interés hortícola en la actualidad cuentan con escasas poblaciones, debido a la colecta inmoderada que han sufrido en el pasado.

ESPECIES MEXICANAS CONSIDERADAS EN PELIGRO DE EXTINCION

Especie
Encyclia kienastii
Galeandra greenwoodiana
Galeottia grandiflora
Laelia anceps ssp. dawsonii
L. gouldiana
L. speciosa
L. superbiens
Lycaste lassioglossa
L. skinneri
Marmodes sotoana
M. uncia
Palumbina candida
Phragmipedium exstaminodium
P. xerophyticum
Rhynchostele majalis
R. uroskinneri
Rossioglossum grande
R. williamsianum
Trichopilia galeottiana"

Clasificación científica
Reino: Plantae
División: Magnoliophyta
Clase: Liliopsida

Orden: Asparagales

Familia: Orchidaceae

Subfamilia: Orchidoideae

Apéndice B*: estudio de aspectos no tecnológicos para el proceso de IR*

Este apéndice es un extracto de lo presentado en [88], dentro del marco de la presente investigación.

Tomando como base una encuesta GVU 7[39], de Abril de 1997 realizada por la GVU (entidad dedicada al estudio de problemáticas relacionadas con la Web), se analizaron las relaciones entre variables colectadas relacionando la satisfacción de un usuario respecto a otros factores culturales y emocionales.

El formulario original fue colocado en la Web y completado de manera anónima por miles de personas. De todos los datos, se filtraron sólo los correspondientes a personas que manifestaron tener algún dominio de Internet. En total quedaron 4232 registros para procesamiento.

La información correspondía a un formulario destinado a captar tres aspectos importantes reflejados en tres secciones:

-SECCION I, grado de uso de la Web: determina con dos preguntas si es o no un usuario frecuente y de experiencia.

-SECCION II. Actitudes y opiniones acerca del uso de la Web: se divide en tres grupos de preguntas:

Grupo i. Con respuestas Si/NO se está de acuerdo con lo vertido en la pregunta

Grupo ii. Con respuesta graduada, según el grado de apreciación de cierta cualidad o característica. En estos casos siempre se debe elegir entre dos opciones opuestas.

Grupo iii. Con respuesta graduada, según la apreciación de la existencia entre dos opciones o características que no son opuestas. En estos casos un valor intermedio equidistante a ambas alternativas significa ausencia total de las dos.

- SECCION III. Datos de referencia: acerca del sexo, edad y nivel educacional.

En la Tabla XLV se sumarizan las variables del formulario y su descripción.

[39] Copyright 1994-1998 Georgia Tech Research Corporation. All rights Reserved. Source: GVU's WWW User Survey www.gvu.gatech.edu/user_surveys.

Tabla XLV.Variables del formulario correspondientes a los atributos de la base.

Variable	Pregunta relacionada
Sección I	
V1	¿Cuán tiempo estima que usa la Web?
V2	¿Cuándo comenzó a usar la Web?
Sección II	
V3	Soy hábil manejando la Web
V4	Me siento sin imaginación cuando uso la Web
V5	Me siento preocupado cuando uso la Web
V6	Es difícil para mí encontrar información en la Web
V7	Usar la Web me hace pensar
V8	Me siento sin inventiva cuando uso la Web
V9	La Web me posibilita hacer muchas cosas
V10	Me siento en “onda” cuando uso la Web
V11	Cuando uso la Web el tiempo de respuesta de la PC es muy corto
V12	Cuando uso la Web me siento estimulado
V13	Me puedo concentrar mucho cuando uso la Web
V14	Sigo links sólo cuando me parece interesante
V15	Usar la Web me hacer realizar lo mejor de mis habilidades
V16	Cuando uso la Web me olvido de todo lo que me rodea
V17	Cuando uso la Web me siento espontáneo
V18	Cuando uso la Web me siento más en el “mundo de la computadora” que en el real
V19	La Web me posibilita distintas oportunidades de acción
V20	Disfruto navegando para ver qué hay
V21	Cuando uso la Web me siento apático
V22	Cuando uso la Web pienso en otras cosas
V23	Cuando uso la Web me distraigo fácilmente
V24	Sé cómo hallar lo que busco
V25	Cuando uso la Web me siento relajado
V26	Cuando uso la Web me siento flexible
V27	Cuando uso la Web el tiempo se va rápido
V28	Cuando uso la Web me siento en una “realidad virtual”
V29	Cuando uso la Web mis capacidades se estiran hasta su límite
V30	Cuando uso la Web me siento juguetón
V31	Usar la Web es lento y tedioso
V32	Usar la Web me cambia
V33	Usar la Web prueba mis habilidades
V34	Cuando uso la Web me resulta sencillo bajar software
V35	Cuando uso la Web me siento creativo
V36	Cuando uso la Web me siento absorto en lo que hago
V37	Las alternativas de interacción con la Web hoy son limitadas
V38	Creo que soy un entendido en técnicas de búsqueda

Variable	Pregunta relacionada
V39	Cuando uso la Web me siento aburrido
V40	Sé menos que la mayoría sobre cómo usar la Web
V41	Cuando uso la Web me siento poco original
V42	Cuando uso la Web me agrada experimentar
V43	Cuando uso la Web me siento en control
V44	Creo que la Web es fácil de usar
V45	Cuando uso la Web me siento ansioso
V46	Navegar con los browsers es:
V47	Cuando uso la Web normalmente:
V48	Interactuar con la Web es:
V49 - V58	Aspectos sobre cómo se siente en gral. acerca de la Web
V59 - V74	Aspectos sobre cómo se siente en gral. cuando usa la Web
Sección III	
V75	Sexo
V76	Edad
V77	¿Cuál es el mayor nivel de educación que tiene?

Luego de un adecuado preprocesamiento, se realizaron los siguientes estudios:

1.-Se seleccionaron las variables más representativas:

Con un estudio de dependencias (gráficos de dispersión, análisis de correspondencias, correspondencia entre los valores de variables[40]). Aquellas que resultaron redundantes, se interpretaron como una validación cruzada de la información obtenida por otras, y fueron desplazadas para el estudio y usadas sólo a los fines verificatorios de los resultados que se iban obteniendo.

2.-Se descartaron variables innecesarias:

4 variables de las 77 originales, debido a que su información no es relevante para este estudio.

3.-Se determinó que la variable V6 como el reflejo del grado de satisfacción del usuario luego de una interacción con Internet.

[40] Algunas variables realizan la misma que otra, pero al revés, entonces los valores graduados tienen una correspondencia inversa. Otras variables realizan preguntas similares pero pertenecen a distintos subgrupos y por lo tanto su graduación debe evaluarse de distinta manera.

4.-Con un árbol de inducción (específicamente algoritmo J48[41]) se determinaron las variables críticas para inferir si un usuario quedará satisfecho luego de una búsqueda. Para ello se procesaron las instancias y se construyó el árbol. Luego se derivaron las reglas equivalentes. Por último se filtraron sólo las reglas con soporte[42] mayor a 20. El resultado fue sorprendentemente sencillo: con sólo 4 reglas y un total de 5 variables (o atributos de la base de datos), se puede inferir razonablemente bien el éxito en una búsqueda. Las reglas se transcriben en la Tabla XLVI.

Tabla XLVI.Reglas con soporte > 20.

id	regla	significado
r1	V24 >6 $\Rightarrow$ No	un usuario que no sabe cómo buscar información no la halla
r2	V24 < 5 $\Rightarrow$ Yes	un usuario que declara saber cómo buscar información la halla
r3	V24 = 5 $\wedge$ V4 = 5 $\Rightarrow$ No	el usuario que no sabe demasiado cómo buscar y se siente preocupado, no halla lo que busca
r4	V24 = 6 $\wedge$ [(V9 = 8 $\wedge$ V27 = 8) $\vee$ (V9 = 7 $\wedge$ V42 = 6) $\vee$ (V9 = 6)] $\Rightarrow$ Yes	Quien no sabe cómo buscar pero se siente con habilidad de explorar y experimentar, halla lo que busca.

5.- Para comprobar si estas variables realmente reflejan factores decisivos para las personas con dificultades en hallar información en la Web, se procesaron los datos con el agrupamiento jerárquico COBWEB[43]: bajo la hipótesis de que realmente son buenas clasificadoras de estos casos, entonces debieran ser buenas para agrupar los

[41] Algorítmica que permite descubrir disposiciones naturales en los datos sin información previa. Es una implementación del conocido C4.5 [96]. Entre sus características: trabaja con bifurcaciones binarias de los valores, bifurcaciones ponderadas cuando hay valores faltantes, trabaja con intervalos de confianza del error estimado y realiza post-pruning (podado a posteriori).

[42] Soporte: término empleado para denominar la cantidad de registros que verifican cierta regla.

[43] CobWeb: método incremental de aprendizaje conceptual. [147] A diferencia de métodos como k-means, no particiona los clusters en grupos disjuntos. Fue desarrollado en los 80 para atributos nominales. Realiza un agrupamiento jerárquico de las instancias y uisa una medida para cualificar los grupos formados. El nodo raíz representa todos los datos y las hojas a cada registro. Los nodos pueden dividirse o fusionarse para reestructurarse a conveniencia durante el entrenamiento. Este algoritmo no requiere una definición previa de la cantidad degrupos a obtener.

casos en grupos donde el éxito o fracaso de los resultados sea similar. Los resultados (ver Tabla XLVII) muestran la formación de 5 grupos (o clusters) de los cuales el grupo 0 y 1 corresponden a los hallazgos exitosos, y los grupos 2 y 3 corresponden al fracaso en el resultado obtenido, sin mezclas en los clusters. Éstos resultados confirmarían que las variables V4, V9, V24, V27 y V42 son atributos que describen las características de los individuos con éxito o no en la búsqueda de información.

Tabla XLVII.Asignación de instancias a los grupos

Cluster	Number instances	Percentage
1	1136	27
2	1788	42
3	880	21
4	908	21
5	1298	31

6.-Se probaron otros algoritmos de clasificación para indagar si existe otro conjunto de variables que sirvan para perfilar a los individuos con éxito o no en las búsquedas. Del análisis surgió que con BayesNet[44] se perfila otro conjunto de variables óptimas: V59, V60, V63, V64, V65, V67, V69, V70 y V73, que participan en la red de la Fig. 53. De esta red es importante rescatar algunas observaciones que se derivan:

-La dificultad en hallar información (V6) se relaciona con la falta de conocimiento en la navegación (V67).

-La falta de conocimiento en la navegación (V67) incide en la sensación de control (V64) y en la necesidad de conocimientos (V73).

-La felicidad al usar la Web incide en el estímulo para navegar (V65), socialización (V69) y exploración de los propios límites (V60).

[44] BayesNet [155]: El aprendizaje Bayesiano se basa en que existe una distribución de probabilidad seguida por los datos. El Naive Bayes, por ejemplo asume que toda instancia tiene una cantidad de aributos y trata de aprender una función objetivo capaz de clasificar una nueva instancia sobre la base de las características de las demás. Estos algoritmos tienen buena representación gráfica cuando la distribución de nodos es relativamente sencilla y pueden sobrellevar fácilmente el problema de la replicación de nodos presente en los árboles de inducción. BayesNet es una extensión de Naive Bayes.

7.-Dado que la búsqueda dentro del espacio de soluciones puede realizarse con distintas estrategias cuando se usa BayesNet, se compararon y evaluaron diversas alternativas. El método de búsqueda óptimo resultó ser TAN[45].

8.-De todas las redes BayesNet, es posible extraer el subgrafo en común que se muestra en laFig. 54, del que se deduce que las siguientes preguntas son claves para determinar el éxito:

-¿Se siente felíz cuando usa la Web?
-¿Se siente estimulado cuando usa la Web?
-¿Se siente excitado cuando usa la Web?

9.-Como conclusiones del trabajo se puede afirmar que, para hallar lo que se busca es importante saber cómo buscar, estar confiado (no preocupado) y sentir habilidad para explorar y experimentar.

[45] BayesNet tiene una etapa exploratoria donde se define el mecanismo de crecimiento y reorganización de la red. Entre los mecanismos está TAN (Tree Augmented Naïve Bayes), que comienza con el nodo de clase como en nodo padre. Luego evalúa si agrega un nuevo nodo padre en cada nodo. El árbol se forma calculando el máximo balanceo de pesos con un algoritmo provisto por Chow y Liu [24].

Fig. 53. BayesNet con búsqueda TAN.

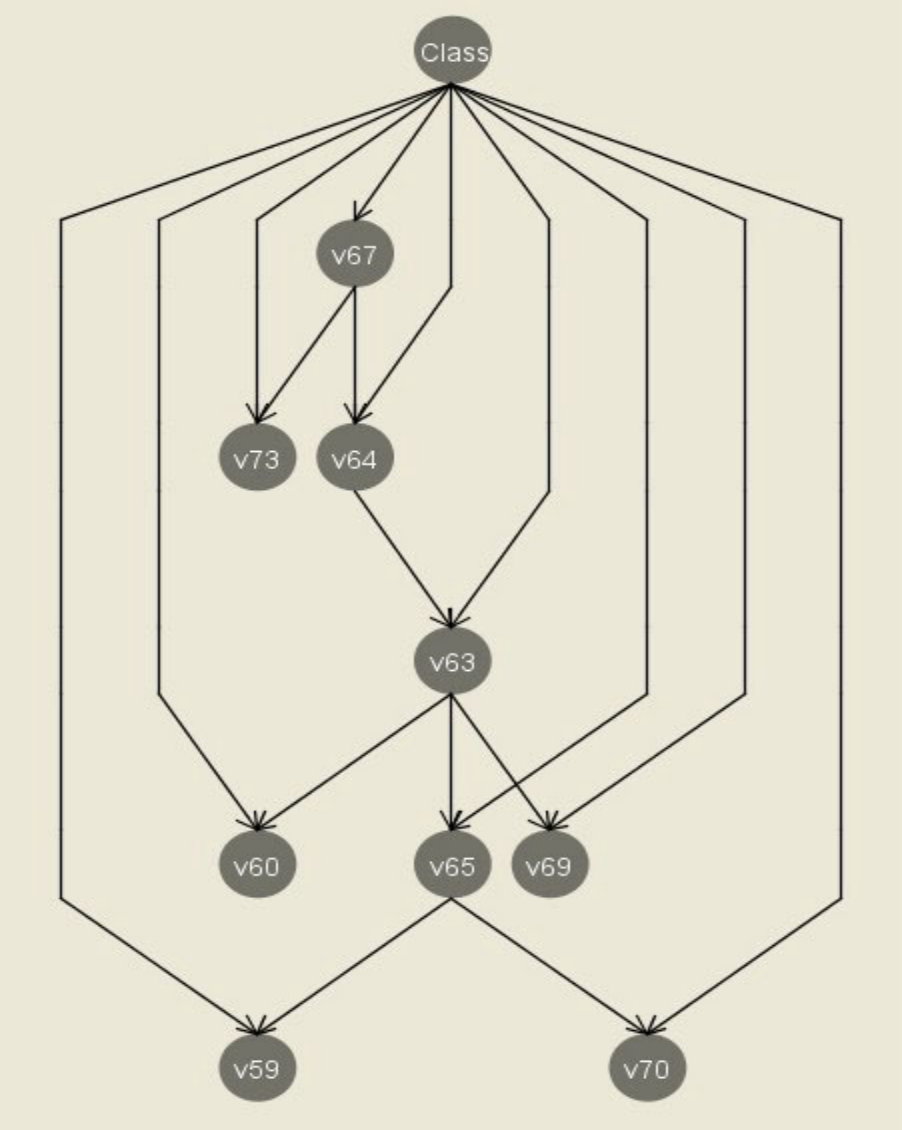
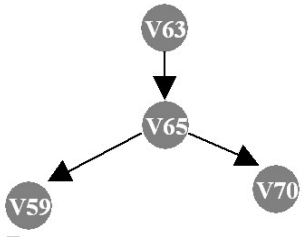


Fig. 54. Subgrafo común en BayesNet.



Apéndice C: relevancia p_o

Encuesta para validar resultados obtenidos con el uso de po.

Formulario de captura de datos.

El formulario original empleado contiene el texto del Apéndice A y luego el formulario que se reproduce a continuación:

Encuesta

nro. protocolo:_____

¿Qué palabras del texto usaría para consultar en un buscador para buscar este artículo en la Web? (complete un máximo de 10 y un mínimo de 2 palabras).

palabra1:_____
palabra2:_____
palabra3:_____
palabra4:_____
palabra5:_____
palabra6:_____
palabra7:_____
palabra8:_____
palabra9:_____
palabra10:_____

Describa muy brevemente los tópicos esenciales de los que habla el texto, usando frases muy cortas. (Complete un mínimo de 1 y máximo de 4 tópicos).

tópico1:_____

_____

_____

tópico2:_____

_____

_____

tópico3:_____

_____

_____

tópico4:_____

_____

_____

Datos obtenidos

A continuación se resumen los resultados obtenidos por la encuesta. Para la correcta evaluación de los mismos se ha definido:

Palabras tipo I: palabras respondidas en la encuesta que pertenecen al texto original y que fueron evaluadas por WIH con p_o cercano a 0.0.

Palabras tipo II: palabras respondidas en la encuesta que pertenecen al texto original y que fueron evaluadas por WIH con p_o lejano a 0.0.

Otras: palabras respondidas en la encuesta que pertenecen al texto pero no pertenecen a las categorías anteriores.

Fuera de texto: palabras respondidas en la encuesta que no pertenecen al texto original.

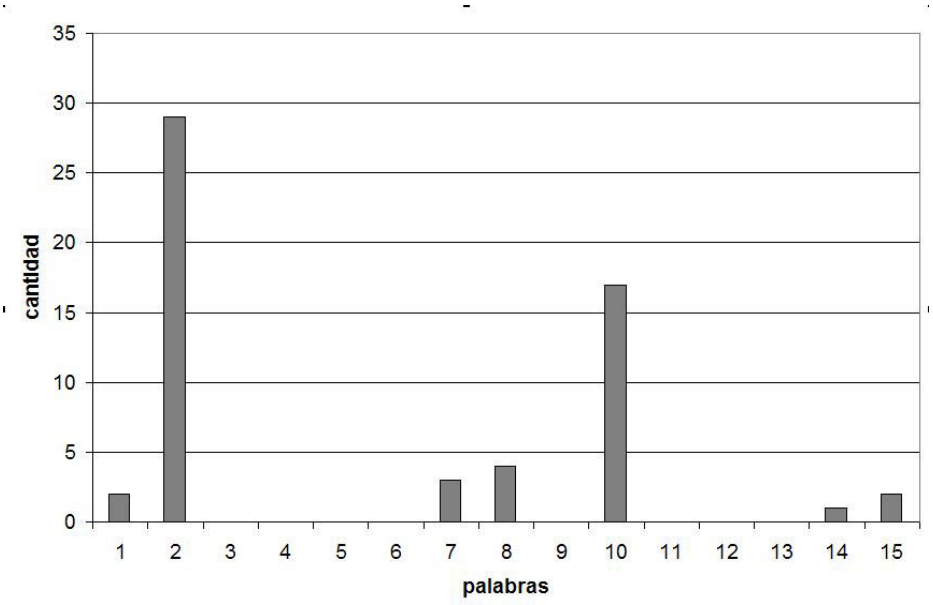
En la Tabla XLVIII se transcriben las frecuencias absolutas obtenidas para las palabras respondidas a la primera pregunta del formulario (palabras que se usarían en

el buscador). La Fig. 55 y Fig. 56 muestran el correspondiente histograma para los tipos de palabras I y II respectivamente.

Tabla XLVIII.Totales por tipo de palabra

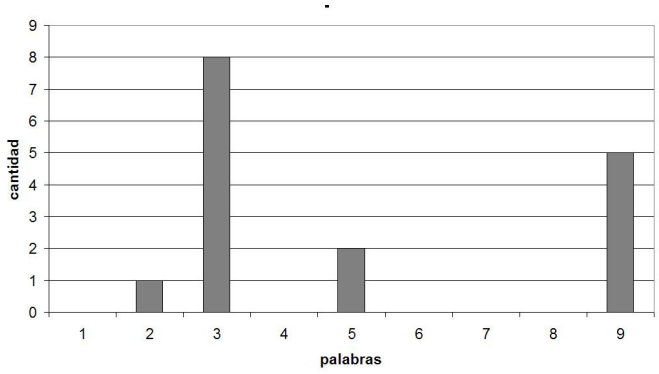
totales	
total tipo I	44
total tipo II	13
otras del texto	33
fuera del texto	18

Fig. 55. Histograma palabras tipo I



1	*orquidacea*
2	orquidal
3	planta
4	reino
5	seccion
6	semilla
7	distribucion
8	variedad
9	division
10	especie
11	familia
12	genero
13	orden
14	estudio
15	subespecie

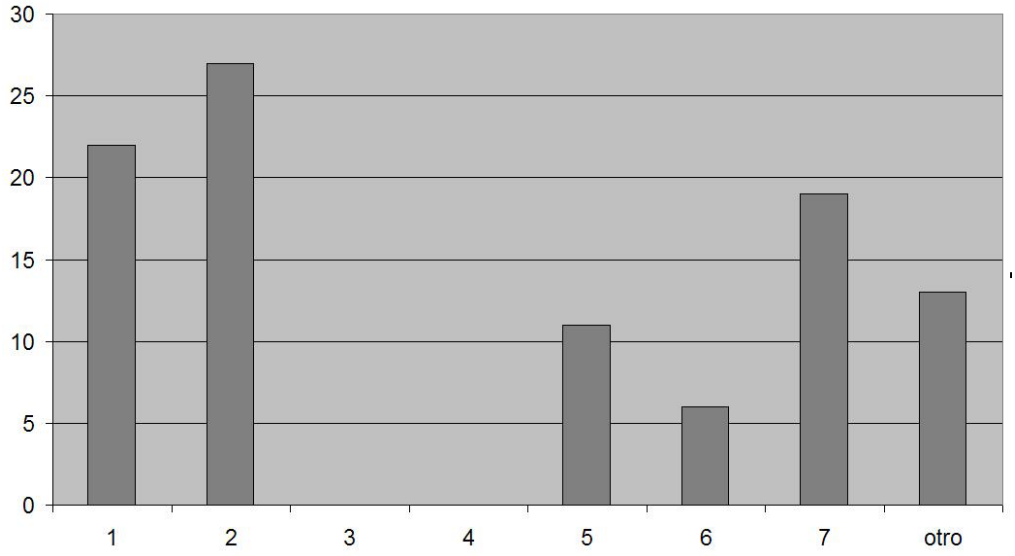
Fig. 56. Histograma palabras tipo II



1	*ser*
2	orquídeas
3	crecimiento
4	tamaño
5	distrib
6	potencial
7	semana
8	caj
9	envases

A continuación se presenta el histograma de personas que han puesto como tópico alguno de los listados. En otros se consideran los tópicos no relacionados.

Fig. 57. Histograma temas



1	*estudio de germinacion de planta orquidacea desde semilla*
2	estudio de especies, subrespecies, distribucion de los orquidales en México, variedades , clasificación científica (Reino, División, Génerp Orden, Familia, subfamilia)
3	experiencia de una persona que pretende germinar en envases de aceitunas dentro de una caja, unas orquídeas para producir keikis
4	se hace un seguimiento semanal del crecimiento
5	relata que en México hay varias especies
6	describe la distribución geográfica,nombres y usos
7	describe las especies, sus amenazas y lista especies en riesgo de extinción

Apéndice D: Análisis de datos p_o

En este apéndice se realiza el estudio numérico de p_o (con muestras). Se observará que:

-Existe una correlación progresivamente lineal entre p_o medio (promedio de p_o para todo el documento) y ***n*** (cantidad de valores p_o promediados por documento).

-El comportamiento de medianas y variabilidad de p_o es esencialmente igual para los tipos y perfiles. La única excepción es la variabilidad en los perfiles.

-El comportamiento de medianas y variabilidad de ***n*** es esencialmente distinto para los tipos y perfiles.

-Cada subgrupo de tipo de documento y perfil de narración tiene un comportamiento Binomial distinguible cuando se considera el valor p_o contra cero.

-Existe una relación entre el nivel de importancia de una sentencia y la cercanía de p_o a cero.

Tipos Texto

Las estadísticas descriptivas

Población total:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
po	150	.10	-.04	.06	-.0001	.01015	.000
n	150	636.00	2.00	638.00	43.7800	70.78134	5009.998
Valid N (listwise)	150						

En cambio de cada subconjunto son:

Literario:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
po	50	.02	-.01	.01	-.0004	.00293	.000

n	50	199.00	8.00	207.00	61.7400	41.37524	1711.911
Valid N (listwise)	50						

mensajes:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
po	50	.10	-.04	.06	.0011	.01661	.000
n	50	42.00	2.00	44.00	4.1400	5.93506	35.225
Valid N (listwise)	50						

Técnico:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
po	50	.03	-.02	.01	-.0010	.00519	.000
n	50	636.00	2.00	638.00	65.4600	105.23874	11075.192
Valid N (listwise)	50						

Se puede apreciar que los máximos y mínimos de los tres conjuntos están solapados para p_o, aunque las medias son ligeramente distintas. Los valores de n muestran también solapamientos pero con desvíos muy distintos en cada caso.

Prueba de normalidad (Shapiro-Wilks modificado) para p_o y n

Dado H0: la distribución es normal, Ha: no se puede asegurar normalidad.

class	Variable	n	Media	D.E.	W*	p (una cola)
literario	po	50	-3.5E-04	2.9E-03	0.73	<0.0001
literario	n	50	61.74	41.38	0.89	0.0005
mensajes	po	50	1.1E-03	0.02	0.66	<0.0001
mensajes	n	50	4.14	5.94	0.34	<0.0001
tecnico	po	50	-1.0E-03	0.01	0.86	<0.0001
tecnico	n	50	65.46	105.24	0.56	<0.0001

En todos los casos $p<0.05$ por lo que se puede rechazar la presunción de normalidad.

Dado que las poblaciones no siguen una distribución normal, estos valores sólo se toman como orientación para el comportamiento genérico de cada muestra.

Análisis de correlación de Pearson entre p_o y n

Es interesante ver para cada una de las categorías el valor de Pearson.

class= literario
Correlacion de Pearson: coeficientes\probabilidades

	p_o	n
po	1.00	0.43
n	0.11	1.00

class= mensajes
Correlacion de Pearson: coeficientes\probabilidades

	p_o	n
po	1.00	0.69
n	0.06	1.00

class= tecnico
Correlacion de Pearson: coeficientes\probabilidades

	p_o	n
po	1.00	0.94
n	-0.01	1.00

Esto estaría indicando que el nivel de correlación lineal entre ambas variables aumenta progresivamente para las categorías literario, mensajes y técnico, siendo significativa en este último caso (suponiendo un umbral de r=.85).

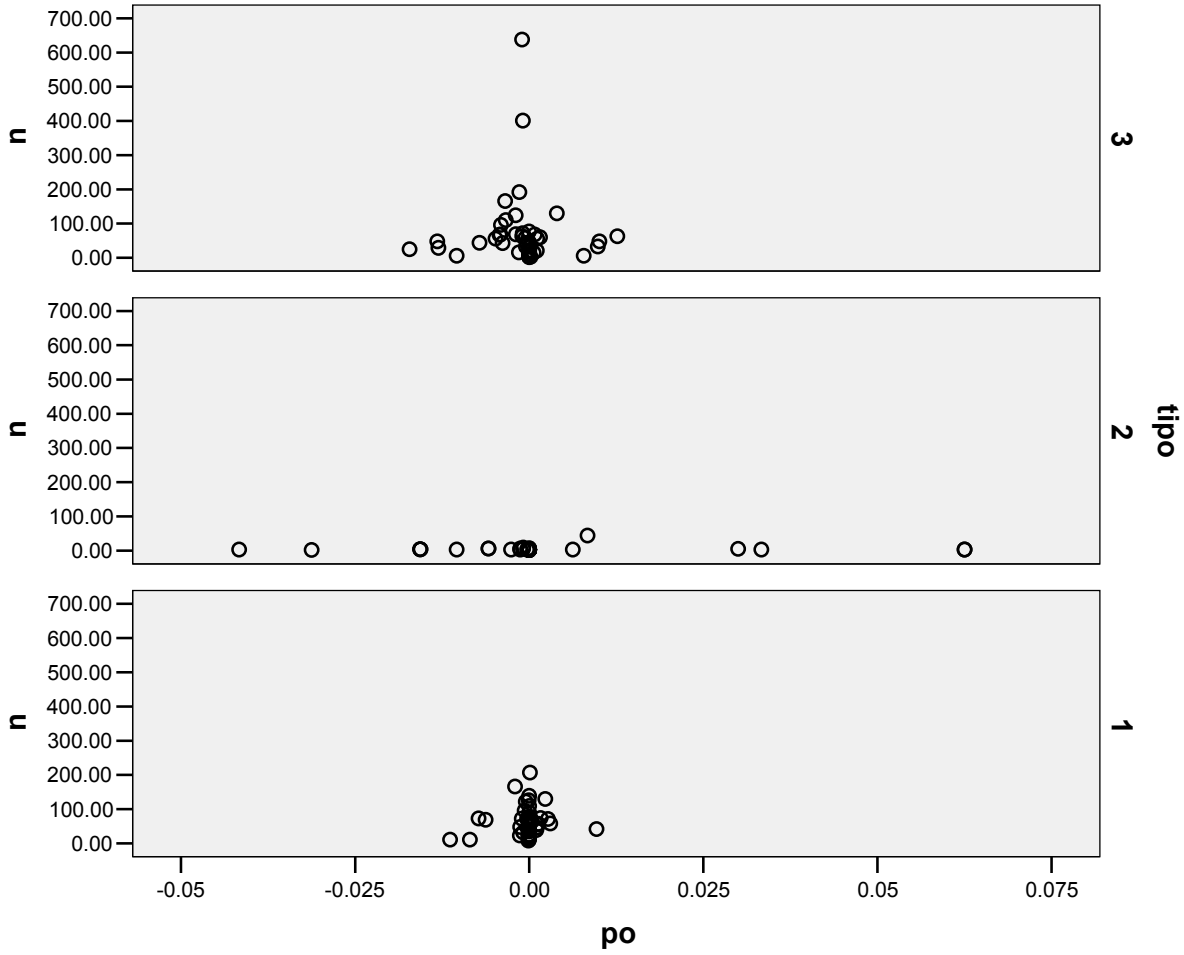
Diagramas de puntos y outliers para p_o

Inicialmente se realizó un diagrama de puntos con los tres tipos de texto considerando:

tipo	descripción

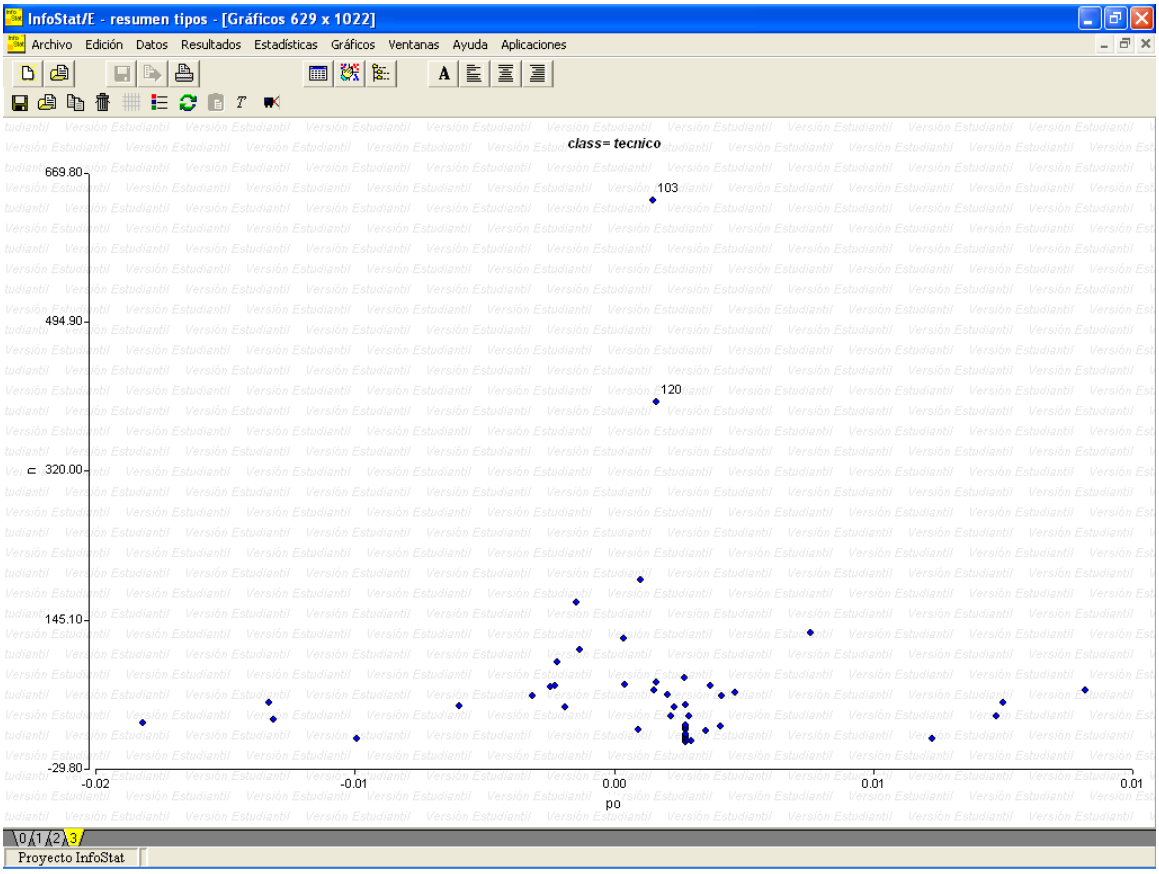
1	literario
2	mensajes
3	técnico

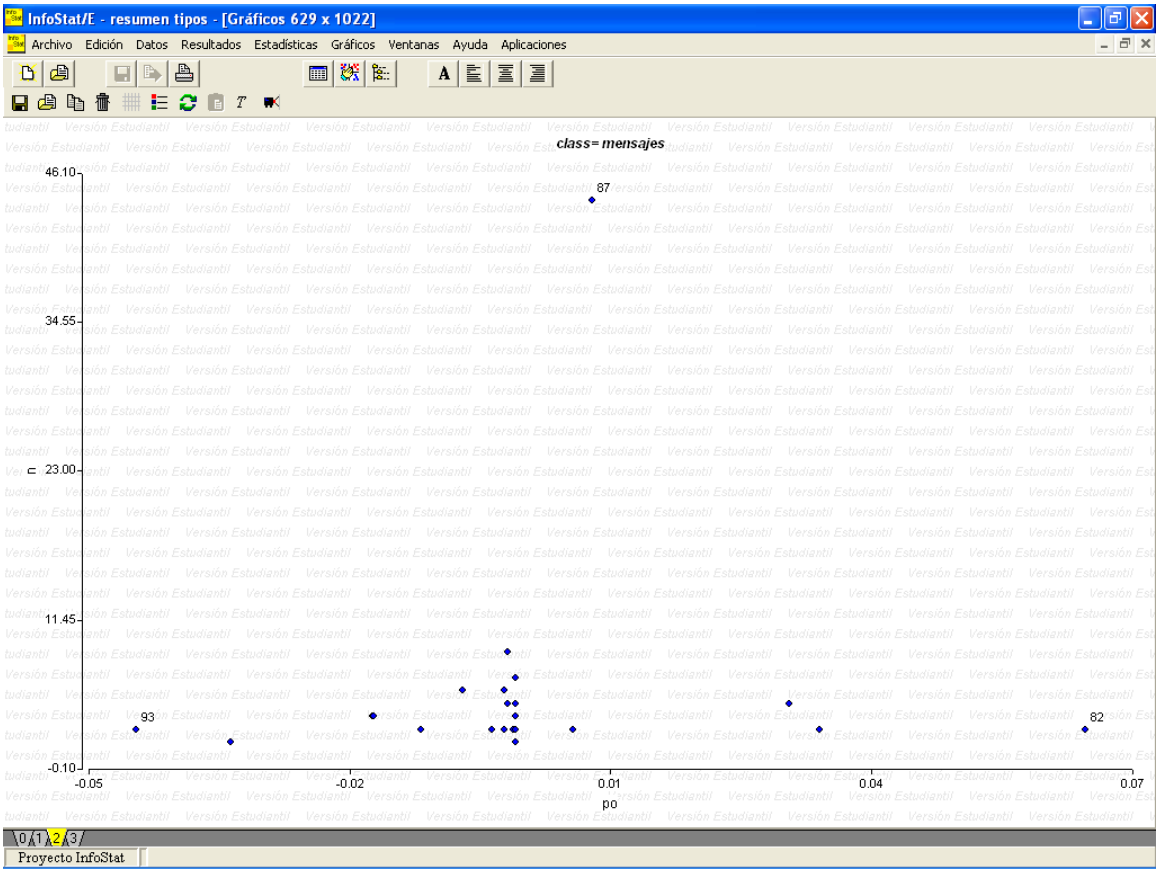
La muestra incluye 50 casos de cada uno de estos tipos mencionados, un total de 150 casos. Se obtuvo el siguiente conjunto de diagramas:

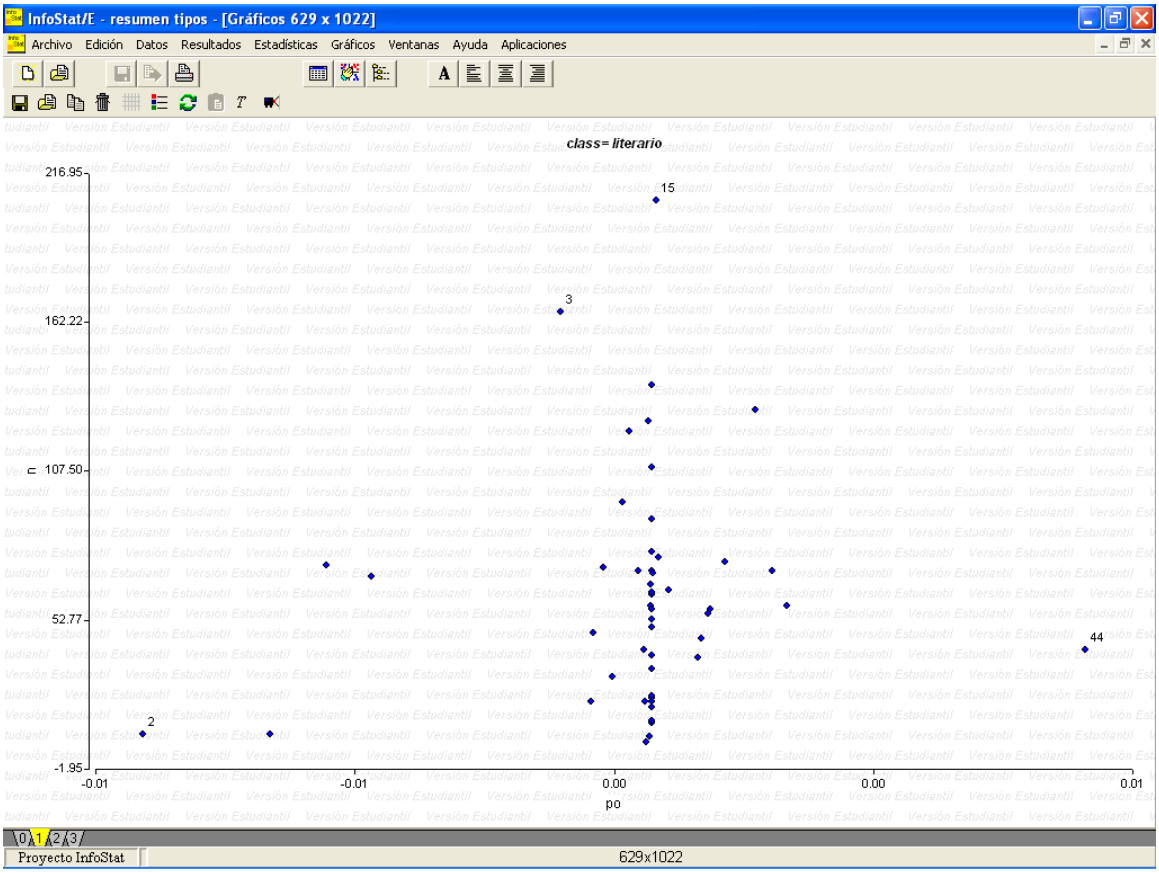


Donde n es la frecuencia de palabras especiales (EBH opuestas y contradictorias) dentro del texto y, p_o es el valor promedio de hallado de aplicar las ponderaciones respectivas según la fórmula $p_o^{i+1} = (p_o^{i} + p_o^{i+1})/2$.

Hay ciertos valores que podrían ser considerados outliers. Los mismos se identifican en las siguientes gráficas:


InfoStat/E - resumen tipos - [Gráficos 629 x 1022]
Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Ayuda Aplicaciones
class= tecnico
669.80
494.90
320.00
145.10
-29.80
103
120
-0.02
-0.01
0.00
0.01
0.01
po
Proyecto InfoStat

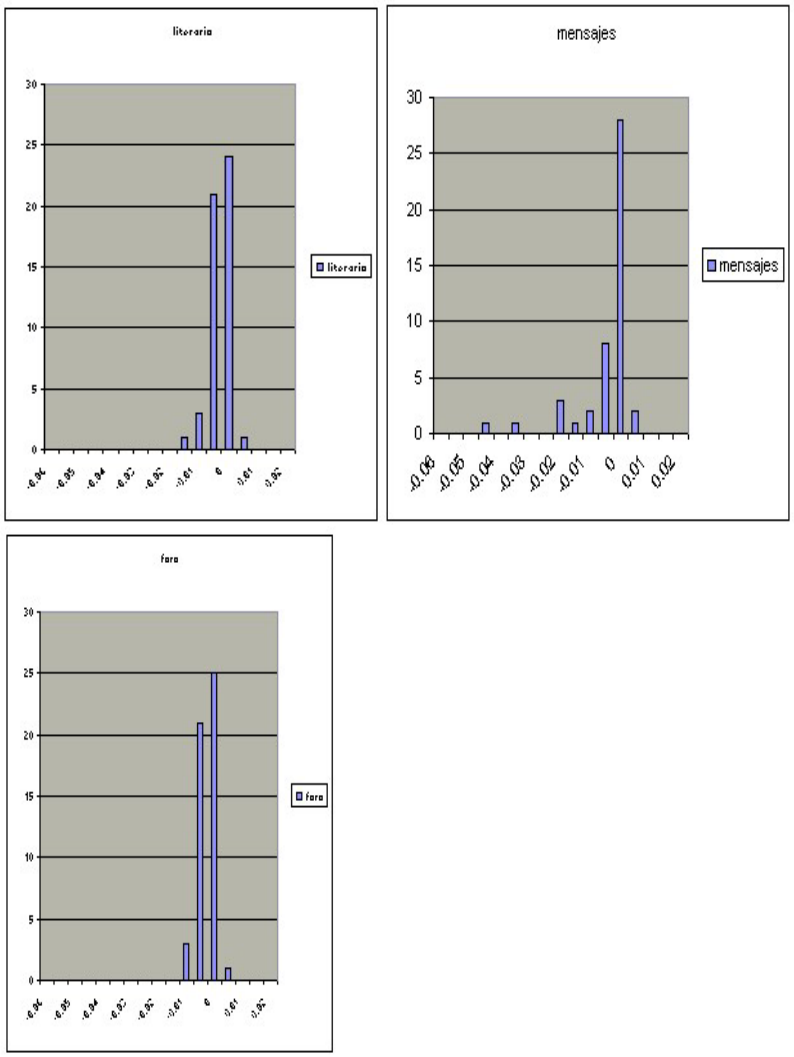

InfoStat/E - resumen tipos - [Gráficos 629 x 1022]
Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Ayuda Aplicaciones
class= mensajes
46.10
34.55
23.00
11.45
-0.10
87
93
82
-0.05
-0.02
0.01
0.04
0.07
po
Proyecto InfoStat



Sin embargo no se tratarán como outliers dado que la información procesada no permite considerarlos como tales con justificación razonable y que se considera podría corresponderse con el tipo de comportamiento poblacional.

Histogramas de frecuencias para p_o

Se calculó el p_o promedio para cada documento, luego se levantaron los siguientes histogramas.



Los gráficos muestran una clara asimetría en todos los casos.

Dado que las muestras no tienen formas similares, su distribución (varianza poblacional) no será similar y por lo tanto no se podrán aplicar tests no paramétricos como Kruskal-Wallis.

Estudio de variabilidad Kruskal-Wallis para p_o

Anova no paramétrico. Se usa porque no respeta precondiciones de t y ANOVA (cantidad de casos, normalidad, varianzas similares). Además las muestras tienen desigualdad notable de varianzas y distribución muy distinta a normal. Toma como hipótesis nula que las muestras son de la misma población. Los resultados sobre los p_o promedio de cada tipo son:

Ranks

	tipo	N	Mean Rank
po	1	50	79.75
	2	50	75.62
	3	50	71.13
	Total	150	

Test Statistics

	po
Chi-Square	1.014
df	2
Asymp. Sig.	.602

a Kruskal Wallis Test

b Grouping Variable: tipo

Como puede verse, la significación es 0.602 > 0.05, y no puede afirmarse que las poblaciones tienen variabilidad similar.

Estudio de medianas poblacionales para p_o

Anova no paramétrico. Se usa porque no respeta precondiciones de t y ANOVA (cantidad de casos, normalidad, varianzas. Se usa para verificar si las medianas son iguales en las poblaciones de los subgrupos.

Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
po	150	-.0009	.0000	.0000
tipo	150	1.00	2.00	3.00

Frequencies

		tipo		
		1	2	3
po	> Median	18	8	15
	<= Median	32	42	35

La mediana poblacional se ubica en $p_o = 0.0$, y puede observarse que para los 3 grupos la cantidad de individuos en ningún caso se acerca al 50%.

Test Statistics(b)

	po
N	150
Median	.0000
Chi-Square	5.303(a)
df	2
Asymp. Sig.	.071

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 13.7.

b Grouping Variable: tipo

El valor da una significación 0.071 > 0.05, con lo que no puede rechazarse la hipótesis original y por lo tanto p_o no tiene una mediana poblacional distinta en cada subgrupo.

Estudio de medianas y variabilidad Kruskal Wallis para n

En cuanto a las medianas, los valores para n son:

Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
n	150	3.0000	22.5000	63.0000
tipo	150	1.00	2.00	3.00

El valor de mediana es 22.5: Para este valor la cantidad de registros que cumplen con ella son:

Frequencies

		tipo		
		1	2	3
n	> Median	43	1	31
	<= Median	7	49	19

con un estadístico:

Test Statistics(b)

	n
N	150
Median	22.5000
Chi-Square	74.880(a)
df	2
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0.

b Grouping Variable: tipo

Donde p = 0.0 indica que realmente se trata de poblaciones distintas según el subgrupo. Esto indicaría que n es buen discriminante de tipos de archivo.

Se realizó el estudio de variabilidad Kruskal Wallis para n.

Ranks

	tipo	N	Mean Rank
n	1	50	105.96
	2	50	28.48
	3	50	92.06
	Total	150	

Test Statistics(a,b)

	n
Chi-Square	90.848
df	2
Asymp. Sig.	.000

a Kruskal Wallis Test

b Grouping Variable: tipo

Indicaría que la variabilidad es distinta en cada subgrupo.

Por lo tanto las poblaciones tienen un comportamiento que permitiría afirmar que son tres subgrupos distintos.

Estudio de p_o a nivel documento

Es interesante estudiar el comportamiento de p_o en relación con los 3 subgrupos de tipos de documento. Tomando el valor promedio de p_o calculado para cada frase (tal como se lo usa en WIH para ponderar los EBH) y se los clasifica como menores,

mayores e iguales a cero, es posible observar un comportamiento perfectamente distinguible por cada uno de los subgrupos. A continuación se muestran los resultados de estimar la curva de comportamiento como Binomial y la bondad de ajuste resultante en cada caso.

Las columnas respectivamente indican:

class: subgrupo

variable: conteo considerado

Clase: subclase dentro del grupo

LI: límite inferior

LS: límite superior

MC: marca de clase

FA: frecuencias absolutas observadas

FR: frecuencias relativas

E (FA): frecuencias absolutas según el modelo distribución al propuesto

E (FR): frecuencias relativas según el modelo distribucional propuesto

Chi-cuadrado: valor de estadístico Chi.

p: prueba de bondad de ajuste

(i) Como Binomial

Dadas las características del experimento se puede decir que la distribución poblacional del mismo será Binomial (también llamada de Bernoulli), donde se considera que:

- En cada prueba del experimento sólo son posibles dos resultados: el suceso A (éxito) y su contrario $\bar{A}$ (fracaso). Ej: la variable MENOR_CERO tiene evento A=el p_o promedio del documento es menor a 0, y $\bar{A}$ =el p_o promedio del documento no es menor a 0.
- El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente, dado que se trata de documentos distintos.
- La probabilidad del suceso A es constante (dado que se trata de la misma subpoblación) representada por p, y no varía de una prueba a otra (se mostró en las secciones Estudio de variabilidad Kruskal-Wallis para y Estudio de medianas poblacionales para , que el comportamiento poblacional de p_o como variable

contínua no varía en el análisis no paramétrico de varianzas y de la mediana). La probabilidad de $\bar{A}$ es 1- p y se representa con q .

- El experimento consta de un número $n=50$ de pruebas.

Los parámetros de la distribución Binomial son tres:

media: $\mu = n.p$

varianza: $\sigma^2 = n.p.q$

Desv. Típica: $\sigma = \sqrt{n.p.q}$

Siendo la función de la variable Binomial:

$$F(x_i) = p(X \le x_i) = \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \ldots + \binom{n}{k} p^k q^{n-k}$$

Siendo k el mayor número entero menor o igual a x_i. Esta función de distribución proporciona, para cada número real x_i, la probabilidad de que la variable X tome valores menores o iguales que x_i.

A continuación se presentan las tablas de prueba de bondad de ajuste donde los coeficientes p (bondad de ajuste), permiten indicar que se trata de una distribución Binomial.

Ajuste: Binomial con estimación de parámetros: p= 0.00720 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	mayor_cero	1	0.00	0.20	0.10	32	0.64	34.84	0.70	0.23	
literario	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.23	
literario	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.23	
literario	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.23	
literario	mayor_cero	5	0.80	1.00	0.90	18	0.36	15.16	0.30	0.76	**0.9434**

Ajuste: Binomial con estimación de parámetros: p= 0.01000 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	menor_cero	1	0.00	0.20	0.10	25	0.50	30.25	0.61	0.91	

literario	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.91	
literario	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.91	
literario	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.91	
literario	menor_cero	5	0.80	1.00	0.90	25	0.50	19.75	0.39	2.31	0.6795

Ajuste: Binomial con estimación de parámetros: p= 0.00280 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	igual_cero	1	0.00	0.20	0.10	43	0.86	43.46	0.87	4.9E-03	
literario	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	4.9E-03	
literario	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	4.9E-03	
literario	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	4.9E-03	
literario	igual_cero	5	0.80	1.00	0.90	7	0.14	6.54	0.13	0.04	0.9998

Ajuste: Binomial con estimación de parámetros: p= 0.00320 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	mayor_cero	1	0.00	0.20	0.10	42	0.84	42.60	0.85	0.01	
mensajes	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	5	0.80	1.00	0.90	8	0.16	7.40	0.15	0.06	0.9996

Ajuste: Binomial con estimación de parámetros: p= 0.00640 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	menor_cero	1	0.00	0.20	0.10	34	0.68	36.27	0.73	0.14	
mensajes	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.14	
mensajes	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.14	
mensajes	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.14	
mensajes	menor_cero	5	0.80	1.00	0.90	16	0.32	13.73	0.27	0.52	0.9718

Ajuste: Binomial con estimación de parámetros: p= 0.01040 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	igual_cero	1	0.00	0.20	0.10	24	0.48	29.65	0.59	1.07	
mensajes	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.07	
mensajes	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.07	
mensajes	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.07	
mensajes	igual_cero	5	0.80	1.00	0.90	26	0.52	20.35	0.41	2.64	0.6196

Ajuste: Binomial con estimación de parámetros: p= 0.00600 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p

tecnico	mayor_cero	1	0.00	0.20	0.10	35	0.70	37.01	0.74	0.11	
tecnico	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	5	0.80	1.00	0.90	15	0.30	12.99	0.26	0.42	0.9809

Ajuste: Binomial con estimación de parámetros: p= 0.01000 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	menor_cero	1	0.00	0.20	0.10	25	0.50	30.25	0.61	0.91	
tecnico	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.91	
tecnico	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.91	
tecnico	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.91	
tecnico	menor_cero	5	0.80	1.00	0.90	25	0.50	19.75	0.39	2.31	0.6795

Ajuste: Binomial con estimación de parámetros: p= 0.00400 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	igual_cero	1	0.00	0.20	0.10	40	0.80	40.92	0.82	0.02	
tecnico	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	5	0.80	1.00	0.90	10	0.20	9.08	0.18	0.11	0.9984

(ii) Como Poisson

El experimento que la genera debe cumplir las siguientes condiciones:

1. El número de éxitos que ocurren en cada región del tiempo o del espacio es independiente de lo que ocurra en cualquier otro tiempo o espacio disjunto del anterior.
2. La probabilidad de un éxito en un tiempo o espacio pequeño es proporcional al tamaño de este y no depende de lo que ocurra fuera de él.
3. La probabilidad de encontrar uno o más éxitos en una región del tiempo o del espacio tiende a cero a medida que se reducen las dimensiones de la región en estudio.

Como consecuencia de estas condiciones, las variables Poisson típicas son variables en las que se cuentan sucesos raros. La función de probabilidad de una variable Poisson es:

$$p(x,\lambda) = \frac{e^{-\lambda}.\lambda^{x}}{x!}, x = 0,1,...\infty, e = 2.71828$$

El parámetro de la distribución es λ que es igual a la media y a la varianza de la variable: $\lambda = \mu = \sigma^2$. Esta característica puede servir para identificar a una variable Poisson en casos en que se presenten serias dificultades para verificar los postulados de definición.

La distribución de Poisson se puede considerar como el límite al que tiende la distribución Binomial cuando n tiende a ∞ y p tiende a 0, siendo n x p constante (y menor que 7); en esta situación sería difícil calcular probabilidades en una variable Binomial y, por tanto, se utiliza una aproximación a través de una variable Poisson con media $\lambda = n.p$.

Se observará que, efectivamente, se cumple esta relación con los datos del caso ya que (ver las tablas de frecuencias correspondientes). Por ejemplo, para el subgrupo literario, cuando p_o es MAYOR_CERO, la Binomial tiene n=0.0072 y n=50. En correspondencia el valor de λ estimado para Poisson es 0.36, que coincide con el producto de n y p Binomiales.

La varianza de la variable aproximada es ligeramente superior a la de la variable Binomial:

$$\sigma^2 = n.p.(1-p) = (1-p).\mu = (1-p).\lambda = (1-p).\sigma_p^2$$

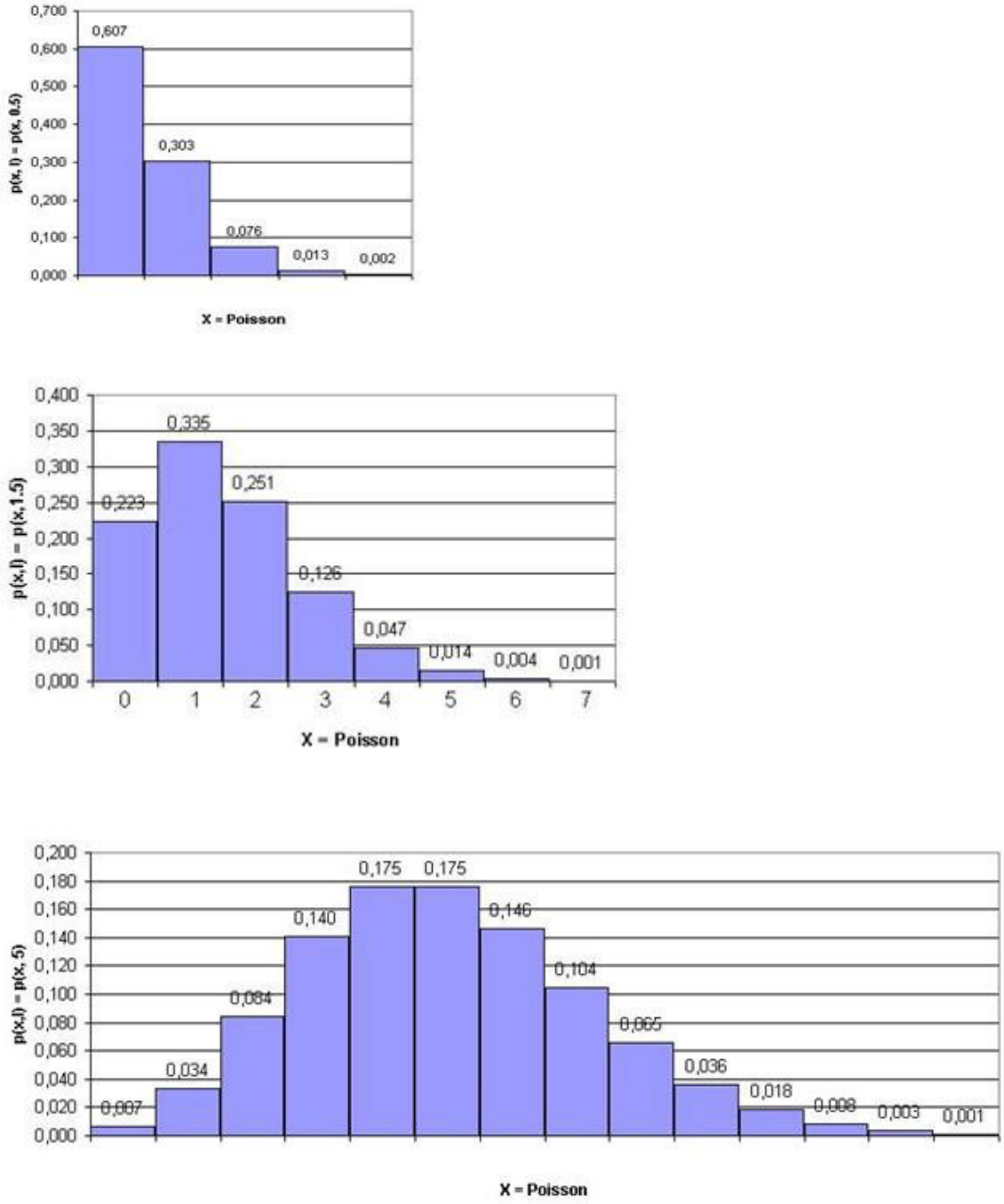
Si $p \approx 0 \Rightarrow 1-p \approx 1 \Rightarrow \sigma_B^2 \approx \sigma_P^2$

Las variables Poisson cumplen la propiedad de que la suma de variables Poisson independientes es otra Poisson con media igual a la suma las medias.

El aspecto de la distribución depende muchísimo de la magnitud de la media. Como ejemplo, en la Fig. 58 se muestran tres casos con λ = 0,5 (arriba a la izquierda), λ = 1,5 (arriba a la derecha) y λ = 5 (abajo) Obsérvese

que la asimetría de la distribución disminuye al crecer λ y que, en paralelo, la gráfica empieza a tener un aspecto acampanado.

Fig. 58. Histograma Poisson con λ=0.5, 1.5 y 5



A continuación se presentan los resultados numéricos que muestran que Poisson es una aproximación buena pero no tanto como Binomial ya que los

valores de p son permanentemente menores que los correspondientes a los de ésta.

Tablas de frecuencias Poisson

Ajuste: Poisson con estimación de parámetros: Lambda= 0.36000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	mayor_cero	1	0.00	0.20	0.10	32	0.64	34.88	0.70	0.24	
literario	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.24	
literario	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.24	
literario	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.24	
literario	mayor_cero	5	0.80	1.00	0.90	18	0.36	15.12	0.30	0.79	0.8522

Ajuste: Poisson con estimación de parámetros: Lambda= 0.50000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	menor_cero	1	0.00	0.20	0.10	25	0.50	30.33	0.61	0.94	
literario	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.94	
literario	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.94	
literario	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.94	
literario	menor_cero	5	0.80	1.00	0.90	25	0.50	19.67	0.39	2.38	0.4978

Ajuste: Poisson con estimación de parámetros: Lambda= 0.14000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	igual_cero	1	0.00	0.20	0.10	43	0.86	43.47	0.87	0.01	
literario	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.01	
literario	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.01	
literario	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.01	
literario	igual_cero	5	0.80	1.00	0.90	7	0.14	6.53	0.13	0.04	0.9980

Ajuste: Poisson con estimación de parámetros: Lambda= 0.16000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	mayor_cero	1	0.00	0.20	0.10	42	0.84	42.61	0.85	0.01	
mensajes	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.01	
mensajes	mayor_cero	5	0.80	1.00	0.90	8	0.16	7.39	0.15	0.06	0.9963

Ajuste: Poisson con estimación de parámetros: Lambda= 0.32000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	menor_cero	1	0.00	0.20	0.10	34	0.68	36.31	0.73	0.15	
mensajes	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.15	
mensajes	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.15	
mensajes	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.15	
mensajes	menor_cero	5	0.80	1.00	0.90	16	0.32	13.69	0.27	0.54	0.9110

Ajuste: Poisson con estimación de parámetros: Lambda= 0.52000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	igual_cero	1	0.00	0.20	0.10	24	0.48	29.73	0.59	1.10	
mensajes	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.10	
mensajes	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.10	
mensajes	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.10	
mensajes	igual_cero	5	0.80	1.00	0.90	26	0.52	20.27	0.41	2.72	0.4368

Ajuste: Poisson con estimación de parámetros: Lambda= 0.30000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	mayor_cero	1	0.00	0.20	0.10	35	0.70	37.04	0.74	0.11	
tecnico	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.11	
tecnico	mayor_cero	5	0.80	1.00	0.90	15	0.30	12.96	0.26	0.43	0.9332

Ajuste: Poisson con estimación de parámetros: Lambda= 0.50000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	menor_cero	1	0.00	0.20	0.10	25	0.50	30.33	0.61	0.94	
tecnico	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.94	
tecnico	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.94	
tecnico	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.94	
tecnico	menor_cero	5	0.80	1.00	0.90	25	0.50	19.67	0.39	2.38	0.4978

Ajuste: Poisson con estimación de parámetros: Lambda= 0.20000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	igual_cero	1	0.00	0.20	0.10	40	0.80	40.94	0.82	0.02	
tecnico	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.02	
tecnico	igual_cero	5	0.80	1.00	0.90	10	0.20	9.06	0.18	0.12	0.9896

(iii) Aproximación normal

Al considerar distribuciones Binomiales, tipo B(n,p), para un mismo valor de p y valores de n cada vez mayores, se ve que sus polígonos de frecuencias se aproximan a una curva en "forma de campana".

Dados los valores inferidos para po, no es posible afirmar que esta aproximación a normalidad tenga lugar. De hecho las pruebas realizadas confirman que tienen comportamiento claramente Binomial, con cierta tendencia a parecerse a una Poisson.

A continuación se reproducen las tablas de prueba de normalidad. Todos los valores p<<0.5, confirmando lo afirmado previamente.

Tablas de frecuencias

Ajuste: Normal con estimación de parámetros: Media= 0.36000 y varianza= 0.23510

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuad	p
literario	mayor_cero	1	0.00	0.20	0.10	32	0.64	18.54	0.37	9.78	
literario	mayor_cero	2	0.20	0.40	0.30	0	0.00	8.11	0.16	17.89	
literario	mayor_cero	3	0.40	0.60	0.50	0	0.00	7.84	0.16	25.73	
literario	mayor_cero	4	0.60	0.80	0.70	0	0.00	6.41	0.13	32.14	
literario	mayor_cero	5	0.80	1.00	0.90	18	0.36	9.10	0.18	40.83	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.50000 y varianza= 0.25510

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	menor_cero	1	0.00	0.20	0.10	25	0.50	13.81	0.28	9.06	
literario	menor_cero	2	0.20	0.40	0.30	0	0.00	7.26	0.15	16.32	
literario	menor_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	24.17	
literario	menor_cero	4	0.60	0.80	0.70	0	0.00	7.26	0.15	31.43	
literario	menor_cero	5	0.80	1.00	0.90	25	0.50	13.81	0.28	40.49	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.14000 y varianza= 0.12286

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
literario	igual_cero	1	0.00	0.20	0.10	43	0.86	28.40	0.57	7.51	
literario	igual_cero	2	0.20	0.40	0.30	0	0.00	10.15	0.20	17.65	
literario	igual_cero	3	0.40	0.60	0.50	0	0.00	6.72	0.13	24.38	
literario	igual_cero	4	0.60	0.80	0.70	0	0.00	3.24	0.06	27.62	
literario	igual_cero	5	0.80	1.00	0.90	7	0.14	1.49	0.03	47.94	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.16000 y varianza= 0.13714

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	mayor_cero	1	0.00	0.20	0.10	42	0.84	27.15	0.54	8.12	
mensajes	mayor_cero	2	0.20	0.40	0.30	0	0.00	9.93	0.20	18.05	
mensajes	mayor_cero	3	0.40	0.60	0.50	0	0.00	7.05	0.14	25.10	
mensajes	mayor_cero	4	0.60	0.80	0.70	0	0.00	3.77	0.08	28.87	
mensajes	mayor_cero	5	0.80	1.00	0.90	8	0.16	2.10	0.04	45.46	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.32000 y varianza= 0.22204

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	menor_cero	1	0.00	0.20	0.10	34	0.68	19.97	0.40	9.85	
mensajes	menor_cero	2	0.20	0.40	0.30	0	0.00	8.40	0.17	18.24	
mensajes	menor_cero	3	0.40	0.60	0.50	0	0.00	7.82	0.16	26.06	
mensajes	menor_cero	4	0.60	0.80	0.70	0	0.00	6.10	0.12	32.16	
mensajes	menor_cero	5	0.80	1.00	0.90	16	0.32	7.71	0.15	41.08	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.52000 y varianza= 0.25469

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
mensajes	igual_cero	1	0.00	0.20	0.10	24	0.48	13.15	0.26	8.95	
mensajes	igual_cero	2	0.20	0.40	0.30	0	0.00	7.15	0.14	16.10	
mensajes	igual_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	23.95	
mensajes	igual_cero	4	0.60	0.80	0.70	0	0.00	7.38	0.15	31.32	
mensajes	igual_cero	5	0.80	1.00	0.90	26	0.52	14.48	0.29	40.50	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.30000 y varianza= 0.21429

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	mayor_cero	1	0.00	0.20	0.10	35	0.70	20.72	0.41	9.83	
tecnico	mayor_cero	2	0.20	0.40	0.30	0	0.00	8.55	0.17	18.39	
tecnico	mayor_cero	3	0.40	0.60	0.50	0	0.00	7.80	0.16	26.19	
tecnico	mayor_cero	4	0.60	0.80	0.70	0	0.00	5.92	0.12	32.11	
tecnico	mayor_cero	5	0.80	1.00	0.90	15	0.30	7.00	0.14	41.24	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.50000 y varianza= 0.25510

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	menor_cero	1	0.00	0.20	0.10	25	0.50	13.81	0.28	9.06	
tecnico	menor_cero	2	0.20	0.40	0.30	0	0.00	7.26	0.15	16.32	
tecnico	menor_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	24.17	
tecnico	menor_cero	4	0.60	0.80	0.70	0	0.00	7.26	0.15	31.43	
tecnico	menor_cero	5	0.80	1.00	0.90	25	0.50	13.81	0.28	40.49	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.20000 y varianza= 0.16327

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
tecnico	igual_cero	1	0.00	0.20	0.10	40	0.80	25.00	0.50	9.00	
tecnico	igual_cero	2	0.20	0.40	0.30	0	0.00	9.48	0.19	18.48	
tecnico	igual_cero	3	0.40	0.60	0.50	0	0.00	7.46	0.15	25.95	
tecnico	igual_cero	4	0.60	0.80	0.70	0	0.00	4.62	0.09	30.56	
tecnico	igual_cero	5	0.80	1.00	0.90	10	0.20	3.44	0.07	43.08	<0.0001

(iv) Conclusiones

Un valor de p < valor de significación nominal de la prueba conduce a un rechazo del modelo distribucional propuesto. En todos los casos que se presentan valores de p que permiten indicar que las muestras son modelizables como Binomiales con los parámetros especificados en el encabezado de cada tabla.

Las distribuciones tienen significación respectivamente superior para Binomiales que para Poisson, en cada subgrupo. Es posible que en el caso de Poisson, la curva en estos rangos y parámetros sea similar a la Binomial, pero no represente el comportamiento real poblacional.

En consecuencia podría afirmarse que las distribuciones son Binomiales con las siguientes características:

tipo texto	rango	p
literario	>0	0.00720
	<0	0.01000
	=0	0.00280
mensajes	>0	0.00320
	<0	0.00640
	=0	0.01040
tecnico	>0	0.00600
	<0	0.01000
	=0	0.00400

Estudio de p_o al nivel de significación

De los diversos tipos de archivos, se puede decir, basándose en los resultados previos, que cada uno tiene características propias en cuanto a cantidad de sentencias típicas, y distribución de p_o. Por ese motivo se seleccionaron dos muestras para estudiar el nivel de significación de frases en relación con p_o. Una muestra es el subgrupo de mensajes y la otra es el perfil "documento", que se estudia en la sección Estudio de po a nivel de significación de forma análoga.

Inicialmente se observaron todos los documentos. Hay un porcentaje que presenta el 100% de valores p_o en cero (ver Fig. 59). Estos casos corresponden a respuestas cortas a un mensaje anterior.

Fig. 59. Mensajes con 100% de valores p_o en cero

De los documentos con valores de p_o distintos de cero, se tomaron los valores de p_o mínimo y máximo para cada documento. No se consideraron otros valores cercanos (en la práctica es necesario el uso de los mismos como se muestra en 5.4. Práctica de la estrategia con po). Se extrajeron las frases correspondientes a estos p_o seleccionados y se estudió si la misma se relaciona al tema principal o a un tema secundario del mensaje. El criterio seguido para considerar las frases es:

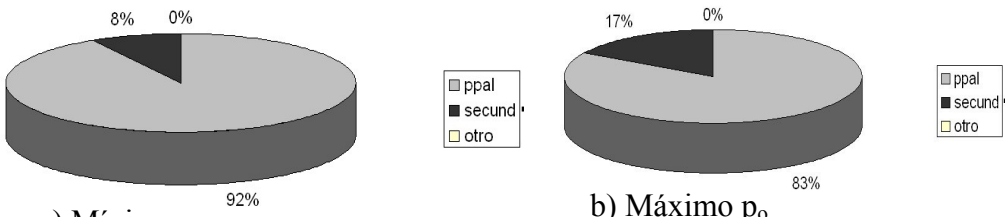
-Si el mensaje en estudio parte de una pregunta inicial: la pregunta en sí misma es considerada el tema principal

-Si no hay pregunta inicial en el mensaje: el tópico descrito en el mensaje es el tema principal.

-Las frases que no atiendan al tema principal son consideradas secundarias.

En la Fig. 60a) se muestran los porcentajes como gráfico de tortas cuando se consideran los p_o mínimos. En la parte b) se muestra el correspondiente a los p_o máximos.

Fig. 60. Significado de las sentencias con mínimo y máximo p_o

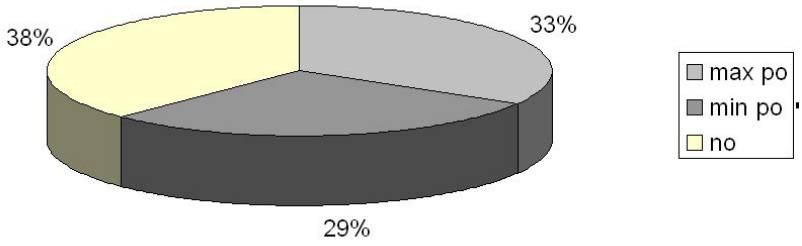


a) Mínimo p_o b) Máximo p_o

Puede apreciarse que en ambos casos el tema principal se ve reflejado y en parte el/los tema(s) secundario(s) también. No se detectaron frases de otro tipo.

Un fenómeno especial es la aparición del tema principal en la primera frase del texto. En el caso de los mensajes tiene una frecuencia muy alta que no se observa, por ejemplo en los documentos estudiados en la sección de perfiles de narración (Estudio de po a nivel de significación). En la Fig. 61 se aprecia que casi la mitad de los casos se trata de primera frase.

Fig. 61. Primer sentencia.



Perfiles narración

Estadísticas descriptivas

De la población:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
po	200	-.050512	.015000	-.00120462	.004538066
n	200	3	2060	147.84	229.440
Valid N (listwise)	200				

En cambio para cada uno de los subconjuntos es:

foro:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
po	50	-.008890	.005500	-.00067571	.002348381
n	50	10	329	80.00	63.025
Valid N (listwise)	50				

web index:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
po	50	-.015625	.000610	-.00090731	.002630066
n	50	3	2060	167.54	336.362
Valid N (listwise)	50				

doc:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
po	50	-.008899	.001686	-.00120591	.001995082
n	50	16	1227	246.08	273.518
Valid N (listwise)	50				

blog:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation

po	50	-.050512	.015000	-.00202953	.008133543
n	50	5	329	97.74	66.093
Valid N (listwise)	50				

Puede observarse que existe solapamiento en los valores de p_o y de n, pero los desvíos estándar son bastante distintos. De todas maneras, dado que la distribución conjunta de valores no es normalizada estos valores sólo son tomados como dato orientativo de las aparentes diferencias entre las muestras. Dado que existen solapamientos no se puede realizar análisis de varianzas.

Dado que n no es una variable continua no se puede hacer análisis discriminante.

Prueba de normalidad (Shapiro-Wilks modificado) para p_o y n

Shapiro-Wilks (modificado)

class	Variable	n	Media	D.E.	W*	p (una cola)
blog	po	50	-2.0E-03	0.01	0.64	<0.0001
blog	n	50	97.74	66.09	0.92	0.0108
doc	po	50	-1.2E-03	2.0E-03	0.77	<0.0001
doc	n	50	246.08	273.52	0.78	<0.0001
foro	po	50	-6.8E-04	2.3E-03	0.85	<0.0001
foro	n	50	80.00	63.03	0.86	<0.0001
webindex	po	50	-9.1E-04	2.6E-03	0.47	<0.0001
webindex	n	50	167.54	336.36	0.51	<0.0001

En todos los casos p<0.05 por lo que se puede rechazar la presunción de normalidad.

Dado que las poblaciones no siguen una distribución normal, estos valores sólo se toman como orientación para el comportamiento genérico de cada muestra. Dado que n no es una variable continua no se puede hacer análisis discriminante.

Análisis de correlación de Pearson entre p_o y n

class= blog

Correlacion de Pearson: coeficientes\probabilidades

	p_o	n

```
po     1.00 0.97
n     -0.01 1.00
```

class= doc

Correlacion de Pearson: coeficientes\probabilidades

```
        po     n
po     1.00 0.44
n     -0.11 1.00
```

class= foro

Correlacion de Pearson: coeficientes\probabilidades

```
        po     n
po     1.00 0.49
n     -0.10 1.00
```

class= webindex

Correlacion de Pearson: coeficientes\probabilidades

```
        po     n
po    1.00  0.84
n     0.03  1.00
```

Es interesante ver para cada una de las categorías el valor de Pearson aumenta desde las clases doc, y foro a webindex y blog, siendo en el último caso significativo (considerando un umbral de 0.85).

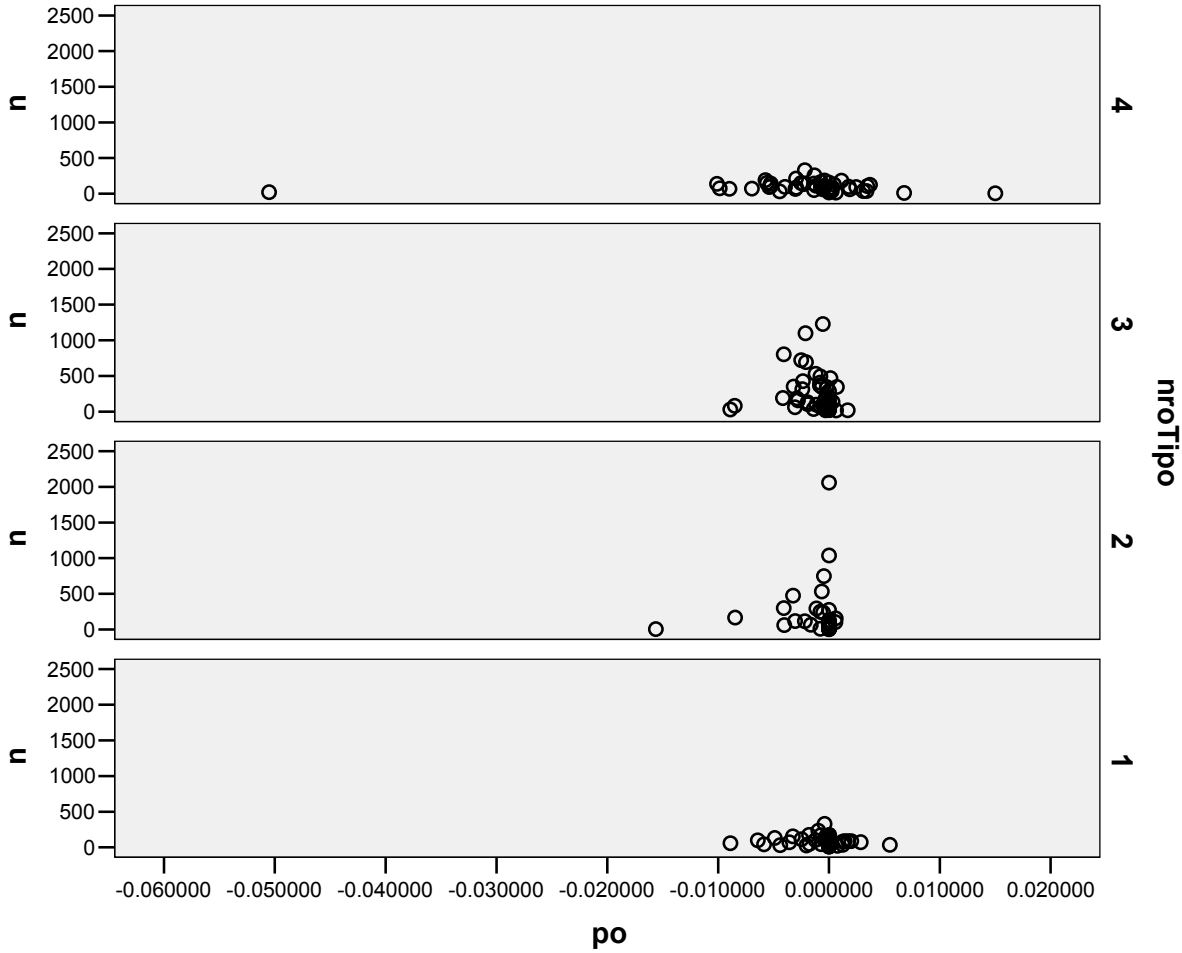
Diagramas de puntos y outliers para p_o

Inicialmente se realizó un diagrama de puntos con los tres tipos de texto considerando:

tipo	descripción
1	foro
2	webindex
3	doc

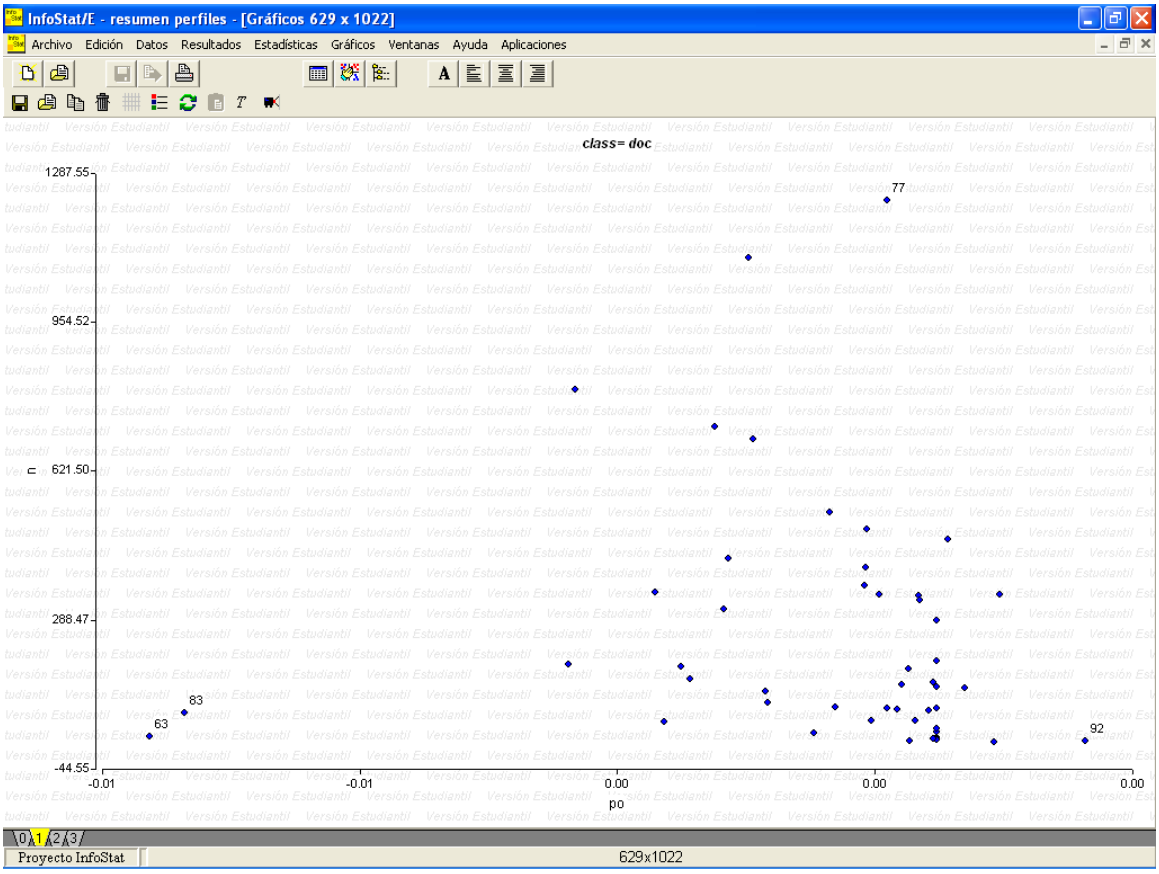
4	blog

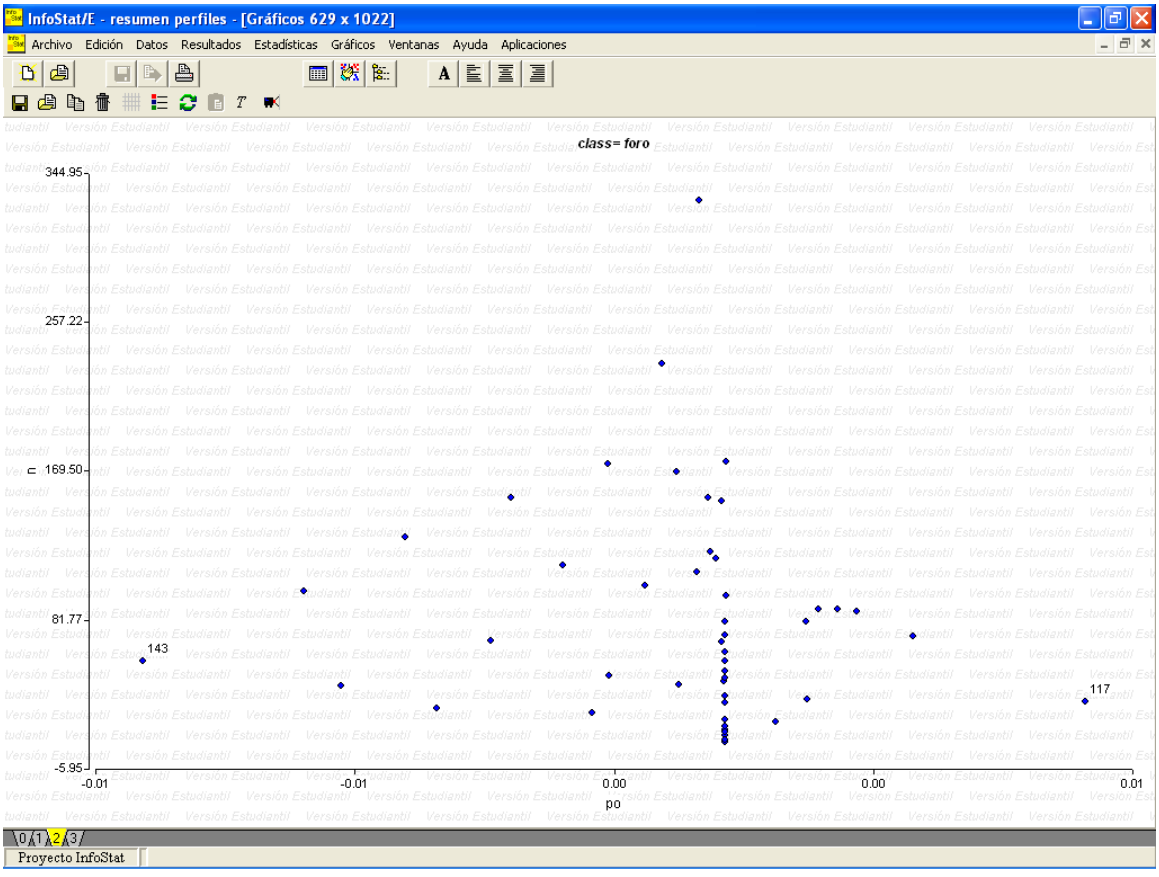
La muestra incluye 50 casos de cada uno de estos tipos mencionados, un total de 200 casos. Se obtuvo el siguiente conjunto de diagramas:

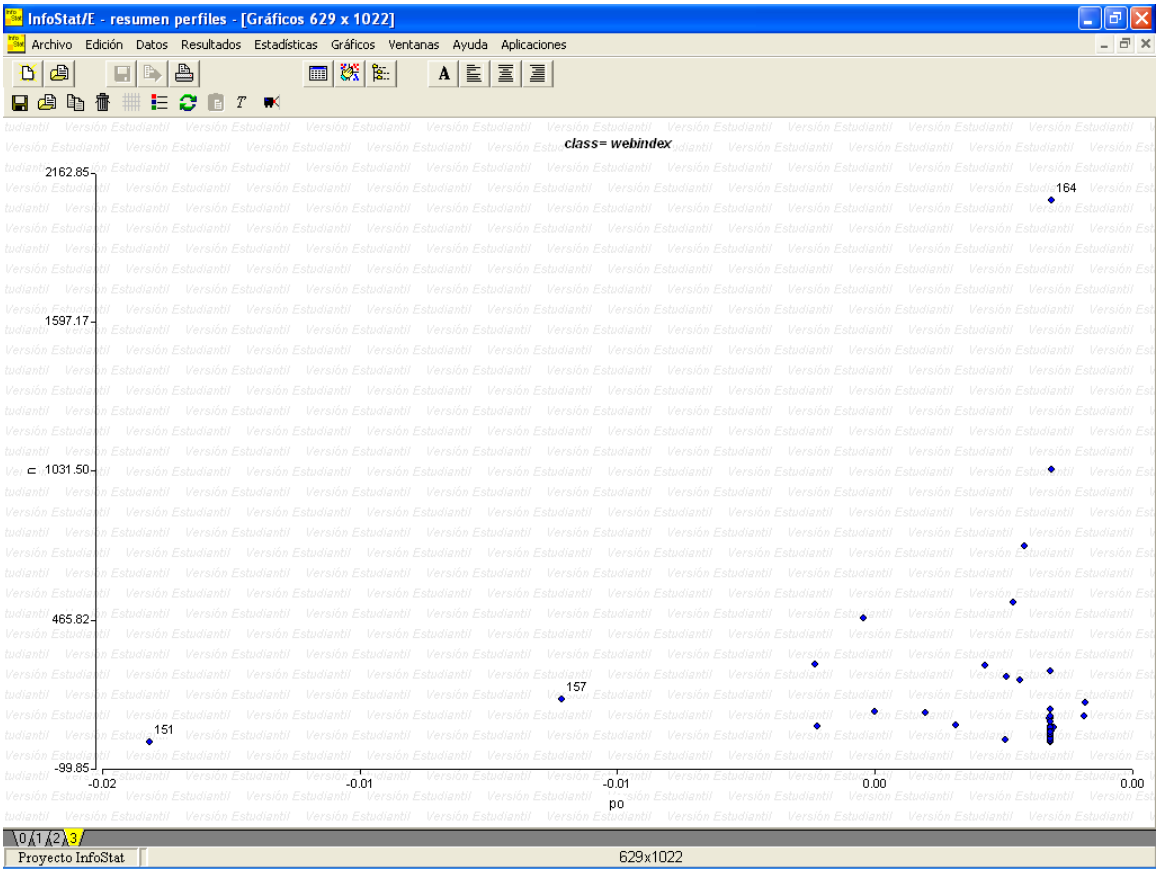


Donde n es la frecuencia de palabras especiales (EBH opuestas y contradictorias) dentro del texto y, p_o es el valor promedio de hallado de aplicar las ponderaciones respectivas según la fórmula $p_o^{i+1} = (p_o^{i} + p_o^{i+1})/2$.

Hay ciertos valores que podrían ser considerados outliers. Los mismos se identifican en las siguientes gráficas:


InfoStat/E - resumen perfiles - [Gráficos 629 x 1022]
Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Ayuda Aplicaciones
class= doc
1287.55
954.52
621.50
288.47
-44.55
77
83
63
92
-0.01
-0.01
0.00
0.00
0.00
po
Proyecto InfoStat
629x1022

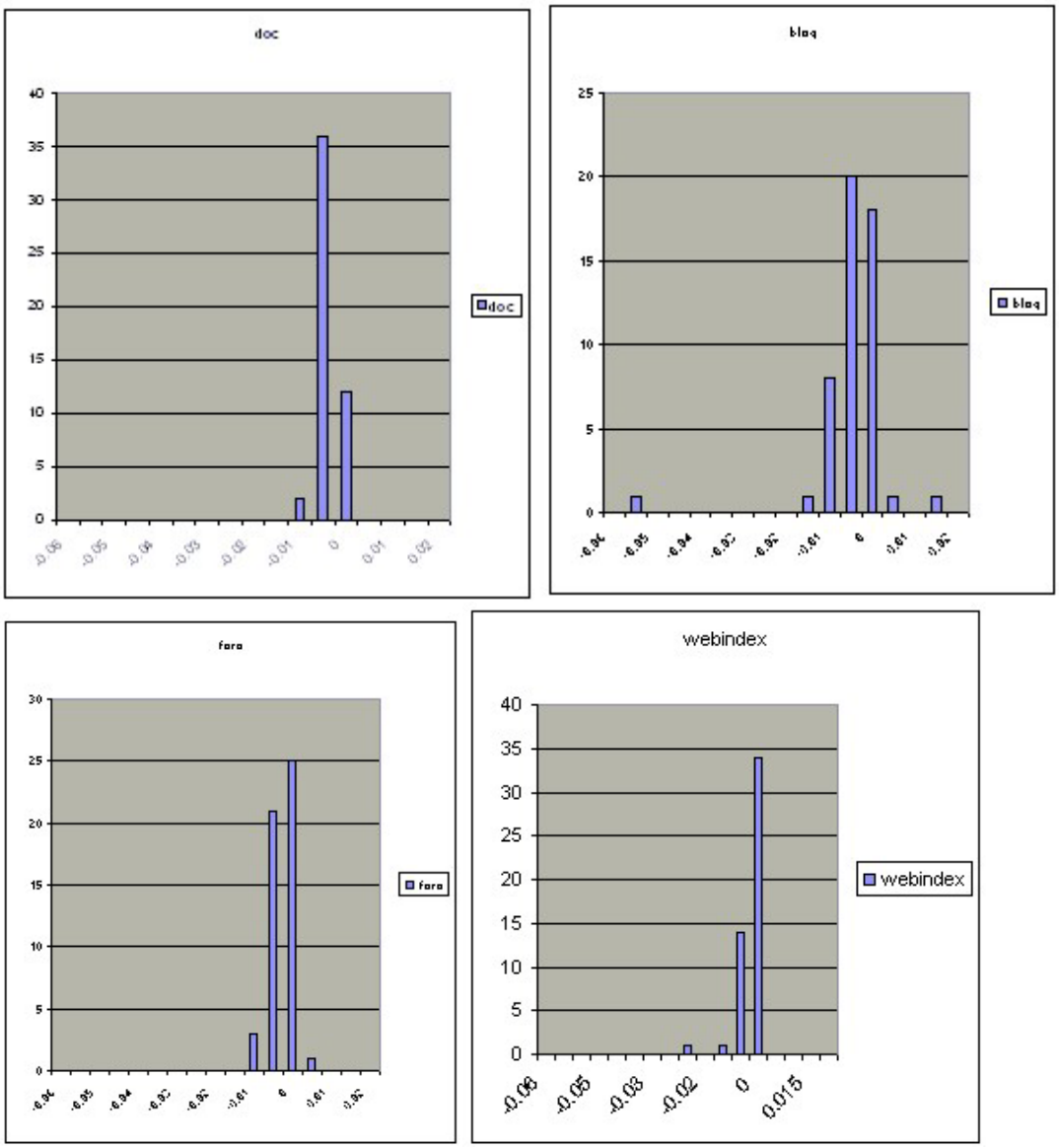

InfoStat/E - resumen perfiles - [Gráficos 629 x 1022]
Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Ayuda Aplicaciones
class= foro
344.95
257.22
169.50
81.77
-5.95
143
117
-0.01
-0.01
0.00
0.00
0.01
po
Proyecto InfoStat



Sin embargo no se tratarán como outliers dado que la información procesada no permite considerarlos como tales con justificación razonable y que se considera podría corresponderse con el tipo de comportamiento poblacional.

Histogramas de frecuencias para p_o

Se calculó el p_o promedio para cada documento, luego se levantaron los siguientes histogramas.



Los gráficos muestran una clara asimetría en todos los casos.

Estudio de variabilidad Kruskal-Wallis para p_o

Anova no paramétrico. Se usa porque no respeta precondiciones de t y ANOVA (cantidad de casos, normalidad, varianzas similares). Además las muestras tienen desigualdad notable de varianzas y distribución muy distinta a normal. Toma como hipótesis nula que las muestras son de la misma población. Los resultados sobre los p_o promedio de cada tipo son:

Ranks

	nroTipo	N	Mean Rank
po	1	50	107.84
	2	50	113.11
	3	50	83.99
	4	50	97.06
	Total	200	

Test Statistics(a,b)

	po
Chi-Square	7.555
df	3
Asymp. Sig.	.056

a Kruskal Wallis Test

b Grouping Variable: nroTipo

Como puede verse, la significación es 0.056 > 0.05, y *no puede afirmarse que las poblaciones tienen variabilidad similar.*

Estudio de medianas poblacionales para p_o

Anova no paramétrico. Se usa porque no respeta precondiciones de t y ANOVA (cantidad de casos, normalidad, varianzas. Se usa para verificar si las medianas son iguales en las poblaciones de los subgrupos.

Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
po	200	-.00178793	-.00002123	.00000000
nroTipo	200	1.25	2.50	3.75

Frequencies

		nroTipo			
		1	2	3	4
po	> Median	27	36	15	22
	<= Median	23	14	35	28

La mediana poblacional se ubica en p_o = -.00002123, y puede observarse que para los 3 grupos la cantidad de individuos en ningún caso se acerca al 50%.

Test Statistics(b)

	po
N	200
Median	-.00002123
Chi-Square	18.720(a)
df	3
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0.

b Grouping Variable: nroTipo

El valor da una significación p=0.0< 0.05, con lo que puede rechazarse la hipótesis original y por lo tanto p_o tiene una mediana poblacional distinta en cada subgrupo, sin embargo no puede afirmarse que se trate de poblaciones distintas dado que las pruebas de variabilidad no dan diferencias significativas.

Estudio de medianas y variabilidad Kruskal Wallis para n

Los percentiles y valores para la variabilidad son:

Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
n	200	36.00	83.50	153.50
nroTipo	200	1.25	2.50	3.75

Ranks

	nroTipo	N	Mean Rank
n	1	50	85.98
	2	50	88.04
	3	50	126.55
	4	50	101.43
	Total	200	

Test Statistics(a,b)

	n
Chi-Square	15.607
df	3
Asymp. Sig.	.001

a Kruskal Wallis Test

b Grouping Variable: nroTipo

Puede observarse que tiene una significancia de 0.001<0.05, por lo que podría afirmarse *que la variabilidad no es la misma para cada subgrupo*.

En cuanto a la prueba de medianas, los valores correspondientes son:

Frequencies

		nroTipo			
		1	2	3	4
n	> Median	19	20	34	27
	<= Median	31	30	16	23

Test Statistics(b)

	n
N	200
Median	83.50
Chi-Square	11.680(a)
df	3
Asymp. Sig.	.009

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0.

b Grouping Variable: nroTipo

Donde la significación nuevamente es buena, de 0.009 < 0.05. Las medianas serían significativamente distintas en los subgrupos.

Estudio de p_o a nivel documento

Es interesante estudiar el comportamiento de p_o en relación con los 4 subgrupos de perfiles.

Tomando el valor promedio de p_o calculado para cada frase (tal como se lo usa en WIH para ponderar los EBH) y se los clasifica como menores, mayores e iguales a

cero, es posible observar un comportamiento perfectamente distinguible por cada uno de los subgrupos. A continuación se muestran los resultados de estimar la curva de comportamiento como Binomial y la bondad de ajuste resultante en cada caso.

class: subgrupo

variable: conteo considerado

Clase: subclase dentro del grupo

LI: límite inferior

LS: límite superior

MC: marca de clase

FA: frecuencias absolutas observadas

FR: frecuencias relativas

E (FA): frecuencias absolutas según el modelo distribucional propuesto

E (FR): frecuencias relativas según el modelo distribucional propuesto

Chi-cuadrado: valor de estadístico Chi.

p: prueba de bondad de ajuste

(v) Como Binomial:

Tomando en consideración lo descrito en la sección correspondiente a Como Binomial, se verifican esencialmente los mismos resultados que para tipos de documentos.

Tablas de frecuencias

Ajuste: Binomial con estimación de parámetros: p= 0.00720 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	mayor_cero	1	0.00	0.20	0.10	32	0.64	34.84	0.70	0.23	
blog	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.23	
blog	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.23	
blog	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.23	
blog	mayor_cero	5	0.80	1.00	0.90	18	0.36	15.16	0.30	0.76	0.9434

Ajuste: Binomial con estimación de parámetros: p= 0.01200 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	menor_cero	1	0.00	0.20	0.10	20	0.40	27.34	0.55	1.97	
blog	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.97	
blog	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.97	
blog	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.97	

blog	menor_cero	5	0.80	1.00	0.90	30	0.60	22.66	0.45	4.35	0.3608

Ajuste: Binomial con estimación de parámetros: p= 0.00080 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	igual_cero	1	0.00	0.20	0.10	48	0.96	48.04	0.96	3.1E-05	
blog	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	3.1E-05	
blog	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	3.1E-05	
blog	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	3.1E-05	
blog	igual_cero	5	0.80	1.00	0.90	2	0.04	1.96	0.04	7.9E-04	>0.9999

Ajuste: Binomial con estimación de parámetros: p= 0.00200 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	mayor_cero	1	0.00	0.20	0.10	45	0.90	45.24	0.90	1.2E-03	
doc	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.2E-03	
doc	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.2E-03	
doc	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.2E-03	
doc	mayor_cero	5	0.80	1.00	0.90	5	0.10	4.76	0.10	0.01	>0.9999

Ajuste: Binomial con estimación de parámetros: p= 0.01520 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	menor_cero	1	0.00	0.20	0.10	12	0.24	23.25	0.46	5.44	
doc	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	5.44	
doc	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	5.44	
doc	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	5.44	
doc	menor_cero	5	0.80	1.00	0.90	38	0.76	26.75	0.54	10.17	0.0577

Ajuste: Binomial con estimación de parámetros: p= 0.00280 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	igual_cero	1	0.00	0.20	0.10	43	0.86	43.46	0.87	4.9E-03	
doc	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	4.9E-03	
doc	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	4.9E-03	
doc	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	4.9E-03	
doc	igual_cero	5	0.80	1.00	0.90	7	0.14	6.54	0.13	0.04	0.9998

Ajuste: Binomial con estimación de parámetros: p= 0.00400 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	mayor_cero	1	0.00	0.20	0.10	40	0.80	40.92	0.82	0.02	
foro	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.02	
foro	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.02	

foro	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.02	
foro	mayor_cero	5	0.80	1.00	0.90	10	0.20	9.08	0.18	0.11	0.9984

Ajuste: Binomial con estimación de parámetros: p= 0.00960 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	menor_cero	1	0.00	0.20	0.10	26	0.52	30.87	0.62	0.77	
foro	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.77	
foro	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.77	
foro	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.77	
foro	menor_cero	5	0.80	1.00	0.90	24	0.48	19.13	0.38	2.01	0.7347

Ajuste: Binomial con estimación de parámetros: p= 0.00640 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	igual_cero	1	0.00	0.20	0.10	34	0.68	36.27	0.73	0.14	
foro	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.14	
foro	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.14	
foro	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.14	
foro	igual_cero	5	0.80	1.00	0.90	16	0.32	13.73	0.27	0.52	0.9718

Ajuste: Binomial con estimación de parámetros: p= 0.00280 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	mayor_cero	1	0.00	0.20	0.10	43	0.86	43.46	0.87	4.9E-03	
webindex	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	4.9E-03	
webindex	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	4.9E-03	
webindex	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	4.9E-03	
webindex	mayor_cero	5	0.80	1.00	0.90	7	0.14	6.54	0.13	0.04	0.9998

Ajuste: Binomial con estimación de parámetros: p= 0.00640 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	menor_cero	1	0.00	0.20	0.10	34	0.68	36.27	0.73	0.14	
webindex	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.14	
webindex	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.14	
webindex	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.14	
webindex	menor_cero	5	0.80	1.00	0.90	16	0.32	13.73	0.27	0.52	0.9718

Ajuste: Binomial con estimación de parámetros: p= 0.01080 n= 50

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	igual_cero	1	0.00	0.20	0.10	23	0.46	29.05	0.58	1.26	
webindex	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.26	

webindex	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.26	
webindex	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.26	
webindex	igual_cero	5	0.80	1.00	0.90	27	0.54	20.95	0.42	3.01	0.5563

(vi) Como Poisson:

Tomando en consideración lo descrito en la sección correspondiente a Como Poisson , se verifican esencialmente los mismos resultados que para tipos de documentos.

Tablas de frecuencias

Ajuste: Poisson con estimación de parámetros: Lambda= 0.36000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	mayor_cero	1	0.00	0.20	0.10	32	0.64	34.88	0.70	0.24	
blog	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.24	
blog	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.24	
blog	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.24	
blog	mayor_cero	5	0.80	1.00	0.90	18	0.36	15.12	0.30	0.79	0.8522

Ajuste: Poisson con estimación de parámetros: Lambda= 0.60000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	menor_cero	1	0.00	0.20	0.10	20	0.40	27.44	0.55	2.02	
blog	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	2.02	
blog	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	2.02	
blog	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	2.02	
blog	menor_cero	5	0.80	1.00	0.90	30	0.60	22.56	0.45	4.47	0.2148

Ajuste: Poisson con estimación de parámetros: Lambda= 0.04000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	igual_cero	1	0.00	0.20	0.10	48	0.96	48.04	0.96	3.2E-05	
blog	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	3.2E-05	
blog	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	3.2E-05	
blog	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	3.2E-05	
blog	igual_cero	5	0.80	1.00	0.90	2	0.04	1.96	0.04	8.3E-04	>0.9999

Ajuste: Poisson con estimación de parámetros: Lambda= 0.10000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	mayor_cero	1	0.00	0.20	0.10	45	0.90	45.24	0.90	1.3E-03	
doc	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.3E-03	
doc	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.3E-03	
doc	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.3E-03	
doc	mayor_cero	5	0.80	1.00	0.90	5	0.10	4.76	0.10	0.01	0.9996

Ajuste: Poisson con estimación de parámetros: Lambda= 0.76000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	menor_cero	1	0.00	0.20	0.10	12	0.24	23.38	0.47	5.54	
doc	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	5.54	
doc	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	5.54	
doc	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	5.54	
doc	menor_cero	5	0.80	1.00	0.90	38	0.76	26.62	0.53	10.41	0.0154

Ajuste: Poisson con estimación de parámetros: Lambda= 0.14000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	igual_cero	1	0.00	0.20	0.10	43	0.86	43.47	0.87	0.01	
doc	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.01	
doc	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.01	
doc	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.01	
doc	igual_cero	5	0.80	1.00	0.90	7	0.14	6.53	0.13	0.04	0.9980

Ajuste: Poisson con estimación de parámetros: Lambda= 0.20000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	mayor_cero	1	0.00	0.20	0.10	40	0.80	40.94	0.82	0.02	
foro	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.02	
foro	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.02	
foro	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.02	
foro	mayor_cero	5	0.80	1.00	0.90	10	0.20	9.06	0.18	0.12	0.9896

Ajuste: Poisson con estimación de parámetros: Lambda= 0.48000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	menor_cero	1	0.00	0.20	0.10	26	0.52	30.94	0.62	0.79	
foro	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.79	
foro	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.79	

foro	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.79	
foro	menor_cero	5	0.80	1.00	0.90	24	0.48	19.06	0.38	2.07	0.5583

Ajuste: Poisson con estimación de parámetros: Lambda= 0.32000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	igual_cero	1	0.00	0.20	0.10	34	0.68	36.31	0.73	0.15	
foro	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.15	
foro	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.15	
foro	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.15	
foro	igual_cero	5	0.80	1.00	0.90	16	0.32	13.69	0.27	0.54	0.9110

Ajuste: Poisson con estimación de parámetros: Lambda= 0.14000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	mayor_cero	1	0.00	0.20	0.10	43	0.86	43.47	0.87	0.01	
webindex	mayor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.01	
webindex	mayor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.01	
webindex	mayor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.01	
webindex	mayor_cero	5	0.80	1.00	0.90	7	0.14	6.53	0.13	0.04	0.9980

Ajuste: Poisson con estimación de parámetros: Lambda= 0.32000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	menor_cero	1	0.00	0.20	0.10	34	0.68	36.31	0.73	0.15	
webindex	menor_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	0.15	
webindex	menor_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	0.15	
webindex	menor_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	0.15	
webindex	menor_cero	5	0.80	1.00	0.90	16	0.32	13.69	0.27	0.54	0.9110

Ajuste: Poisson con estimación de parámetros: Lambda= 0.54000

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	igual_cero	1	0.00	0.20	0.10	23	0.46	29.14	0.58	1.29	
webindex	igual_cero	2	0.20	0.40	0.30	0	0.00	0.00	0.00	1.29	
webindex	igual_cero	3	0.40	0.60	0.50	0	0.00	0.00	0.00	1.29	
webindex	igual_cero	4	0.60	0.80	0.70	0	0.00	0.00	0.00	1.29	
webindex	igual_cero	5	0.80	1.00	0.90	27	0.54	20.86	0.42	3.10	0.3767

(vii) Aproximación normal

Tomando en consideración lo descrito en la sección correspondiente a Aproximación normal, se verifican esencialmente los mismos resultados que para tipos de documentos.

Tablas de frecuencias

Ajuste: Normal con estimación de parámetros: Media= 0.36000 y varianza= 0.23510

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	mayor_cero	1	0.00	0.20	0.10	32	0.64	18.54	0.37	9.78	
blog	mayor_cero	2	0.20	0.40	0.30	0	0.00	8.11	0.16	17.89	
blog	mayor_cero	3	0.40	0.60	0.50	0	0.00	7.84	0.16	25.73	
blog	mayor_cero	4	0.60	0.80	0.70	0	0.00	6.41	0.13	32.14	
blog	mayor_cero	5	0.80	1.00	0.90	18	0.36	9.10	0.18	40.83	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.60000 y varianza= 0.24490

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	menor_cero	1	0.00	0.20	0.10	20	0.40	10.47	0.21	8.67	
blog	menor_cero	2	0.20	0.40	0.30	0	0.00	6.68	0.13	15.35	
blog	menor_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	23.19	
blog	menor_cero	4	0.60	0.80	0.70	0	0.00	7.85	0.16	31.04	
blog	menor_cero	5	0.80	1.00	0.90	30	0.60	17.15	0.34	40.66	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.04000 y varianza= 0.03918

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
blog	igual_cero	1	0.00	0.20	0.10	48	0.96	39.53	0.79	1.82	
blog	igual_cero	2	0.20	0.40	0.30	0	0.00	8.75	0.17	10.57	
blog	igual_cero	3	0.40	0.60	0.50	0	0.00	1.61	0.03	12.17	
blog	igual_cero	4	0.60	0.80	0.70	0	0.00	0.11	2.3E-03	12.29	
blog	igual_cero	5	0.80	1.00	0.90	2	0.04	3.1E-03	6.2E-05	1304.99	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.10000 y varianza= 0.09184

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	mayor_cero	1	0.00	0.20	0.10	45	0.90	31.46	0.63	5.82	
doc	mayor_cero	2	0.20	0.40	0.30	0	0.00	10.48	0.21	16.30	
doc	mayor_cero	3	0.40	0.60	0.50	0	0.00	5.58	0.11	21.88	
doc	mayor_cero	4	0.60	0.80	0.70	0	0.00	1.95	0.04	23.84	
doc	mayor_cero	5	0.80	1.00	0.90	5	0.10	0.52	0.01	62.22	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.76000 y varianza= 0.18612

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	menor_cero	1	0.00	0.20	0.10	12	0.24	4.86	0.10	10.51	
doc	menor_cero	2	0.20	0.40	0.30	0	0.00	5.24	0.10	15.75	
doc	menor_cero	3	0.40	0.60	0.50	0	0.00	7.67	0.15	23.42	
doc	menor_cero	4	0.60	0.80	0.70	0	0.00	9.08	0.18	32.50	
doc	menor_cero	5	0.80	1.00	0.90	38	0.76	23.15	0.46	42.02	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.14000 y varianza= 0.12286

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
doc	igual_cero	1	0.00	0.20	0.10	43	0.86	28.40	0.57	7.51	
doc	igual_cero	2	0.20	0.40	0.30	0	0.00	10.15	0.20	17.65	
doc	igual_cero	3	0.40	0.60	0.50	0	0.00	6.72	0.13	24.38	
doc	igual_cero	4	0.60	0.80	0.70	0	0.00	3.24	0.06	27.62	
doc	igual_cero	5	0.80	1.00	0.90	7	0.14	1.49	0.03	47.94	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.20000 y varianza= 0.16327

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	mayor_cero	1	0.00	0.20	0.10	40	0.80	25.00	0.50	9.00	
foro	mayor_cero	2	0.20	0.40	0.30	0	0.00	9.48	0.19	18.48	
foro	mayor_cero	3	0.40	0.60	0.50	0	0.00	7.46	0.15	25.95	
foro	mayor_cero	4	0.60	0.80	0.70	0	0.00	4.62	0.09	30.56	
foro	mayor_cero	5	0.80	1.00	0.90	10	0.20	3.44	0.07	43.08	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.48000 y varianza= 0.25469

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	menor_cero	1	0.00	0.20	0.10	26	0.52	14.48	0.29	9.18	
foro	menor_cero	2	0.20	0.40	0.30	0	0.00	7.38	0.15	16.55	
foro	menor_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	24.40	
foro	menor_cero	4	0.60	0.80	0.70	0	0.00	7.15	0.14	31.55	
foro	menor_cero	5	0.80	1.00	0.90	24	0.48	13.15	0.26	40.50	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.32000 y varianza= 0.22204

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
foro	igual_cero	1	0.00	0.20	0.10	34	0.68	19.97	0.40	9.85	
foro	igual_cero	2	0.20	0.40	0.30	0	0.00	8.40	0.17	18.24	
foro	igual_cero	3	0.40	0.60	0.50	0	0.00	7.82	0.16	26.06	
foro	igual_cero	4	0.60	0.80	0.70	0	0.00	6.10	0.12	32.16	
foro	igual_cero	5	0.80	1.00	0.90	16	0.32	7.71	0.15	41.08	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.14000 y varianza= 0.12286

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	mayor_cero	1	0.00	0.20	0.10	43	0.86	28.40	0.57	7.51	
webindex	mayor_cero	2	0.20	0.40	0.30	0	0.00	10.15	0.20	17.65	
webindex	mayor_cero	3	0.40	0.60	0.50	0	0.00	6.72	0.13	24.38	
webindex	mayor_cero	4	0.60	0.80	0.70	0	0.00	3.24	0.06	27.62	
webindex	mayor_cero	5	0.80	1.00	0.90	7	0.14	1.49	0.03	47.94	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.32000 y varianza= 0.22204

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	menor_cero	1	0.00	0.20	0.10	34	0.68	19.97	0.40	9.85	
webindex	menor_cero	2	0.20	0.40	0.30	0	0.00	8.40	0.17	18.24	
webindex	menor_cero	3	0.40	0.60	0.50	0	0.00	7.82	0.16	26.06	
webindex	menor_cero	4	0.60	0.80	0.70	0	0.00	6.10	0.12	32.16	
webindex	menor_cero	5	0.80	1.00	0.90	16	0.32	7.71	0.15	41.08	<0.0001

Ajuste: Normal con estimación de parámetros: Media= 0.54000 y varianza= 0.25347

class	Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
webindex	igual_cero	1	0.00	0.20	0.10	23	0.46	12.49	0.25	8.85	
webindex	igual_cero	2	0.20	0.40	0.30	0	0.00	7.04	0.14	15.89	
webindex	igual_cero	3	0.40	0.60	0.50	0	0.00	7.85	0.16	23.74	
webindex	igual_cero	4	0.60	0.80	0.70	0	0.00	7.49	0.15	31.23	
webindex	igual_cero	5	0.80	1.00	0.90	27	0.54	15.14	0.30	40.52	<0.0001

(viii) Conclusiones

Un valor de p < valor de significación nominal de la prueba conduce a un rechazo del modelo distribucional propuesto. En todos los casos que se presentan p>0.05 permite indicar que las muestras son modelizables como Binomiales con los parámetros especificados en el encabezado de cada tabla.

Las distribuciones tienen significación respectivamente superior para Binomiales que para Poisson, en cada subgrupo. Es posible que en el caso de Poisson, la curva en estos rangos y parámetros sea similar a la Binomial, pero no represente el comportamiento real poblacional. De hecho la distribución parece no ser

significativamente similar para el caso de Poisson en el subgrupo de perfiles doc, cuando el valor es MENOR que cero (p=0.015).

En consecuencia podría afirmarse que las distribuciones son Binomiales con las siguientes características:

perfil narración	rango	p
blog	>0	0.00720
	<0	0.01200
	=0	0.00080
doc	>0	0.00200
	<0	0.01520
	=0	0.00280
foro	>0	0.00400
	<0	0.00960
	=0	0.00640
webindex	>0	0.00280
	<0	0.00640
	=0	0.01080

Estudio de p_o a nivel de significación

De los diversos perfiles, se puede decir, basándose en los resultados previos, que cada uno tiene características propias en cuanto a cantidad de sentencias típicas, y distribución de p_o. Pero el comportamiento poblacional de p_o no ha mostrado estar influenciado por el subgrupo en cuestión. Por ese motivo se seleccionó la muestra de mensajes para estudiar el nivel de significación de frases en relación a p_o. Un estudio similar se muestra en la sección 2.g) Estudio de po al nivel de significación para el tipo de documento "mensajes".

Inicialmente se observaron todos los documentos, y prácticamente no hay documentos que presenten el 100% de valores p_o en cero (ver Fig. 62). Estos casos corresponden a cuatro documentos de temas diversos, con el siguiente contenido:

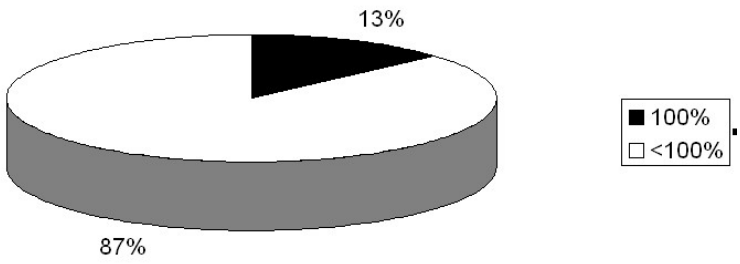
-ayudas para un viaje (comentarios). 25 sentencias.

-Currículum Vitae (el encabezado). 288 sentencias.

-decreto 243-01 (una parte). 24 sentencias.

-formulario de presentación de proyectos (un formulario a completar por un usuario). 197 sentencias.

Fig. 62. Frases con 100% de $p_o = 0.0$



De los documentos con valores de p_o distintos de cero, se tomaron los valores de p_o mínimo y máximo para cada documento. No se consideraron otros valores cercanos (en la práctica es necesario el uso de los mismos como se muestra en 5.4. Práctica de la estrategia con po). Se extrajeron las frases correspondientes a estos p_o seleccionados y se estudió si la misma se relaciona al tema principal o a un tema secundario del mensaje. El criterio seguido para considerar las frases es distinto según el contenido del documento:

Consideraciones de tema principal:

-Si es un documento en general: título o subtítulos.

-Si es un contrato: cláusulas.

-Si es un documento explicativo del trabajo de una comisión: descripción de la comisión, su actividad, objetivo, miembros, alcance.

-Si es una circular legal: descripción de la misma.

-Si es un documento que refleja la sesión de un tribunal: dictámenes.

-Si es un documento con el texto de una canción: repetición de un estribillo o título.

-Si es un formulario para completar: requisitos o condiciones.

-Si es un documentos que habla de una persona: la descripción de lo principal que dijo/hizo.

Consideraciones de tema secundario:

-Si es documento general: frase sin relación directa al título o subtítulos (comentarios, anécdotas, etc.).

-Si es un contrato: comentario acerca de la(s) cláusulas, consecuencias de aplicación, explicación con profundidad.

-Si es un documento explicativo del trabajo de una comisión: comentarios sobre la actividad.

-Si es una circular legal: implicancias y consecuencias.

-Si es un documento que refleja la sesión de un tribunal: comentarios acerca de su historia, antecedentes, planes, comentarios en general.

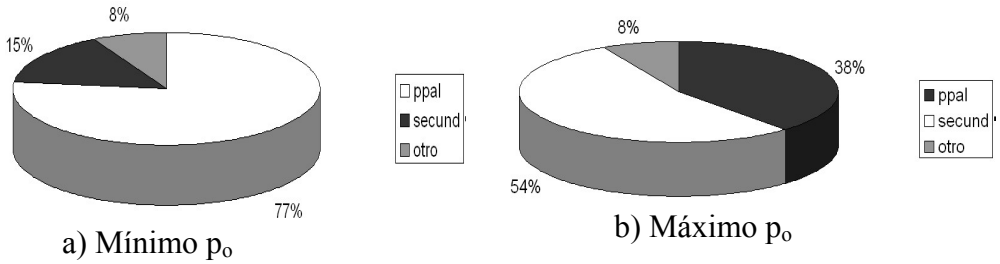
-Si es un documento con la letra de una canción: cualquier frase que no haga al tema principal.

-Si es un formulario a completar: explicación, temas relacionados.

-Si es un documento que habla de una persona: comentarios, anécdotas, opiniones acerca de la persona.

En la Fig. 63a) se muestran los porcentajes como gráfico de tortas cuando se consideran los p_o mínimos. En la parte b) se muestra el correspondiente a los p_o máximos.

Fig. 63. Significación de sentencias



a) Mínimo p_o b) Máximo p_o

Se estudió la aparición del tema principal en la primera frase del texto. En el caso de los mensajes tiene una frecuencia muy alta que no se observa, en los documentos aquí estudiados. Esto se refleja en la Tabla XLIX.

Tabla XLIX.Posición sentencia

es primera frase	
max p_o	0
min p_o	0
no	26

Apéndice E: Texto base para codificación en símbolos

Topacio
De Wikipedia

Cristales de topacio. El topacio es un mineral perteneciente al grupo de los aluminosilicatos. Su nombre deriva, según Plinio el Viejo, de la isla Topazos que se halla en el Mar Rojo. Sin embargo, los yacimientos de esta isla constan de olivina, frecuentemente confundida con el topacio.

Se utiliza a menudo como piedra preciosa y algunas veces ha sido confundido con el diamante: el llamado Diamante de Braganza, incluido como diamante en la corona portuguesa, es un topacio.

Se trata de un mineral cristalizado en el sistema ortorómbico con la fórmula general Al2SiO4(OH, F)2, indicando el paréntesis alrededor de HO y F que la proporción entre fluoruros F e hidroxi OH puede variar en un amplio rango, aunque su suma siempre será constante.

Su densidad es de 3,5 - 3,6 gr/cm³

Topacio azul, pulido el color es variable; a menudo se encuentran tonos de ocre, azul, violeta, rojo o incoloro. Además, puede ser variado fácilmente con medios artificiales: aplicando rayos gama o haces de electrones se consiguen tonalidades pardas o ligeramente verdosas y calentándolo se obtienen tonalidades azules o rojizas.

En la escala de Mohs le corresponde dureza de 8. Sin embargo, fractura fácilmente y por esta razón es difícil de trabajar.

Comercialmente se intentan vender algunas variedades de cuarzo con denominación de topacio.

Yacimientos

Se encuentra habitualmente en forma de cristales prismáticos crecidos en huecos que están unidos con la roca madre. Además existe una variante masiva o granulosa. A menudo se halla acompañado de berilio, turmalina y apatita en rocas ácidas magmáticas como las permatitas. También se encuentre en gneises.

Algunos de los yacimientos más importantes se encuentran en Brasil, República Checa, Sajonia, Noruega, Suecia, Japón, México, Sri Lanka, Birmania, Pakistan y los Estados Unidos.

Apéndice F: Formulario de encuesta para representación simbólica

Edad: ____ Sexo:____ Ocupación[46]:_____

Cantidad de horas aproximadas que usa la Web por semana: _____(h/sem)

1.Suponga que tiene una biblioteca de textos con diversos temas. Ud. dispone de un catálogo para buscar textos. Si desea buscar sobre una enfermedad sobre el sistema linfático denominada linfedema:

1.i)Escriba 3 preguntas que usted desearía realizar. Escríbalas en orden de importancia comenzando por la más importante para usted.

P1: _____

P2: _____

P3: _____

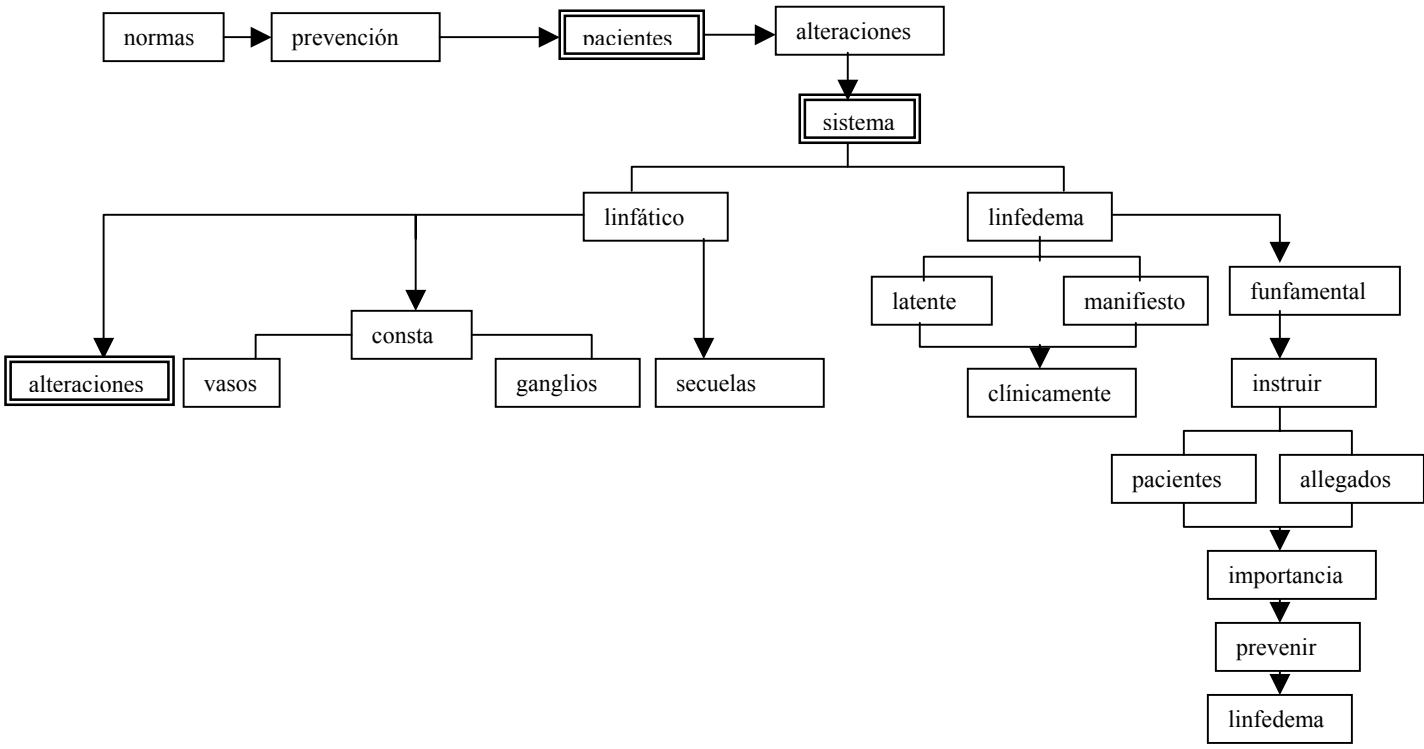
1.ii)Escriba a lo sumo 4 palabras que usted usaría para representar cada una de las preguntas que escribió en el inciso anterior (i).

P1: _____ _____ _____ _____

P2: _____ _____ _____ _____

P3: _____ _____ _____ _____

1.iii)Si usted viera el siguiente diagrama:



[46] Si es estudiante especifique la carrera.

Sabiendo que es un esquema resumido de un texto (**T**), qué tema le sugiere el texto?

_____

_____

1.iv)Usando el esquema de a.iii, escriba 3 preguntas que usted supone se podrían contestar con el texto **T** representado.

P1: _____

P2: _____

P3: _____

2.Suponga que la siguiente secuencia de palabras representa otro texto ($\mathbf{T_2}$):

-Ley

-Cámara

-Aporte normativa fundamental

2.i) Qué temas le sugiere?. Qué 3 nombres le colocaría al texto $\mathbf{T_2}$?

N1: _____

N2: _____

N3: _____

2.ii) Si averigua que el texto está en la siguiente dirección de internet: www.democracia-diario.com.ar

Qué tema le sugiere? Qué 3 nombres le colocaría al texto $\mathbf{T_2}$?

N1: _____

N2: _____

N3: _____