Una técnica para la extracción automática de resúmenes basada en una gramática de estilo

Hilda Yelitza Contreras Zambrano
hyelitza@ula.ve

Tutor: Jacinto A. Dávila Quintero
jacinto@ula.ve

www.bdigital.ula.ve

Trabajo de Grado Presentado ante la ilustre Universidad de Los Andes como requisito final para optar al título de:
MAGISTER SCIENTIAE EN COMPUTACIÓN

UNIVERSIDAD DE LOS ANDES
CONSEJO DE ESTUDIOS DE POSTGRADO
FACULTAD DE INGENIERÍA
Postgrado de Computación

Diciembre 2002

C.C.Reconocimiento

Resumen: Este documento presenta un experimento lingüístico que consiste en resumir textos escritos en español. El resumen es realizado con una herramienta computacional que aplica técnicas simbólicas basadas en una "gramática de estilos". Esta gramática modela las reglas de estilo para la escritura propuestas por Williams (1990). El programa puede obtener desde los tópicos de las oraciones de cada párrafo y reconocer elementos sintáctico-estructurales de cohesión y coherencia textual, hasta los tópicos más importantes del párrafo. Estos resultados son aprovechados para construir un resumen con oraciones asociadas a dichos tópicos. Esta versión de nuestro resumidor muestra cómo sobre la base de reglas lógicas para definir estilos y "tópicos" se pueden obtener resúmenes de textos "aceptables" por evaluadores humanos. Con esto se sugiere que aplicando estos modelos podemos obtener resultados reduciendo la complejidad del tradicional procesamiento morfológico, sintáctico y semántico. Para finalizar se realiza una descripción de las estrategias planteadas para extender y continuar este proyecto.

Palabras clave: Resumen Automático de textos, Extracción de Información, Lingüística Textual, Procesamiento del Lenguaje Natural.

www.bdigital.ula.ve

C.C.Reconocimiento

A ti, Abuela querida
A mis padres, por ser el comienzo de lo que soy

www.bdigital.ula.ve

C.C.Reconocimiento

Agradecimientos

Al profesor Jacinto Dávila, autor del concepto de esta tesis y perfecto motivador de estas páginas, quien supo guiarme hasta el final. A la profesora Melva Márquez, responsable de la hazaña intelectual de incorporarme al mundo de la Lingüística. A Rodrigo Martínez y Lucía Fabbri por aterrizar en Mérida y honrar con su presencia la defensa de esta tesis en el marco de un simposio internacional.

Quisiera extender mi más sincero agradecimiento a los queridos amigos de siempre, quienes supieron mantener mi animo. Especialmente a Chino, Melva, Ana Lía y Néstor por leer partes del manuscrito y artículos derivados de esta investigación.

También hago un reconocimiento a los estudiantes del curso de "Lógica y Matemáticas" semestre B-2002 de los postgrados de "Computación" y "Modelado y Simulación" de la Facultad de Ingeniería, por la contribución en la evaluación del presente proyecto.

www.bdigital.ula.ve

C.C.Reconocimiento

Tabla de Contenido

www.bdigital.ula.ve
C.C.Reconocimiento

Índice de Ilustraciones

Índice de Tablas

www.bdigital.ula.ve

C.C.Reconocimiento

1 Capítulo I: Procesamiento del Lenguaje Textual

En los últimos años se ha presentado un crecimiento explosivo de la información impulsado por los avances tecnológicos en los medios de almacenamiento y las redes de comunicación. Este hecho ha incrementado los problemas de almacenaje y extracción de grandes volúmenes de contenidos textuales y, además, ha hecho más difícil el acceso exacto y rápido a la información requerida por los usuarios. Como consecuencia, la información relevante generalmente es difícil de encontrar y suele requerir una búsqueda exhaustiva.

La descripción compacta del contenido relevante de un documento puede permitir incrementar la eficiencia en el procesamiento, recuperación y clasificación del material textual. Las técnicas de análisis automático de contenido textual procesan el lenguaje natural de los documentos para obtener desde descriptores y palabras claves hasta resúmenes. Estos extractos de información proveen al hombre y a los sistemas informáticos de mayores elementos para enjuiciar la importancia de un documento y reemplazar, al menos en un primer momento, la lectura exhaustiva.

Este capítulo realiza una revisión teórica del procesamiento del lenguaje textual que enmarca el contexto de este trabajo. Para ello, se comienza expresando el objetivo general de esta investigación y se continúa con algunos modelos del lenguaje escrito estudiados en la Lingüística Textual. A partir de allí se presenta una revisión del procesamiento humano e informático del texto, se continúa describiendo una de las características específicas de los textos (el estilo) y se culmina con la descripción y análisis del problema específico de esta investigación.

1.1 Objetivo General de la Investigación

Esta investigación tiene como objetivo general experimentar una técnica de procesamiento del lenguaje natural basada en la teoría lingüística de estilos formulada por Williams (1990). Esta técnica es aplicada a cierto tipo de textos para extraer información de sus contenidos. Con esta investigación se pretende ofrecer como resultado una herramienta informática para facilitar el procesamiento de grandes volúmenes de textos en español. La herramienta consiste en una extracción automática de resúmenes y tópicos, que ayuden a decidir sobre la relevancia de los textos y anticipen al lector sobre su contenido.

www.bdigital.ula.ve

El texto es el objeto de estudio de esta investigación. El texto es entendido como un medio de comunicación que dirige la actividad interpretativa de los usuarios textuales. La descripción y explicación de este objeto de estudio ha requerido investigaciones exhaustivas por parte de las diversas disciplinas, obteniendo como resultado varios enfoques sobre los procesos de producción y recepción textual de los humanos (Beaugrande y Dressler, 1997). Lo que se pretende en este trabajo es ofrecer un modelo simple de procesamiento computacional para extraer información relevante de los textos.

1.2 Lingüística Textual

La lingüística textual según Beaugrande y Dressler (1997) es un área interdisciplinaria encargada de regular las relaciones entre la lingüística, la ciencia cognitiva y la inteligencia artificial orientadas al procesamiento textual. En el ámbito de la ciencia del lenguaje, el término *lingüística textual* se utiliza para etiquetar cualquier tipo de estudio relacionado con el texto, siempre que este sea el objeto principal de investigación.

El texto es un acontecimiento comunicativo particular, debido a que el productor textual generalmente compone un discurso estructurado y utiliza recursos lingüísticos particulares que no se usan comúnmente en el habla. De esta manera, dentro de la riqueza expresiva del lenguaje natural, el escritor escoge una forma de comunicar sus ideas entre las posibles expresiones gramaticales que le permite su lengua. Beaugrande (1980) propone que el texto sea visto como un sistema, como una serie de elementos que funcionan conjuntamente. Si se considera que una lengua es *un sistema virtual de opciones disponibles susceptibles de ser activadas*, entonces el texto es un sistema real en que se han elegido unas opciones determinadas y se han utilizado en la producción de una estructura concreta.

C.C.Reconocimiento

Existen diferentes raíces históricas de la denominada *ciencia del texto* (Beaugrande y Dressler, 1997), formada por diversas disciplinas que se han encaminado hacia el estudio específico e independiente de los textos. Parte de su origen se remonta a la Grecia antigua con la forma más antigua de preocupación textual, la retórica. De igual manera, las nociones del campo tradicional de la estilística son semejantes a algunos principios de textualidad. Los textos también han sido objeto de estudio prioritario de los estudios literarios y han estado sometidos al escrutinio de la antropología, aunque ambas disciplinas se hayan limitado al estudio de unos tipos de textos específicos con características concretas. Además, la sociología se ha interesado por el estudio de la conversación, llamado generalmente análisis del discurso, de vital importancia para la ciencia del texto. Las disciplinas mencionadas anteriormente comparten, por diversos motivos, intereses en común con el estudio del texto. Sin embargo, debe mencionarse a la filología como una línea de investigación en el campo de la lingüística que antecede a la constitución de la lingüística moderna y de la ciencia del texto.

Las investigaciones en todas estas áreas han permitido que la lingüística textual pueda obtener como resultado diferentes modelos acerca de la expresión del lenguaje en forma escrita. Beaugrande y Dressler (1997) proponen un modelo complejo para el procesamiento textual donde cualquier texto debe cumplir siete normas (ver tabla 1.1) y tres principios reguladores de la comunicación (eficiencia, efectividad y adecuación). Las nociones están centradas en el texto, enfocadas en materiales textuales. También están centradas en el usuario, por que explican el funcionamiento de la actividad comunicativa. Con este modelo se intenta ir más allá de las estructuras y preguntarse cómo y porqué se utilizan los textos.

Norma de Textualidad	Tipo	Centrada en
Cohesión	Lingüístico	Texto
Coherencia		
Intencionalidad	Psicolingüístico	Usuario
Aceptabilidad		
Situacionalidad	Sociolingüístico	
Intertextualidad		
Informatividad	Computacional	

www.bdigital.ula.ve

Tabla 1.1. Modelo de procesamiento textual. Normas o criterios de textualidad (Beaugrande y Dressler, 1997).

Según este modelo, la ***cohesión*** consiste en la interconexión de las secuencias oracionales de la superficie textual a través de relaciones gramaticales. Un texto posee ***coherencia*** cuando los conceptos que componen el universo del discurso están interconectados a través de relaciones conceptuales de diversa naturaleza. La ***intencionalidad*** consiste en la organización cohesiva y coherente del texto siguiendo un plan dirigido hacia el cumplimiento de una meta. La ***aceptabilidad*** se manifiesta cuando un receptor reconoce que una secuencia de enunciados constituye un texto cohesionado, coherente e intencionado con un contenido, a su juicio, relevante. La ***situacionalidad*** se refiere a los factores que hacen que un texto sea pertinente en un determinado contexto. La ***intertextualidad*** indaga en la interpretación de un texto dependiendo del conocimiento que se tenga de textos anteriores. La ***informatividad*** es un factor de novedad que motiva el interés por la recepción de un texto.

Esta disciplina enfocada en el estudio del texto representa un aporte significativo a las investigaciones y avances del procesamiento del lenguaje natural, específicamente al tratamiento de documentos. Algunas de las variables estudiadas en lingüística textual serán consideradas en el desarrollo de este trabajo.

C.C.Reconocimiento

1.3 Procesamiento Humano del Lenguaje Textual

El lenguaje es un proceso comunicativo donde emisor y receptor procesan determinada información en función de un conocimiento lingüístico y un conocimiento del mundo compartido (Winograd, 1983). Los investigadores se han planteado como tarea reflejar la organización y funcionamiento de las estructuras y procesos lingüísticos y de las estructuras y procesos cognitivos (Moreno, 1998). Esto implica la modelización tanto de la competencia como de la actuación lingüística, así como la inclusión de los factores extralingüísticos en el modelo.

En particular, la lingüística textual ha modelado los procesos de producción y recepción textual que llevan a cabo las personas en la actividad comunicativa. El presente estudio se ha enfocado en la recepción textual, que según Beaugrande (1997), puede modelizarse como una serie de fases dominantes del procesamiento que se recorre en dirección contraria a la producción. La recepción textual ocurre en actividades como la lectura, la elaboración de resúmenes o síntesis y la extracción de terminología a partir de textos, las cuales son consideradas en esta sección como capacidades cognitivas humanas. En general, muchas acciones humanas tienen relación con la recepción de algún tipo de texto.

La terminología y el trabajo documental requieren un proceso de recepción textual. Tanto los documentalistas como los terminólogos usan la comunicación especializada existente en los textos especializados. El trabajo terminológico debe partir de una selección y análisis de la documentación especializada del tema a considerar. Esto le permite al terminólogo adquirir competencia cognoscitiva sobre la materia permitiéndole detectar, clasificar e ilustrar unidades terminológicas. El procesamiento de los textos en esta disciplina esta centrada en la identificación de conceptos los cuales han de delimitarse y describirse con medios lingüísticos (Cabré et al, 2000) (Arntz, 1995).

Por otra parte, los especialistas en documentación aplican varios tratamientos sobre un documento con el objetivo fundamental de facilitar su posterior recuperación. Para un documentalista, un documento recuperable es aquel que posee una descripción, una catalogación en una clasificación previamente establecida y un análisis de su contenido (Arntz, 1995). Es precisamente en estas actividades donde interviene el proceso de recepción textual, pues están basadas en el contenido del documento. En particular la descripción del contenido de un documento, que consiste en la explicitación de los elementos más representativos de la información que transmite, comporta dos actividades: la indización y la elaboración de resúmenes. En el fondo, la indización es una operación terminológica, pues identifica explícitamente las unidades y expresiones representativas del contenido. La elaboración de resúmenes por parte del documentalista consiste en una operación de condensación, ya que selecciona la información más relevante del contenido textual y la expresa de manera sintética.

www.bdigital.ula.ve

Los documentalistas y terminólogos procesan los documentos para obtener vocabularios, diccionarios, tesauros, documentos recuperables y catalogados. Sin embargo, la lectura y comprensión de los textos especializados, fase fundamental de ambas áreas, es un proceso que no exige metodología, ni siquiera explicación. De aquí que esta capacidad de recepción textual del ser humano sea motivo de investigación por parte de diferentes disciplinas.

A continuación se mencionan algunos ejemplos de este tipo de investigaciones, comenzando con la propuesta de los investigadores Behrens y Rosen (1982), quienes explican el proceso de escribir resúmenes a través de una serie de pasos:

(Paso 1) Dividir y etiquetar el texto en secciones o fases del pensamiento. Subrayar los términos o las ideas claves.

(Paso 2) Escribir un resumen (de una sola oración) de cada fase del pensamiento, si fuese apropiado de cada párrafo.

(Paso 3) Escribir una oración tópico resumen, es decir, una oración resumen de todo el texto. La oración resumen debe expresar la idea central del texto. El contenido del resumen depende de la intención del autor, por ejemplo de los textos (a) informativos (debe colocarse la información de qué, cómo, cuándo y dónde) de los (b) argumentativos (debe presentarse la conclusión del autor), y (c) descriptivos (debe incluirse la información del objeto descrito y sus características principales).

C.C.Reconocimiento

(Paso 4) Escribir resumen del texto. Combinar las oraciones tópicas (paso 3) con el resumen de cada fase del pensamiento (paso 2). Eliminar repeticiones y combinar las oraciones para lograr una corriente lógica de ideas.

(Paso 5) Revisar el resumen. Insertar palabras y frases de transición donde sea necesario para asegurar la coherencia.

Por otra parte, Beaugrande y Dressler (1997) concluyen que la recepción textual comienza en la "superficie", en la presentación misma del texto, y opera "descendiendo"a las fases "más profundas". La superficie lineal se analiza desde el punto de vista de las relaciones de dependencia gramatical. Los elementos afectados por estas relaciones son las expresiones que activan los conceptos almacenados en la memoria durante una fase denominada *recuperación conceptual*. Tan pronto como la configuración conceptual crece y adopta cierta densidad, pueden extraerse las ideas principales mediante una fase de recuperación de ideas. La posterior extracción de planes[1] que el productor textual intenta seguir, se realiza durante la fase de recuperación del plan textual. Una vez que el lector ha recuperado los conceptos, las ideas y los planes que ocurren en los textos estarán por fin en disposición de ofrecer un tratamiento adecuado a todas las posibles acciones y reacciones suscitadas en el texto.

Las fases de este proceso no se encuentran separadas rígidamente; al contrario, existen movimientos entre ellas de acuerdo a los resultados de una fase en particular. Además, las fases pueden tener variaciones de duración e intensidad en función de factores como: el juicio del receptor sobre la calidad del texto, el conocimiento del receptor sobre el contenido del texto y la implicación cognitiva del receptor en la comunicación. Sobre la base de este planteamiento, la recepción textual incluye un umbral de finalización en el cual el nivel de comprensión del texto se juzga satisfactorio (Beaugrande y Dressler, 1997).

No obstante, a pesar de tales propuestas y resultados, aún existe mucha controversia con relación a cómo el ser humano realiza el procesamiento del lenguaje natural. Los esfuerzos de desarrollo e investigación se han centrado en el estudio de técnicas computacionales que permitan el tratamiento de ciertos fenómenos lingüísticos a través de los computadores. No es casualidad que estos numerosos esfuerzos hayan tenido su origen precisamente en una necesidad de tratamiento textual.

www.bdigital.ula.ve

1.4 Procesamiento del Lenguaje Natural por parte del Computador

El Procesamiento del Lenguaje Natural (PLN), originalmente desarrollado a comienzos de la Guerra Fría (Locke y Booth, 1955) como el mecanismo de los físicos soviéticos para la traducción de documentos, es uno de los primeros objetivos computacionales más investigados. Estos esfuerzos prematuros por analizar y modelar el lenguaje humano fueron caracterizados por una técnica sin conocimiento lingüístico y por el bajo rendimiento computacional de la época.

Según Covington (1994) *"El Procesamiento del Lenguaje Natural es el uso de computadoras para entender lenguajes (naturales) humanos tales como ingles, francés o japonés. Por 'entender' no se quiere decir que el computador tenga pensamientos, sentimientos y conocimientos humanizados, sino que el computador pueda reconocer y usar información expresada en lenguaje humano"*.

Con estas expectativas, la Inteligencia Artificial ha dirigido una parte de su trabajo hacia la programación de un computador para entender el lenguaje natural. Así, se han realizado diversos procedimientos para procesar (y entender) el lenguaje natural. La lingüística teórica también ha aportado con el fruto de su investigación y, en los años sesenta del siglo XX, se comienza a introducir la información lingüística en el procesamiento del lenguaje natural. Moreno (1998) afirma que así se definió un área de conocimiento llamada Lingüística Computacional y se constituyó la *Association for Computational Lingüistics* (ACL).

[1] Los planes se refieren a las metas discursivas que desean alcanzar los escritores a través del texto. Los investigadores en inteligencia artificial se han interesado en analizar las intenciones de los comunicadores, ocultas tras el significado de las palabras que se utilizan en la interacción (Schank y Abelson, 1977).

C.C.Reconocimiento

A partir de estas iniciativas surgen diferentes tecnologías para el procesamiento computacional del lenguaje. En particular, la **extracción de información** (IE *Information Extraction*) consiste en el procesamiento de colecciones de textos para transformarlas en información que pueda ser digerida y analizada más fácilmente. Para ello, se identifican los fragmentos de textos relevantes, se extrae la información relevante de los fragmentos, y con estas piezas, se organiza la información requerida en una estructura coherente. Se trata de reconocer la información importante contenida en los documentos y trasladarla a un formato predefinido para que pueda ser tratada y recuperada con mayor facilidad.

El objetivo de los investigadores de IE es construir sistemas que encuentren ítems que puedan ser de interés para el análisis humano a partir de documentos. Además de la información relevante deben conseguirse las relaciones entre ellos, mientras que se ignora la información irrelevante y extraña. Hoy día, sin embargo, los sistemas de IE tratan solamente con tipos específicos de textos y han logrado resultados parciales (Cowie y Lehnert, 1996).

Manaris y Slator (1996) definen un sistema de PLN como aquel que encapsula un modelo del lenguaje natural en algoritmos apropiados y eficientes, en donde las técnicas de modelado están ampliamente relacionadas con conocimientos de muchos otros campos, incluyendo por ejemplo:

- la ciencia de la computación, que provee métodos para representar modelos, diseñar e implementar algoritmos para herramientas de software;
- la Lingüística, la cual contribuye con nuevos modelos lingüísticos y procesos;
- la Matemática, encargada de proponer modelos formales y métodos de análisis; y finalmente
- la Neurociencia, que explora los mecanismos mentales.

En particular, el área de la lingüística ha contribuido con el PLN aportando el conocimiento lingüístico de las lenguas naturales. Este conocimiento se incorporó a los sistemas de PLN a partir de los años sesenta y se convirtió en uno de sus componentes importantes.

www.bdigital.ula.ve

1.4.1 Conocimiento Lingüístico

El conocimiento lingüístico se puede organizar en diferentes niveles o componentes, ya que según Covington (1994) la estructura de cualquier lenguaje humano se puede dividir naturalmente en niveles. Manaris y Slator (1996) describen los niveles del conocimiento lingüístico dentro de un sistema de PLN desde el punto de vista de la característica declarativa (qué) y procedural (cómo), tal como se muestra en la Tabla 1.2:

	Características del nivel de conocimiento lingüístico	
Nivel	Declarativo (qué)	Procedimental (cómo)
Fonológico	Sonidos hablados	Formar morfemas
Morfológico (Léxico)	Unidades de palabras, Palabras	Formar palabras, Derivar unidades de significado.
Sintáctico	Funciones estructurales de palabras (colección de palabras)	Formar oraciones
Semántico	Significado independiente del contexto	Derivar significado de oraciones
Discurso	Funciones estructurales de oraciones (colección de oraciones)	Formar diálogos
Pragmático	Significado dependiente del contexto	Derivar significado de oraciones relativo al discurso circundante

Tabla 1.2. Niveles de conocimiento en el procesamiento del Lenguaje Natural

C.C.Reconocimiento

- **Nivel Fonológico**: la Fonología estudia cómo los sonidos hablados son usados en el lenguaje. Cada lenguaje tiene un alfabeto de sonidos distinguibles, llamados fonemas. Este nivel analiza las realizaciones acústicas, por tanto, solo aparece en los sistemas de reconocimiento del habla.
  Tecnológicamente, el tratamiento del habla por parte del computador está un poco separado del resto del procesamiento del lenguaje natural, debido a que este tipo de tratamiento tiene relación con el análisis de la forma de la onda del sonido y el reconocimiento de patrones, mientras que el resto de los niveles depende de una programación simbólica y un razonamiento automatizado.

- **Nivel Morfológico**: la morfología es la rama de la lingüística que se preocupa por la descripción de la estructura de las palabras y los procesos de formación de las palabras. La idea general es que los morfemas[2] individuales, pueden ser combinados para formar palabras. Hay tres procesos diferentes en la formación de palabras:

  - La **inflexión**: La morfología inflexional se preocupa por las relaciones gramaticales tales como el número gramatical y el tiempo verbal. Los afijos de inflexión no cambian la categoría sintáctica de las raíces a las que ellos están conectados. Así por ejemplo, tanto árbol como árboles (la raíz "árbol" mas el afijo plural "es") son nombres.
  - La **derivación**: La morfología derivacional describe como son creadas nuevas palabras con la ayuda de afijos. Por ejemplo, el adjetivo "nacional" se deriva del sustantivo "nación".
  - La **composición**: se preocupa por la construcción de palabras nuevas combinando morfemas libres, como en paraguas, de "para" y "agua".

  La morfología es un componente primordial para aquellas lenguas ricas en formas flexionadas (como el español o el alemán). La morfología es útil para evitar la expansión innecesaria de formas completamente flexionadas en el diccionario.

www.bdigital.ula.ve

- **Nivel Sintáctico**: la sintaxis, o construcción de oraciones, es el nivel más bajo en el cual el lenguaje humano es constantemente creativo. Noam Chomsky (1957) fue el primero en hablar sobre este punto porque introdujo las "gramáticas generativas", cuyas oraciones son descritas por reglas dadas. En lugar de listar las oraciones y sus estructuras directamente, se construyen las oraciones a partir de estas reglas.
  El conocimiento sintáctico es un componente básico de cualquier sistema de PLN, pues se encarga de reconocer las oraciones gramaticales y asignarles una estructura. El reconocimiento de la estructura de las oraciones por parte del computador es llevado a cabo por un algoritmo llamado *parsing* o analizador sintáctico.

- **Nivel Semántico**: la semántica, o significado, es el nivel en el cual el lenguaje hace contacto con el mundo real o imaginario. Se trata de la primera tarea del componente interpretativo de un sistema de PLN, la cual consiste en asignar un significado a cada una de las oraciones analizadas independientemente del contexto. Se realiza lo que se denomina composición semántica, que es la composición de significados de palabras para formar significados de oraciones. Por ejemplo "Juan ama a María", se forma del significado de "Juan", "ama" y "María", y se puede representar como fórmulas lógicas así: *ama(Juan,Maria)*. La semántica oracional es una parte

[2] Los morfemas son las unidades distintivas mínimas de la gramática. Hay dos clases de morfemas: *Formas Libres* (pueden ocurrir como palabras separadas) y *Formas de Salto* (que no pueden ocurrir como palabras en si mismas). Estas últimas reciben el nombre de Afijos. Por ejemplo, la palabra en inglés "unselfish" se compone de tres morfemas, "un", "self", y "ish". El morfema "self" es una forma libre mientras "un" y "ish" son formas de salto. En particular, "un" es aquí un prefijo, "ish" es un sufijo y "self" es una raíz.

C.C.Reconocimiento

imprescindible de cualquier sistema, ya que sin ella no podríamos asignar significado a las estructuras analizadas.

- **Nivel Discursivo**: aquí se tratan los aspectos de interpretación afectados por las oraciones emitidas anteriormente. En este nivel se almacena el conocimiento que permite relacionar entre sí el significado de las oraciones aisladas e integrarlo para formar unidades mayores. En concreto, este conocimiento se utiliza para interpretar los pronombres anafóricos, resolver los elementos elididos y los aspectos temporales. Este componente es necesario para que los sistemas tengan y usen el conocimiento del contexto comunicativo en el que se están produciendo los mensajes y tiene en cuenta aspectos pragmáticos como las intenciones del emisor y del receptor (Moreno, 1998).

- **Nivel Pragmático**: se refiere al uso del lenguaje en el contexto. En general la pragmática incluye aspectos del conocimiento conceptual del mundo que van mas allá de las condiciones reales literales de cada oración. Este conocimiento lo tienen en cuenta los hablantes cuando se comunican mediante una lengua. Les sirve para comprender mucha información sobrentendida pero no expresada explícitamente en las oraciones. Mientras la sintaxis y semántica estudian las oraciones, la pragmática estudia "las acciones del discurso" y las situaciones donde el lenguaje es usado.

1.4.2 Modelos del PLN

La característica esencial de un modelo es que nos permite hacer inferencias acerca del objeto modelado. Los fenómenos lingüísticos suelen tener una gran variedad de modelos matemáticos asociados y cada uno de estos modelos proporciona un conocimiento parcial sobre el fenómeno en cuestión (Moreno, 1998). Por tanto, pareciera ser necesario emplear el método más apropiado según el caso y combinar distintas aproximaciones para evitar la omisión de aspectos esenciales.

www.bdigital.ula.ve

Los modelos y métodos de PLN pueden ser clasificados en: simbólicos, empíricos o estadísticos, conexionistas e híbridos. Los dos primeros son llamados modelos matemáticos del lenguaje. El enfoque simbólico esta basado en el conocimiento, emplea reglas y algoritmos operantes con estructuras de datos que representan el conocimiento del lenguaje natural. El enfoque empírico o estadístico involucra colecciones de muestras del lenguaje (corpus), las cuales son etiquetadas y usadas para crear modelos estadísticos para PLN. La técnica conexionista usa redes neuronales para representar el conocimiento lingüístico. Por otra parte, las técnicas híbridas combinan uno o más de los modelos anteriores, con el fin de complementar las ventajas de cada uno y resolver problemas de dominios y aplicaciones específicos. Los primeros tres enfoques son explicados a continuación.

1.4.2.1 Modelos simbólicos

Los sistemas simbólicos se basan en la manipulación de símbolos. Fueron concebidos por los matemáticos para captar de manera rigurosa y sistemática la demostración de teoremas matemáticos y lógicos. Según Moreno (1998), estos modelos son los predominantes en las ciencias cognitivas (lingüística, sicología o inteligencia artificial) entendiendo los procesos mentales, incluyendo el lenguaje, basados en manipulación de símbolos. Dentro de la lingüística, Chomsky (1957) fue el primero en introducir de manera sistemática el paradigma lógico formal.

Típicamente las reglas de inferencia en un sistema formal permiten concentrarse en la sintaxis del modelo, independientemente de su interpretación. Muchos lingüistas piensan que el lenguaje tiene una naturaleza regular o lógica, y ello es lo que tratan de reflejar en sus gramáticas formales. Moreno (1998) afirma que generalmente, estas gramáticas han demostrado ser eficaces en la descripción y explicación de fenómenos relacionados con la competencia[3].

[3] La competencia se refiere al conocimiento que cada hablante tiene de su lengua materna

C.C.Reconocimiento

El fundamento teórico de dichos modelos está enmarcado en las gramáticas formales. Existen diferentes tipos de gramáticas que están formalizadas rigurosamente. Entre ellas tenemos: las gramáticas generativas, gramáticas categoriales, gramáticas de dependencia, gramáticas de cadenas lingüísticas de Harris y gramáticas de adjunción de árboles (Moreno, 1998). Sin embargo, las más conocidas son las gramáticas generativas, también conocidas como gramáticas de estructura de frase o sintagmáticas, propuestas por Chomsky (1957). Según Bach (1974), cualquier gramática que defina precisa y explícitamente las oraciones de una lengua es una gramática generativa. Son las más extendidas en Lingüística Computacional.

Los modelos simbólicos son el paradigma predominante en Lingüística Computacional, su repertorio de conceptos y métodos es muy amplio y ha sido aplicado sobre múltiples problemas y lenguas. De ellos los mas usados son los autómatas de estados finitos (que sirven por su sencillez y eficiencia de procesamiento, pero que tienen un poder expresivo muy limitado) y las gramáticas independientes del contexto complementadas con gramáticas de unificación y rasgos (por su poder expresivo para dar cuenta de fenómenos lingüísticos) (Moreno, 1998).

1.4.2.2 Modelos estocásticos

La aplicación de la probabilidad y la estadística al estudio del lenguaje tiene una tradición al menos tan antigua como la de los modelos formales. La idea general es inferir conocimiento directamente de los datos, buscando regularidades significativas. Aplicando la estrategia general de contar con la mayor cantidad posible de datos para poder establecer una probabilidad lo más cercana posible a la frecuencia relativa estable.

Además de la teoría de las probabilidades y la estadística, la Teoría de la Información (Lyons, 1968) es uno de los fundamentos teóricos de los modelos estadísticos. La teoría de la información tiene por objetivo descubrir las leyes matemáticas que gobiernan los sistemas diseñados para comunicar y manipular información. Lyons (1968) señala los conceptos de *rendimiento funcional* (depende de la frecuencia de aparición de elementos en la misma posición) y *contenido informático*[4] (depende inversamente de la probabilidad de aparición de una unidad en un contexto determinado) como los conceptos básicos en la teoría de la información.

www.bdigital.ula.ve

Los modelos estadísticos, también llamados métodos cuantitativos, proporcionan una solución al gran problema de los modelos simbólicos: la ambigüedad. Cuando una oración presenta varias estructuras o interpretaciones posibles para escoger, se elegirá la más probable en función de las probabilidades de cada opción.

Las técnicas básicas consisten en calcular las frecuencias de las palabras que aparecen en un conjunto de textos, y deducir probabilidades. Aunque el uso de la probabilidad tiene escasa utilidad en la vida real, se busca predecir acontecimientos a partir de cierta información incompleta, por ejemplo, el análisis más probable de una oración en un texto a partir de análisis anteriores.

Moreno (1998) nos reseña el método de estimación más sencillo, que emplea frecuencias relativas extraídas de un conjunto de datos, llamado corpus lingüístico. La anotación de las unidades del corpus: el texto debe estar marcado con información para inferir estadísticas más útiles. Las marcas más habituales son morfosintácticas: para cada unidad se especifica su categoría, concordancia, etc. Aunque se puede etiquetar el texto con cualquier tipo de información pertinente, por ejemplo sintagmática o semántica-léxica.

Dentro de las limitaciones de los modelos estadísticos podemos destacar varias. La representatividad del corpus es probablemente el problema más importante de todo modelo estadístico en general, pues son totalmente dependientes del corpus, de manera que si se intenta aplicar el modelo a otro dominio los resultados son pobres. Otra limitación tiene que ver con la localidad. Es muy eficiente con las relaciones locales, pero incapaz con las relaciones a larga distancia, mientras las gramáticas sintagmáticas tienen

[4] Lyons (1968), señala que cuanto más previsible es una unidad, menos información aporta (contenido informático), aunque resulte útil a los hablantes (rendimiento funcional). Sin embargo, debe decirse que hay unidades más o menos informativas de acuerdo al contexto.

C.C.Reconocimiento

medios para tratar con elementos discontinuos, la estadística parece que no ha encontrado soluciones (Lyons, 1968).

1.4.2.3 Modelos conexionistas

La mayoría de estos sistemas de PLN usan una estrategia lineal que no se corresponde con el procesamiento simultáneo y en paralelo que realiza nuestro cerebro. Si al procesar una emisión lingüística también consultamos simultáneamente diferentes componentes lingüísticos, entonces las aproximaciones inspiradas en el funcionamiento neuronal pueden ser una manera de acercarse no sólo a una simulación del cerebro sino a un procesamiento del lenguaje más eficiente (Moreno, 1998). Con estas ideas surgen entonces los modelos biológicos aplicados al PLN.

Los modelos conexionistas se dividen en dos grandes grupos. Los inspirados en el cerebro -el conexionismo (redes neuronales)- y los inspirados en la vida y la evolución -la computación evolutiva (algoritmo genético)-.

Las redes neuronales pretenden imitar el comportamiento del cerebro. En un modelo conexionista no están explícitos ni una gramática ni un lexicón. En su lugar existen unas estructuras formadas por nodos y conexiones en forma de red. Sobre esta red se realiza un reconocimiento de semejanzas en los patrones de activación de nodos (fonemas, morfemas, oraciones): dos estructuras son similares si excitan los mismos nodos.

Una de las características más destacadas de las redes neuronales es el procesamiento en paralelo: los diferentes procesos de reconocimiento de palabras, análisis morfosintáctico y semántica actúan simultáneamente.

Los algoritmos genéticos son algunas de las técnicas más utilizadas en la simulación de los sistemas complejos adaptativos. La computación evolutiva se basa en la capacidad de la naturaleza de resolver problemas. La idea central en esta técnica es reproducir el entorno en el cual se produce una evolución. Los sistemas están compuestos por individuos que interactúan estableciendo relaciones muy complejas de competencia y colaboración. Una parte importante de los sistemas de un modelo de simulación evolutiva son los mecanismos de evaluación, selección y reproducción. Se debe escoger según el problema la técnica o estrategia que mejor se adapte.

www.bdigital.ula.ve

Según Moreno, cualquier nivel lingüístico puede someterse a experimentación en una simulación evolutiva. Sin embargo, los niveles fonológico y léxico son los que mejor se prestan para esta representación (Moreno, 1998).

1.4.3 Alcances y aplicaciones del PLN

El conocimiento del mundo es un factor importante en los sistemas de PLN. Un sistema PLN debe incorporar, incluso en forma restringida, el conocimiento externo y de la experiencia humana. Covington (1994) dice, además, que el PLN depende de otros dos factores. El primero se refiere al poder de las computadoras, pues la aparición de microcomputadores en 1980 ha marcado la diferencia. Previamente, el PLN fue tan costoso que las personas exigían resultados perfectos, que nunca fueron alcanzados. Esta situación ha cambiado y las aplicaciones, aunque imperfectas, son más económicas y los usuarios encuentran buenos usos para ellas.

El segundo factor y quizá el más importante, es que el PLN depende del conocimiento exacto de cómo el lenguaje humano trabaja -lo cual, en estos momentos, no se conoce suficientemente-. Hasta hace pocos años, el lenguaje fue estudiado casi exclusivamente para la enseñanza de lenguas. Por otra parte, la ciencia de la lingüística tiene solamente unas pocas décadas de antigüedad, y no hay todavía consenso en relación con algunos hechos básicos.

Según Charniak (1993), el estancamiento de los sistemas basados en el conocimiento se debe a que en ellos se asume que la comprensión de las lenguas naturales depende básicamente de una gran cantidad de "conocimiento del mundo". Por lo tanto los sistemas de PLN tienen que contar con dicho conocimiento para tener éxito en su simulación de la facultad lingüística. Parece un hecho indiscutible que, a pesar de los múltiples intentos realizados en IA, no se dispone de un modelo o marco formal para

C.C.Reconocimiento

representar con éxito dicho conocimiento de "sentido común" que todo hablante parece emplear para entender los mensajes que recibe.

Los sistemas de PLN deben atacar una variedad de problemas relacionados con el lenguaje natural (Manaris y Slator, 1996):

- Inexactitud: errores ortográficos, signos de puntuación incorrectos, palabras transpuestas y oraciones agramaticales.
- Incompletitud: construcciones elípticas, anáforas, etc.
- Imprecisión: el uso de términos relativos sin un punto específico de referencia y el uso de términos cualitativos.
- Ambigüedad, debido a que pueden surgir múltiples interpretaciones en cualquier nivel del conocimiento lingüístico (ver tabla 1.2). La ambigüedad puede ser resuelta usando el conocimiento de un nivel más alto.

Estos fenómenos lingüísticos característicos de los lenguajes naturales han requerido experimentaciones exhaustivas de los diferentes métodos y modelos. A principio de los años noventa, se producen debates entre los diferentes modelos y a finales de esta década parece surgir una impresión generalizada de recurrir a la combinación de diferentes modelos para obtener resultados aceptables (Moreno, 1998).

Sin embargo, esta simbiosis no parece ser suficiente. Según, la apreciación de Moreno (1998) actualmente no se cuenta con aplicaciones PLN robustas, de gran cobertura y de fácil manejo. Por tanto, sugiere aprovechar el carácter experimental de la simulación computacional para avanzar nuestro conocimiento teórico sobre el lenguaje. Si se quiere que los programas se aproximen a las capacidades lingüísticas humanas, su diseño debe reflejar sus características dinámicas. Lo cual supone un gran cambio teórico con una aproximación global.

A pesar de estas perspectivas teóricas, las técnicas de PLN han logrado en la práctica desarrollar aplicaciones tales como: consultas a bases de datos, extracción de información de textos, recuperación de documentos relevantes de una colección, traducciones de un lenguaje a otro, reconocimiento de palabras habladas. En todas estas áreas existen aplicaciones útiles, pero ellas no pueden trabajar sin límites en cualquier dominio (Russell y Norving, 1995).

www.bdigital.ula.ve

Parece importante concluir esta sección aceptando las limitaciones de esta área interdisciplinaria. Los resultados de las investigaciones en PLN obligan a entender que con el conocimiento y la tecnología actual no ha sido posible conseguir una aplicación completa, sin errores y de ámbito general pues el dominio de aplicación es fundamental para ampliar el alcance del procesamiento a términos aceptables. Por ello, las consideraciones prácticas tienen un peso decisivo en la elaboración de aplicaciones PLN (Dávila et. al. 2002).

1.4.4 Estado del arte de las técnicas de resumen automático

Según Hahn y Mani (2000) actualmente están disponibles algunas herramientas de resumen automático[5]. Estas aplicaciones son útiles pero están limitadas a la **extracción**- selección de piezas originales del documento fuente concatenadas para producir un texto corto. A diferencia de la **abstracción**, la cual parafrasea el contenido del texto en términos más generales.

El método de concatenación para la extracción no asegura la coherencia del resumen, lo que hace que el texto sea difícil de leer. Además, ninguna de estas herramientas manejan múltiples fuentes, ni la información no textual de los documentos (como por ejemplo: tablas estadísticas, datos económicos).

Para tratar estas limitaciones, los investigadores están revisando una variedad de técnicas clasificadas en dos categorías. El primer grupo es denominado técnicas pobres en conocimiento, las cuales dependen

[5] Herramientas tales como AutoSummarize de Microsoft® a partir de Office 97®, Minería de texto de IBM®, el Context de Oracle® y el resumidor de Inxight® (parte de las herramientas de búsqueda de AltaVista), entre otras.

C.C.Reconocimiento

de agregar nuevas reglas para cada nuevo dominio de aplicación o lenguaje. El otro grupo son las técnicas ricas en conocimiento, las cuales asumen que la comprensión del significado del texto permite un resumen más efectivo. Los dos tipos de métodos no son mutuamente excluyentes, pues algunas aplicaciones son híbridas (Hahn y Mani, 2000).

En ambos métodos la principal limitación es el requerimiento de compresión. La técnica de extracción pretende obtener un resumen entre el 5 y el 30 por ciento de la longitud fuente del documento. Sin embargo, los objetivos de la compresión sobre múltiples fuentes son mucho menores. Estas altas tasas de reducción representan un reto, debido a que son difíciles de alcanzar, según Hahn y Mani (2000), sin manejar cierta cantidad de información sobre el conocimiento del mundo.

Por otra parte, Maña, Buenaga y Gómez (1998) clasifican las técnicas empleadas en la generación de resúmenes en estadísticas o simbólicas, utilizadas en aplicaciones generales o específicas, respectivamente. Los sistemas estadísticos independientes del dominio generan resúmenes inconsistentes e incompletos. Sin embargo, se han mejorado sus resultados al incorporar reglas semánticas y discursivas, agregándoles también cierta dependencia al contexto y tipología textual específica.

Los resumidores automáticos pueden tener otras clasificaciones (Acero et. al., 2000). Según su propósito, es decir, atendiendo al uso o tarea al que están destinados, los resúmenes se clasifican en:

- Indicativos, en donde el objetivo es anticipar al lector el contenido del texto y ayudarle a decidir sobre la relevancia del documento original,
- Informativos, cuando pretenden sustituir al texto completo incorporando toda la información nueva o trascendente, y
- Críticos, al incorporar opiniones o comentarios que no aparecen en el texto original.

Finalmente, atendiendo al enfoque, podemos distinguir entre resúmenes:

- Genéricos, si recogen los temas principales del documento y van destinados a un grupo amplio de personas, y
- Adaptados al usuario, en el caso del resumen que se confecciona de acuerdo con los intereses (i.e. conocimientos previos, ámbitos de interés o necesidades de información) del lector o grupo de lectores al que va dirigido.

www.bdigital.ula.ve

Hasta hace poco, los resumidores genéricos fueron los más populares, pero esta prevaleciendo la recuperación de información personalizada. Por tanto, los resúmenes adaptados a los usuarios están ganando importancia. Algunas herramientas soportan tanto resumidores generales como los enfocados al usuario. Por otra parte, los trabajos más recientes han desarrollado su tecnología en la síntesis de un único documento (Hahn y Mani, 2000).

Cualquiera sea el tipo de resumidor automático, los procesos para obtener el resumen tienen tres fases: **análisis del texto fuente**, **determinación de los puntos relevantes** y **sintetización de la salida apropiada**.

Los métodos de extracción hacen énfasis en el segundo componente, determinar las unidades relevantes del texto. Estas unidades son generalmente oraciones y son encontradas por la relevancia estadística y léxica o por la coincidencia con patrones de frases. Todos estos valores se agrupan en un sistema lineal de pesos, pues la suma de estos valores individuales es el peso de la unidad (oración). Este valor será considerado en la tercera fase del proceso donde se realiza la síntesis con las unidades de mayor peso del texto original.

Este método sencillo de resumen automático resulta fácil de implementar y ofrece resultados similares a pesar de existir diferentes variantes (Hahn y Mani, 2000). Los resúmenes generados pueden resultar incoherentes, con vacíos y presentan referencias anafóricas sin resolución. Sin embargo, se pueden lograr los deseados niveles de compresión (Hahn y Mani, 2000).

La investigación presentada por Kupiec (1995) muestra un sistema clasificador basado en redes neuronales que aprende cómo resumir, a partir del entrenamiento con un corpus de resúmenes generados

C.C.Reconocimiento

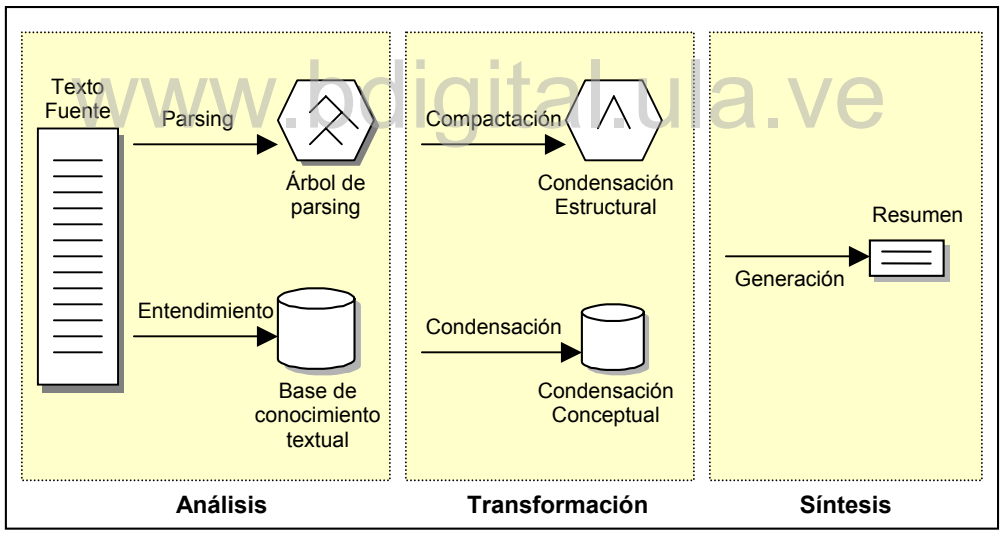
por humanos. Estos métodos basados en corpus, usados también por el resumidor de *Inxight*, son útiles para los usuarios si estos disponen de un corpus de textos y sus correspondientes resúmenes.

Por su parte, los métodos de abstracción requieren un procesamiento simbólico del lenguaje natural. Esto incluye gramáticas y lexicones para realizar el *parsing* y la generación de texto. Estos métodos también pueden requerir ontologías, que representan el sentido común y el conocimiento del dominio específico, con la finalidad de razonar durante el análisis.

En la figura 1.1 se esquematizan dos estrategias básicas. La primera usa un método lingüístico tradicional analizando sintacticamente las oraciones y generando un árbol de *parsing*. Este método también puede usar la información semántica. El proceso de compactación se realiza sobre el árbol de *parsing*, eliminado y reagrupando sus partes de acuerdo a criterios estructurales.

La segunda estrategia de los métodos de abstracción tiene sus raíces en la IA y está enfocado en el entendimiento del lenguaje natural. Un analizador sintáctico también es parte de este método, pero el resultado son las estructuras de representación conceptual del contenido del texto fuente, las cuales son posteriormente ensambladas en una base de conocimiento textual. El proceso de transformación usa técnicas de abstracción ricas en conocimiento y en inferencia lógica (Dijk, 1977) (Hahn y Reimer, 1999). Estas propuestas reducen la redundancia y eliminan la información irrelevante del texto origen, realizando en esencia una condensación conceptual. En ambos casos, se desemboca en un resumen.

Actualmente, según Hahn y Mani (2000), existen cuatro áreas de resumen automático que atraen el interés de los investigadores. Estas áreas son el desarrollo de resumidores que puedan abordar múltiples documentos, documentos multimedia, documentos con múltiples lenguajes y documentos híbridos que mezclen el idioma y el tipo (multimedia o no). Además, trabajan sobre documentos con varios formatos como HTML y XML, y explotan la información de las etiquetas asociadas.



www.bdigital.ula.ve

Figura 1.1. Arquitectura de los métodos de abstracción en resumen automático.

En general, la investigación en técnicas de resumen automático es relativamente joven y requiere más evaluación y desarrollo. Sin embargo, hay evidencia de que muchas de las técnicas actuales son realmente útiles (Hahn y Mani, 2000).

1.5 El estilo en el lenguaje escrito

Los estudios estilísticos se han realizado desde diversos puntos de vista. A pesar de la variedad de enfoques, casi todos los trabajos se basan en la convicción de que el estilo es el resultado de una

C.C.Reconocimiento

determinada elección entre opciones diversas que se ha realizado mediante el proceso de producción de un texto (Beaugrande y Dressler, 1997). De acuerdo con este planteamiento estilístico, los receptores esperan razonablemente que aparezca en el texto una determinada serie de secuencias dominantes con mayor frecuencia que otras.

Según los trabajos de DiMarco y Hirst (1993), un escritor usa varias construcciones sintácticas con un objetivo estilístico que ellos llaman de "alto nivel". Para capturar esta clase de intuición lingüística, estos investigadores desarrollaron la idea de una "gramática de estilos", la cual relaciona las estructuras sintácticas de un lenguaje con un conjunto de objetivos estilísticos independientes del lenguaje. Para expresar estos objetivos se requiere de una estructura sintáctica.

Por su parte, Williams ha propuesto reglas de estilo para ayudar a mejorar la claridad de la escritura (Williams, 1990). Sus libros están dirigidos a los anglohablantes pero parecen admitir una buena adaptación al español. Las recomendaciones de Williams consideran cuidadosamente las necesidades del lector. Las siguientes secciones explican en detalle las reglas de estilo que se utilizan como base en el desarrollo de este trabajo.

1.5.1 Reglas de estilo de Williams

El profesor Joseph Williams, de la Universidad de Chicago, sugiere reglas de estilo particularmente útiles debido a que están orientadas al lector. En Williams (1990) se presentan elementos indispensables para obtener un estilo de escritura legible: claridad, cohesión, coherencia, énfasis, elegancia, concisión y longitud. Los primeros tres son los más útiles y mejor planteados por Williams. Según Williams son estos precisamente los elementos básicos de una escritura legible, que sirve a un lenguaje útil en la comunicación.

Las recomendaciones comienzan en el ámbito de la oración. Williams cree que los lectores encuentran a las oraciones fáciles de leer y entender cuando su forma de razonar sigue la lógica de la oración: los sujetos de la oración deberían ser los actores, y los verbos de las oraciones deberían ser las acciones cruciales (regla de claridad oracional. El comienzo de una oración debería retomar el pasado y conectar al lector con las ideas que se habían mencionado antes. El final de la oración debería inducir y es el lugar para colocar nuevas ideas y nueva información (regla de cohesión entre oraciones). En la propuesta de estilo de Williams, a cada oración de un texto se le asocia un tópico y estos tópicos son usados para generar una secuencia coherente de oraciones que constituye un párrafo.

www.bdigital.ula.ve

Su propuesta continúa en el nivel del párrafo. Las oraciones que constituyen un párrafo deberían tener tópicos consistentes y coherentes entre sí. Nuevos tópicos y nuevos temas deberían encontrarse al final de las oraciones introductorias del párrafo. Los lectores encontrarán un párrafo como coherente si tiene solo una oración que exprese el resumen, la cual casi siempre se encuentra o al final del párrafo o como la última de las oraciones introductorias del párrafo (esta es la clave para la coherencia).

1.5.2 Claridad

La claridad nos permite identificar de manera precisa los actores y acciones del relato. Williams enuncia los primeros principios de la escritura clara:

- Los sujetos de las oraciones enuncian el reparto de personajes.
- Los verbos que van con estos sujetos enuncian las acciones cruciales de los cuales aquellos personajes son parte.

Para cumplir con estos principios es suficiente usar una forma clara y transparente al presentar los verbos y los sujetos. Una manera común de ocultar los actores es usar la "voz pasiva" y la

C.C.Reconocimiento

"sustantivación"[6] . Sin embargo, el uso de la sustantivación puede ser adecuado y válido en ciertos momentos. Williams se refiere a la tendencia de abusar de su uso con lo cual se va en detrimento de la claridad de la escritura. Para ser claros, es necesario saber que los lectores identifican dos niveles de estructura dentro de una oración o frase la cual debemos integrar:

(1) su secuencia gramatical predecible: sujeto + verbo + complemento
(2) su historia, un nivel de significado cuyas partes tiene un orden fijo: caracteres + acciones.

Estructura fija	Sujeto	Verbo	Complemento
Contenido variable	Caracteres	Acciones	----

Tabla 1.3. Claridad de Williams

Los elementos fijos aparecen en toda oración o frase que posea un sentido completo. Se trata de una estructura fija a la cual debemos tratar de asignar agentes y acciones. Los elementos variables, por su misma condición, se pueden mover de cualquier manera o incluso pueden no aparecer dentro de la oración. Lo que propone Williams es hacer coincidir los caracteres con el sujeto y las acciones con los verbos.

El personaje es un elemento variable del nivel histórico de la oración. En general, hay muchos tipos de personajes. Los más importantes son los agentes -el origen directo de cualquier acción o condición - es decir, son los causantes de las acciones. Es importante destacar que cuando se hace coincidir el personaje con el sujeto de la oración, éste se convierte en el agente, es decir, el causante directo de la acción.

Las acciones también se identifican como elementos variables de la estructura histórica de la oración. Entendemos por acciones a aquellas palabras que describen un estado de alteración (física, psicológica o espiritual) del personaje respecto a su ambiente. En contraste, entendemos por verbos aquellas acciones que afectan, no al personaje, sino al agente de la oración.

www.bdigital.ula.ve

Cuando se trata de interpretar un párrafo completo es importante definir una *"cadena lógica y consistente de sujetos"* (Williams, 1990) (Ochoa, 1999). Esto permite que el lector identifique los agentes de las acciones de cada una de las oraciones, realizando conexiones lógicas entre ellas. La idea es que al inicio se logre "anclar" al lector en un concepto que le es familiar o conocido para introducir luego un concepto nuevo. Otra recomendación es usar al principio una oración que oriente al lector en que es lo que sigue. La cadena consistente de sujetos debe corresponder con los argumentos de una inferencia lógica.

Lo anterior se refiere a la claridad local interna de cada oración independientemente de un contexto o de una intención. Existen dos elementos más, aparte de la claridad, que deben considerarse para alcanzar un texto legible. Ellos son la cohesión y la coherencia.

1.5.3 Cohesión y Coherencia

La cohesión y la coherencia son propiedades intrínsecas de los textos y responsabilidad absoluta de quien los produce. Encargadas de orientar los procesos cognoscitivos interpretativos que han de poner en funcionamiento los receptores. La cohesión es la interconexión de secuencias oracionales que componen la superficie textual a través de las relaciones gramaticales, como la repetición, las formas pronominales, la correferencia, la elisión o la conexión (Beaugrande y Dressler, 1997).

Williams (1990) entiende por cohesión a "la manera cómo las diversas oraciones que conforman un texto escrito permanecen unidas bajo un mismo contexto o discurso". Con cada oración que escribimos

[6] La sustantivación es un recurso del lenguaje que consiste de la transformación de acciones en sujetos. Se usa para ocultar los verbos (acciones). También existe cuando se transforma un adjetivo en nombre (nominaliza un adjetivo). Por ejemplo el verbo "descubrir", es substantivado con la palabra "descubrimiento".

C.C.Reconocimiento

debemos establecer el mejor encuentro entre los principios de la claridad local y los principios de cohesión que unen oraciones separadas dentro de un mismo discurso.

Las recomendaciones de Williams para lograr la cohesión dentro del texto son:

1. Usar conectores lógicos, para conectar una oración a la precedente y predisponer al lector a capturar nuevos conceptos (denominado metadiscurso transicional). Por ejemplo: *además*, *como resultado*, *pero*, etc.
2. Usar expresiones para evaluar lo que sigue. Por ejemplo: *quizá*, *afirmativamente*, *bajo estas circunstancias*, etc.
3. Ubicar el tiempo y el espacio dentro de la narración. Por ejemplo: entonces, *mas tarde*, *en América*.
4. Enunciar el tópico al inicio de la oración, generalmente en el sujeto de la oración.

Las primeras tres recomendaciones son llamadas "frases de transición". Su objetivo es permitir mantener la correcta cohesión y coherencia entre las oraciones. No contiene información relevante.

Ahora bien, Williams dice que es más importante el efecto acumulativo de la secuencia de tópicos, que los tópicos individuales de las oraciones individuales. Es aquí donde él sugiere poder escribir tanto oraciones pasivas como activas mientras esto genere una cadena de tópicos consistente, dando prioridad a las oraciones activas.

Las cadenas consistentes de tópicos deben enfocar la atención sobre un conjunto de conceptos circunscritos. Para lograr un foco consistente y un estilo cohesivo Williams ofrece dos principios:

1. Colocar en el sujeto / tópico las ideas que hemos mencionado antes (familiares al lector). Colocar al inicio de una oración aquellas ideas que ya hayan sido enunciadas, referidas e implicadas, o bien aquellos conceptos que pueden asumirse como familiares y conocidos por el lector.
2. El grupo de oraciones del párrafo debe mantener consistente los tópicos. Colocar al final de la oración lo más nuevo o reciente, lo más sorprendente, la información más significativa, es decir, la información que desea extender y desplegar.

www.bdigital.ula.ve

Basándonos en lo anterior, las cadenas de tópicos deben reflejar una secuencia de tópicos entre las oraciones del párrafo. Según los niveles de Williams, la "nueva información" está referida al tópico, que ha sido enunciado en la "vieja información" de la misma oración.

El ***tópico*** es un concepto que juega un papel preponderante en la búsqueda de la cohesión. Williams define el tópico como "el sujeto psicológico de la oración". Es decir, el tópico es el elemento que lleva la carga lógica del discurso, tanto oral como escrito. Desde el punto de vista lógico los tópicos son los conceptos emitidos o involucrados en cada una de las proposiciones que posee un argumento, ya sea en sus premisas o en su conclusión.

El tópico es "casi siempre una frase nominal de cualquier tipo que el resto de la oración caracteriza o comenta, sobre la que se dice cualquier cosa. Son ideas que definen de qué trata el texto. En forma incremental estas ideas topicalizadas proporcionan avisos temáticos que enfocan la atención del lector hacia un conjunto bien definido y limitado de ideas conectadas" (Williams, 1990).

Las cadenas lógicas y consistentes de sujetos (claridad) tienden a solaparse con las cadenas de tópicos (cohesión). Ambas cuentan con un criterio similar en su construcción, y serán las mismas en la medida que nuestros tópicos sean los sujetos de nuestras oraciones. Sin embargo, cuando existan puntos donde estos dos criterios sean divergentes, debe prevalecer el criterio de cohesión.

Gráficamente se puede representar la cohesión como sigue, manteniendo la consistencia con la representación de la claridad:

Estructura fija	Tópico	Énfasis
Contenido variable	Vieja inf. familiar	Nueva inf. No familiar

Tabla 1.4. Cohesión de Williams.

C.C.Reconocimiento

En la tabla 1.4. se observa que la vieja información y la nueva información poseen relación entre sí. A esta relación es lo que Williams llama *ideas topicalizadas*. Es decir, la nueva información referida al tópico que ha sido mencionado en la vieja información. La vieja información son aquellos conceptos contenidos en las oraciones que son conocidos por el lector y donde potencialmente encontraremos al tópico. En este caso, la vieja información es análoga a *personajes*, y el concepto seleccionado como tópico es análogo a *agente*. La nueva información son aquellas oraciones que poseen nuevos y más complejos conceptos. Estos poseen nuevos tópicos que nos permitirán movernos, "lógicamente", entre diversos contextos asegurando así la cohesión del discurso escrito.

Por otra parte, Williams define párrafo cohesivo como aquel que tiene una cadena consistente de tópicos (denominadas cadenas temáticas). Un párrafo cohesivo induce un nuevo tópico y cadena temática en una posición predecible, al final de la(s) oración(es) introductoria(s) del párrafo. Este principio nos permite identificar dos nuevos elementos dentro del párrafo: "salida o arranque" y "discusión". Es decir, que un párrafo coherente tendrá usualmente una oración sencilla, o varias, que claramente articule su tema o idea principal. De la misma manera un párrafo coherente ubicará típicamente aquella idea principal o punto en uno de los dos lugares siguientes: en el arranque o en la discusión del párrafo.

Williams introduce la definición de coherencia a través de las reglas anteriores, pero se debe entender que la coherencia va más allá de lo que se ha escrito explícitamente en un texto. De ello se deduce que la coherencia no es un simple rasgo que aparezca en los textos, sino que se trata más bien de un producto de los procesos cognitivos puestos en funcionamiento por los lectores (Beaugrande y Dressler, 1997). Por tanto, Williams plantea que los lectores, conscientes o no, tratan de dividir un discurso (párrafo, sección o texto completo) en dos secciones:

- El arranque que corresponde a un segmento corto de apertura o introducción del discurso. Puede tener una longitud de una, dos, tres o más oraciones. Se recomienda que sea corto para facilitar el proceso de lectura.
- La discusión, resto del párrafo, es el segmento largo donde los escritores desarrollan nuevas ideas en torno a los temas o tópicos mencionados en el arranque.

www.bdigital.ula.ve

Además, los lectores esperan conseguir en el arranque el tópico (denominado el "punto") tratado en la discusión. Específicamente, esperan que al final del arranque se encuentre el tema que se repetirá a lo largo de la sección de discusión. En este punto, Williams hace analogía con el contenido de un telegrama que captura la idea esencial del discurso. La mayoría de los escritores omiten esta información porque confían en que sus lectores tendrán las mimas suposiciones, conocimientos, actitudes y valores. Generalmente esto no es así, y el punto puede no ser obvio para los lectores. Por lo tanto, se recomienda exponer explícitamente esta idea principal del discurso en el texto.

El modelo de Williams fue presentado de manera ascendente (claridad, cohesión y coherencia). Sin embargo, su aplicación es descendente y se produce simultáneamente. Por tanto, es necesario establecer qué hacer cuando hay conflicto entre las reglas. En el modelo de Williams existen prioridades donde la coherencia prevalece a la cohesión y ésta, a su vez, al principio de la claridad.

Según Williams, los elementos básicos de una escritura legible que encuentra un lenguaje útil en la comunicación son la claridad, cohesión y coherencia, porque logran una calidad en la información, así como una comunicación clara, precisa y efectiva.

1.6 Problema de la investigación

Como se mencionó al inicio de este capítulo, el problema general de esta investigación consiste en abordar las dificultades de un computador al realizar el procesamiento de textos escritos por humanos en lenguaje natural. En particular, interesa extraer el contenido relevante del documento.

El inconveniente del procesamiento del lenguaje natural por parte del computador es significativamente grande por su diferencia con los lenguajes formales. Los formalismos lógicos de los

C.C.Reconocimiento

lenguajes formales tienen una serie de características, técnicas y herramientas para su procesamiento. Al tratar de aplicar estas técnicas y herramientas al lenguaje que usan los humanos surgen una serie de limitaciones. Estas limitaciones, según Quesada y Amores (2000), están dadas por fenómenos lingüísticos exclusivos de los lenguajes naturales, entre los que destacamos: la magnitud del tamaño del léxico, la complejidad de la semántica y pragmática del conocimiento lingüístico, la ambigüedad léxica, la ambigüedad sintáctica y estructural y los fenómenos contextuales.

En el ámbito del lenguaje textual, Miller y Johnson-Laird (1976) destacan como importante que la expresión lingüística de una escena o secuencia de acontecimientos está compuesta por elementos discretos, mientras que su imagen mental es un contenido continuo. Surge entonces un problema de asimetría entre contenido y expresión En particular, el productor textual relaciona las expresiones lingüísticas mediante dependencias gramaticales, organizándolos en formatos lineales que permiten construir la superficie textual. El repertorio de dependencias gramaticales de una lengua es mucho más reducido que el repertorio de relaciones conceptuales que necesitan activar los receptores textuales (Beaugrande y Dressler, 1997).

En esta investigación hemos reducido el gran problema de procesar el lenguaje natural, al procesamiento de textos, es decir, al lenguaje en forma escrita. Dentro del amplio universo de textos nos interesa aquellos que se adapten a ciertas características. En las siguientes secciones se expondrán las variables o características que hemos consideramos para seleccionar los textos. Comenzando por factores como el idioma, la gramática para expresar el lenguaje, los tipos de textos y su adecuación al tratamiento computacional. Se discutirá en particular sobre el estilo como un componente significativo de la tipología textual, se realizará una revisión de algunas investigaciones sobre semántica y pragmática en el procesamiento computacional del lenguaje. Se culmina expresando el problema específico en términos de estas variables.

1.6.1 El idioma y la gramática

www.bdigital.ula.ve

Este trabajo se enfoca en aquellos textos escritos en el idioma español. La selección del idioma es importante porque cada lengua tiene sus propias peculiaridades que merecen especial atención, sobre todo al pretender emplear un tratamiento computacional. Según Moreno (1998), la gran mayoría de los avances en el PLN se han enfocados en el inglés, marginando aspectos esenciales del español: la morfología, el orden casi libre de los constituyentes y la omisión del sujeto.

La ambigüedad sintáctica es el principal problema de la flexibilidad del español. Esto obliga a tratar el problema de seleccionar el análisis semántico y pragmático correcto para una oración determinada de entre un número (con frecuencia grande) de análisis sintácticos posibles. Además, como señala Moreno (1998), el lenguaje español presenta algunos fenómenos particulares:

- Coordinación de las estructuras en género y número.
- Orden de constituyentes. Las gramáticas son generalmente desarrolladas para el inglés, en donde el orden de los constituyentes es prácticamente fijo. En español hay más posibilidades de ordenación de los constituyentes oracionales.
- Existencia de elementos nulos o vacíos. El español es una lengua donde habitualmente se omite el sujeto si esta implícito en el contexto o se conoce por las desinencias del verbo.

Desde la perspectiva de los modelos simbólicos, el idioma es un factor determinante para escribir las gramáticas que tratan computacionalmente una lengua natural. Una gramática formal es una especificación rigurosa y explícita de la estructura de una lengua. Las gramáticas formales se escriben siguiendo una convención o lengua artificial creadas parea describir lenguas naturales. Las gramáticas son muy útiles porque están bien definidas, son rigurosas, facilitan la evaluación de hipótesis, permiten hacer predicciones y permiten el desarrollo de aplicaciones (Moreno, 1998). Las gramáticas formales más conocidas y utilizadas son las gramáticas generativas, propuestas por Chomsky (1957), también conocidas como gramáticas de estructura de frase o sintagmáticas.

C.C.Reconocimiento

Una de las gramáticas de la jerarquía de Chomsky más conocida y utilizada en computación son las gramáticas tipo 2 independientes del contexto. El término Gramática Independientes del Contexto (CFG *Context Free Grammar*) indica un modelo particular para describir la sintaxis del lenguaje conforme a algunas limitaciones bastante básicas sobre las reglas y la relación de adyacencia entre los componentes de dichas reglas. La mayoría de los lenguajes naturales parecen seguir un comportamiento libre del contexto, con la excepción posible del idioma alemán suizo (Gazdar y Mellish, 1989). La mayoría de las teorías contemporáneas de sintaxis de lenguaje natural se derivan del esquema de CFG.

Una CFG es una descripción formal de un conjunto de "Reglas de Producción", que definen las oraciones bien formadas del lenguaje. Las reglas, por supuesto, son en sí mismas también escritas en un lenguaje formal definido por un vocabulario y una sintaxis. En la tabla 1.5. se muestra la definición formal de las CFG.

Gramática Independiente del Contexto	
Vocabulario	Las reglas contienen tres tipos de símbolos: • **No Terminales**: Corresponde a los componentes del lenguaje a describir. Uno de los no terminales tiene una posición especial. A este se le llama el símbolo distintivo. • **Terminales**: Corresponde a las palabras del lenguaje a describir. • → : (El símbolo de la flecha) Delimita el lado izquierdo de una regla de su lado derecho.
Sintaxis	Las oraciones en el lenguaje del CFG son las reglas de producción (la sintaxis aquí se refiere al formato de las reglas en sí mismas, no al lenguaje que ellos describen). Una regla de producción tiene las siguientes propiedades: 1. Se compone de un lado izquierdo (LHS *Left Hand Side*) y un lado derecho (RHS *Right Hand Side*) separados por una flecha: LHS → RHS. 2. El LHS se compone de un solo símbolo no terminal. 3. El RHS consiste de uno o más no terminales, o un solo terminal.

www.bdigital.ula.ve

Tabla 1.5. Gramática Independiente del Contexto

Estas gramáticas proporcionan la estructura jerárquica interna de las oraciones. Se pueden describir construcciones recursivas que no podían ser tratadas con las gramáticas regulares. También permite expresar la alternancia y la opcionalidad. Además, las gramáticas independientes del contexto tienen propiedades formales que facilitan el diseño de algoritmos de *parsing*. Sin embargo, hay problemas con el tratamiento de ciertos fenómenos lingüísticos como los constituyentes discontinuos[7], la subcategorización[8] y la concordancia[9]. Moreno (1998) afirma que en la practica, ningún sistema de PLN de cierta cobertura utiliza la versión pura de este tipo de gramáticas.

[7] Constituyentes discontinuos son constituyentes que se pueden encontrar en más de una posición estructural (Moreno, 1998)

[8] Subcategorización es un fenómeno básicamente léxico-semántico en donde la estructura oracional se predice en función de la semántica verbal. Tiene una importancia esencial en la sintaxis porque especifica las posibilidades de combinación de las palabras. Los verbos y algunos adjetivo admiten una estructura de complementos, la subcategorización se refiere al numero y a la categoría de los complementos de cada verbo.

[9] Concordancia es un fenómeno de muchos lenguajes en donde las palabras toman ciertas inflexiones dependiendo de relación que guardan con las otras palabras de una oración. Esta relación se refiere al género y número.

C.C.Reconocimiento

1.6.2 Tipos de textos y estilo

La intertextualidad del modelo textual de Beaugrande y Dressler (1997) es, en un sentido general, la responsable de la evolución de los tipos de textos, entendiéndose por 'tipo' una clase de texto que presenta ciertos patrones característicos. Los tipos de textos son marcos globales que controlan la serie de opciones disponibles que pueden utilizarse en la interacción.

De esta manera, la ciencia del texto describe y hace explícitos los rasgos que identifican los tipos de textos, sin llegar a una descripción completa. Estos rasgos son diversos pues varían desde las características gramaticales hasta los factores comunicativos e intencionales. La función que cumplen los textos, es decir, el hecho de que los textos puedan utilizarse de diversas maneras indica que pertenecen a tipos de textos distintos. Incluso, una propiedad simple como el formato en el que se presenta el texto sobre la página impresa, activa por sí mismo una serie de expectativas acerca del tipo de texto de que se trata.

Los autores mencionados anteriormente también describen algunos de los tipos de textos establecidos tradicionalmente por la tipología lingüística. Estos pueden definirse mediante procedimientos funcionales, es decir, examinando la contribución que realiza cada tipo textual a la interacción comunicativa. Además, pueden definirse por la aplicación de un **patrón global**, utilizado por escritores y lectores para representar los bloques de conocimiento del contenido textual (Beaugrande y Dressler, 1997) (Miller, 1956). Desde esta perspectiva se podrían identificar algunas tendencias dominantes en la tipología textual:

- Los textos **descriptivos** son empleados para enriquecer los espacios de conocimiento cuyos centros de control son las situaciones o los objetos. Es previsible que se dé una elevada frecuencia de aparición de relaciones conceptuales de atribuciones de características, de estados, de ejemplificación y de especificación. La superficie textual descriptiva refleja una elevada densidad de modificadores y complementos. El patrón global más aplicado es el marco[10].
- Los textos **narrativos** se utilizan para organizar discursivamente las acciones y los acontecimientos en un orden secuencial determinado. Seguramente tendrán una elevada frecuencia de aparición las relaciones conceptuales para marcar la causa, la razón, el propósito, la posibilidad y la proximidad temporal. La superficie textual narrativa reflejará una elevada densidad de estructuras subordinadas. El patrón global más aplicado es el esquema[11].
- Los textos **argumentativos** son usados para persuadir al receptor textual de que determinadas creencias o ideas son verdaderas o falsas, favorables o desfavorables para su interés. En este tipo de texto aparecen con mucha frecuencia relaciones conceptuales para expresar la razón, la significación, la volición, el valor y la oposición. La superficie textual argumentativa reflejará una elevada densidad de mecanismos cohesivos que expresan el énfasis y la insistencia (por ejemplo: repetición, paralelismo, paráfrasis). El patrón global que se aplica normalmente es el plan[12] (encaminado a la deducción de creencias).

www.bdigital.ula.ve

La propuesta precedente se refiere a las características y a los usos prototípicos de los tipos textuales, por lo que constituye una primera aproximación al fenómeno. La tipología textual en la realidad comunicativa es un tema complejo que la ciencia del texto ha tratado de simplificar elaborando clasificaciones de tipos puros e ideales de textos. En la practica esta simplificación ha originado la dificultad de que los ejemplos reales no encajan completamente en las características de un tipo específico. Recientemente, la investigación de Adam (1992), demuestra que no existen textos

[10] Los marcos son patrones globales que contienen conocimiento de sentido común sobre algunos conceptos prototípicos (Beaugrande y Dressler, 1997) (Charniak, 1975).

[11] Los esquemas son patrones globales de acontecimientos y de estados integrados en secuencias vinculadas por relaciones de causalidad y de proximidad temporal (Beaugrande y Dressler, 1997) (Mandler y Johnson, 1997).

[12] Los planes son patrones globales de acontecimientos y de estados que conducen a una meta intencionada (Beaugrande y Dressler, 1997) (Schank y Abelson, 1977).

C.C.Reconocimiento

tipológicamente puros, sino textos en donde se integran secuencias prototípicas de naturaleza diversa (explicativa, descriptiva, argumentativa, narrativa, etc.).

Por ejemplo, los textos científicos intentan incrementar y transmitir el conocimiento aceptado comúnmente acerca del "mundo real". Cumplen la finalidad de explorar, ampliar o clarificar el conocimiento almacenado por la sociedad en un campo específico de hechos. Estos textos contienen la presentación y el análisis de la evidencia a la que se ha llegado a partir de la observación directa de la documentación.

Una característica importante de los textos científicos se presenta en el estudio de Grice (1975). Sobre el problema de la intencionalidad en la producción textual propone una serie de preceptos para comunicar información. Grice plantea la máxima calidad que tiene que ver con la sinceridad, "no diga aquello que cree que es falso, ni aquello de lo que carezca de prueba", que se aplica rigurosamente en los textos científicos.

Por otra parte, las áreas de conocimiento especializadas, en su mayoría científicas, se comunican a través de un lenguaje especializado. Esta área de la lengua pretende una comunicación unívoca y libre de contradicciones basada en una terminología establecida. La función lingüística más frecuente que cumplen los textos especializados es denominativa y expositiva y, además, de carácter referencial. Los términos surgen en la comunicación especializada cuando los especialistas precisan denominar un concepto de su disciplina (Cabré, 2000).

El lenguaje especializado no se agota en el vocabulario especializado sino que se caracteriza por una serie de rasgos, sobre todo sintácticos y estructurales a nivel de texto. No obstante, el discurso especializado está determinado por la dimensión léxica de los textos especializados, es decir, por su terminología (Arntz y Picht, 1995).

Para identificar las características generales de la estructura de textos especializados se debe partir del análisis de un amplio corpus que tenga en cuenta la mayor variedad de tipos de textos posibles (Arntz y Picht, 1995). Hoffmann (1985) realizó un estudio considerando los enfoques de la estilística funcional y un conjunto de textos especializados de diferentes lenguas; precisó características comunes a un gran número de lenguajes especializados. Los principales rasgos morfosintácticos de estos textos son los siguientes:

www.bdigital.ula.ve

- El verbo pierde su referencia temporal concreta, generalmente aparece en tiempo presente y sobre todo en tercera persona del singular.
- A menudo el verbo está en voz pasiva (o pasiva-refleja).
- El verbo común como categoría léxica desempeña un papel relativamente poco importante.
- El sustantivo juega un papel importante.
- El singular se emplea con mucha mayor frecuencia que el plural.
- El adjetivo aparece con relativa frecuencia.

Resulta interesante el hecho de que la comparación realizada por Hoffmann (1985) reflejara numerosas coincidencias aunque se aplicara a distintas áreas especializadas. La comparación de los textos entre áreas tan distintas como las Ciencias Sociales y las Ciencias Naturales no dio resultados divergentes importantes en sus características morfosintácticas.

1.6.3 Semántica y Pragmática

Muchas palabras y oraciones pueden ser ambiguas y tener mas de un significado, su significado puede ser falso o producir implicaciones falsas. El significado depende de los principios que usan las personas cuando hablan, por ejemplo ser relevantes y hacer énfasis en las oraciones verdaderas (Grice, 1975). En este sentido la pragmática tiene dos conceptos importantes: la implicación y la presuposición de las oraciones. La implicación de una oración comprende la información que no es parte de su significado, pero que debe ser inferida por un oyente razonable. Las presuposiciones de una oración son las cosas que deben ser verdaderas para que la oración sea verdadera o falsa. Es decir, sobre la base de las

C.C.Reconocimiento

presuposiciones (conocimiento verdadero de un dominio) las personas interpretan las oraciones y derivan conocimiento (implicaciones) que pueden ser o no verdaderos.

Investigaciones como la de Wiebe, Hirst y Horton (1996), sugieren que todo texto está asociado a un contexto lingüístico particular que determina el significado de todas sus palabras y oraciones. Según esos trabajos, el escritor de un texto se dirige a un lector con algún propósito y los lectores pueden inferir la intención subyacente y usarla para comprender el texto. En los trabajos de (Wiebe, Hirst y Horton, 1996) se concluye que el uso del lenguaje involucra mucho más que creación y comprensión de palabras aisladas. Estos autores entienden que incorporar el contexto significa expresar el matiz y el estilo en el lenguaje. La exacta escogencia de palabras, frases y estructura de las oraciones afectan el significado y el efecto preciso de una palabra. Los aspectos de estilo o enfoque del escritor son mucho más parte del mensaje que pretende el hablante que su propio significado literal.

El conocimiento del mundo se ha logrado expresar con técnicas de representación del conocimiento, tales como modelos, redes semánticas y ontologías. El conocimiento del mundo del hablante, se puede entender entonces como una base de conocimiento. Este conocimiento del mundo interviene en la interpretación oracional y del discurso, pues en base a dicho conocimiento se maneja el conocimiento implícito de los hablantes, que permiten resolver las ambigüedades, y también se resuelven las anáforas y elipsis.

Investigaciones como la de Wiebe, Hirst y Horton (1996) sugieren que todo texto está asociado a un contexto lingüístico particular que determina el significado de todas sus palabras y oraciones. Según esos trabajos, el escritor de un texto se dirige a un lector con algún propósito y los lectores pueden inferir la intención subyacente y usarla para comprender el texto. En los trabajos de Wiebe, Hirst y Horton (1996) se concluye que el uso del lenguaje involucra mucho más que creación y comprensión de palabras aisladas. Estos autores entienden que incorporar el contexto significa expresar el matiz y el estilo en el lenguaje. La exacta escogencia de palabras, frases y estructura de las oraciones afectan el significado y el efecto preciso de una palabra. Los aspectos de estilo o enfoque del escritor son mucho más parte del mensaje que pretende el hablante que su propio significado literal.

www.bdigital.ula.ve

Los trabajos de DiMarco y Hirst (1993), utilizan los estilos con el fin de asegurar que una traducción automática retenga los objetivos de comunicación. Para ello se requiere una estructura sintáctica diferente en el lenguaje destino. Para capturar esta clase de intuición lingüística, estos investigadores desarrollaron la idea de una "gramática de estilos", la cual relaciona las estructuras sintácticas de un lenguaje con un conjunto de objetivos estilísticos independientes del lenguaje. En las tareas de traducción, este objetivo puede ser determinado en el texto origen y ser usado en la generación del nuevo texto.

Como se mencionó en la sección anterior, los diferentes tipos de textos tienen una intención comunicativa caracterizada por ciertos patrones lingüísticos. Los investigadores se ayudan de estas propiedades de los textos para obtener resultados en el procesamiento semántico y pragmático del contenido. Por ejemplo, el trabajo de Mani et. al. (1998) muestra varias técnicas para determinar la información relevante de textos usando modelos de cohesión y coherencia. Estos autores modelan los textos en términos de relaciones entre palabras o expresiones para determinar la conexión del texto. También se describen los textos en base a su coherencia, usando relaciones de alto nivel entre cláusulas y oraciones para determinar la estructura argumentativa del texto. Los resultados muestran resúmenes generados automáticamente evaluados positivamente por humanos, y revela que los métodos de coherencia son más efectivos para obtener la información relevante del contenido textual.

1.6.4 Problema específico

El problema específico de esta investigación consiste en el procesamiento del lenguaje natural de textos escritos en el idioma español. Se restringe a tratar textos escritos en el marco de un área especializada, cuyos autores sigan el estilo de escritura que propone Williams (1990). Las sugerencias de Williams dan fundamento a ciertas reglas para la escritura de documentos denominadas “reglas de estilo”. Se piensa poder identificar los tópicos de cada oración en los textos escritos con algún apego a esas reglas. Por

C.C.Reconocimiento

tanto, la intención es usar los tópicos, según estas reglas, como información básica para extraer descriptores significativos de los párrafos en un texto y posteriormente un resumen.

La solución que se propone a este problema emplea técnicas y modelos de procesamiento del lenguaje natural, específicamente las estrategias de los modelos simbólicos. Se utilizará una gramática formal generativa, concretamente una gramática independiente del contexto que emplea cláusulas definidas. Además, la solución propuesta en esta investigación se limita a emplear el párrafo como unidad de tratamiento de los textos. Según la clasificación de Pereira (1996) se trata de una gramática orientada a tarea, pues la intención es buscar conceptos y relaciones mediante reglas que especifican posibles estructuras con conceptos pertinentes. En el ámbito de las técnicas de resumen automático, se trata de un enfoque híbrido, pues se usan técnicas de abstracción para el análisis del texto y estrategias de reducción en la sintetización de la salida.

Esta limitación y especificación del problema permite mantener como hipótesis principal de esta investigación la siguiente: "Esta aplicación computacional, aplicada a textos especializados escritos en base a las reglas de Williams en el idioma español, produce resúmenes de esos textos aceptables para los humanos". Creemos que en este contexto se pueden obtener resultados aceptables.

Otra hipótesis importante es la siguiente: "la gramática basada en estilos reduce la complejidad de análisis sintáctico del texto, sin afectar su alcance". Al final obtendremos un texto cuyos elementos comprenderemos y podremos ordenar en un resumen.

www.bdigital.ula.ve

C.C.Reconocimiento

2 *Capítulo II: Un Resumidor Simbólico. Un experimento para evaluar las gramáticas basadas en estilo*

Este capítulo contiene la descripción del resumidor basado en los estilos de Williams. En la primera sección se proporciona una visión global del sistema y luego se abordarán los detalles de cada módulo.

2.1 Descripción general de la estrategia del Resumidor Simbólico

El resumidor implementado tiene seis componentes básicos: tokenizador, gramática, claridad, cohesión/coherencia, tópico común y salida. En la tabla 2.1 mostramos un esquema de estos componentes con sus entradas y salida.

Entrada	Componente	Salida
Texto	Tokenizador	Texto segmentado
Texto segmentado	Gramática	Texto etiquetado
Texto etiquetado	Claridad	Lista de tópicos [1]
Lista de tópicos [1]	Cohesión/ Coherencia	Lista de tópicos [2]
Lista de tópicos [2] + factor de resumen	Tópico común	Lista de tópicos [3]
Lista de tópicos [3] + Texto segmentado	Salida	Texto Resumen

www.bdigital.ula.ve

Tabla 2.1: La estructura del resumidor simbólico.

El tokenizador textual segmenta el texto en unidades de procesamiento. Es decir, identifica el párrafo como unidad de extracción y separa sus oraciones y palabras.

El siguiente componente implementa un *parser* para el idioma español en base a una gramática simplificada. La salida que se obtiene al aplicar la gramática es denominada "texto etiquetado" porque separa cada oración en sujeto, verbo y complemento. Esto se realiza en una revisión superficial de la gramática de cada oración (superficial porque apenas separa esos componentes). En dicha revisión, el verbo principal es identificado utilizando un diccionario de verbos para el español, que hemos circunscrito a ciertos dominios procurando mejor cobertura para los textos procesados (ver anexo B.1).

Es importante aclarar que nuestra intención no es realizar un análisis exhaustivo de todos los constituyentes gramaticales de la oración. El objetivo es identificar solamente aquellos constituyentes que nos permitan aplicar las reglas lógicas de los siguientes niveles.

El componente "claridad" se refiere a la claridad oracional que plantea Williams (1990). En esta fase se identifica el tópico de cada oración por separado. También se filtran los marcadores discursivos o conectores gramaticales de cada oración empleando un diccionario de conectores. Williams define tópico como "el sujeto psicológico de la oración". Es decir, al parecer, el tópico es la parte de la oración que lleva la carga lógica del discurso, tanto oral como escrito.

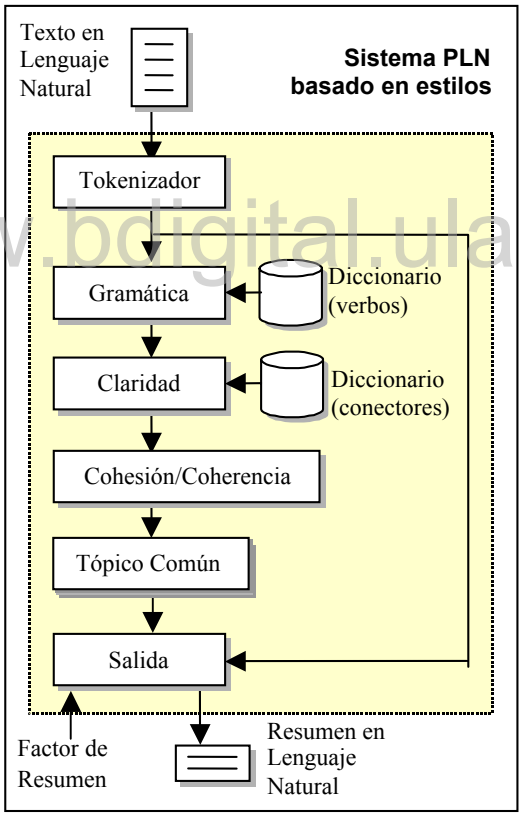
Hemos preferido considerar en los tópicos los conceptos emitidos o involucrados en cada una de las proposiciones que posee un argumento. Desde el punto de vista sintáctico, el tópico generalmente es expresado en una frase nominal que el resto de la oración explica o caracteriza. Por lo tanto, escogemos como tópico las frases nominales contenidas en el sujeto y en el complemento de la oración dependiente del tipo de verbo.

C.C.Reconocimiento

En el ámbito de la semántica del discurso encontramos el siguiente componente que usa las reglas de estilo definidas en el ámbito de los párrafos -reglas de cohesión y coherencia de Williams (1990). Aquí se consideraron los elementos superficiales para eliminar la ambigüedad como la repetición parcial o total de tópicos. También se tomaron en cuenta los elementos para compactar la superficie como las formas pronominales (anáforas) y las elipsis (Beaugrande y Dressler, 1997). El procesamiento de estos elementos está basado en la lista de tópicos de claridad (Lista de tópicos [1]), para obtener una lista de tópico de cohesión (lista de tópicos [2]) de menor cardinalidad. En resumen, a través de este proceso se trata de compactar la lista de tópicos de entrada en función de las reglas de estilo.

Hasta este punto, los componentes del resumidor representan una teoría de apoyo para caracterizar a los tópicos relevantes. Es decir, son una axiomatización que nos dice que ciertas frases de ciertos discursos son tópicos relevantes o adecuados de esos discursos (lista de tópicos [2]). Hasta este punto intervienen las reglas de estilo de Williams.

Adicionalmente hemos podido extender esta teoría de apoyo para incluir un mecanismo de ponderación de tópicos en cada párrafo. Esto consiste básicamente en aumentar la "prioridad" de cada tópico de acuerdo con la ocurrencia de los elementos superficiales de la cohesión textual (Beaugrande y Dressler, 1997) resueltos en el componente anterior. Luego se considera el factor de resumen (varía de 0 a 2, de menor a mayor capacidad de síntesis respectivamente) para escoger los tópicos que representarán a cada párrafo.



www.bdigital.ula.ve

Figura 2.1. Un diagrama del resumidor simbólico.

Para finalizar, el último componente genera un resumen por párrafo del texto original. Esto consiste en mostrar como resumen las oraciones del texto que contienen los tópicos de la lista de tópicos [3]. Aquellas oraciones que no contienen ningún tópico en la lista de tópicos final son eliminadas de la

C.C.Reconocimiento

salida. Cada oración resumen tiene una marca o etiqueta en el tópico para diferenciarlo del resto de la oración.

La estrategia descrita permite que nuestro resumidor sea capaz de extraer del texto original aquellas oraciones que contienen los tópicos del párrafo, al cual denominamos *texto resumen*. Esto se puede observar gráficamente en la figura 2.1.

En el capítulo anterior, se mencionaron algunos modelos que la lingüística textual y otras disciplinas han propuesto para expresar los procesos de producción y recepción textual. Con mayor énfasis fueron tratados algunos modelos de recepción textual en la sección 1.3. Partiendo de estas referencias podemos culminar esta sección haciendo una comparación entre las reglas que propone Behrens y Rosen (1982) para obtener el resumen de un texto y las reglas que proponemos para que un computador obtenga resúmenes a partir de párrafos (ver Tabla 2.2.).

Pasos para obtener resúmenes de textos (Behrens y Rosen, 1982)	**Pasos para obtener resúmenes de párrafos (resumidor simbólico)**
(Paso 1) Dividir y etiquetar el texto en secciones o fases del pensamiento. Subrayar los términos o las ideas claves.	(1) Dividir y etiquetar el texto en palabras, oraciones y párrafos (Tokenizador). Para cada oración identificar el trío [sujeto, verbo, complemento] (Gramática).
(Paso 2) Escribir un resumen (de una sola oración) de cada fase del pensamiento, si fuese apropiado de cada párrafo.	(2) Seleccionar tópico de cada oración (Módulo Claridad). Seleccionar tópicos por párrafo (Cohesión / Coherencia).
(Paso 3) Escribir una oración tópico resumen, es decir, una oración resumen de todo el texto. La oración resumen debe expresar la idea central del texto. El contenido del resumen depende la intención del autor, por ejemplo de los textos (a) informativos debe colocarse la información de qué, cómo, cuándo, dónde y cómo, de los (b) argumentativos debe presentarse la conclusión del autor (c) descriptivos debe incluirse la información del objeto descrito y sus características principales.	
(Paso 4) Escribir resumen del texto. Combinar las oraciones tópicas (paso 3) con el resumen de cada fase del pensamiento (paso 2). Eliminar repeticiones y combinar las oraciones para lograr una corriente lógica de ideas.	(3) Escribir un resumen con todas las oraciones que contengan a los tópicos del párrafo (Salida).
(Paso 5) Revisar el resumen. Insertar palabras y frases de transición donde sea necesario para asegurar la coherencia.	

www.bdigital.ula.ve

Tabla 2.2. Cuadro comparativo entre los pasos para escribir un resumen de Behrens y Rosen (1982) y los pasos del resumidor simbólico.

Aunque existe cierta analogía entre ambos planteamientos, se puede observar que el Paso 3 de la propuesta de Behrens y Rosen no tiene un equivalente en los componentes del resumidor. Esto se debe a que este paso involucraría manejar información conceptual (idea central del texto) y pragmática (intención del autor), que aún no incorpora el resumidor. Por razones de alcance de este proyecto no contamos con ningún tipo de representación de conocimiento de los textos (Ver la sección 3.3 sobre la extensión del resumidor).

C.C.Reconocimiento

De igual manera, el Paso 5 tampoco tiene un equivalente en las reglas de nuestro resumidor simbólico. Esto representaría para un programa informático poder manejar información semántica para establecer relaciones entre oraciones. En nuestro caso puede no ser necesario porque nos aprovechamos del hecho de partir de textos basados en los estilos de Williams que usan conectores y marcadores discursivos. Sin embargo, estos conectores pueden no ser los más adecuados en las oraciones que permanecen en el resumen resultante. Estos puntos sobre la extensión del resumidor se discutirán con mayor profundidad en la sección 3.3.

2.2 Descripción detallada e implementación del resumidor

En esta sección mostramos los pormenores involucrados en la implementación del resumidor. Se comienza con una breve descripción y justificación de las herramientas computacionales utilizadas. Posteriormente se continua con las especificaciones técnicas de cada módulo y se finaliza con la descripción de los diccionarios empleados en el procesamiento.

2.2.1 Metodología y herramientas empleadas en la implementación

Según Quesada y Amores (2000), la mayor parte de los trabajos en PLN, y en Inteligencia Artificial, se han implementado usado lenguajes basados en los paradigmas funcionales (Lisp) y lógicos (Prolog). Esto se debe a que en PLN se aborda el estudio de un problema que carece de un modelo formal bien conocido y, si este estudio pretende resultados computacionales, el modelo recomendable es el de ensayo y error. Es decir, se implementa una primera versión a partir de las intuiciones de partida y, en una espiral de prueba, correcciones y reinicios, se trabaja simultáneamente en la implementación y en el modelo formal que pretende representar y resolver el problema.

Este trabajo ha sido implementado usando la metodología de "prueba y error" un modelo muy genérico de desarrollo de software (Sommerville, 1992). Este modelo de trabajo (ver figura 2.2), junto con las características del problema hace aconsejable el uso de un lenguaje de programación interpretado y con estructuras semánticas cercanas al problema. No interesa tanto la eficiencia y generalización de las técnicas y modelos usados, como demostrar la viabilidad de esas técnicas y modelos.

www.bdigital.ula.ve

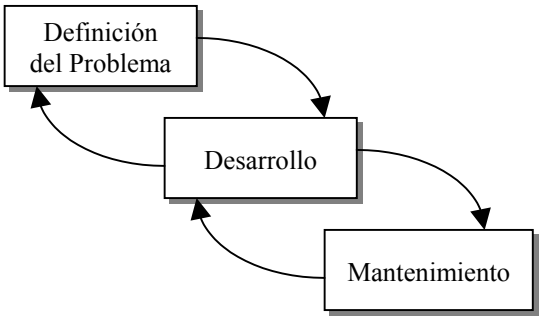


Figura 2.2. Modelo de prueba y error.

Las reglas del resumidor se codificaron en PROLOG, un lenguaje de programación de alto nivel que permite al programador concentrarse en la lógica de su problema, antes que en los medios de ejecución particulares del computador. La herramienta utilizada es el software llamado SWI-Prolog[13] (Bowen y Byrd, 1983). El SWI-Prolog es muy popular y tiene una amplia comunidad de usuarios que ha guiado su desarrollo hasta obtener una herramienta con características de compatibilidad, portabilidad,

[13] SWI-Prolog es una herramienta desarrollada por el departamento SWI (Social Science Informatics) de la Universidad de Amsterdam. Se trata de una implementación de Prolog basada en un subconjunto de la WAM (Warren Abstract Machine (Warren, 1983)). http://www.swi-prolog.org

C.C.Reconocimiento

escalabilidad y estabilidad. Además, tiene la ventaja de ser una implementación basada en una especificación ISO (Deransart et al., 1996).

El interpretador SWI-Prolog ha sido diseñado para obtener implementaciones que puedan ser usadas para experimentar con programación lógica y las relaciones entre programación lógica y otros paradigmas. SWI-Prolog tiene un amplio conjunto de predicados y un rendimiento razonable, lo cual permiten desarrollar aplicaciones importantes. Las versiones actuales de este software ofrecen características favorables como sistema de módulos, recolector de basura e interfaces con otros lenguajes de programación.

Otra herramienta computacional que ha sido usada en la implementación de esta investigación es el TACT (*Textual Analysis Computing Tools*)[14]. Es una herramienta poderosa para realizar experimentos lingüísticos y análisis de discurso. El TACT permitió procesar un corpus textual y extraer la información necesaria para construir los diccionarios de esta propuesta (ver sección 2.8.3). Además, algunas estadísticas generadas por la herramienta fueron usadas en la comparación y evaluación de los resultados (Ver sección 3.2).

El TACT puede realizar tareas como crear diccionarios, obtener colocaciones, y calcular estadísticas. También permite definir parámetros de búsqueda de varias maneras a través de archivos de consulta. El resultado de estas búsquedas puede ser expresado de diferentes maneras.

TACT esta organizado en 15 programas integrados en una interfaz principal, cada programa realiza una única tarea. Un grupo de programas esta orientado a ayudar al usuario en el marcaje de texto. Otro grupo a crear bases de datos textuales (TDB *Textual Database*) en archivos de texto. Además, otro grupo provee maneras para realizar análisis textual: búsqueda por palabras, frases y categorías de palabras, y muestra los resultados en varios formatos. Por otra parte, el sistema TACT produce estadísticas de longitud y frecuencia de palabras, mostrados en orden alfabético y de frecuencia. Esta herramienta realiza comparación entre archivos y cálculo de colocaciones.

2.2.2 Módulos del resumidor

www.bdigital.ula.ve

El resumidor esta organizado en módulos independientes entre sí, pero relacionados a través de los datos que comparten. En esta sección se explica el funcionamiento de cada módulo y se muestran ejemplos de su funcionamiento. Con estos detalles se pretende introducir al lector en la revisión del código Prolog que se despliega en el Anexo A.

2.2.2.1 Tokenizador

El tokenizador permite proveer una entrada en la forma que lo necesita el resto de los módulos del sistema. Este módulo funciona como un traductor, que recibe como entrada el texto en lenguaje natural y obtiene la salida en la sintaxis de Prolog. Este proceso divide la entrada en unidades (tokens) o átomos que serán en este caso las palabras. Se debe tener en cuenta que existen símbolos de separación y de puntuación y que algunos signos de puntuación son considerados como palabras debido a que asumen un comportamiento sintáctico. Cada párrafo del texto original es representado como una lista de oraciones y las oraciones como una lista de palabras.

El diseño y la programación de este módulo tiene como referencia la implementación de un tokenizador presentada en Covington (1994). Este define un procesamiento por niveles y una clasificación de caracteres por tipos que usamos en nuestra implementación. Este tokenizador clasifica los caracteres en los siguientes tipos:

- Caracteres *Fin*, marcan fin de línea, de párrafo, de entrada del archivo.
- Caracteres *Blanco*, que separan palabras.

[14] TACT es un sistema de recuperación y análisis textual sobre bases de datos textuales en idiomas europeos. Este sistema se comenzó a desarrollar como una iniciativa de cooperación hacia las Humanidades por parte de IBM y la Universidad de Toronto durante los años 1986-89. http://www.chass.utoronto.ca/cch/tact.html

C.C.Reconocimiento

- Caracteres *Alfabéticos*, letras que forman parte de las palabras.
- Caracteres *Numéricos*, dígitos que forman parte de fechas, cifras, etc.
- Caracteres *Especiales*, marcas de puntuación, separadores de palabras.

Hay tres niveles de procesamiento: leer un párrafo, leer una oración y leer una palabra. La función de leer párrafo, construye la lista de oraciones hasta conseguir un caracter de *Fin*. Por su parte leer una oración significa construir la lista de palabras y terminar cuando se consiga el caracter de *Fin* o el caracter *Especial* del punto. Leer una palabra consiste en construir un átomo con los caracteres que la forman. La palabra, numérica y alfabética, esta delimitada por caracteres tipo *Fin*, *Blanco* y *Especiales*.

Exponemos un modelo del funcionamiento de este componente en la tabla 2.3. Hemos realizado modificaciones al tokenizador original (Covington, 1994), guiadas por requerimientos particulares del resumidor y el tipo de corpus usado en la implementación y prueba. Los cambios son los siguientes:

1. El tokenizador por cada párrafo genera como salida dos listas de tokens, la primera con los tokens en *lower-case* y la segunda con los tokens originales del texto (incluye los *upper-case*). La segunda lista se usa para efectos de salida. Se conservar las mayúsculas correspondientes a las palabras al inicio de oraciones, a las siglas y nombres propios.
2. Hay excepciones en los caracteres Especiales, ya que no todos son separadores de palabras; algunos funcionan como alfabéticos pues se les agregan algunos tokens según el contexto[15].

Texto Entrada	La oferta mundial de este tipo de grano ha disminuido desde principios del siglo XX entre 40 y 50%, llegando a niveles de 5% de participación mundial (120.000 tm/año). Sin embargo, las exportaciones venezolanas han representado en los últimos años el 8% de la oferta mundial de cacao fino.
Uso del Tokenizador	`tokenizador(Parrafo,ParrafoU,ProximoCaracter)`
Salida	`Parrafo` = [[la, oferta, mundial, de, este, tipo, de, grano, ha, disminuido, desde, principios, del, siglo, xx, entre, 40, y, 50%, (,), llegando, a, niveles, de, 5%, de, participación, mundial, (, 120.000, tm/año,)], [sin, embargo, (,), las, exportaciones, venezolanas, han, representado, en, los, últimos, años, el, 8%, de, la, oferta, mundial, de, cacao, fino]] `ParrafoU`= [[La, oferta, mundial, de, este, tipo, de, grano, ha, disminuido, desde, principios, del, siglo, XX, entre, 40, y, 50%, (,), llegando, a, niveles, de, 5%, de, participación, mundial, (, 120.000, tm/año,)], [Sin, embargo, (,), las, exportaciones, venezolanas, han, representado, en, los, últimos, años, el, 8%, de, la, oferta, mundial, de, cacao, fino]]

Tabla 2.3. Ejemplo del funcionamiento del módulo Tokenizador.

Tenemos dos condiciones necesarias para realizar el procesamiento de segmentación del texto: (1) la entrada debe estar en formato ASCII extendido para considerar los caracteres especiales del español y (2) es importante que el texto no tenga saltos de líneas entre las oraciones del párrafo, ya que el párrafo es identificado por el fin de línea.

[15] Por ejemplo los caracteres especiales %, $, / y °, en tokens que indican porcentajes, costos, temperatura, grados, etc. Los textos especializados contienen datos científicos en su discurso, por eso se hace necesario considerarlos en el procesamiento para tomarlos como un solo token.

www.bdigital.ula.ve
C.C.Reconocimiento

2.2.2.2 Gramática

Este módulo expresa en forma de una gramática de cláusulas definidas, DCG (ver sección 1.6.1), las reglas para procesar una oración. El objetivo de este procesamiento es separar la oración en tres componentes fundamentales: sujeto, verbo y complemento. Para esta tarea partimos de una gramática sencilla que muestra la expresión general de una oración:

```
s --> np, vp.
```

Una frase nominal o sujeto es una categoría gramatical de la frase, abreviada como np. Las frases nominales tienen un nombre, con algunas categorías como adjetivos, determinantes, modificadores, etc. Las frases verbales (vp) contienen un verbo y el resto, si existe, es una frase nominal con algunas categorías adicionales como modificadores y adverbios.

En las lenguas naturales la cantidad de adverbios, nombres, adjetivos y determinantes es suficientemente grande. Por tanto, esperar que una gramática pueda reconocer estos elementos y las posibles combinaciones entre ellos no es una tarea eficiente, ni menos posible si consideramos el problema de la ambigüedad del lenguaje. Para evitar este problema y aprovechando que el objetivo no es ser exhaustivo, se programó una gramática sencilla basada en el verbo. Esto quiere decir que sólo es necesario identificar el verbo como la frontera de los otros dos componentes. El fragmento de oración que antecede al verbo es el sujeto y el que se ubica después es el complemento. El funcionamiento de este módulo se ejemplifica en la tabla 2.4.

Entrada	`Parrafo =` [[en, el, período, comprendido, entre, 1975, y, 1991, (,), foncacao, suministró, asistencia, técnica, y, financiera, a, los, productores], [a, partir, de, 1991, (,), dicho, organismo, cumplió, estas, funciones, en, forma, irregular, debido, fundamentalmente, a, deficiencias, gerenciales, y, administrativas, (, foncacao, (,), 1996,)], [en, consecuencia, (,), a, finales, de, 1999, se, plantea, el, proceso, de, liquidación, del, fondo, nacional, del, cacao], [en, este, contexto, (,), actualmente, las, funciones, de, asistencia, técnica, y, fitosanitaria, están, fundamentalmente, a, cargo, de, instituciones, como, el, fonaiap, (,), la, fundación, ciara, y, la, colaboración, de, algunas, estaciones, experimentales, (,), como, por, ejemplo, chama, y, ocumare, de, la, costa]]
Uso de la Gramática	`gramatica(Parrafo,ParrafoG)`
Salida	`ParrafoG =` [[sujeto([en, el, período, comprendido, entre, 1975, y, 1991, (,), foncacao]), verbo([suministró]), complemento([asistencia, técnica, y, financiera, a, los, productores])], [sujeto([a, partir, de, 1991, (,), dicho, organismo]), verbo([cumplió]), complemento([estas, funciones, en, forma, irregular, debido, fundamentalmente, a, deficiencias, gerenciales, y, administrativas, (, foncacao, (,), 1996,)])], [sujeto([en, consecuencia, (,), a, finales, de, 1999]), [sujeto_verbo([se]), verbo([plantea])], complemento([el, proceso, de, liquidación, del, fondo, nacional, del, cacao])], [sujeto([en, este, contexto, (,), actualmente, las, funciones, de, asistencia, técnica, y, fitosanitaria]), verbo([están]), complemento([fundamentalmente, a, cargo, de, instituciones, como, el, fonaiap, (,), la, fundación, ciara, y, la, colaboración, de, algunas, estaciones, experimentales, (,), como, por, ejemplo, chama, y, ocumare, de, la, costa])]]

Tabla 2.4. Ejemplo del funcionamiento del módulo Gramática.

www.bdigital.ula.ve
C.C.Reconocimiento

Este módulo etiqueta como verbo el primero que aparece en la oración en el caso de oraciones compuestas. Sin embargo, las oraciones generalmente comienzan con conectores o marcadores discursivos[16] para expresar la relación con otras oraciones. Estos conectores pueden ser desde una palabra hasta frases elaboradas que pueden contener verbos. Por ejemplo, las frases como "a partir", "es importante señalar", "es decir", contiene los verbos *partir*, *señalar*, *ser* y *decir*. Por tanto, se presenta un conflicto entre el verbo del marcador discursivo y el verbo de la oración. En este componente se incluyeron reglas para descartar el verbo del marcador discursivo (aunque está primero) y considerar el segundo verbo como el de la oración. Un ejemplo de esto lo podemos observar en la segunda oración de la tabla.2.4.

Por otra parte, parece importante exponer algunos detalles sobre las reglas de definición de los verbos de una oración programados en este componente. En primer lugar hemos considerado cuatro tipos de frases verbales (vp), mostrados en las siguientes reglas:

```
vp(V,C) --> verb_compuesto(V), comp(C).   % verbo compuesto (más de un verbo)
vp(V,C) --> verb_impersonal(V), comp(C). % verbo de una oración impersonal
vp(V,C) --> verb_aux(V), comp(C).         % verbo auxiliar (haber, ser, estar)
vp(V,C) --> verb(V), comp(C).             % los demás verbos
```

El verbo en impersonal es aquel verbo que forma parte de una oración impersonal[17], es decir, aquella que usa el pronombre "se", en lugar de un sujeto general. A este verbo lo llamamos impersonal y lo representamos como un verbo complejo. Un ejemplo de esto lo podemos ver en la tercera oración de la tabla 2.4, en donde se expresa el verbo en una lista como la siguiente: `[sujeto_verbo([se]), verbo([plantea])]`. Incorporar el pronombre como parte del verbo tiene la finalidad de facilitar a los siguientes módulos el tratamiento de las oraciones impersonales.

Como se explicará en la sección 2.2.3, la definición de la gramática requiere tener un diccionario de verbos clasificados en: verbos auxiliares (verb_aux), verbos participios (verb_part) y resto de los verbos (verb). Esta clasificación se justifica por las características de los textos del corpus usado para las pruebas (ver sección 1.6.2). En estos textos predominan las oraciones impersonales y también son frecuentes los verbos compuestos con participios. Además, existe la necesidad de diferenciar el participio, antecedido por un verbo auxiliar, y posibles adjetivos o adverbios del resto de la oración. Por lo tanto, los verbos compuestos quedan definidos en las siguientes reglas:

www.bdigital.ula.ve

```
verb_compuesto([V1,V2]) --> verb_aux(V1), verb_part(V2).
verb_compuesto([V1,V2]) --> verb_aux(V1), verb(V2).
verb_compuesto([V1,V2]) --> verb_aux(V1), verb_aux(V2).
verb_compuesto([V1,V2]) --> verb(V1), verb(V2).
```

Es importante aclarar el comportamiento de este módulo al realizar el análisis de una oración cuyo verbo no está en el diccionario. En este caso, el resultado del *parsing* será la lista vacía (`[]`). Esto es evidentemente un problema de cobertura del diccionario verbal, mas no será un error para el resto de los módulos. El siguiente módulo tiene contemplado el tratamiento de una lista vacía. En términos prácticos dicha oración no estará en el resumen (independientemente de su relevancia).

2.2.2.3 Claridad

Este módulo aplica la claridad oracional de Williams (ver sección 1.5.2) que consiste en identificar el agente en el ámbito oracional según las reglas de estilo. Williams se refiere a la claridad del párrafo

[16] Los conectores o marcadores discursivos son denominados por Williams "frases de transición" (Ver sección 1.4.2. sobre Cohesión y Coherencia).

[17] La tercera persona del singular también es llamada impersonal.

C.C.Reconocimiento

como una "cadena lógica consistente de sujetos" (conformada por los agentes), razón por la cual menciona la influencia de las oraciones activas, pasivas, y la sustantivación en la manifestación del agente. Según Williams el tópico se encuentra generalmente en el sujeto de las oraciones activas, coincidiendo así con el agente según las reglas de claridad (ver sección 1.5.3).

La Tabla 2.5 ilustra el comportamiento de este módulo. Allí observamos la lista resultante de tópicos formada por los sujetos de las tres oraciones del párrafo de entrada.

Entrada	`ParrafoG = [[sujeto([la, epidemia]), [verbo([ha]), verbo([sido])], complemento([desastrosa, para, europa, (,), especialmente, para, el, reino, unido, (,), en, muchos, sentidos])], [sujeto([algunos, columnistas]), [verbo([han]), verbo([hablado])], complemento([incluso, de, annus, horribilis, británico, (,), por, los, problemas, de, inundaciones, (,), accidentes, ferroviarios, (,), vacas, locas, y, aftosa, (,), que, ha, debido, afrontar])], [sujeto([particularmente, el, mal, de, las, vacas, locas]), [verbo([ha]), verbo([puesto])], complemento([muchas, cosas, en, tela, de, juicio, (,), tanto, en, los, aspectos, políticos, (,), como, en, los, técnicos, y, económicos])]]`
Uso de Claridad	`claridad_w(ParrafoG, Topicos)`
Salida	`Topicos = [[la, epidemia], [algunos, columnistas], [el, mal, de, las, vacas, locas]]`

Tabla 2.5. Ejemplo de la Claridad Oracional de Williams.

El resultado de la Tabla 2.5 es válido para un párrafo donde todas las oraciones son activas, pero no todos los párrafos son tan claros. Las oraciones con voz pasivas tienen su agente en el complemento. En ese caso, para Williams es más importante el efecto acumulativo de la secuencia de agentes o tópicos, que los tópicos individuales de cada oración. Por tanto, se pueden escribir oraciones activas y pasivas, dando prioridad a las activas considerando las otras reglas de estilo (cohesión y coherencia).

www.bdigital.ula.ve

En la definición del módulo de claridad entonces será importante considerar las reglas de cohesión y coherencia, porque predominan sobre las de claridad. En este sentido, se debe razonar sobre la definición de cohesión, la cual presenta la vieja información al inicio de la oración y después expone la nueva información. La nueva información está relacionada con la vieja a través de relaciones conceptuales y sintácticamente por medio del verbo. Por tanto, la vieja información define el concepto sobre el que se dice algo en la oración (nueva información), y al que ya hemos denominado tópico (ver tabla 1.4). Por tanto, siguiendo estas reglas parece lógico decir que el tópico es la vieja información de cada oración.

También es importante notar que el párrafo del ejemplo anterior no contiene sujetos cohesionados, es decir, con repetición parcial o total, elipsis o con referencias anafóricas. En el caso de haber este tipo de expresión lingüística en los sujetos, la lista de tópicos formada de esta manera sería insuficiente para procesar posteriormente la cohesión. Por ejemplo, la tabla 2.6 muestra el procesamiento que realiza este módulo, sobre otro párrafo, considerando algunos tópicos extraídos de complementos. Además, lo comparamos con la definición de claridad oracional de Williams.

En el resultado anterior se observa la lista de tópicos formada por los elementos de la lista de la claridad oracional. La salida `TopicosWilliams` (en negritas) está contenida en la lista `Topicos` (lista completa), como se observa a continuación:

> [**[el, pequeño, caficultor]**, [otra, estrategia, distinta, a, la, de, la, mera, búsqueda, de, productividad, (,), desarrollando, tecnologías, tendentes, a, acrecentar, el, uso, de, los, factores, de, producción, más, productivos, (,), tales, como, la, mano, de, obra, (,), mientras, procura, para, su, familia, un, cierto, nivel, de, bienestar, económico, y, social], **[la, unidad, de, producción]**, **[los, estudios, especializados]**, [una, correlación, estrecha, entre, el,

C.C.Reconocimiento

tamaño, de, la, explotación, y, el, rendimiento, obtenido, (,), ya, que, (,), al, parecer, (,), los, resultados, dependen, del, tipo, de, cultivo], **[los, productos, tradicionales, de, exportación, (,), como, el, café, (,), las, unidades, de, menor, tamaño], [un, rendimiento, competitivo, en, comparación, con, unidades, de, mayor, tamaño]**, **[más, intensivamente, los, patrones, tecnológicos, modernos]**]

Entrada	`ParrafoG` = [[sujeto([el, pequeño, caficultor]), [verbo([deberá]), verbo([aplicar])], complemento([otra, estrategia, distinta, a, la, de, la, mera, búsqueda, de, productividad, (,), desarrollando, tecnologías, tendentes, a, acrecentar, el, uso, de, los, factores, de, producción, más, productivos, (,), tales, como, la, mano, de, obra, (,), mientras, procura, para, su, familia, un, cierto, nivel, de, bienestar, económico, y, social])], [sujeto([el, hecho, de, que, la, unidad, de, producción]), verbo([sea]), complemento([pequeña, no, es, un, gran, impedimento])], [sujeto([los, estudios, especializados, no]), [verbo([han]), verbo([encontrado])], complemento([una, correlación, estrecha, entre, el, tamaño, de, la, explotación, y, el, rendimiento, obtenido, (,), ya, que, (,), al, parecer, (,), los, resultados, dependen, del, tipo, de, cultivo])], [sujeto([en, el, caso, de, los, productos, tradicionales, de, exportación, (,), como, el, café, (,), las, unidades, de, menor, tamaño]), verbo([tienen]), complemento([un, rendimiento, competitivo, en, comparación, con, unidades, de, mayor, tamaño])], [sujeto([pero, en, la, medida, en, que, la, unidad, de, producción, aumenta, de, tamaño, (,)]), [sujeto_verbo([se]), verbo([usan])], complemento([más, intensivamente, los, patrones, tecnológicos, modernos])]]
Uso de Claridad	`claridad_williams(ParrafoG, TopicosWilliams)` `claridad(ParrafoG, Topicos)`
Salida Claridad Oracional Williams	`TopicosWilliams` = [[el, pequeño, caficultor], [la, unidad, de, producción], [los, estudios, especializados], [los, productos, tradicionales, de, exportación, (,), como, el, café, (,), las, unidades, de, menor, tamaño], [más, intensivamente, los, patrones, tecnológicos, modernos]]
Salida	`Topicos` = [[el, pequeño, caficultor], [otra, estrategia, distinta, a, la, de, la, mera, búsqueda, de, productividad, (,), desarrollando, tecnologías, tendentes, a, acrecentar, el, uso, de, los, factores, de, producción, más, productivos, (,), tales, como, la, mano, de, obra, (,), mientras, procura, para, su, familia, un, cierto, nivel, de, bienestar, económico, y, social], [la, unidad, de, producción], [los, estudios, especializados], [una, correlación, estrecha, entre, el, tamaño, de, la, explotación, y, el, rendimiento, obtenido, (,), ya, que, (,), al, parecer, (,), los, resultados, dependen, del, tipo, de, cultivo], [los, productos, tradicionales, de, exportación, (,), como, el, café, (,), las, unidades, de, menor, tamaño], [un, rendimiento, competitivo, en, comparación, con, unidades, de, mayor, tamaño], [más, intensivamente, los, patrones, tecnológicos, modernos]]

Tabla 2.6. Ejemplo del módulo claridad comparado con claridad oracional de Williams.

www.bdigital.ula.ve

También se debe considerar aquellos párrafos que contienen elementos de cohesión superficial como las anáforas. Vemos en la tabla 2.7 un ejemplo de un párrafo en donde la segunda oración hace referencia a la nueva información de la primera oración.

En esta oportunidad el resultado de `TopicosWilliams` también está contenido en `Topicos`:

C.C.Reconocimiento

[**[un, informe, de, un, comité, científico, de, la, unión, europea]**, [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas], **[[esta_antes], [singular, G2703]]**, [en, experimentos, de, laboratorio]].

Entrada	`ParrafoG` = [[sujeto([un, informe, de, un, comité, científico, de, la, unión, europea]), verbo([reveló]), complemento([que, las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas])], [sujeto([pero, que, hasta, ahora, esto, solo]), [verbo([ha]), verbo([ocurrido])], complemento([en, experimentos, de, laboratorio])]]
Uso de Claridad	`claridad_williams(ParrafoG, TopicosWilliams)` `claridad(ParrafoG, Topicos)`
Salida Claridad Oracional Williams	`TopicosWilliams` = [[un, informe, de, un, comité, científico, de, la, unión, europea], [[esta_antes], [singular, _G2643]]]
Salida	`Topicos` = [[un, informe, de, un, comité, científico, de, la, unión, europea], [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas], [[esta_antes], [singular, _G2703]], [en, experimentos, de, laboratorio]]

Tabla 2.7. Ejemplo del módulo claridad comparado con claridad oracional de Williams.

Aquí se observa la referencia `[esta_antes]`, relacionada con la nueva información anterior, es decir `[las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas]`. Si sólo se toma en cuenta la salida del `TopicosWilliams` se asociaría de manera errada la referencia `[esta_antes]` con `[un, informe, de, un, comité, científico, de, la, unión, europea]`.

www.bdigital.ula.ve

En vista de esta situación se decidió, para esta versión del resumidor, incluir tanto los sujetos como los complementos en la lista de tópicos del párrafo. Es decir, no manejaremos la lista de sujetos que plantea Williams; en su lugar, se genera una lista con los "tópicos sujetos" y "tópicos complementos" del párrafo. Obviamente, se convierte en una complicación del módulo de claridad, pero simplificará el tratamiento posterior de los elementos de cohesión.

Como se ha observado en los ejemplos anteriores (Figuras 2.7 a 2.9), los tópicos se obtienen eliminando las "frases de transición" de la oración a través de un diccionario de conectores y un filtro de expresiones (ver sección 2.2.3.2). Aplicando la siguiente regla general:

```
claridad([OracionW|ParrafoW],[Topico|TopicosParrafo]) :-
      sujeto(OracionW,Sujeto),
      filtrar_expresion(Topico,Sujeto,[]),
      claridad(ParrafoW,TopicosParrafo).
```

Para determinar el tópico de cada oración se tomaron en cuenta las siguientes variables: el tipo de verbo, la existencia de sujeto (sujeto vacío) y la complejidad del complemento. Con estas variables en cuenta tenemos los siguientes casos:

- Caso 1: La oración contiene un verbo auxiliar. El tópico es el sujeto.
- Caso 2: El sujeto contiene una referencia anafórica. El tópico es el complemento. Además, a la lista de tópicos se le agrega una marca con la información sintáctica de la referencia anafórica del sujeto (género y número del pronombre).
- Caso 3: Oración impersonal. El tópico es el sujeto y el complemento. El sujeto puede ser vacío.
- Caso 4: Defecto. El tópico está en el sujeto y en el complemento.

C.C.Reconocimiento

El "tópico complemento" se obtiene procesando sus oraciones subordinadas. Ello requiere someter nuevamente el complemento por el módulo de gramática, con el objeto de obtener su estructura (sujeto, verbo, complemento). A partir de esta estructura se aplicará recursivamente el proceso de claridad hasta que no se puedan identificar más oraciones secundarias. La siguiente regla muestra la manera en que se reutiliza el componente de gramática y claridad:

```
claridad([OracionW|ParrafoW],[TopicoComplemento|TopicosParrafo]) :-
      complemento(OracionW,Complemento),
      gramatica([Complemento],[ComplementoW]),
      claridad(ComplementoW,TopicoComplemento),
      claridad(ParrafoW,TopicosParrafo).
```

2.2.2.4 Cohesión / Coherencia

Este componente recibe la lista de tópicos del módulo anterior, con el objeto de aplicar algunas reglas de cohesión y coherencia propuestas por Williams. En general, se puede describir este procesamiento a través de la siguiente regla:

```
cohesion_coherencia (Topicos,ListaTopicos) :-
  identificar_topico_arranque(Topicos,ListaPonderada),
  resolver_anafora(ListaPonderada,ListaSinAnafora),
  simplificar_topicos_identicos(ListaSinAnafora,ListaTopicos).
```

Antes de comenzar con la explicación de este módulo, se debe mencionar el cambio estructural de la lista de tópicos. A partir de este punto, los tópicos tendrán un número asociado que hemos denominado *ponderación*. Esta modificación permitirá establecer criterios en el siguiente módulo para valorar la relevancia de los tópicos en función de su ponderación. La tarea, entonces, consiste en simplificar los tópicos a partir de las reglas de cohesión y coherencia, y reajustar sus ponderaciones. Por esta razón, los procedimientos que se explican a continuación procesan esta estructura y la información de ponderación.

www.bdigital.ula.ve

El predicado `identificar_topico_arranque(Topicos,ListaPonderada)`, identifica el arranque (primer tópico) y el discurso del párrafo (resto de los tópicos). Cada elemento de la lista `ListaPonderada` está formado por el número ponderado y la lista de palabras del tópico. Los tópicos de cada párrafo tendrán al inicio la misma ponderación (1), excepto el primero (al cual se le asocia 0). En este punto se están aplicando las reglas de estilos relacionadas con el tópico de arranque (inicio del párrafo), que incluye información conocida o mencionada anteriormente (principio de cohesión de Williams). Es importante notar que Williams no menciona la longitud del arranque de un párrafo. Por tanto, nos limitamos a considerar que el primer tópico del párrafo es por lo menos el primer tópico. Luego de tener inicializada la estructura de los tópicos del párrafo, se hace un recorrido para reconocer y relacionar las anáforas del texto.

El siguiente predicado, `resolver_anafora(ListaPonderada,ListaSinAnafora)`, determina las asociaciones anafóricas del párrafo utilizando sus tópicos. Es decir, este módulo pretende solucionar el problema de resolver las referencias anafóricas del texto. Sobre este problema específico se han realizado muchas investigaciones para identificar o revelar el objeto antecedente de la referencia anafórica debido a su dependencia contextual y pragmática (Covington, 1994) (Palomar, 2001). La mayoría de estos algoritmos son muy complejos y requieren un sistema con un amplio conocimiento lingüístico, además de la manipulación de contenido semántico. Por razones de alcance del proyecto, hemos implementado un algoritmo muy sencillo para realizar esta tarea sin pretender cubrir todos los casos y manteniendo un margen de error en el resultado. Nuestra principal simplificación consiste en considerar la lista de tópicos como los posibles objetos referenciados por la anáfora y en resolver la

C.C.Reconocimiento

anáfora en función de la sintaxis. Lo que se propone es buscar hacia atrás un tópico con el mismo número y género del pronombre demostrativo que representa la referencia anafórica.

Surge entonces otro problema: cómo identificar el número y el género de un tópico sin incluir procesamiento morfológico de los sustantivos (mayor conocimiento lingüístico). Para este cuestionamiento también hemos realizado una simplificación en función de las características del idioma español. Las palabras de una oración en español aparecen relacionadas entre sí por medio de fenómenos como la concordancia (género, número, persona). Por tanto, se tiene la posibilidad de establecer el género y número del tópico a través del determinante que, normalmente, lo acompaña.

Entrada	`Topicos` = [[un, informe, de, un, comité, científico, de, la, unión, europea], [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas], [[esta_antes], [singular, _G2703]], [en, experimentos, de, laboratorio]]
Uso Cohesión y Coherencia	`cohesion_coherencia(Topicos,ListaTopicos)`
Tópico Arranque	`identificar_topico_arranque(Topicos,ListaPonderada)` `ListaPonderada` = [[0, [un, informe, de, un, comité, científico, de, la, unión, europea]], [1, [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas]], [1, [[esta_antes], [singular, _G2703]]], [1, [en, experimentos, de, laboratorio]]]
Resolver Anáfora	`resolver_anafora(ListaPonderada,ListaSinAnafora)` `ListaSinAnafora` = [[0, [un, informe, de, un, comité, científico, de, la, unión, europea]], [2, [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas]], [1, [en, experimentos, de, laboratorio]]]

Tabla 2.8 Ejemplo de las reglas "tópico arranque" y "resolver anáfora" del módulo cohesión / coherencia.

www.bdigital.ula.ve

Para identificar las referencias anafóricas se ubica la marca `[esta_antes]` y las características de género y número. A partir de esta marca se divide la lista de tópicos en dos listas: `Antes` y `Despues`. Para resolver la anáfora debe buscarse, en sentido contrario, un determinante en la lista `Antes` que concuerde con las características de la marca. La siguiente regla describe formalmente este procedimiento:

```
resolver_anafora(ListaPonderada,ListaSinAnafora):-
  buscar_anafora(ListaPonderada,Antes,Despues,DatoLinguistico),
  invertir_lista(Antes,AntesInv),
  relacionar_topico(AntesInv,DatoLinguistico,TopicoRelacionado),
  invertir_lista(TopicosRelacionado,TopicoRelacionadoInv),
  concatenar_listas([TopicoRelacionadoInv,Despues],ListaNueva,[]),
  resolver_anafora(ListaNueva,ListaSinAnafora).
```

El predicado `relacionar_topico` elimina la referencia anafórica y modifica la ponderación del tópico relacionado, sumando las ponderaciones de ambos. Para ilustrar este procedimiento obsérvese el ejemplo de la Tabla 2.8.

El último procedimiento trata la cohesión de los tópicos por su repetición parcial o total, `simplificar_topicos_identicos(ListaSinAnafora, ListaSinDuplicado)`. Este procedimiento se aplica a la lista de tópicos resultante de aplicar la resolución anafórica. El predicado mencionado busca los tópicos que contengan palabras idénticas en orden secuencial en la lista `ListaSinAnafora`.

C.C.Reconocimiento

En esta etapa, es necesario recordar el concepto de "cadena consistente de tópicos" que menciona la cohesión de Williams, el cual se refiere a las relaciones que existen entre los tópicos de cada una de las oraciones de un párrafo. Estas relaciones pueden ser de diferente naturaleza, sintácticas o conceptuales. En nuestro caso sólo nos hemos ocupado de las relaciones sintácticas de la superficie textual, propias de la cohesión. Por ejemplo, se implementó una función considerando que los tópicos no aparecen sintacticamente idénticos, pues el autor del texto puede usar sinónimos o repetición parcial del tópico. Si un tópico aparece más de una vez en el texto, es más probable considerarlo como un buen representante del texto que si solo aparece una vez. Por tanto, se utiliza una definición trivial, pero práctica, para decir que dos tópicos son el mismo. Esta definición se utiliza dentro de la regla `simplificar_topicos_identicos` que se muestra a continuación:

```
simplificar_topicos_identicos([TopicoP|Resto],[TF|ListaSinDup]):-
  topico(TopicoP,Topico),
  filtrar_sustantivo_adjetivo(Topico,ListaSustantivoAdjetivo),
  fusionar_palabra(ListaSustantivoAdjetivo,TopicoP,Resto,[TF,ListaF])
  ,
  simplificar_topicos_identicos(ListaF,ListaSinDup).
```

Entrada	`Topicos` = [[venezuela], [el, café], [un, vuelco, en, la, actividad, agrícola, (,), ya, que, este, rubro, comienza, a, desplazar, al, cacao, disminuyendo, su, importancia, en, los, mercados, internacionales, (,), y, consecuentemente, declinando, su, producción, y, valor, exportado], [la, actividad, petrolera], [escenario, económico, venezolano], [[esta_antes], [singular, femenino]], [un, desplazamiento, total, de, las, exportaciones, de, café, (, líder, entonces, de, las, exportaciones, venezolanas,), por, las, exportaciones, petroleras]]
Uso de Claridad	`cohesion_coherencia(Topicos,ListaTopicos)`
Tópico Arranque	`identificar_topico_arranque(Topicos,ListaPonderada)` `ListaPonderada` = [[0, [venezuela]], [1, [el, café]], [1, [un, vuelco, en, la, actividad, agrícola, (,), ya, que, este,rubro, comienza, a, desplazar, al, cacao, disminuyendo, su, importancia, en, los, mercados, internacionales, (,), y,consecuentemente, declinando, su, producción, y, valor, exportado]], [1, [la, actividad, petrolera]], [1, [escenario, económico, venezolano]], [1, [[esta_antes], [singular, femenino]]], [1, [un, desplazamiento, total, de, las, exportaciones, de, café, (, líder, entonces, de, las, exportaciones, venezolanas,), por, las, exportaciones, petroleras]]]
Resolver Anáfora	`resolver_anafora(ListaPonderada,ListaSinAnafora)` `ListaSinAnafora` = [[0, [venezuela]], [1, [el, café]], [1, [un, vuelco, en, la, actividad, agrícola, (,), ya, que, este, rubro, comienza, a, desplazar, al, cacao, disminuyendo, su, importancia, en, los, mercados, internacionales, (,), y, consecuentemente, declinando, su, producción, y, valor, exportado]], [2, [la, actividad, petrolera]], [1, [escenario, económico, venezolano]], [1, [un, desplazamiento, total, de, las, exportaciones, de, café, (, líder, entonces, de, las, exportaciones, venezolanas,), por, las, exportaciones, petroleras]]]
Simplificar Tópicos idénticos	`simplificar_topicos_identicos(ListaSinAnafora,ListaSinDuplicado)` `ListaSinDuplicado` = [[0, [venezuela]], [2, [el, café]], [3, [la, actividad, petrolera]], [1, [escenario, económico, venezolano]]]

Tabla 2.9 Ejemplo del módulo cohesión / coherencia.

www.bdigital.ula.ve
C.C.Reconocimiento

La regla anterior muestra un proceso simple. Donde se asume que un buen escritor piensa bien cada tópico cuando lo presenta por primera vez y mantiene su forma igual durante el resto del texto. Por lo tanto, se ha implementado un procedimiento para buscar tópicos en donde se repiten palabras (no se toman en cuenta determinantes, pronombres, conectores, etc.). Si la misma palabra (sustantivo, adjetivo) se repite en algunos de los tópicos restantes, entonces se modifica (se suman) la ponderación del primer tópico y se elimina el segundo tópico de la lista (`fusionar_palabra`). Así, cada vez obtendremos menos tópicos por cada párrafo. A través de esta estrategia se asume que el autor sabe ordenar sus ideas y coloca primero el mejor de los tópicos. Hace la excepción de escoger el segundo tópico en el orden del texto si tiene mejor ponderación que el primero. En la tabla 2.9 se muestra un ejemplo del módulo completo y, en particular, de la simplificación de tópicos idénticos.

Como puede observarse esta estrategia es muy simple y la cobertura de los elementos de cohesión no es completa. Sin embargo, se ha tratado de conservar la robustez cuando no se puedan resolver las anáforas o identificar los tópicos idénticos. En esas circunstancias la lista resultante de tópicos consistirá en los tópicos de entrada sin las marcas de referencias anafóricas. Pensamos que el componente cohesión / coherencia del resumidor es la parte más susceptible a modificación y expansión por su relación más directa con el nivel semántico. Por ejemplo, se puede incorporar el "conocimiento del dominio" para identificar otras similitudes entre otros tópicos y ayudar en la resolución de las anáforas.

2.2.2.5 Tópico común

La idea del componente *tópico común* es reducir la cardinalidad del conjunto de tópicos según una necesidad de síntesis del usuario. El usuario plantea un factor de resumen para indicar el nivel de reducción requerido del texto. De esta manera el predicado recibe como tercer parámetro el `Factor` de resumen `seleccionar_topico_comun(ListaTopicos, TopicoComun, Factor)`.

En el módulo anterior se generó como salida una lista donde cada tópico tiene asociado un número indicando su ponderación. Toda esta información es recibida como insumo para el procesamiento del tópico común. El primer paso consiste en ordenar dicha lista por el valor de la ponderación de cada tópico y descartar algunos de ellos. El descarte está condicionado al nivel de resumen que el usuario necesita en la salida. Se han considerado tres niveles de resumen:

www.bdigital.ula.ve

- El Nivel 0 elimina los tópicos con ponderación igual a 0.
- El Nivel 1 descarta los tópicos con ponderación 0 y 1.
- El Nivel 2 desecha todos los tópicos excepto el mejor: aquel que presenta mayor ponderación.

Ninguno de estos niveles toma en cuenta el porcentaje de tópicos a descartar; simplemente los elimina. De esta manera, si se tienen varios tópicos con ponderación 0 y 1 y sólo un tópico con ponderación mayor que 1, el resumen del nivel 1 y 2 será exactamente igual. Estas sencillas reglas para definir el factor de resumen pueden no resultar las más adecuadas en algunos casos, tal y como ya lo hemos visto.

2.2.2.6 Salida

Este módulo tiene el objetivo de procesar la lista de tópico común para generar como salida un resumen del párrafo original. En la definición del alcance de este proyecto se determinó el tipo de resumen que se pretendía generar. Aunque se están usando técnicas de abstracción para obtener el resumen automático, el producto no pretende ser un resumen constructivo. En este caso la generación del resumen consiste en la extracción de las oraciones del párrafo original consideradas más relevantes. Las oraciones relevantes son aquellas que contienen el tópico de la lista *tópico común*.

En este punto es importante conservar la ortografía del texto original en relación con el uso de las mayúsculas, como el comienzo de las oraciones, los acrónimos y nombres propios. Como se mencionó en el módulo de tokenizador se generaron dos listas separadas en *tokens*. La lista *resultado* esta en

C.C.Reconocimiento

minúscula y la segunda lista contiene las mayúsculas del texto fuente. En este módulo se utiliza la segunda lista para generar la salida con la siguiente regla:

```
salida_html(Texto,TextoUpper,TopicoComun):-
   etiquetar_topico(TopicoComun,Texto,TextoUpper,TextoEtiquetado),
   imprimir_etiquetas_encabezado,
   imprimir_texto(TextoEtiquetado),
   imprimir_etiquetas_cierre.
```

El predicado `etiquetar_topico` resalta los tópicos del texto original (`TextoUpper`) con una etiqueta HTML en la salida estándar. De esta manera, el resultado sería el que se muestra en la Tabla 2.10.

Texto Fuente	Un informe de un comité científico de la unión Europea reveló que las ovejas y las cabras pueden contraer, teóricamente, el mal de las vacas locas. Pero que hasta ahora esto solo ha ocurrido en experimentos de laboratorio.
Entrada	`TopicoComun` = [[1, [en, experimentos, de, laboratorio]], [2, [las, ovejas, y, las, cabras, pueden, contraer, (,), teóricamente, (,), el, mal, de, las, vacas, locas]]]
Uso de Salida	`salida(Texto, TextoUpper, TopicoComun)` con factor 0
Salida	Un informe de un comité científico de la unión Europea reveló que **las ovejas y las cabras pueden contraer, teóricamente, el mal de las vacas locas**. Pero que hasta ahora esto solo ha ocurrido **en experimentos de laboratorio**.

Tabla 2.10 Ejemplo del módulo salida.

www.bdigital.ula.ve

En esta primera versión del resumidor se empleó el formato HTML para generar la salida a través de una interfaz estándar. Puede considerase el uso de un conjunto de etiquetas que aporte mayor información como XML y aprovechar las flexibilidades y beneficios de esta tecnología.

2.2.3 Diccionarios empleados

Cualquier sistema de PLN requiere un lexicón o diccionario con información morfológica, sintáctica o semántica de las palabras de la lengua. Estos diccionarios computacionales se caracterizan por una clara división de los tipos de información, expresada de manera formalizada y estructurada. El lexicón siempre depende de la gramática, pues las entradas léxicas no son más que los elementos terminales que se insertan en las reglas gramaticales. Los lexicones más sencillos son los de las gramáticas sintagmáticas independientes del contexto. Su complejidad depende de la cantidad de información sintáctica y semántica (Moreno, 1998).

La mayoría de los datos de los diccionarios empleados en esta investigación se obtuvo usando la herramienta TACT, descrita en la sección 2.2.1. Luego se aumentaron los datos progresivamente con nuevas entradas según la experiencia en el desarrollo de la propuesta. Otro aspecto importante de mencionar es la estructura lineal de estos diccionarios. Este tipo de organización es la forma más elemental de estructurar información, puesto que sigue el orden en que se han añadido las entradas. Este tipo de estrategia es generalmente utilizada en gramáticas experimentales.

A continuación se explican los diccionarios empleados, su generación, utilización y lineamientos para futuras expansiones.

C.C.Reconocimiento

2.2.3.1 Diccionario de verbos

La herramienta TACT permitió obtener a partir del corpus de prueba los verbos en sus formas infinitivo y participio. La extracción de estos verbos se realizó a través de la definición de expresiones regulares en base a la morfología de las desinencias verbales.

Como se determinó en el objetivo de este proyecto, el tipo de textos que se pretende atacar presenta ciertas características morfosintácticas que también limitan la cobertura de este diccionario. Los rasgos particulares son la alta frecuencia de aparición de verbos en tiempo presente y, sobre todo, en tercera persona del singular. Además, es muy común que el verbo esté en voz pasiva (o pasiva-refleja). Esta restricción de alcance nos redujo el trabajo a un diccionario verbal muy simple. Por tanto, no fue necesario considerar la implementación de un conjugador verbal porque hubiera incrementado el tiempo de desarrollo, la complejidad y el rendimiento de la aplicación[18].

Por lo tanto, el lexicón verbal es incompleto es este estudio porque no posee todos los verbos del corpus de prueba. Por ejemplo, faltan los verbos irregulares en participio y los verbos en tiempo pasado y tercera persona. La tabla 2.11 muestra los dos grupos de verbos que conforman el lexicón.

Grupo obtenidos con el TACT	
verbos infinitivos	Los verbos infinitivos que aparecen en el texto, reconocidos por las desinencias verbales indicativas del modo infinitivo (ar, er, ir). Se revisó manualmente la lista obtenida para eliminar sustantivos como *lugar*, *primer*, etc.
	`verb(verbo([abastecer])) --> [abastecer].` `verb(verbo([morir])) --> [morir].`
verbos con enclíticos	Los verbos cuasirreflejos identificados por su terminación (-se).
	`verb(verbo([abastecerse])) --> [abastecerse].` `verb(verbo([agruparse])) --> [agruparse].`
verbos participios regulares	Los verbos en participio pasado terminados en los sufijos ado y ido
	`verb_part(verbo([acometido])) --> [acometido].` `verb_part(verbo([acompañado])) --> [acompañado].`
verbos gerundios	Los verbos gerundios regulares terminados en los sufijos ando y endo.
	`verb(verbo([simplificando])) --> [simplificando].` `verb(verbo([surgiendo])) --> [surgiendo].`
Grupo obtenido manualmente	
verbos auxiliares	Se incluyó la conjugación completa de los verbos auxiliares: ser, estar y haber.
	`verb_aux(verbo([haber])) --> [haber].` `verb_aux(verbo([habiendo])) --> [habiendo].`
verbos modales perifrásticos[19]	Se incluyó la conjugación completa de los verbos: tener, poder y deber.
	`verb(verbo([debiendo])) --> [debiendo].` `verb(verbo([debo])) --> [debo].`
verbos irregulares	La lista de verbos irregulares del corpus fue obtenida manualmente.
	`verb(verbo([llevan])) --> [llevan].` `verb(verbo([suministró])) --> [suministró].`

Tabla 2.11. Tipos de verbos que conforman el diccionario verbal del resumidor.

[18] Esta complicación se basa en que la conjugación verbal del español está compuesta por 55 formas (descartando las formas obsoletas del futuro imperfecto del subjuntivo). Se trata de un número elevado en comparación con otros sistemas morfológicos, pero en el español existen formas diferentes que utilizan el mismo morfema flexivo, con lo cual se produce una redundancia en las descripciones tradicionales. La morfología verbal del español se puede representar en pocas jerarquías de desinencias verbales (Moreno, 1998).

[19] Bosque y Demonte (1999).

www.bdigital.ula.ve
C.C.Reconocimiento

En función de la descripción anterior del diccionario se pueden expresar las vías para aumentar el tamaño del lexicón. Para ello es necesario observar los predicados, pues sólo hay tres opciones: `verb_aux`, `verb` y `verb_part`. Los verbos auxiliares `verb_aux` están completos, no así los verbos participios `verb_part` y los verbos irregulares, en infinitivo y en gerundio etiquetados como `verb`. Debe escribirse la cláusula `verb_part(verbo[X]) --> [X]` para agregar un verbo X en participio y una entrada como `verb(verbo[Y]) --> [Y]` para añadir un verbo `Y` del resto de las formas verbales.

Por otra parte, el módulo de cohesión y coherencia requiere un diccionario de pronombres demostrativos y determinantes para el proceso de resolución de anáforas (explicado en la sección 2.2.2.4). Este pequeño diccionario contiene las entradas léxicas de los pronombres demostrativos del español extendidos con información sintáctica (específicamente género y número), como se muestra a continuación:

```
pronombre(singular,_) --> [eso].
pronombre(plural,masculino) --> [estos].
```

Los artículos o determinantes se expresaron como hechos. Además, se les agregó información sintáctica:

```
es_articulo(el,singular,masculino).
es_articulo(los,plural,masculino).
```

2.2.3.2 Diccionarios de conectores

Se entiende por conectores aquellas expresiones que nos permiten engranar los diferentes niveles del discurso. Generalmente son utilizados para orientar al lector y garantizar una adecuada interpretación. En ese sentido, se conciben como elementos de metainformación sobre el contenido textual.

www.bdigital.ula.ve

En el principio de cohesión de Williams se menciona el uso de frases de transición para conectar las oraciones de un párrafo antes de introducir el tópico. Las frases son de varios tipos: conectores lógicos, expresiones de evaluación y expresiones de tiempo-espacio. Las entradas de este lexicón simplemente identifican la cadena de caracteres que compone cada una de estas frases de la siguiente manera:

```
expresion([por, lo, tanto]) --> [por, lo, tanto].
expresion([como, resultado]) --> [como, resultado].
```

La lista de palabras de la frase es la unidad o elemento terminal de la gramática que permite eliminar esta frase de una oración. Esta gramática es denominada “filtro de expresiones” en el componente Claridad del resumidor. Las reglas que lo definen son las siguientes:

```
filtrar_expresion([]) --> [].
filtrar_expresion(Topico) --> expresion(_),
filtrar_expresion(Topico).
filtrar_expresion(Topico) --> filtrar_expresion_resto(Topico),
expresion(_).
filtrar_expresion(Resto) --> filtrar_expresion_resto(Resto).
filtrar_expresion_resto([]) --> [].
filtrar_expresion_resto([X|Resto]) --> [X|Resto].
```


C.C.Reconocimiento

Para la generación de este diccionario se utilizaron fuentes distintas. Una parte estuvo constituida por los ejemplos de la propuesta de Williams. Otra parte estuvo constituida por las colocaciones[20] presentes en los textos de prueba que se obtuvieron con la aplicación TACT. Las demás entradas del diccionario fueron ingresadas manualmente en las pruebas informales del sistema.

Conclusión

En este capítulo se ha explicado detalladamente el funcionamiento del resumidor y superficialmente el código que lo implementa. Para una revisión más completa se sugiere revisar el código fuente en Prolog del Anexo A.

www.bdigital.ula.ve

[20] Una colocación es una secuencia de dos o más palabras repetidas en un texto.

C.C.Reconocimiento

3 *Capítulo III: Evaluación de los Resultados del Experimento*

El resumidor expuesto en el capítulo anterior fue implementado con una interfaz web en Internet[21] (figura 3.1) para facilitar su uso y prueba.

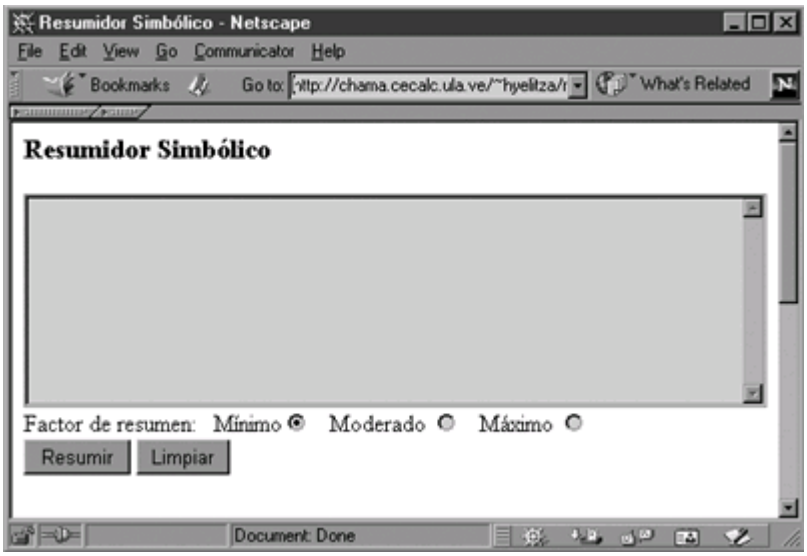


Figura 3.1. Resumidor simbólico.

www.bdigital.ula.ve

A continuación, se ilustra su funcionamiento con un ejemplo. En (1) se muestra un párrafo y en (2), el resumen producido por este programa:

(1) El deterioro de la actividad agrícola se tradujo en una menor participación del sector en la actividad económica, debido principalmente al surgimiento de la industria petrolera. En consecuencia, dentro del mencionado sector, la declinación de la actividad cacaotera fue incontenible. El Estado crea el Fondo Nacional del Café y del Cacao (FNCC) en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros. Esto formó parte de las políticas proteccionistas gestadas en esa época. En 1975 se reestructura el Fondo Nacional del Café y del Cacao, dividiéndose en dos organismos autónomos e independientes: el Fondo Nacional del Café (FONCAFE) y el Fondo Nacional del Cacao (FONCACAO). De esta forma, se inicia el monopolio de la compra, distribución y exportación del cacao, ejercido por el Estado a través de FONCACAO como empresa comercializadora.

(2) En consecuencia, dentro del mencionado sector, **la declinación de la actividad cacaotera** fue incontenible. **El Estado** crea **el Fondo Nacional del Café y del Cacao (FNCC)** en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros. Esto formó **parte de las políticas proteccionistas gestadas en esa época**. De esta forma, se inicia **el monopolio de la compra**, distribución y exportación del cacao, ejercido por el Estado a través de FONCACAO como empresa comercializadora.

[21] Página con interfaz web para el resumidor http://chama.cecalc.ula.ve/~hyelitza/resumidor .html o http://cesimo.ing.ula.ve/INVESTIGACION/PROYECTOS/GIL/resumidor.html

C.C.Reconocimiento

El texto anterior de 140 palabras es reducido a las 83 mostradas en el ejemplo (2). Observe que el párrafo conserva algo de sentido y contiene material relevante, aún cuando el resumidor no posee conocimiento experto sobre el dominio.

Las reglas de estilo de Williams nos permiten explicar el resultado anterior. En primer lugar, las reglas de coherencia indican que las ideas que se colocan al inicio de un párrafo deben contener información conocida por el lector o referida anteriormente en el texto. Esta información se denomina arranque de un párrafo coherente (Williams, 1990). Las características del arranque nos lleva a considerarlo como información menos relevante para el resumen del párrafo. Por esta razón, la primera oración del texto (1) es eliminada del texto (2).

Luego del arranque del párrafo se tiene una serie de oraciones que introducen cadenas temáticas con relación al tópico arranque. Este conjunto de oraciones se denomina discusión de un párrafo coherente (Williams, 1990). Los tópicos (destacados en negritas) contenidos en la discusión se obtienen al aplicar las reglas de claridad y cohesión. En base a estos tópicos y a las reglas de "tópico común", se eliminó la quinta oración del texto original. Como se puede observar, el tópico de esta oración es mencionado anteriormente (Fondo Nacional del Café y del Cacao), y se le agregó información nueva. En particular, la repetición parcial o total de tópicos nos permite reducir el contenido del resumen escogiendo solamente su primera ocurrencia.

Por otra parte, nuestro resumidor puede recibir como parámetro un mayor nivel de síntesis (factor de resumen igual a 2), permitiendo obtener sobre el mismo ejemplo la siguiente salida:

(3) El Estado crea **el Fondo Nacional del Café y del Cacao (FNCC)** en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros.

Se puede observar que la oración del ejemplo (3) es apropiada para transmitir el sentido de ese párrafo.

www.bdigital.ula.ve

3.1 Evaluación

Un resumidor, en la intuición de cualquier lingüista, debería tener en cuenta el concepto de relevancia o de un substituto apropiado, respecto a algún contexto o solicitud de información. Una noción de la relevancia ha sido propuesta por Cooper (1971) y la denomina "relevancia lógica". La relevancia está definida en términos de la consecuencia lógica: "un documento es relevante a una necesidad de información, si y solamente si, contiene por lo menos una sentencia que sea relevante a esa necesidad".

Los investigadores del área de PLN han destacado la importancia de producir metodologías de evaluación de estos sistemas porque pueden resultar muy costosos y complejos (King, 1996). King dice que los resultados varían enormemente en función del propósito, alcance, y naturaleza de los objetos que están siendo evaluados. En el ámbito específico de las técnicas de resumen automático, parece que la calidad y efectividad de los resúmenes es difícil de medir debido a lo complicado que es determinar las propiedades de un buen resumen (Maña, Buenaga y Gómez, 1998). En las siguientes secciones se plantea un método sencillo de evaluación para demostrar la calidad de los resultados del resumidor simbólico.

3.1.1 Método de evaluación

Los métodos de evaluación de resúmenes automáticos están dirigidos a determinar la adecuación y utilidad de un resumen con relación a su fuente. Según Hahn y Mani (2000) hay dos tipos de métodos de evaluación:

- **Métodos intrínsecos** (o normativos): En este método los usuarios juzgan la calidad del resumen a través de su análisis directo. Los usuarios califican la naturalidad: la capacidad del resumen para cubrir las ideas claves, o bien la similitud con los resúmenes escritos por humanos (expertos en el área o no).

  Ninguna de estas medidas son enteramente satisfactorias. El resumen ideal, en particular, es difícil de construir y raramente único. De la misma forma como hay muchas maneras para

C.C.Reconocimiento

describir un evento o una escena, los usuarios pueden producir muchos resúmenes (extractos o abstractos) considerados aceptables.
En efecto, la evidencia empírica de Salton et al. (1997) muestra que las personas raramente llegan a un acuerdo con relación a cuales oraciones o párrafos deben incluirse en el sumario. Esto quiere decir que el nivel de coincidencia entre diferentes resúmenes realizados por personas y los generados automáticamente no es muy alto, por lo que muchos investigadores consideran que éste no parece ser un sistema de evaluación adecuado.

- **Métodos extrínsecos**: Los usuarios juzgan la calidad del resumen de acuerdo a la manera como afecta la culminación de alguna otra tarea: ayudar a determinar la relevancia de los tópicos de interés del texto o permitir responder ciertas preguntas relacionadas con el contenido.
  Este enfoque, propuesto por Hand (1997), evalúa la calidad del resumen respecto a una tarea determinada, con el objetivo de medir su utilidad. Esta orientación en la evaluación requiere ciertas condiciones de prueba como corpus de textos y resúmenes, sistemas de recuperación de información, y conjuntos de consultas y resultados relevantes evaluados por expertos. Un ejemplo de este tipo de evaluación indirecta es el propuesto por Maña, Buenaga y Gómez (1998), donde no se empleó ningún tipo de usuario y, en su lugar, se usó un sistema de recuperación de información. Les permitió luego obtener medidas cuantitativas importantes como la capacidad de retención de información relevante de los resúmenes en función de la efectividad de recuperación de los documentos a partir de sus resúmenes correspondientes.

Para efectos de esta investigación, se aplicó un método intrínseco. Se considera que este proyecto no amerita usar métodos complejos en una versión inicial. Por otra parte, no se cuenta con una infraestructura (corpus, sistemas de recuperación de información) para aplicar un método extrínseco. Pensamos que cuando se logre un refinamiento de las reglas aplicadas y se corrijan las fallas actuales, se puede pensar en métodos evaluativos más sistemáticos que permitan mostrar sus avances. Sin embargo, suponemos que el procedimiento de prueba empleado puede mostrar la apreciación del sistema por parte de los usuarios, además de contribuir en el desarrollo y continuación del proyecto. Un buen escenario para esa prueba es la aplicación de clasificación bibliohemerográfica. Este proyecto fue originalmente inspirado para ese propósito (Contreras y Dávila, 2001; Dávila et. al. 2002; Dávila y Contreras, 2002).

www.bdigital.ula.ve

3.1.2 Descripción del experimento y resultados

Para el experimento se seleccionó un dominio de conocimiento particular y se escogió un conjunto de textos especializados de una revista académica local[22]. La revista, denominada "Agroalimentaria", es una publicación científica, arbitrada, de distribución internacional y de frecuencia semestral, especializada en el área de las ciencias sociales relacionadas con los estudios sobre la alimentación. La referencia de dichos artículos se encuentra en el Anexo B.

Para efectos de esta evaluación se realizó la selección de cinco artículos de esta publicación en base a los siguientes criterios:

(1) Textos narrativos y descriptivos: Se seleccionaron aquellos artículos cuyo contenido implicara algún tipo de análisis de hechos en el tiempo (históricos) o también descripción de situaciones y objetos. Se leyó el *abstract* o resumen de cada artículo.

(2) Contenido textual: Se prefieren aquellos artículos cuyo contenido no se fundamente en análisis numéricos, con la finalidad de garantizar la mayor cantidad de información textual posible. Para ello se revisó el texto completo verificando la cantidad de graficos, tablas, etc.

(3) Estilo del autor: A través de la lectura de la introducción del artículo se revisó rápidamente el estilo del autor. Se tomaron en cuenta factores como el tamaño de las oraciones y de los

[22] Revista "Agroalimentaria" del (CIAAL) Centro de Investigaciones Agroalimentarias, Facultad de Ciencias Económicas y Sociales de la Universidad de Los Andes. Mérida – Venezuela.
http://www.saber.ula.ve/ciaal/agroalimentaria/

C.C.Reconocimiento

párrafos, el número de oraciones por párrafo y el uso de marcadores discursivos entre las oraciones.

(4) Descriptores del autor: Se escogieron aquellos artículos que tuviesen palabras asociadas por su autor, con la finalidad de disponer de información descriptiva de un experto. Esta información puede servir para evaluaciones futuras aplicadas al texto completo.

Con estos criterios se intentaba garantizar que los artículos tuviesen información textual básicamente y presentaran, por lo menos parcialmente, los elementos de estilos de Williams. A partir de estos artículos se escogió un grupo de párrafos de diferentes secciones para someterlos al experimento.

Párrafo Original				Resumen Mínimo			Resumen Máximo		
N	**ID-Párrafo**	**#Pal.**	**#Ora.**	**#Pal.**	**%Red.**	**Opinión**	**#Pal.**	**%Red.**	**Opinión**
[1]	[*TESIS-A]	160	8	143	89,38	SI	32	20,00	SI
[2]	[N11-A2]	194	10	123	63,40	SI	15	7,73	SI
[3]	[N9-A4]	232	7	209	90,09	SI	31	13,36	SI
[4]	[N9-A6-A]	183	5	183	100,00	NO	43	23,50	SI
[5]	[N9-A4]	159	5	159	100,00	NO	64	40,25	SI
[6]	[N12-A1]	200	9	174	87,00	SI	14	7,00	NO
[7]	[N12-A1]	208	8	162	77,88	SI	32	15,38	SI
[8]	[N9-A4]	133	5	109	81,95	SI	23	17,29	SI
[9]	[N11-A2-I]	150	5	120	80,00	SI	37	24,67	NO
[10]	[N11-A4]	268	8	95	35,45	NO	19	7,09	SI
[11]	[N9_A4]	232	7	189	81,47	SI	27	11,64	SI
[12]	[N11-A2]	168	8	137	81,55	NO	16	9,52	SI
[13]	[N11-A4]	94	4	55	58,51	SI	17	18,09	SI
[14]	[N11-A2-I]	200	6	163	81,50	SI	33	16,50	NO
[15]	[*JACINTO]	144	13	66	45,83	SI	8	5,56	NO
[16]	[N9-A4]	197	5	118	59,90	SI	68	34,52	SI
[17]	[N11-A2-C]	147	6	120	81,63	SI	30	20,41	SI
[18]	[*N14-A6]	237	6	162	68,35	SI	21	8,86	SI
[19]	[N11-A4-I]	139	4	73	52,52	NO	44	31,65	NO
[20]	[N9-A4-I]	137	4	94	68,61	NO	13	9,49	SI
[21]	[N12-A1]	65	3	65	100,00	NO	26	40,00	NO
[22]	[N12-A1]	221	8	213	96,38	SI	47	21,27	NO
[23]	[N12-A1]	159	4	115	72,33	SI	17	10,69	SI
[24]	[N9-A6]	143	5	143	100,00	SI	30	20,98	SI
[25]	[N9-A6]	355	10	220	61,97	SI	29	8,17	NO
[26]	[N11-A2]	161	5	109	67,70	SI	25	15,53	SI

Tabla 3.1. Resultados de la evaluación del resumidor.

www.bdigital.ula.ve

C.C.Reconocimiento

Luego de haber cumplido con los criterios de selección, se entregó la propuesta a 26 estudiantes que cursaban estudios de postgrado en el área de computación, modelado y simulación de sistemas de la Facultad de Ingeniería de la Universidad de Los Andes. La evaluación fue realizada como una actividad académica del curso de Lógica y Matemáticas. A estos estudiantes les fue asignado un párrafo, con la instrucción de someterlo al resumidor para evaluar la salida. La evaluación exigida debía estar formada por un argumento apoyando su opinión[23] sobre el resumidor y por ideas claras sobre sus limitaciones y alcances. Además, se les notificó el interés por saber si la técnica propuesta efectivamente resume los textos seleccionados.

Los estudiantes entregaron la evaluación individual y las evaluaciones aparecen en el Anexo B. Las opiniones de cada estudiante sobre el resumidor aplicado al párrafo asignado fueron catalogadas en dos conjuntos diferentes "Si Resume" y "No Resume". Además, se muestran los datos de cantidad de palabras y oraciones del párrafo original y del resumen con el fin de mostrar el nivel de reducción del contenido del texto. A partir de estos datos se muestra como resultado la Tabla 3.1.

En esta tabla se muestra por cada párrafo original un identificador asociado al artículo de la revista. Allí se pueden observar tres párrafos que poseen un identificador con un signo asterisco "*", indicando que son párrafos obtenidos de otra fuente y no pertenecen a la revista del dominio. Estos párrafos fueron agregados al experimento con la finalidad de observar su comportamiento y compararlos con el resto. Además, se muestran los datos sobre el número de palabras y oraciones del párrafo, con lo cual podemos tener una idea de la estructura del párrafo y la longitud de sus oraciones.

Por cada nivel de resumen, mínimo y máximo, se muestra el número de palabras del resumen, el porcentaje de palabras en función del texto original y su evaluación por parte del estudiante. Esta evaluación consistió en decir, por cada nivel, si la herramienta resume el párrafo asignado. En base a estos datos se calcularon datos claves como los mostrados en la Tabla 3.2.

Tipo de Resumen	Opinión		Porcentaje de reducción del texto	
	NO Resume	SI Resume	Promedio	Desviación estándar
Mínimo	27 %	73 %	76%	17
Máximo	31 %	69 %	18%	10

www.bdigital.ula.ve

Tabla 3.2. Datos relevantes de la evaluación y uso del resumidor.

Con estos resultados podemos concluir que, para un 70% de los 26 estudiantes la herramienta resume el texto que les fue asignado. Este resultado verifica positivamente nuestra hipótesis de investigación planteada en el capítulo I.

Por otra parte, en relación con la segunda asignación pedida a los estudiantes, algunos mostraron criterios sobre lo que puede considerarse un buen resumidor; además, ofrecieron información sobre las limitaciones. Otros dieron sugerencias en base a su texto particular. Los comentarios que consideramos importantes para trabajos futuros también se encuentran en el Anexo B.

3.1.3 Comentarios sobre los resultados

Los datos mostrados anteriormente nos indican que los usuarios perciben la herramienta como satisfactoria. Sin embargo, como se mencionó anteriormente, debemos considerar que la apreciación de un resumen varía significativamente según las expectativas personales. Esta situación se pudo constatar en las respuestas de este grupo de estudiantes, donde hubo diferencias en cuanto a la importancia de factores como:

- El porcentaje de reducción del texto.

[23] Considerando que un argumento correcto no depende de su conclusión.

C.C.Reconocimiento

- La selección de ideas principales del resumen.
- La presencia de frases de transición entre oraciones.
- El resaltado de los tópicos de las oraciones.
- La coherencia del resumen.

Lo importante es que independientemente de las razones y criterios de los estudiantes, el 70% informó sobre la pertinencia de los resúmenes para tener una idea preliminar del contenido completo del párrafo. Algunos evaluadores argumentaron que el tipo de párrafo que estaban probando era propicio para obtener esos resultados, pues un párrafo con la idea principal seguido de las ideas secundarias y cerrado con la conclusión era el tipo de párrafo más adecuado para este procesamiento. Como vemos estos resultados no se alejan mucho de lo que Williams recomienda para garantizar la coherencia. Con esta información verificamos que el "estilo del autor" es un factor determinante de los resultados.

Sin embargo, el resumidor actual presenta fallas. Algunos de los estudiantes destacaron la existencia de oraciones aisladas que no parecían coherentes. En otros casos, existe información temporal que no es tomada en cuenta dentro de la secuencia de las oraciones. Además, ciertos resúmenes presentaron referencias anafóricas y elipsis que no permiten la total comprensión del contenido textual. Otros detectaron fallas en la selección de los tópicos en las oraciones.

Por otra parte, los niveles de reducción moderado y mínimo arrojaron los mismos resultados. Estos datos no se muestran en este documento pues se trató de una falla en el programa que se usó en el experimento. Con este error los resultados obtenidos para el párrafo asignado con el nivel mínimo y el nivel moderado son exactamente iguales; por lo tanto consideramos la revisión y análisis de la definición de estos niveles.

La mayoría de los evaluadores destacó como positivo la manera en que se destacaron los tópicos dentro del texto resumen. Inclusive, unos consideraron que el resumen ideal podría formarse a través de la disposición adecuada de dichos tópicos en una o varias oraciones. Precisamente, ésta es una de las ideas planteadas como resúmenes constructivos en los siguientes apartados.

www.bdigital.ula.ve

3.2 Como extender el Resumidor

El principal objetivo de la siguiente sección es marcar la vía de expansión del resumidor en futuras investigaciones.

3.2.1 Limitaciones y escalabilidad

Nuestro programa resumidor está dirigido a textos especializados. Estos textos tienen generalmente rasgos morfosintácticos que involucran las referencias verbales, nominales, y adjetivales descritas en Arntz y Picht (1995). En general, los textos con estas características también se ajustan a las sugerencias de estilo de Williams. Sin embargo, los textos no especializados pueden incluirse si consideran estas reglas de estilo. Más aún, el resumidor puede procesar un texto que no se ajuste al estilo Williams. Nuestro objetivo de robustez es que, en estos casos, el programa no colapse y genere una salida, aún si no es muy útil como resumen.

Otra limitación está relacionada con los diccionarios empleados en el resumidor. Se usa un diccionario de verbos en español para el componente gramatical y un diccionario de marcadores discursivos para las reglas de claridad. Ambos diccionarios fueron generados a partir del corpus de prueba. Por tanto, el sistema descarta aquellas oraciones cuyos verbos no están en el diccionario verbal. Además, las oraciones que contengan marcadores discursivos que no estén en el diccionario serán incluidas como parte del tópico.

Sería muy importante considerar las reglas de estilo a partir de textos completos o secciones de texto, pues todas las reglas de estilo aplicadas parten del párrafo. Tanto la claridad como la cohesión se ubican en el ámbito de oración y párrafo. La coherencia va mas allá y considera las relaciones entre los párrafos, pero estas relaciones son generalmente conceptuales. Nuestro sistema debe representar y manipular conocimiento de cada dominio si queremos que se aplique para estas tareas.

C.C.Reconocimiento

En particular, el resumidor propuesto tiene poco alcance en el tratamiento del discurso, específicamente para resolver anáforas y elipsis del texto. Su naturaleza compleja no aconseja una presentación simplificadora, como la presentada en este documento. La ampliación de la cobertura en este sentido, requiere considerar algunas teorías semánticas formales que abordan estos aspectos. En concreto, proponemos revisar la teoría de Representaciones Discursivas (Kamp y Reyle, 1993) y la Semántica de Situaciones (Barwise y Perry, 1983). Ambas intentan ampliar las posibilidades de la lógica de primer orden y de superar la semántica oracional.

Se considera que los retos consecuentes deben enfocarse en afinar el resumen con criterios de relevancia más precisos, sin desmejorar los logros que ya se obtienen. Creemos que esto es posible, pues la representación se presta a la elaboración y a la integración con otros mecanismos.

3.2.2 Hacia resúmenes constructivos

Uno de los objetivos pendientes es un resumidor constructivo más general. Es decir, un resumidor capaz de obtener como resultado oraciones o frases nuevas que no estén literalmente en el texto, pero que representen un resumen del texto con frases u oraciones correctas en español. Desde el punto de vista de las técnicas de resumen automático (sección 1.4.4), se trata de aplicar la abstracción para parafrasear el contenido del texto en términos más generales.

Según Maña, Buenaga y Gómez (1998), las técnicas de resumen constructivas suelen estar circunscritas a dominios muy concretos. Una idea que proponemos es explotar la semántica de los verbos, que presenta mayor independencia del dominio de conocimiento, con el fin de definir el significado de una oración. Esto quiere decir que además de determinar el tópico de las oraciones según el tipo de verbo (componente de claridad), se puede usar el significado del verbo. Una implementación de esta idea consistiría en usar un tesauro de verbos con sus representaciones semánticas; según tales significados escoger cierta parte de la oración como tópico. De esta manera, podemos relacionar los tópicos de las oraciones con ciertos conectores contando con la semántica verbal y sus tópicos.

www.bdigital.ula.ve

Para ilustrar esta idea, considere el siguiente texto de una noticia internacional, tomada de "*The Wall Street Journal Americas*" (15-02-2001).

El texto del discurso se presenta en la Tabla 2.7, en el módulo de claridad: "*Un informe de un comité científico de la unión Europea reveló que las ovejas y las cabras pueden contraer, teóricamente, el mal de las vacas locas. Pero que hasta ahora esto solo ha ocurrido en experimentos de laboratorio*".

Las reglas de extracción son las siguientes:

R1- `T` es un tópico del discurso `D` si en el discurso `D`, un `Agente` *revela* `T`.

R2- `T` es un tópico del discurso `D` si en el discurso `D`, un `Agente` *revela* `T'` **y** `T'` contiene la información que `Agente2` *puede* hacer `T`.

R3- `T` es un tópico del discurso `D` si en el discurso `D`, un `Agente` *revela* `T'` **y** `T'` contiene la información que `Agente2` *puede contraer* `Algo` y `T` = `Algo` **en** `Agente2`.

R4- Un tópico `T` es el común más específico en `D` si es un tópico del discurso `D` **y no** existe otro tópico `T'` de `D` tal que `T'` más general que `T`.

R5- `T'` es más general que `T` si `T'` es más breve que `T`.

Con estas reglas el tópico resultante es: "*mal de las vacas locas **en** ovejas y cabras*". La frase anterior puede considerarse como la *síntesis* del discurso `D`.

Un conjunto de reglas como las anteriores pueden ser incorporadas al resumidor para producir la frase resumen. Observen que, si bien estas reglas fueron inspiradas por ese texto en particular (y permiten resolverlo), las reglas son generales. Pueden utilizarse en otras oraciones donde se cumpla la peculiar relación de un tópico "en" otro que modelan esas reglas.

C.C.Reconocimiento

3.2.3 Paradigmas en las relaciones entre tópicos

Esta última observación nos ha llevado a considerar una estrategia para ampliar la cobertura del resumidor y permitir la generación "constructiva" de resúmenes. La estrategia consiste en definir "paradigmas" en las posibles relaciones entre tópicos en una misma oración, como el caso del último ejemplo, y relaciones entre tópicos en oraciones distintas (en un mismo párrafo, para comenzar).

Las relaciones **inter-tópico** estarían definidas entre las oraciones de un párrafo. Aquí, los marcadores discursivos pueden ser una guía que establece relaciones sugeridas por el escritor. Las relaciones conceptuales de los términos que contienen los tópicos pueden identificarse a través de relaciones conceptuales del área de conocimiento.

En cuanto a los paradigmas interoracionales existe el paradigma "Tópico1 igual a Tópico2", incorporado al resumidor actual en una primera forma. Este paradigma nos permite decidir si un tópico dado se repite a lo largo de un párrafo. Como se puede apreciar, tal igualdad no es sintáctica, salvo en casos triviales. La comparación debe apelar a la semántica puesto que quizás por razones de elegancia, los escritores rara vez repiten exactamente la misma frase tópico.

Por otra parte, las relaciones **intra-tópicos** consideran los tópicos dentro de una misma oración, es decir, entre el tópico principal y los tópicos de las oraciones subordinadas. En este caso, pueden considerarse las preposiciones gramaticales (si existen) y las características conceptuales que el verbo aporta al tópico. Esta idea se apoya en el formalismo semántico de Schank (1975), denominado **dependencia conceptual**, el cual intenta representar el evento expresado en la oración, a través de un grupo de palabras que lo describen. Para eso, se trasladan todos los verbos a un pequeño conjunto de acciones primitivas, a través de una clasificación verbal, donde se identifican actores, objetos, origen y objetivo según el verbo. Con este formalismo, Schank propone una representación del lenguaje dependiente del conocimiento.

Por su parte, la lingüística textual cree probable que cualquier lector supone, en virtud de la configuración verbal, que las acciones descritas intentan ser una pista de la caracterización de los agentes. Esta operación de enriquecimiento del mundo textual mediante la aportación del propio conocimiento que el lector tiene del mundo se denomina **hacer referencias** (Beaugrande y Dressler, 1997). Esta operación exige complementar los conceptos y las relaciones que se manifiestan en el texto con el fin de rellenar sus discontinuidades.

Considerar un resumidor capaz de hacer estas referencias requiere un procesamiento semántico con mayor cobertura. Esto significa tener en cuenta que un texto no tiene sentido en sí mismo, sino gracias a la interacción establecida entre el conocimiento presentado en el texto y el conocimiento almacenado en la memoria de los lectores.

www.bdigital.ula.ve

C.C.Reconocimiento

4 Capítulo IV: Conclusiones

La presente investigación ha utilizado como base los lineamientos propuestos por Williams (1990) para escribir textos claros en el idioma inglés. En primer lugar, nos planteamos usar estas reglas diseñadas para escribir textos, como una estrategia invertida para leer y procesar dichos textos. En segundo lugar, tuvimos la inquietud de verificar en que medida se acercan estas reglas, propuestas originalmente para el idioma inglés, a la escritura en el idioma español. Nos propusimos realizar un procesamiento automático del contenido de textos en español con las reglas de estilo de Williams.

Las reglas de Williams comienzan por caracterizar el principio de "claridad" localizado en una oración aislada. Se trata de la capacidad para identificar los sujetos y las acciones de la oración. Luego, las recomendaciones permiten la "cohesión" del texto haciendo énfasis en el tópico, idea generalmente enunciada en el sujeto de una oración. Por último, consideran el principio de "coherencia" que permite estructurar conceptualmente el contenido del texto, en *arranque* y *discusión*.

Nuestra principal hipótesis de trabajo establece que, si se aplica un procesamiento adecuado a ciertos estilos, entonces se pueden extraer sistemáticamente los tópicos adecuados y relevantes de un documento escrito sobre la base de esos estilos. A partir de estos tópicos se aplica una técnica para resumir textos, generando otro texto con información más comprimida sobre el contenido textual. Para verificar la hipótesis, en una primera evaluación esperamos que los tópicos y resúmenes derivados fueran aceptados por humanos como válidos o útiles en alguna medida

En esta investigación se ha instrumentado nuestra hipótesis sobre la posibilidad de explotar criterios basados en lógica para asociar tópicos. Además, hemos ilustrado el argumento con las salidas de un programa codificado en Prolog. Este programa está enmarcado dentro de los modelos simbólicos de procesamiento de lenguaje natural, instrumentando el análisis del lenguaje español. Si bien este programa se concentra en la exploración de oraciones y párrafos con los criterios de claridad, cohesión y coherencia de Williams, se le puede considerar una versión preliminar de un resumidor automático.

www.bdigital.ula.ve

La evaluación de esta herramienta nos permitió verificar nuestra hipótesis de trabajo. Aproximadamente el 70% de los evaluadores (26) de resúmenes generados por el sistema reportó un resultado aceptable y útil. Admitimos, no obstante, que resta todavía mucho trabajo para aproximar esta descripción compacta (resumen) del contenido del documento a una descripción realizada por humanos.

Sin embargo, es motivador informar que el formalismo para la representación del conocimiento lingüístico empleado (lógica), ha sido tolerante a la elaboración de reglas de análisis basadas en criterios poco ortodoxos en el procesamiento lingüístico (reglas de estilo). Cabe recalcar que dichas reglas procesan solamente la superficie textual del contenido (a través de claridad, cohesión y coherencia). Teniendo en cuenta que la estructura oracional afecta el significado y efecto de las palabras, los aspectos de estilo del escritor aportan un mensaje para el lector que puede complementar el significado literal del texto. Esta consideración nos ha permitido obtener resultados aceptables sin la presencia y manejo de conocimiento del mundo del hablante, que influye significativamente en la semántica del lenguaje natural.

La herramienta de resumen automático presentada en este documento es la aplicación de esta investigación. Esta aplicación permite al usuario revisar el contenido de un texto rápidamente permitiendo enjuiciar la relevancia del contenido para un requerimiento particular. El aporte teórico es la validación de la teoría lingüística de estilo de Williams para extraer tópicos de textos.

La producción de contenidos en nuestro idioma motiva la prolongación de este proyecto. Recomendamos que tales extensiones se concentren en el refinamiento de las reglas de estilo ya existentes, en la experimentación con otras técnicas de resumen y en la incorporación de manejo del conocimiento. Creemos que estos factores permitirán producir resúmenes constructivos de mayor utilidad y calidad.

C.C.Reconocimiento

Anexo A: Código Prolog del resumidor simbólico

```
%% Resumidor simbólico %%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% resumir(+Tipo_Salida,+Factor_Resumen)
%% Resumidor con entrada y salida web. Captura los datos del formulario y los coloca
%% en archivos (factor y test).

:- use_module(library(cgi)).

    main :-
        format('<HTML>~n', []),
        format('<HEAD>~n', []),
        format('<TITLE>Summarization SWI-Prolog CGI</TITLE>~n', []),
        format('</HEAD>~n~n', []),
        format('<BODY>~n', []), format('<P>', []),
            see('c:/https/tesis/cgi-bin/datos/factor.txt'),
            get0(PrimerC),
            name(C,[PrimerC]),
            format('Texto Resumen - factor ',[]),
            write(C), format('<BR><HR>',[]),
            see('c:/https/tesis/cgi-bin/datos/test.txt'),
            resumir(1,C),
        format('</BODY>~n</HTML>~n', []),halt.

resumir :- leer_texto.

resumir(TipoSalida,Factor) :- leer_texto(TipoSalida,Factor).

leer_texto :-
        tokenizador(Parrafo,ParrafoU,ProximoC),
        gramatica(Parrafo,ParrafoW), !,
        claridad(ParrafoW,Topicos),
        topico_comun(Topicos,Topico),
        salida_html(Parrafo,ParrafoU,Topico),
        leer_resto_texto(ProximoC).

leer_texto(1,Factor) :-
        tokenizador(Parrafo,ParrafoU,ProximoC),
        format('PÁRRAFO = '), write(Parrafo), nl, nl,
        format('<BR><BR>'), format('PÁRRAFO U= '),
        write(ParrafoU), nl, nl,
        gramatica(Parrafo,ParrafoW),
        format('<BR><BR>'), format('ParrafoW = '),
        write(ParrafoW), nl, nl,
        claridad(ParrafoW,Topicos),
        format('<BR><BR>'), write('TÓPICOS = '),
        write(Topicos), nl, nl,
        topico_comun_con_salida(Topicos,Topico,Factor),
        format('<BR><BR>'), write('TÓPICO COMÜN = '),
        write(Topico), nl, nl,
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
        salida_html(Parrafo,ParrafoU,Topico),
        leer_resto_texto(ProximoC,1).

leer_resto_texto(ProximoC) :-
        tipo_caracter(ProximoC,fin,-1),
        !.

leer_resto_texto(ProximoC) :-
        tipo_caracter(ProximoC,_,_),
        leer_texto.

leer_resto_texto(ProximoC,_) :-
        tipo_caracter(ProximoC,fin,-1),
        !.

leer_resto_texto(ProximoC,TipoSalida) :-
        tipo_caracter(ProximoC,_,_),
        leer_texto(TipoSalida).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% salida_html(+Texto,+TextoUpper,+Topico).
% Dado el texto de un párrafo, resalta los tópicos del texto original
% en negrita (TextoUpper) en una salida HTML (salida estándar).

salida_html(Texto,TextoUpper,Topico):-
        etiquetar_topicos(Topico,Texto,TextoUpper,TextoEtiquetado),
        imprimir_etiquetas_encabezado,
        imprimir_texto(TextoEtiquetado),
        imprimir_etiquetas_cierre.

etiquetar_topicos([],_,TextoUpper,TextoUpper).

etiquetar_topicos([Topico|Resto],Texto,TextoUpper,TextoEtiquetado):-
        topico(Topico,TopicoLista),
        etiquetar_topico_en_texto(TopicoLista,Texto,TextoUpper,TextoTopico),
        etiquetar_topicos(Resto,Texto,TextoTopico,TextoEtiquetado).

etiquetar_topico_en_texto(_,[],[],[]):- fail,!.

etiquetar_topico_en_texto(TopicoLista,[Oracion|_],[OracionUpper|TextoUpper],[OracionEtiqueta|Texto
Upper]):-
        subset(TopicoLista,Oracion),
        insertar_etiqueta(TopicoLista,Oracion,OracionUpper,OracionEtiqueta).

etiquetar_topico_en_texto(TopicoLista,[_|Texto],[OracionUpper|TextoUpper],[OracionUpper|TextoTopi
co]):-
        etiquetar_topico_en_texto(TopicoLista,Texto,TextoUpper,TextoTopico).

insertar_etiqueta(TopicoLista,Oracion,OracionUpper,OracionEtiqueta):-
        topico_en_oracion(OracionDividida,TopicoLista,Oracion,[]),
        insertar_bold(OracionDividida,OracionUpper,OracionEtiqueta).
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
insertar_bold(OracionDivida,OracionUpper,OracionEtiqueta):-
        dividir_oracion_upper(OracionDivida,OracionUpper,[Antes,Topico,Despues]),
        append(Antes,[[60,66,62]],AntesEtiqueta),
        append(AntesEtiqueta,Topico,AntesEtiquetaTopico),
        append(AntesEtiquetaTopico,[[60,47,66,62]],AntesTopicoEtiqueta),
        append(AntesTopicoEtiqueta,Despues,OracionEtiqueta).

dividir_oracion_upper([Antes,Topico,Despues],OracionUpper,[AntesUpper,TopicoUpper,DespuesUppe
r]) :-
        lista_upper(Antes,OracionUpper,RestoAntes,AntesUpper),
        lista_upper(Topico,RestoAntes,RestoTopico,TopicoUpper),
        lista_upper(Despues,RestoTopico,[],DespuesUpper).

lista_upper([],L,L,[]).

lista_upper(SubListaLower,[P1|ListaUpper],RestoUpper,[P1|SubListaUpper]) :-
        es_especial(P1),
        lista_upper(SubListaLower,ListaUpper,RestoUpper,SubListaUpper).

lista_upper([_|SubListaLower],[P1|ListaUpper],RestoUpper,[P1|SubListaUpper]) :-
        lista_upper(SubListaLower,ListaUpper,RestoUpper,SubListaUpper).

topico_en_oracion([T1,TopicoLista,T2],TopicoLista) -->
        antes(T1), igual(TopicoLista), despues(T2).

www.bdigital.ula.ve

antes([X|Resto]) --> [X|Resto].
antes([]) --> [].
despues([]) --> [].
despues([X|Resto]) --> [X|Resto].
igual(Topico,TopicoResto,Resto):- append(Topico,Resto,TopicoResto).

imprimir_etiquetas_encabezado :-
        write('<P>'),nl.

imprimir_texto([]).

imprimir_texto([Oracion|Resto]):-
        member([60,66,62],Oracion),
        imprimir_oracion(Oracion),
        write('.'),
        imprimir_texto(Resto).

imprimir_texto([_|Resto]):-
        imprimir_texto(Resto).

imprimir_oracion([]).

imprimir_oracion([Palabra|Resto]):-
        imprimir_especial(Palabra),
        imprimir_oracion(Resto).
```


C.C.Reconocimiento

```
imprimir_oracion([Palabra|Resto]):-
        write(' '),
        write(Palabra),
        imprimir_oracion(Resto).

imprimir_especial(Palabra):-
        es_coma(Palabra).

imprimir_especial(Palabra):-
        es_especial(Palabra).

es_coma([44]) :- write(',').
es_especial([60,66,62]) :- write('<B>').
es_especial([60,47,66,62]) :- write('</B>').

imprimir_etiquetas_cierre :-
        nl,write('</P>').

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Paso 4 : Tópico Común %%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Cohesión y Coherencia %%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Tópico Común segun factor de resumen %%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% topico_comun(+ListaTopicosParrafo, -TopicoComun).
%% Dado una lista de tópicos de un párrafo retorna su tópico común
%% El tópico común identifica el arranque y el discurso del párrafo
%% resuelve las referencias anafóricas simples y la repetición de tópicos
%% del texto, además fusiona la ponderación de relevancia de cada tópico.
%% Al final se selecciona según la ponderación de la lista los tópicos más
%% frecuentes y más específicos.

topico_comun_con_salida(ListaTopicosParrafo, TopicoComun,Factor) :-
        identificar_topico_arranque(ListaTopicosParrafo, ListaPonderada),
        write('<BR><BR>TOPICOS PONDERADOS = '),nl,write(ListaPonderada),nl,
        resolver_anafora(ListaPonderada,ListaSinAnaforas),
        nl,write('<BR>TOPICOS SIN ANAFORAS = '),nl,write(ListaSinAnaforas),nl,
        simplificar_topicos_identicos(ListaSinAnaforas, ListaSinDuplicados),
        nl,write('<BR>TOPICOS SIN DUPLICADOS = '),nl,write(ListaSinDuplicados),nl,nl,
        seleccionar_topico_comun(ListaSinDuplicados, TopicoComun,Factor).

topico_comun(ListaTopicosParrafo, TopicoComun,Factor) :-
        identificar_topico_arranque(ListaTopicosParrafo, ListaPonderada),
        resolver_anafora(ListaPonderada,ListaSinAnaforas),
        simplificar_topicos_identicos(ListaSinAnaforas, ListaSinDuplicados),nl,nl,
        seleccionar_topico_comun(ListaSinDuplicados, TopicoComun,Factor).

%% identificar_topico_arranque(+ListaTopicosParrafo, -ListaPonderada).
%% Identifica el arranque y el discurso del párrafo, asocia un 0 al
%% arranque (primer tópico) y un 1 al discurso (resto de los tópicos).
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
%% Cada elemento de la lista ListaPonderada tiene dos elementos el número
%% ponderado y la lista de palabras de la oracion.

identificar_topico_arranque([], []).
identificar_topico_arranque([Topico|RestoTopicos], [[0,Topico]|RestoLista]):-
        identificar_discurso(RestoTopicos, RestoLista).

identificar_discurso([], []).
identificar_discurso([[]|RestoTopicos], RestoLista):-
        identificar_discurso(RestoTopicos, RestoLista).
identificar_discurso([Topico|RestoTopicos], [[1,Topico]|RestoLista]):-
        identificar_discurso(RestoTopicos, RestoLista).

%% resolver_anafora(+ListaPonderada,-ListaSinAnafora).
%% Realiza las asociaciones anafóricas del párrafo, simplifica los tópicos
%% y reajusta la frecuencia de los tópicos. Busca la ocurrencia de la(s)
%% anáfora(s) del Párrafo, previamente identificadas como [esta_antes]
%% con los datos de género y número del pronombre correspondiente.
%% La anáfora divide la lista de tópicos en dos listas: Antes y Después.
%% Para resolver la anáfora debe buscarse en la lista Antes un determinante
%% del tópico semejante al del [esta_antes], se recorre la lista en sentido
%% contrario. Se continua en la lista Despues hasta el final del párrafo.

resolver_anafora([Topico|[]],Topico).

resolver_anafora(ListaPonderada,ListaSinAnafora):-
        buscar_anafora(ListaPonderada,Antes,Despues,DatosLinguisticosAnafora),
        invertir_lista(Antes,AntesInvertida),
        relacionar_topico_anterior(AntesInvertida,DatosLinguisticosAnafora,TopicosRelacionado),
        invertir_lista(TopicosRelacionado,TopicosRelacionadoInvertido),
        concatenar_listas([TopicosRelacionadoInvertido,Despues],ListaNueva,[]),
        resolver_anafora(ListaNueva,ListaSinAnafora).

resolver_anafora(ListaPonderada,ListaPonderada).

%% buscar_anafora(ListaPonderada,Antes,Despues,DatosLinguisticosAnafora).
%%

buscar_anafora([],[],[],[]):- fail,!.

buscar_anafora([Topico|RestoTopico],_,RestoTopico,DatosLinguisticosAnafora):-
        topico_es_anafora(Topico),
        obtener_datos_linguisticos(Topico, DatosLinguisticosAnafora).

buscar_anafora([Topico|RestoTopico],[Topico|Antes],Despues,DatosLinguisticosAnafora):-
        buscar_anafora(RestoTopico,Antes,Despues,DatosLinguisticosAnafora).

topico_es_anafora([_,[[esta_antes],_]]).

obtener_datos_linguisticos([_,[[esta_antes],[Numero,Genero]]],[Numero,Genero]).
```


www.bdigital.ula.ve

C.C.Reconocimiento

```
%% invertir_lista(Lista,ListaInvertida)
%%

invertir_lista(Lista,ListaInvertida):-
        reverse(Lista,ListaInvertida).

%%% frecuencia(Topico,Frecuencia).
%% dado un topico retorna su frecuencia o numero de ponderacion

frecuencia([Frecuencia,_],Frecuencia).

%% topico(TopicoPonderado,Topico).
%% dado un tópico retorna la lista que contiene la oración del tópico

topico([_,Topico],Topico).

topico_ponderado([Frecuencia,Topico],Frecuencia,Topico).

%% relacionar_topico_anterior(Antes,DatoLinguistico,TopicoRelacionado)
%% Busca en la lista Antes invertida un determinante con los mismos
%% datos lingüísticos de la anáfora, cuando lo encuentre aumenta la
%% ponderación del tópico.

relacionar_topico_anterior([],_,[]).

relacionar_topico_anterior([Topico|Resto],DatosLinguisticosAnafora,[TopicoRelacionado|Resto]):-
        topico(Topico,TopicoLista),
        contiene_articulo_relacionado(TopicoLista,DatosLinguisticosAnafora),
        relacionar_topico(Topico,TopicoRelacionado).

relacionar_topico_anterior([Topico|Antes],DatosLinguisticosAnafora,[Topico|TopicosRelacionado]):-
        relacionar_topico_anterior(Antes,DatosLinguisticosAnafora,TopicosRelacionado).

relacionar_topico([Frecuencia,Topico],[FrecuenciaNueva,Topico]):-
        FrecuenciaNueva is Frecuencia + 1.

%% contiene_articulo_relacionado(Topico,DatosLinguisticosAnafora).
%%

contiene_articulo_relacionado([],_):- fail,!.

contiene_articulo_relacionado([Palabra|_],[NumeroAnafora,GeneroAnafora]):-
        es_articulo(Palabra,NumeroAnafora,GeneroAnafora).

contiene_articulo_relacionado([Palabra|_],_):-
        es_articulo(Palabra,_,_),
        fail,!.

contiene_articulo_relacionado([_|Resto],DatosLinguisticosAnafora):-
        contiene_articulo_relacionado(Resto,DatosLinguisticosAnafora).
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
%% es_articulo(Palabra,Numero,Genero).

es_articulo(el,singular,masculino).
es_articulo(los,plural,masculino).
es_articulo(la,singular,femenino).
es_articulo(las,plural,femenino).
es_articulo(un,singular,masculino).
es_articulo(unos,plural,masculino).
es_articulo(una,singular,femenino).
es_articulo(unas,plural,femenino).
es_articulo(alguna,singular,femenino).
es_articulo(algunas,plural,femenino).
es_articulo(algún,singular,masculino).
es_articulo(algunos,plural,masculino).

%% concatenar_listas([TopicosRelacionado,Despues],ListaNueva,[]).
%% ListaNueva = TopicosRelacionado + Despues

concatenar_listas([TopicosRelacionado,Despues]) -->
        lista(TopicosRelacionado),
        lista(Despues).

lista([X|Resto]) --> [X|Resto].
lista([])--> [].

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% simplificar_topicos_identicos(ListaSinAnaforas, ListaSinDuplicados).
%% Busca los tópicos que contengan palabras idénticas en orden secuencial
%% en la lista anterior.

simplificar_topicos_identicos([], []).

simplificar_topicos_identicos([Topico|[]], [Topico|[]]).

simplificar_topicos_identicos([Topico|RestoListaSinAnaforas],
[TopicoFusionado|ListaSinDuplicados]):-
        fusionar_topicos(Topico,RestoListaSinAnaforas,[TopicoFusionado|ListaFusionada]),
        simplificar_topicos_identicos(ListaFusionada,ListaSinDuplicados).

%% fusionar_topicos(Topico,RestoListaSinAnaforas,ListaFusionada),
%% Aplica al primer tópico de la lista un filtro para obtener solo los
%% sustantivos y adjetivos de la frase nominal que representa al tópico.
%% Se obtiene una lista de palabras (en donde no hay: determinantes,
%% pronombres, ni conectores) y aplica fusionar palabra a cada uno de los
%% elementos.

fusionar_topicos(TopicoPonderado,RestoListaSinAnaforas,ListaFusionada):-
        topico(TopicoPonderado,Topico),
        filtrar_sustantivo_adjetivo(Topico,ListaSustantivoAdjetivo),
        fusionar_palabra(ListaSustantivoAdjetivo,TopicoPonderado,RestoListaSinAnaforas,ListaFusion
ada).
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
filtrar_sustantivo_adjetivo([], []).

filtrar_sustantivo_adjetivo([Topico|Resto], ListaSustantivoAdjetivo):-
        es_articulo(Topico,_,_),!,
        filtrar_sustantivo_adjetivo(Resto,ListaSustantivoAdjetivo).

filtrar_sustantivo_adjetivo([Topico|Resto], ListaSustantivoAdjetivo):-
        es_pronombre(Topico),!,
        filtrar_sustantivo_adjetivo(Resto,ListaSustantivoAdjetivo).

filtrar_sustantivo_adjetivo([Topico|Resto], ListaSustantivoAdjetivo):-
        es_conjuncion(Topico),!,
        filtrar_sustantivo_adjetivo(Resto,ListaSustantivoAdjetivo).

filtrar_sustantivo_adjetivo([Topico|Resto], [Topico|ListaSustantivoAdjetivo]):-
        filtrar_sustantivo_adjetivo(Resto,ListaSustantivoAdjetivo).

es_pronombre(lo).
es_pronombre(le).
es_pronombre(les).
es_pronombre(esa).
es_pronombre(ese).
es_pronombre(se).
es_pronombre(X) :- pronombre(_,_,[X],[]).

es_conjuncion(a).
es_conjuncion(e).
es_conjuncion(y).
es_conjuncion(de).
es_conjuncion(del).
es_conjuncion(en).
es_conjuncion(con).
es_conjuncion(por).
es_conjuncion(como).
es_conjuncion(para).
es_conjuncion(que).
es_conjuncion(qué).

%% fusionar_palabra(Palabra,Topico, Lista,ListaFusionada).
%% Recorre la lista de palabras del tópico, previamente filtradas hasta el final

fusionar_palabra([],Topico,Lista,[Topico|Lista]).

fusionar_palabra([Palabra|Resto],Topico,ListaTopicos,ListaFusionada):-
        buscar_palabra_en_lista(Palabra,Topico,TopicoResultante,ListaTopicos,ListaResultante),
        fusionar_palabra(Resto,TopicoResultante,ListaResultante,ListaFusionada).

fusionar_palabra([_|RestoListaPalabras],Topico,ListaTopicos,ListaFusionada):-
        fusionar_palabra(RestoListaPalabras,Topico,ListaTopicos,ListaFusionada).
```

www.bdigital.ula.ve

C.C.Reconocimiento

```
%% buscar_palabra_en_lista(Palabra,Topico,ListaTopicos,ListaResultante).
%% Recorre la lista del resto de los tópicos en donde hay que buscar una
%% palabra que pertenece a un tópico anterior. Si la búsqueda tiene éxito
%% entonces se construye una lista con los tópicos en donde aparezca dicha
%% palabra. Se escoge de dicha lista mas el tópico original el más
%% representativo.

buscar_palabra_en_lista(_,Topico,Topico,[],[]).

buscar_palabra_en_lista(Palabra,Topico,TopicoResultante,[TopicoPonderado|Resto],Resultado):-
        topico(TopicoPonderado,TopicoResto),
        member(Palabra,TopicoResto),
        escoger_topico_representativo(Topico,TopicoPonderado,TopicoRepresentativo),
        buscar_palabra_en_lista(Palabra,TopicoRepresentativo,TopicoResultante,Resto,Resultado).

buscar_palabra_en_lista(Palabra,Topico,TopicoResultante,[TopicoResto|Resto],[TopicoResto|ListaResu
ltante]):-
        buscar_palabra_en_lista(Palabra,Topico,TopicoResultante,Resto,ListaResultante).

%% El mejor de los tópicos que se estan fusionando es el primer tópico
%% que aparece en el texto. Se asume que el escritor del texto sabe ordenar
%% sus ideas y colocó el mejor tópico antes. Se escoge el segundo tópico en
%% el orden del texto si este tiene mejor ponderacion.

escoger_topico_representativo(Topico1,Topico2,TopicoResultado):-
        frecuencia(Topico1,Frecuencia1),
        frecuencia(Topico2,Frecuencia2),
        topico(Topico2,TopicoR),
        Frecuencia2 > Frecuencia1,
        FrecuenciaR is Frecuencia1 + Frecuencia2,
        topico_ponderado(TopicoResultado,FrecuenciaR,TopicoR).

escoger_topico_representativo(Topico1,Topico2,TopicoResultado):-
        frecuencia(Topico1,Frecuencia1),
        frecuencia(Topico2,Frecuencia2),
        topico(Topico1,TopicoR),
        FrecuenciaR is Frecuencia1 + Frecuencia2,
        topico_ponderado(TopicoResultado,FrecuenciaR,TopicoR).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% seleccionar_topico_comun(+ListaTopicosSinDuplicados, -TopicoComun).
% Selecciona de la lista resultante de tópicos aquellos mejor poderados.
% Toma en cuenta la cardinalidad de la lista de tópicos y un factor de
% resumen, el cual esta entre [0,1,2]. Donde 0 son todos los Tópicos menos
% los que tiene 0 de ponderación. El 1 indica que descarta los 0 y los 1.
% El 2 escoje solo el mejor tópico por cada parrafo.

seleccionar_topico_comun(ListaSinDuplicados, TopicoComun,Factor):-
        ordenar_descendentemente(ListaSinDuplicados,ListaOrdenada),
        obtener_relevantes(ListaOrdenada,TopicoComun,Factor).
```


www.bdigital.ula.ve

C.C.Reconocimiento

```
ordenar_descendentemente(ListaSinDuplicados,ListaOrdenada):-
        sort(ListaSinDuplicados,ListaOrdenada).

%% factor 1 por defecto

obtener_relevantes(ListaOrdenada,TopicoComun,Factor):-
        factor(0,Factor),
        filtrar_ceros(ListaOrdenada,TopicoComun).

obtener_relevantes(ListaOrdenada,TopicoComun,Factor):-
        factor(2,Factor),
        mejor_topico(ListaOrdenada,TopicoComun).

obtener_relevantes(ListaOrdenada,TopicoComun,Factor):-
        filtrar_ceros_unos(ListaOrdenada,TopicoComun).

factor(Factor,Factor).

filtrar_ceros([],[]).

filtrar_ceros([Topico|Resto],TopicoComun):-
        frecuencia(Topico,0),
        filtrar_ceros(Resto,TopicoComun).

filtrar_ceros(ListaOrdenada,ListaOrdenada).

filtrar_ceros_unos([],[]).

filtrar_ceros_unos([Topico|Resto],TopicoComun):-
        frecuencia(Topico,0),
        filtrar_ceros_unos(Resto,TopicoComun).

filtrar_ceros_unos([Topico|Resto],TopicoComun):-
        frecuencia(Topico,1),
        filtrar_ceros_unos(Resto,TopicoComun).

filtrar_ceros_unos(ListaOrdenada,ListaOrdenada).

mejor_topico([],[]).

mejor_topico(ListaOrdenada,[TopicoComun]):-
        last(TopicoComun,ListaOrdenada).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Paso 3 : Claridad %%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% obtener tópicos de cada oración %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% Primitivas para obtener de la Oración Williams el sujeto, verbo y
%% complemento.
```


www.bdigital.ula.ve

C.C.Reconocimiento

```
%% sujeto(+OracionW, -Sujeto)
%% obtener el sujeto de la oracion Williams

sujeto([sujeto(S)|_],S).

%% verbo(+OracionW, -Verbo)
%% obtener el verbo de la oracion Williams

% cuando el verbo es simple
verbo([sujeto(_),verbo(V)|_],V).
% cuando el verbo es una lista de verbos, devuelve el primer verbo
% del verbo compuesto
verbo_compuesto([sujeto(_),[verbo(V)|_]|_],V).
% cuando el verbo es binario, se devuelve los dos verbos
verbo_binario([sujeto(_),[verbo(V1),verbo(V2)]|_],V1,V2).
% cuando la oración es impersonal, pronombre "se"
verbo_impersonal([sujeto(_),[sujeto_verbo(V)|_]|_],V).

%% complemento(+OracionW, -Complemento)
%% obtener el complemento de la oracion Williams

complemento([sujeto(_),_,complemento(C)],C).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% claridad(+ParrafoWilliams,-TopicosParrafo)
%% Dado un párrafo Williams (con el criterio de claridad local), retorna
%% sus tópicos (de que se habla en cada oración del párrafo).

claridad([],[]).

%% Caso Tópico 1
%% Se obtiene el verbo de la oración, si es auxiliar se obtiene el sujeto
%% de la oración, al sujeto se le aplica el filtro de expresiones y el
%% resultado es el tópico de la oración.
%% Los verbos auxiliares expresan la idea de esencia o sustancia.
%% El tópico de una oración con verbo copulativo es el sujeto de la oración.
%% Para saber si un verbo es auxiliar se revisa la lista de verbos
%% verb_aux. No interesa cuando el verbo es compuesto.

claridad([OracionWilliams|ParrafoWilliams],[Topico|TopicosParrafo]) :-
        verbo(OracionWilliams,Verbo),
        verb_aux(_,Verbo,[]),
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(Topico,Sujeto,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

%% Caso particular de ciertos verbos ("revelar" y "puede contraer")
%% para el ejemplo de las vacas locas.
%% Tópico compuesto en el complemento.

claridad([OracionWilliams|ParrafoWilliams],[Topico|TopicosParrafo]) :-
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
        %% si el verbo de la oración es uno de este grupo
        verbo(OracionWilliams,[reveló]),
        complemento(OracionWilliams,Complemento),
        %% Obtener Tópico a partir del Complemento
        filtrar_expresion(TopicoComplemento,Complemento,[]),
        gramatica([TopicoComplemento],[ComplementoWilliams]),
        sujeto(ComplementoWilliams,Sujeto),
        filtrar_expresion(Topico1,Sujeto,[]),
        verbo_binario(ComplementoWilliams,[pueden],[contraer]),
        complemento(ComplementoWilliams,Complemento2),
        filtrar_expresion(Topico2,Complemento2,[]),
        construir_topico([Topico2, [en], Topico1],Topico,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

%% Caso Tópico con anáforas.
%% el sujeto se refiere a un tópico mencionado anteriormente. Se agrega el
%% tópico del complemento de la oración que contiene esta referencia
%% anafórica.

claridad([OracionWilliams|ParrafoWilliams],[[[esta_antes],[Numero,Genero]],TopicoComplemento|Top
icosParrafo]) :-
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(Topico_posible,Sujeto,[]),
        contiene_anafora(Numero,Genero,Topico_posible,[]),
        complemento(OracionWilliams,Complemento),
        gramatica([Complemento],[ComplementoWilliams]),
        extraer_topico_complemento(ComplementoWilliams,TopicoComplemento),
        claridad(ParrafoWilliams,TopicosParrafo).

claridad([OracionWilliams|ParrafoWilliams],[[[esta_antes],[Numero,Genero]],Topico|TopicosParrafo])
:-
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(Topico_posible,Sujeto,[]),
        contiene_anafora(Numero,Genero,Topico_posible,[]),
        complemento(OracionWilliams,Complemento),
        filtrar_expresion(Topico,Complemento,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

%% Caso Tópico Impersonal
%% verbo con pronombre impersonal "se", la oración es impersonal.
%% Generalmente estas oraciones tienen su tópico en el complemento.

claridad([OracionWilliams|ParrafoWilliams],[Topico|TopicosParrafo]) :-
        %% si el verbo contiene el pronombre impersonal "se"
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion([],Sujeto,[]),
        verbo_impersonal(OracionWilliams,[se]),
        complemento(OracionWilliams,Complemento),
        %% Obtener Tópicos a partir del Complemento de la oración
        gramatica([Complemento],[ComplementoWilliams]),
        extraer_topico_complemento(ComplementoWilliams,Topico),
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
        claridad(ParrafoWilliams,TopicosParrafo).

claridad([OracionWilliams|ParrafoWilliams],[Topico|TopicosParrafo]) :-
        %% si el verbo contiene el pronombre impersonal "se"
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion([],Sujeto,[]),
        verbo_impersonal(OracionWilliams,[se]),
        complemento(OracionWilliams,Complemento),
        %% Obtener Tópicos a partir del Complemento de la oración
        filtrar_expresion(Topico,Complemento,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

claridad([OracionWilliams|ParrafoWilliams],[TopicoSujeto,Topico|TopicosParrafo]) :-
        %% si el verbo contiene el pronombre impersonal "se"
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(TopicoSujeto,Sujeto,[]),
        verbo_impersonal(OracionWilliams,[se]),
        complemento(OracionWilliams,Complemento),
        %% Obtener Tópicos a partir del Complemento de la oracion
        gramatica([Complemento],[ComplementoWilliams]),
        extraer_topico_complemento(ComplementoWilliams,Topico),
        claridad(ParrafoWilliams,TopicosParrafo).

claridad([OracionWilliams|ParrafoWilliams],[TopicoSujeto,Topico|TopicosParrafo]) :-
        %% si el verbo contiene el pronombre impersonal "se"
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(TopicoSujeto,Sujeto,[]),
        verbo_impersonal(OracionWilliams,[se]),
        complemento(OracionWilliams,Complemento),
        %% Obtener Tópicos a partir del Complemento de la oración
        filtrar_expresion(Topico,Complemento,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

%% Caso Tópico por defecto
%% los demás verbos (predicativos) expresan estado o acción

claridad([OracionWilliams|ParrafoWilliams],[Topico|TopicosParrafo]) :-
        complemento(OracionWilliams,[]),
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(Topico,Sujeto,[]),
        claridad(ParrafoWilliams,TopicosParrafo).

claridad([OracionWilliams|ParrafoWilliams],[TopicoSujeto,TopicoComplementoFiltrado|TopicosParraf
o]) :-
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(TopicoSujeto,Sujeto,[]),
        %% Obtener Tópicos a partir del Complemento de la oracion
        complemento(OracionWilliams,Complemento),
        gramatica([Complemento],[]),
        filtrar_expresion(TopicoComplementoFiltrado,Complemento,[]),
        claridad(ParrafoWilliams,TopicosParrafo).
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
claridad([OracionWilliams|ParrafoWilliams],[TopicoSujeto,TopicoComplemento|TopicosParrafo]) :-
        sujeto(OracionWilliams,Sujeto),
        filtrar_expresion(TopicoSujeto,Sujeto,[]),
        %% Obtener Tópicos a partir del Complemento, cuando éste es una oración con verbo
        complemento(OracionWilliams,Complemento),
        gramatica([Complemento],[ComplementoWilliams]),
        extraer_topico_complemento(ComplementoWilliams,TopicoComplemento),
        claridad(ParrafoWilliams,TopicosParrafo).

extraer_topico_complemento([],[]).

extraer_topico_complemento(ComplementoWilliams,Topico):-
        verbo(ComplementoWilliams,Verbo),
        verb_aux(_,Verbo,[]),
        sujeto(ComplementoWilliams,Sujeto),
        filtrar_expresion(Topico,Sujeto,[]).

extraer_topico_complemento(ComplementoWilliams,Topico):-
        verbo_impersonal(ComplementoWilliams,[se]),
        complemento(ComplementoWilliams,Complemento),
        filtrar_expresion(Topico,Complemento,[]).

extraer_topico_complemento(ComplementoWilliams,Topico):-
        sujeto(ComplementoWilliams,Sujeto),
        filtrar_expresion([],Sujeto,[]),
        complemento(ComplementoWilliams,Complemento),
        filtrar_expresion(Topico,Complemento,[]).

extraer_topico_complemento(ComplementoWilliams,Topico):-
        sujeto(ComplementoWilliams,Sujeto),
        filtrar_expresion(TopicoPosible,Sujeto,[]),
        es_conjuncion(TopicoPosible),
        complemento(ComplementoWilliams,Complemento),
        filtrar_expresion(Topico,Complemento,[]).

extraer_topico_complemento(ComplementoWilliams,Topico):-
        sujeto(ComplementoWilliams,Sujeto),
        filtrar_expresion(Topico,Sujeto,[]).

igual(X,X).

%% construir tópicos compuestos (con DCG)

construir_topico([T1,C,T2]) --> topico(T1), conector(C), topico(T2).

topico([X|Resto]) --> [X|Resto].
conector([X|Resto]) --> [X|Resto].

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% filtrar_expresion(-TextoSinExpresion) *DCG
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
%% filtrar_expresion(-TextoSinExpresion, +Texto, +Listavacia) *Prolog
%% Filtra una frase y elimina las expresiones de evaluación, conectores
%% lógicos y de tiempo y espacio de la oración. Estas expresiones no son
%% parte del tópico. Williams define el uso de estas expresiones como
%% cohesion entre las oraciones de un párrafo. pág 49 del libro de Style

filtrar_expresion([]) --> [].

filtrar_expresion(Topico) --> expresion(_), filtrar_expresion(Topico).

filtrar_expresion(Topico) --> filtrar_expresion_resto(Topico), expresion(_).

filtrar_expresion(Resto) --> filtrar_expresion_resto(Resto).

filtrar_expresion_resto([]) --> [].

filtrar_expresion_resto([X|Resto]) --> [X|Resto].

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% chequea si la frase contiene anáforas
% por ahora sólo identifica pronombres.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

contiene_anafora(Numero,Genero) --> pronombre(Numero,Genero).
contiene_anafora(Numero,Genero) --> palabra(_), pronombre(Numero,Genero), palabra(_).
palabra([X|Resto]) --> [X|Resto].
palabra([]) --> [].

pronombre(singular,_) --> [eso].
pronombre(plural,masculino) --> [esos].
pronombre(singular,femenino) --> [esa].
pronombre(plural,femenino) --> [esas].
pronombre(singular,_) --> [esto].
pronombre(plural,masculino) --> [estos].
pronombre(singular,femenino) --> [esta].
pronombre(plural,femenino) --> [estas].
pronombre(singular,masculino) --> [este].
pronombre(singular,masculino) --> [éste].
pronombre(singular,_) --> [ello].
pronombre(plural,femenino) --> [ellas].
pronombre(singular,masculino) --> [aquel].
pronombre(singular,femenino) --> [aquella].
pronombre(plural,femenino) --> [aquellas].
pronombre(plural,masculino) --> [aquellos].
pronombre(singular,femenino) --> [dicha].
pronombre(plural,femenino) --> [dichas].
pronombre(singular,masculino) --> [dicho].
pronombre(plural,masculino) --> [dichos].

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Paso 2 : Gramática %%%%%%%%
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
%%%%%%%%%%%%%% Construir Tabla sujeto, verbo, complemento %%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% gramatica(+Parrafo,-ParrafoWilliams)
% Implementa el concepto de Claridad de Williams. La claridad local de
% oración es la estructura fija del discurso escrito: sujeto, verbo y
% complemento. Procesa el párrafo de entrada (dividido en tokens) y
% retorna un párrafo donde cada oración es una lista de la siguiente forma:
% [sujeto(Sujeto),verbo(Verbo),complemento(Complemento)]

gramatica([],[]).

% Se utiliza un setof para la gramática s con el fin de usar el primer
% término de la lista L, pues es el término donde el sujeto es de menor
% longitud, es decir se tiene el primer verbo de la oración.

gramatica([Oracion|RestoParrafo],[OracionWilliams|RestoParrafoWilliams]) :-
        setof(OracionW,s(OracionW,Oracion,[]),Lista),
        oracion(Lista,OracionWilliams),
        gramatica(RestoParrafo,RestoParrafoWilliams).

gramatica([_|RestoParrafo],RestoParrafoWilliams) :-
        gramatica(RestoParrafo,RestoParrafoWilliams).

% oracion, retorne la oracion con menor sujeto que tenga el verbo mas
% complejo.

www.bdigital.ula.ve

oracion([Oracion|[]],Oracion).

oracion([Oracion1,Oracion2|_],Oracion2) :-
        sujeto(Oracion1,Sujeto1),
        sujeto(Oracion2,Sujeto1).

oracion([Oracion1,_|_],Oracion1).

% s(?Estructura,+Oracion,+ListaVacia)
% Implementa una gramática basada en verbos, sólo reconoce el verbo
% principal de la oración (a través de un diccionario de verbos). Lo que
% esta antes del verbo es el sujeto y lo que esta después es el complemento.
% Usa DCG. Contienen la definición básica de las oraciones: frase nominal,
% frase verbal (verbo y complemento). Es decir s --> np,vp
% Retorna la estructura (Estructura) de la oración (Oracion), en términos
% de sujeto, verbo y complemento (Retorna una lista con estos elementos).

s([NP,V,C]) --> np(NP), vp(V,C).

% frase nominal (np)
% Frase que contiene algún nombre (sustantivo), correspondiente al sujeto
% sintáctico de la oración. Considera algunas frases o marcadores
% discursivos  que tienen conflictos con los verbos.

np(sujeto([C1,C2|Resto])) --> marcador_dis(C1,C2), !, np_resto(Resto).
```


C.C.Reconocimiento

```
np(sujeto([C1,C2,C3|Resto])) --> marcador_dis(C1,C2,C3), !, np_resto(Resto).
np(sujeto([])) --> [].
np(sujeto([X|Resto])) --> [X|Resto].

np_resto([X|Resto]) --> [X|Resto].

marcador_dis(a,partir) --> [a,partir].
marcador_dis(es,decir) --> [es,decir].
marcador_dis(es,importante) --> [es,importante].
marcador_dis(es,importante,señalar) --> [es,importante,señalar].

% frase verbal (vp)
% constituida por el verbo (verb) y el complemento de la oración (comp).

vp(V,C) --> verb_compuesto(V), comp(C).
vp(V,C) --> verb_impersonal(V), comp(C).
vp(V,C) --> verb_aux(V), comp(C).
vp(V,C) --> verb(V), comp(C).

%% oración impersonal: usa el pronombre "se", el sujeto permisible es la
%% generalidad de la gente.

verb_impersonal([V1,V2,V3,V4]) -->  impersonal(V1,V2), verb_compuesto([V3,V4]).
verb_impersonal([V1,V2,V3]) -->  impersonal(V1), verb_compuesto([V2,V3]).
verb_impersonal([V1,V2,V3]) -->  impersonal(V1,V2), resto_impersonal(V3).
verb_impersonal([V1,V2]) -->  impersonal(V1), resto_impersonal(V2).
impersonal(sujeto_verbo([se]), sujeto_verbo([le])) --> [se, le].
impersonal(sujeto_verbo([se]), sujeto_verbo([les])) --> [se, les].
impersonal(sujeto_verbo([se])) --> [se].
%% si el verbo no esta en el diccionario de verbos
resto_impersonal(verbo([X])) --> [X].

%% casos de verbos compuestos

% Los verbos en participio solo pueden estar despues de los verbos
% auxiliares. Si no es asi son sustantivos o adjetivos.

verb_compuesto([V1,V2]) --> verb_aux(V1), verb_part(V2).
verb_compuesto([V1,V2]) --> verb_aux(V1), verb(V2).
verb_compuesto([V1,V2]) --> verb_aux(V1), verb_aux(V2).
verb_compuesto([V1,V2]) --> verb(V1), verb(V2).

% complemento de la frase verbal

comp(complemento([Y|Resto])) --> [Y|Resto].

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Paso 1 : tokenizador %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%% lee archivo de entrada, separa en tokens %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```


www.bdigital.ula.ve

C.C.Reconocimiento

```
%% Esto corresponde al primer paso para escribir resumenes.
%% Consiste en dividir el texto en fases del pensamiento: párrafos,
%% oraciones y palabras.
%% Este tokenizador general lee la entrada estandar y por cada
%% párrafo genera como salida dos listas de token, la primera
%% con los tokens en lower-case y la segunda con los tokens
%% originales del texto (incluye upper-case).

% tokenizador(-Atomos,-AtomosUpper,-ProximoC)
% Lee una línea del texto, separándola en una lista de átomos.
% Atomos = párrafo lower-case, delimitado por el caracter especial
% de fin de línea [10]. AtomosUpper = párrafo original del texto
% (incluye upper-case)

tokenizador(Atomos,AtomosUpper,ProximoC) :-
        leer_caracter(PrimerC, PrimerCUpper, PrimerT),
        leer_resto_p(PrimerC, PrimerCUpper, PrimerT, Atomos, AtomosUpper, ProximoC).

leer_resto_p(46,46,especial,Parrafo, ParrafoUpper,ProximoC) :-
        !,
        leer_caracter(Caracter, CaracterUpper, TipoC),
        leer_resto_p(Caracter,CaracterUpper,TipoC, Parrafo, ParrafoUpper, ProximoC).

leer_resto_p(32,32,blanco,Parrafo,ParrafoUpper,ProximoC) :-
        !,
        leer_caracter(Caracter,CaracterUpper,TipoC),
        leer_resto_p(Caracter,CaracterUpper,TipoC,Parrafo,ParrafoUpper,ProximoC).

leer_resto_p(Caracter,Caracter,fin,[],[],Caracter) :- !.

%% tipo alfanumérico

leer_resto_p(PrimerC,PrimerCUpper,PrimerT,[Oracion|Atomos],[OracionUpper|AtomosUpper],Proxim
oCaracter) :-
        leer_oracion(PrimerC,PrimerCUpper,PrimerT,Oracion,OracionUpper,ProximoC),
        tipo_caracter(ProximoC,ProximoT,PC),
        leer_resto_p(ProximoC,PC,ProximoT,Atomos,AtomosUpper,ProximoCaracter).

% leer_atomos(-Atomos,-AtomosUpper,-ProximoC)
% Lee una línea del texto, separándola en una lista de átomos lower-case
% y upper-case respectivamente.

leer_atomos(Atomos, AtomosUpper, ProximoC) :-
        leer_caracter(PrimerC, PrimerCUpper, PrimerT),
        leer_oracion(PrimerC, PrimerCUpper, PrimerT, Atomos, AtomosUpper, ProximoC).

% leer_oracion(+PrimerC,+PrimerCUpper,+PrimerT,-Lista,-ListaUpper,-ProximoC)
% Dado el primer caracter lower y upper case, respectivamente, además
% del tipo de caracter correspondiente retorna la lista de palabras de
% la oración. La oración esta delimitada por cualquier caracter de fin,
% en especial el punto [46].
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
leer_oracion(Caracter,Caracter,fin,[],[],Caracter) :- !.

leer_oracion(46,46,especial,[],[],46) :- !.

leer_oracion(_,_,blanco,Atomos,AtomosUpper,ProximoC) :-
        !,
        leer_atomos(Atomos,AtomosUpper,ProximoC).

leer_oracion(PrimerC,PrimerCUpper,especial,[A|Atomos],[AUpper|AtomosUpper],ProximoC) :-
        !,
        name(A,[PrimerC]),
        name(AUpper,[PrimerCUpper]),
        leer_atomos(Atomos,AtomosUpper,ProximoC).

%% tipo alfanumérico

leer_oracion(PrimerC,PrimerCUpper,PrimerT,[A|Atomos],[AUpper|AtomosUpper],ProximoCaracter) :-
        palabra_completa(PrimerC,PrimerCUpper,PrimerT,ProximoC,ProximoT,A,AUpper),
        leer_oracion(ProximoC,ProximoC,ProximoT,Atomos,AtomosUpper,ProximoCaracter).

% leer_caracter(-Caracter,-Tipo)
% Lee un caracter de la entrada estándar y obtiene el tipo de caracter
% de la función tipo_caracter

leer_caracter(Caracter,Tipo) :-
        get0(C),
        %% lee un caracter de la entrada estándar
        tipo_caracter(C,Tipo,Caracter).

% leer_caracter(-Caracter,-CaracterUpper,-Tipo)
% Lee un caracter de la entrada estándar, devuelve el caracter en lower-case
% y el caracter original del texto (contiene upper-case), retorna también el
% tipo de caracter de la función tipo_caracter

leer_caracter(CaracterLower,C,Tipo) :-
        get0(C),
        %% lee un caracter de la entrada estándar
        tipo_caracter(C,Tipo,CaracterLower).

% palabra_completa(+PrimerC,+PrimerCUpper,+PrimerT,-Lista,-ListaUpper)
% Dado el primer caracter y el primer tipo de caracter lee el resto de
% la palabra, colocándola en la lista lower-case y obtiene también la
% lista en upper-case.

%% para token alfabéticos (primer caracter alfabético)

palabra_completa(PrimerC,PrimerCUpper,alfa,ProximoC,ProximoT,Palabra,StringUpper) :-
        !,
        leer_caracter(Caracter,CaracterUpper,TipoC),
        palabra_completa_alfa(Caracter,CaracterUpper,TipoC,Lista,ListaUpper,ProximoC,ProximoT),
```


www.bdigital.ula.ve

C.C.Reconocimiento

```
        name(Palabra,[PrimerC|Lista]),
        name(PalabraUpper,[PrimerCUpper|ListaUpper]),
        string_to_atom(StringUpper,PalabraUpper).

%% para tokens numéricos (primer caracter numérico)

palabra_completa(PrimerC,_,num,ProximoC,ProximoT,Palabra,StringUpper) :-
        !,
        leer_caracter(Caracter,_,TipoC),
        palabra_numerica_completa(Caracter,TipoC,Lista,ProximoC,ProximoT),
        append([PrimerC|Lista],[44],ListaP),
        name(A,ListaP),
        atom_chars(A,L),
        append(L2,[(',')],L),
        concat_atom(L2,Palabra),
        string_to_atom(StringUpper,Palabra).

%% NOTA: el "name" de los número redondos por ejemplos 3.000 genera
%% como resultado un átomo con el valor de 3. Debe hacerse un
%% procesamiento adicional para tratar estos casos (concatenar una
%% coma al final, realizar el name, dividir en tokens, extraer la
%% última coma y concatenar para obtener la Palabra).

% palabra_completa_alfa(+PrimerC,+PrimerCUpper,+alfa,-Palabra,
% -PalabraUpper,-ProximoC,-ProximoT)
% Obtiene un token completo cuando el primer caracter del token es
% alfabético. Genera dos tokens uno en lower-case y el segundo según el
% texto original (incluye upper-case).

palabra_completa_alfa(PrimerC,PrimerCUpper,alfa,[PrimerC|Lista],[PrimerCUpper|ListaUpper],Proxim
oC,ProximoT) :-
        !,
        leer_caracter(Caracter,CaracterUpper,TipoC),
        palabra_completa_alfa(Caracter,CaracterUpper,TipoC,Lista,ListaUpper,ProximoC,ProximoT).

palabra_completa_alfa(PrimerC,PrimerCUpper,num,[PrimerC|Lista],[PrimerCUpper|ListaUpper],Proxi
moC,ProximoT) :-
        !,
        leer_caracter(Caracter,CaracterUpper,TipoC),
        palabra_completa_alfa(Caracter,CaracterUpper,TipoC,Lista,ListaUpper,ProximoC,ProximoT).

palabra_completa_alfa(PrimerC,_,PrimerT,[],[],PrimerC,PrimerT).

% palabra_numerica_completa(+PrimerC,+PrimerT,-AtomoNumerico,-ProximoC,
% -ProximoT)
% Obtiene un token completo (palabra) con caracteres numéricos, cuando el
% primer caracter no es alfabético. Comtempla los casos de números
% decimales, porcentajes, años y último tokens de una oración.

palabra_numerica_completa(PrimerC,PrimerT,[PrimerC|Lista],ProximoC,ProximoT) :-
        member(PrimerT,[num,alfa]),
```


www.bdigital.ula.ve
C.C.Reconocimiento

```
        !,
        leer_caracter(Caracter,TipoC),
        palabra_numerica_completa(Caracter,TipoC,Lista,ProximoC,ProximoT).

palabra_numerica_completa(PrimerC,_,[PrimerC|Lista],ProximoC,ProximoT) :-
        member(PrimerC,[46,44]),
        leer_caracter(Caracter,TipoC),
        member(TipoC,[num]),
        palabra_numerica_completa(Caracter,TipoC,Lista,ProximoC,ProximoT).

palabra_numerica_completa(PrimerC,PrimerT,[],PrimerC,PrimerT).

% tipo_caracter(+Codigo,?Type,-NuevoCodigo)
% Dado un código ASCII, clasifica el caracter en "fin" (de linea/archivo/palabra),
% "alfa" (alfabético y numéricos), "especiales" al resto de los caracteres y "blanco"

%tipo_caracter(10,fin,10) :- !. % fin de línea en DOS
%tipo_caracter(13,fin,13) :- !. % fin de línea en UNIX
tipo_caracter(-1,fin,-1) :- !. % fin de archivo

%% blanco y otros caracteres de control

tipo_caracter(Codigo,blanco,Codigo) :-
        Codigo =< 32,
        !.

%% dígitos numéricos

tipo_caracter(Codigo,num,Codigo) :-
        48 =< Codigo, Codigo =< 57,
        !.

%% letras lower-case, alfabéticos

tipo_caracter(Codigo,alfa,Codigo) :-
        97 =< Codigo, Codigo =< 122,
        !.

%% letras upper-case, alfabéticos

tipo_caracter(Codigo,alfa,NuevoCodigo) :-
        65 =< Codigo, Codigo =< 90,
        !,
        %%NuevoCodigo is Codigo. % NO trasladar a lower-case
        NuevoCodigo is Codigo + 32. % trasladar a lower-case

%% vocales acentuadas y tilde en minúsculas
%% la lista representa respectivamente L = [á,é,í,ó,ú,ñ]

tipo_caracter(Codigo,alfa,Codigo) :-
        member(Codigo,[225,233,237,243,250,241]),
```

www.bdigital.ula.ve

C.C.Reconocimiento

```
        !.

%% vocales acentuadas y tilde en mayúsculas
%% la lista representa respectivamente L = [Á,É,Í,Ó,Ú,Ñ]
%% L = [[193,225],[201,233],[205,237],[211,243],[218,250],[209,241]]
%% L = [[Á,á],[É,é],[Í,í],[Ó,ó],[Ú,ú],[Ñ,ñ]]

tipo_caracter(Codigo,alfa,NuevoCodigo) :-
        member(Codigo,[193,201,205,211,218,209]),
        !,
        %%NuevoCodigo is Codigo. % NO trasladar a lower-case
        NuevoCodigo is Codigo + 32. % trasladar a lower-case

%% caracteres especiales tratados como alfabéticos
%% la lista representa respectivamente L = [%,$,/,°]

tipo_caracter(Codigo,alfa,Codigo) :-
        member(Codigo,[37,36,47,176]),
        !.

%% todos los especiales

tipo_caracter(Codigo,especial,Codigo).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% diccionario de expresiones %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    expresion([como, resultado]) --> [como, resultado].
    expresion([pero]) --> [pero].

    .....


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% diccionario de verbos %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    verb(verbo([abastecer])) --> [abastecer].
    verb(verbo([acarreando])) --> [acarreando].

    .....
```


C.C.Reconocimiento

Anexo B: Datos de la evaluación del resumidor simbólico

B.1. Artículos evaluados de la revista Agroalimentaria

- **[N9-A4]** Pulido, Nelson. "La organización: Base del éxito de los productores de papa en los Andes Venezolanos". Agroalimentaria No. 9 Diciembre 1999.
- **[N9-A6]** Cartay, Rafael. "Estrategias de sobrevivencia de los pequeños caficultores en tiempos de crisis". Agroalimentaria No. 9 Diciembre 1999.
- **[N11-A2]** Díaz, Katty. "La comercialización del cacao en Venezuela: Un análisis antes y después de la apertura comercial. 1975-1998". Agroalimentaria No. 11 Diciembre 2000.
- **[N11-A4]** Quintero Rizzuto, Liliana; Cartay, Rafael. "El circuito del cacao en Venezuela, 1990-1999: Caracterización y estrategias para mejorar la competitividad". Agroalimentaria No. 11 Diciembre 2000.
- **[N12-A1]** Cartay, Rafael. "Consideraciones sobre los comportamientos del consumidor en situaciones de riesgo e incertidumbre alimentaria". Agroalimentaria No. 12 Junio 2001.

B.2. Tabla de Evaluaciones

La siguiente tabla muestra las evaluaciones del resumidor simbólico realizadas por estudiantes de postgrado. Los textos que se publican a continuación son fragmentos originales escritos por los evaluadores del sistema.

[#] Estudiante / ID Párrafo / Palabras / Oraciones	Factor de resumen mínimo Palabras / Oraciones / Comentario	Factor de resumen máximo Palabras / Oraciones / Comentario
[1] Manuel Vielma / [*TESIS-A] 160 p. / 8 o.	143 p. / 6 o.	32 p. / 1 o.
	Define cuales son las oraciones principales del párrafo a resumir, determinando cuales son las palabras clave de cada oración, eliminando características secundarias o referencias a esas palabras claves. En conclusión tomo como aceptable este tipo de resumen ya que deja las características principales del párrafo.	A mi parecer el resumidor en este nivel toma la idea principal con mayor descripción y toma la parte del párrafo que hace referencia a esa idea principal obviando puntos clave a los cuales también hace referencia el párrafo, por lo tanto este nivel no me parece que sea de buena precisión.
[2] José Torres / [N11_A2] 194 p. / 10 o. <u>Opinión General:</u> El resumir introduce espacios entre palabras y comas. El programa resume los textos de una manera apropiada, pero dependiendo del texto. Cumple con su cometido resumiendo los textos que se le introducen y es muy fácil de utilizar.	123 p. / 6 o.	15 p. / 1 o.
	El resumen minino elimino dos partes del texto que no eran de relevancia para la idea general que se pretende transmitir, por lo que a mi parecer funciona de una manera apropiada	Con respecto al resumen Máximo me parece que la reducción es muy sustancial, y que prácticamente tan solo deja lo que se puede considerar como la idea principal lo cual me parece apropiado.
[3] Thomas López	209 p. / 5 o.	31 p. / 1 o.

www.bdigital.ula.ve
C.C.Reconocimiento

/ [N9_A4] 232 p. / 7 o.	Las palabras subrayadas se consideran innecesarias. A este nivel, el resumen contiene toda la información relevante pero no omite toda la información innecesaria, por lo tanto (según las reglas establecidas) no es un buen resumen. Este texto no incluye la información sobre el Acta Constitutiva y Los Estatutos del Comité de Riego, al omitir información que se considera relevante, no satisface la regla por lo que no se puede considerar un buen resumen.	Este texto resumido omite información no relevante, pero también omite información principal, lo cual conduce (según la regla de la sección anterior) a la conclusión de que no es un buen resumen.
[4] Luis Sosa / [N9-A6-A] 183p. / 5 o. <u>Opinión General:</u> (RP) Con el factor de resumen en máximo, mas bien diría que es una aplicación realista y operativa, que nos ayudaría y nos libera de algunas tareas redundantes y fatigosas.	183 p. / 5 o.	43 p. / 1 o.
	Considero que no se comporto como un resumidor, con lo que presento en negrita no me señalaba lo que trataba el documento y mucho menos lo considero un buen resumen	Me indico exactamente lo que trata el documento, fue tan bueno como el que realice; en conclusión, lo tomaría como una herramienta de prueba en este nivel
[5] Leonardo Segovia / [N9-A4] 159 p. / 5 o.	159 p. / 5 o.	64 p. / 1 o.
	El Resumidor no cumple con el propósito para el cual está diseñado debido a en los dos(2) primeros niveles no sufrió ningún cambio, es decir quedo exactamente igual; por tanto no se cumple el propósito de la aplicación en dichos niveles.	El resultado un párrafo breve del texto evaluado, pero no se cumple con el propósito ya que dicho párrafo no constituye un resumen del texto, sino corresponde a la extracción del párrafo que tiene la idea principal del mismo.
[6] José Sánchez / [N12-A1] 200 p. / 9 o. <u>Opinión General:</u> (RP) Para todos los niveles se observa que el resumidor omite o suprime a lo máximo oraciones del parrafo, pero no llega al nivel de supresión de las palabras prescindibles	174 p. / 7 o.	14 p. / 1 o.
	Los resultados de los 2 primeros niveles muestran 174 palabras, un número muy elevado para el nivel moderado, si lo que se busca es reducir a términos breves y precisos lo esencial del texto. Los niveles mínimo y moderado no omiten articulos, conectores, adverbios que son prescindibles sin que el texto pierda lo esencial.	el nivel máximo reduce a un excelente número de palabras 8% pero no conserva lo esencial del parrafo en cuestión
[7] Roberto Olivar / [N12_A1] 208 p. / 8 o.	162 p. / 6 o.	32 p. / 1 o.
	El resumidor es capaz de presentar la información de forma clara.	
[8] Raul Jose	109 p. / 3 o.	23 p. / 1 o.

www.bdigital.ula.ve
C.C.Reconocimiento

/ [N9_A4] 133 p. / 5 o.	En el nivel mínimo y máximo explica el párrafo asignado en pocas palabras y de una manera coherente	El programa RESUMIDOR es bueno.
[9] Marianela Dávila / [N11_A2-I] 150 p. / 5 o. <u>Opinión General:</u> Un buen resumen en mi opinion tomaria la idea principal, su contexto y su conclusion	120 p. / 4 o.	37 p. / 1o.
	En las dos primeras opciones (minimo y moderado) el resumidor elimina la oracion que hace el contexto del parrafo. Desde mi punto de vista, hace falta esa oracion,	La idea obtenida por el resumidor maximo, no resume el contenido del parrafo. No queda claro el contexto anterior ni las consecuencias de la decision tomada con la empresa.
[10] Jesús Ochoa / [N11_A4] 268 p. / 8 o. (RP)	95 p. / 3 o.	19 p. / 1 o.
	Muestra deficiencia en la forma como presenta los tópicos de una oración a otra. A pesar de tener claridad para identificar los agentes y las acciones asociadas a cada uno de ellos, el manejo de los tópicos no es adecuado.	El texto se orienta a Venezuela como productor de cacao. Nótese que el resumen maneja como tópico principal a Venezuela y al cacao, luego, dos posibles resúmenes pueden hacerse tomando en cuenta a Venezuela como tópico o al cacao como tal.
[11] Jesús Márquez / [N9_A4] 232 p. / 7 o.	189 p. / 5 o.	27 p. / 1 o.
	En el texto resumido se omiten dos frases, además de algunas palabras en negritas: los argumentos principales; en la primera frase (omitida) observo que no hay una conclusión en función de la premisa única, en la segunda sucede lo mismo. Por lo anterior concluyo que el RESUMIDOR en este nivel arroja buenos resultados.	El texto original ha sido suprimido en casi totalidad, solo se muestra la primera frase. Considero que en el texto original existe otra frase importante.
[12] Luis Dávila / [N11_A2] 168 p. / 8 o.	137 p. / 7 o.	16 p. / 1 o.
	Lo evaluó como mínimo ya que el resumen consiste en omitir una parte del texto. Esto nos indica de cierta manera un porcentaje de texto resumido del 10% aprox., lo cual para ser un resumidor con 3 niveles puede ser considerado como bajo	Factor Máximo, mi evaluación es bastante positiva. Presentar fundamentalmente la idea principal del texto. Indica un porcentaje de resumen de 90%
[13] Lisdrelys Dugney / [N11_A4] 94 p. / 4 o.	55 p. / 3 o.	17 p. / 1 o.
	El resumidor tomó la información más importante del texto. Por lo tanto, el resumidor tiene alcance.	El resumidor tiene alcance, logra tomar en el resumen máximo la idea principal del texto. Se entiende sobre el tema que tratará el texto posteriormente.
[14] Klaudia Laffaille	163 p. / 5 o.	33 p. / 1 o.

www.bdigital.ula.ve

C.C.Reconocimiento

/ [N11_A2-I] 200 p. / 6 o. <u>Opinión General:</u> Todos los resúmenes se resalta las palabras o frases importantes, facilitando así la comprensión del texto. Todos los resúmenes carecen de la cualidad de expresar las ideas de forma clara, sólo se omite información con cierto criterio que, en este caso no es el más adecuado.	Lo que realiza es omitir la información (en el caso analizado suprime la oración correspondiente a las causas de la disminución de la oferta mundial)	El resumen máximo omite la idea principal, colocando solo el principio del párrafo
[15] Joe Hernández / [*JACINTO] 144 p. / 13 o. <u>Opinión General:</u> Permanece la idea principal y pierde un poco la coherencia entre oraciones. Por el porcentaje de reducción obtenido me parece muy buen resultado.	66 p. / 7 o.	8 p. / 1 o.
	El resumen es bastante bueno, El resumen elaborado no pierde la idea principal, además permanece coherente con respecto al párrafo original. El resumen en factor mínimo transmite en un nivel muy aceptable, la idea que se describe en el párrafo original. Esto gracias a la exclusión de algunas ideas secundarias presentes en el párrafo.	En este factor el resumen no es favorable, no existe ningún indicio del párrafo original. El resumidor en este factor debería transmitir algo mas concluyente de la idea principal.
[16] Glenda González / [N9_A4] 197 p. / 5 o. <u>Opinión General:</u> El resumidor tiene un buen funcionamiento, si consideramos el tamaño del texto	118 p. / 3 o.	68 p. / 1 o.
	El resumidor permite extraer las ideas principales y secundarias al utilizar la opción moderado y mínimo.	En el caso de la opción máximo, el resumidor extrajo la idea principal del texto, el resultado fue muy satisfactorio
[17] Erasmo Gomez / [N11_A2-C] 147 p. / 6 o. <u>Opinión General:</u> El resumidor de texto es una buena herramienta. Subraya algunas oraciones o palabras, que le dan al usuario una mayor comprensión sobre el texto original	120 p. / 5 o.	30 p. / 1 o.
	El resumen posee estructuras gramaticales que pueden ser más simplificadas. Su parte positiva: el subrayado de el argumento principal del texto e identifica todas las ocasiones en el texto donde se hace referencia al tema o argumento principal. Este resumen es bueno, ya que conserva la idea principal del párrafo. Posee oraciones secundarias que forman claramente parte integral del tema principal. Simplifica algunas estructuras gramaticales sin perder la estructura del texto.	Creo que la idea principal del texto original no se ve reflejada de manera clara, lo que permite que el significado del texto pueda tender a perderse.
[18] Diego Mosquera	162 p. / 4 o.	21 p. / 1 o.

www.bdigital.ula.ve

C.C.Reconocimiento

/ [*N14-A6] 237 p. / 6 o. Opinión General: (RP) Es notable que el resumidor simbólico hace un esfuerzo por "capturar" la idea principal de cada proposición (en este caso separadas por el signo punto).	Es notable en los resumenes en factor 0 y 1, que se enfoca en describir sólo una parte de la idea principal y obvia por completo el resto de la idea. Sin embargo, no deja de ser consistente.	
[19] Daniel Signorelli / [N11_A4-I] 139 p. / 4 o. Opinión General: Mi conclusión general es que el resumidor (al menos con el texto utilizado en mi caso) no funciona. En ninguno de los tres niveles de resumen, se mostró cual es el sujeto del tema	73 p. / 2 o.	44 p. / 1 o.
[20] Johanna Chacin / [N9_A4-I] 137 p. / 4 o. (RP)	94 p. / 3 o.	13 p. / 1 o.
	Este resumidor no piensa igual que nosotros en vista que le pude quitar lineas, y agregar palabras nuevas pero dejando el parrafo más claro, es decir de una manera más clara de entender.	En el ultimo nivel trabaja totalmente diferente utiliza como idea princial si se puede decir al sujeto del parrafo.
[21] Carmen Rodríguez / [N12_A1] 65 p. / 3 o. Opinión General: La salida no cumple con los criterios evaluados, en consecuencia el programa No se considera como un buen resumidor.	65 p. / 3 o.	26 p. / 1 o.
	En este nivel de resumen mínimo, el resumidor solo determina algunas palabras claves del texto pero No lo resume, pues no extrae los aspectos más importantes del tema y además, incluye opiniones y comentarios del autor que deberían ser omitidos.	En el último nivel, el resumidor, resume los comentarios y opiniones del autor, extrae algunos aspectos importantes del tema pero, pierde de vista parte de la idea principal del contenido y considera como esencial sólo una parte de esa idea, que no explica brevemente todo el tema.
[22] Jose Brito / [N12_A1] 221 p, / 8 o. Opinión General: (RP) De acuerdo con esto, el resumidor produjo unos resúmenes aceptables para los factores 0 y 1. No así con el factor 2, donde simplemente no captura la idea principal.	213 p. / 7 o.	47 p. / 1 o.
	La salida con el factor 0 resulta bastante satisfactoria. Ciertamente, palabra por palabra, el texto es idèntico al original (hay una variación insignificante: aparecen unos espacios adyacentes a ciertos signos de puntuaciòn), y por tanto, no podemos hablar de un resumen sintàcticamente manifiesto.	El resumen con factor 2 no es aceptable, porque reduce todo el texto a una cita que ni siquiera contiene la idea principal.
[23] Anny Olivar / [N12_A1] 159 p. / 4 o.	115 p. / 3 o.	17 p. / 1 o.
	Por tal razón el resumidor, para este caso hace una abstracción general respecto al texto original dejando vacios en ciertas oraciones.	En este caso el resumidor abarca lo principal del texto, pero debería exponer a que conclusión se llego ó cual fue el logro de la crisis de seguridad alimentaria.

www.bdigital.ula.ve
C.C.Reconocimiento

[24] Alirio Lozada / [N9_A6] 143 p. / 5 o. Opinión General: (RP) Se puede señalar que el resumidor muestra una síntesis coherente del texto que fue ingresado en este caso. Cumple con las reglas de inferencia lógica, considera los conectores del párrafo para determinar las premisas, también infiere sobre cual es la conclusión del mismo, en este caso la idea principal del texto. Básicamente el resumen que he generado en muy similar al que el programa resumidor ha realizado.	143 p. / 5 o.	30 p. / 1 o.
	En los caso 1 y 2, estos elementos son colocados en negritas para mostrar las ideas más relevantes del mensaje, así mismo se puede apreciar que en ambos casos no disminuye mucho la longitud del texto.	Se produce una disminución considerable de la longitud del texto y solo se muestra la idea principal del texto. La salida es óptima, aunque no contextualiza al lector sobre qué y a quién se hace alusión en el texto, si muestra una idea coherente que expresa la principal noción del párrafo.
[25] Alfredo Vergara / [N9_A6] 355 p. / 10 o. (RP)	220 p. / 7 o.	29 p. / 1 o.
	La modalidad de mínimo, contiene tres de los puntos fundamentales que se encuentran en el pasaje original. Esto, refleja su capacidad para resumir. Sin embargo, el resumen no es realmente coherente.	La modalidad de máximo, es la que a mi juicio menos cumple con las propiedades de un resumen, pues la idea expuesta no da para nada una idea acerca del tema ó asunto tratado.
[26] Alexy Sanchez / [N11_A2] 161 p. / 5 o. Opinión General: (RP) Al comparar los resúmenes que produce el resumidor con los resúmenes que produce un ser humano, yo en este caso, nos damos cuenta de que ambos presentan la misma información.	109 p. / 4 o.	25 p. / 1 o.
	Para estos resultados el resumidor fue capaz de identificar el argumento principal del párrafo	Notamos que el resumidor presenta sólo el argumento principal y la premisa de la que se deriva

www.bdigital.ula.ve

B.3. Criterios y Recomendaciones de los evaluadores del sistema

[3] Thomas López

Para un texto determinado, decimos que otro texto lo resume adecuadamente, si este contiene las ideas principales del texto original y evita la información innecesarias. Desde el punto de vista lógico, esto se puede describir como: P y Q entonces R,

Donde:

P: El resumen contiene las ideas principales del texto original

Q: El resumen omite la información innecesaria del texto original

R: El texto resumido RESUME adecuadamente al texto original

[5] Leonardo Segovia.

Como recomendación podemos indicar que se deben tomar en cuenta la omisión de ciertas palabras en los párrafos, en nuestro texto ejemplo las palabras "Así, por ejemplo" pueden ser omitidas y no pierde sentido el párrafo.

C.C.Reconocimiento

Otras expresiones también pueden ser omitidas como aquellas colocadas como aclaración: "La Cámara Municipal, representada por los concejales, se comprometió a aportar de cuatro a cinco millones" se puede transformar a "La Cámara Municipal se comprometió a aportar de cuatro a cinco millones" y no pierde su propósito la oración.

Por ahora, estas son las recomendaciones que te podemos indicar, debido a que deberíamos conocer más a fondo la forma de las normas aplicadas y estudiar más sobre el uso y redacción del castellano.

Nota: otra observación es que el Resumidor al realizar al proceso, separa la coma(,) de la palabra que la antecede, esto contradice una regla gramatical del castellano.

[6] José Sánchez

Al someter a prueba el RESUMIDOR con el texto que me correspondía observo que:

- Los niveles mínimo y moderado no omiten articulos, conectores, adverbios que son prescindibles sin que el texto pierda lo esencial.
- El nivel máximo aún cuando reduce drásticamente la cantidad de palabras conserva conectores innecesarios para ese nivel de resumen.
- Para todos los niveles se observa que el resumidor omite o suprime a lo máximo oraciones del parrafo, pero no llega al nivel de supresión de las palabras prescindibles, que es fundamental para cumplir con lo descrito en los textos referidos que es reducir a términos breves y precisos lo esencial.

[9] Marianela Dávila

El resumidor maximo refina aun mas las ideas llevandolas a una minima expresion, que puede parecer exagerada. Hay parrafos que tienen una idea principal y el resto de sus oraciones se desarrollan en base a esa idea. Pero hay parrafos que poseen mas de una idea principal, que no se pueden manejar tan a la ligera. Me gustaria saber como maneja el resumidor este tipo de parrafos.

www.bdigital.ula.ve

[10] Jesús Ochoa

Al momento de resumir no se cumple la condición del manejo de tópicos: colocar información vieja o ya conocida al inicio de la oración e información nueva o desconocida al final. En el texto se presenta cuando comienza a hablar del posicionamiento, sin haber introducido previamente tal concepto. Considero que el resumidor debe manejar este criterio, de tal manera que cuando omita algunas premisas y conclusiones verifique que la conexión se produzca tomando en cuenta el manejo de tópicos.

[13] Lisdrelys Rojas

Los párrafos bien estructurados son aquellos que al comenzar tienen, la idea principal del texto, para darle una visión al lector sobre el tema a tratar en el resto del párrafo. Luego, en la mitad tienen una secuencia de lo que se deriva de la idea principal, y finalmente, una conclusión del tema tratado en el párrafo.

En párrafos bien estructurados, se puede apreciar, que este resumidor siempre va a tener alcance, para el resumen mínimo, como para el resumen máximo, puesto que resume lo más importante, tomando en cuenta la idea principal del párrafo.

[14] Klaudia Laffaille

Para evaluar el resumidor simbólico voy a considerar los siguientes atributos como ventajas en la elaboración de un resumen, y posteriormente las comparará con los resúmenes obtenidos. Un buen resumen:

C.C.Reconocimiento

- Toma en cuenta las ideas principales y las expresa de forma clara.
- Conserva la información esencial a menos de que su formato requiera suprimirla (resumen moderado o máximo) y en este caso expresa las ideas principales.
- Proporciona una idea precisa y completa del contenido del texto.
- Es claro y breve (de acuerdo al caso).

Cualidades recomendadas:

- Proporcionarle al programa la posibilidad de buscar los sujetos, las acciones y las ideas principales para que pueda, en función de las reglas de la gramática y la sintaxis, reorganizar textos en función de facilitar la comprensión del trasfondo o idea principal del texto.
- Crear una lista de sinónimos posibles para que el resumidor no esté limitado a utilizar solo las palabras contenidas en el texto sino que pueda cambiarlas (sin modificar el sentido) en pro de una mejor comunicación.
- Mejorar el criterio que aplica para omitir la información. En el caso de un resumen realizado por un humano, se utiliza el criterio de importancia en función de la idea principal, si es posible, es este el que debe incorporarse al resumidor.

[16] Glenda González

Según la evaluación se puede decir que:

-El resumidor permite extraer en forma exacta la idea principal de un texto al usar la opción máximo, y las ideas principales y secundarias al utilizar la opción moderado y mínimo.

-El tamaño del texto es un factor determinante en el empleo del resumidor al momento de escoger alguna de las opciones presentadas por éste, para obtener un resultado satisfactorio.

-El empleo de el resumidor en la opción mínimo para textos muy grandes, la opción de moderado para textos de mediano tamaño y la opción de máximo para párrafos, permite obtener en forma concisa la(s) idea(s) más relevantes de lo que se resume.

www.bdigital.ula.ve

Sugerencia: Pienso que el resumidor tiene un buen funcionamiento, si consideramos como dije anteriormente el tamaño del texto a resumir, sería de gran ayuda, aunque sé que no es fácil, establecer un rango de tamaño del texto (o de palabras) en el cuál el resumidor tenga un funcionamiento óptimo para cada opción y hacerselo saber a los usuarios.

[12] Luis Dávila

Personalmente, se me ocurrió cambiar de lugar algunas de las oraciones contenidas en el párrafo (conservando exactamente las mismas palabras) y observe que el resumidor eliminaba exactamente las misma parte del texto. Esto nos indica que ciertamente el resumidor mantiene la claridad de las oraciones más importantes a conservar. De manera similar al factor mínimo, realice pruebas de intercambio del orden de las oraciones y observé que el resumidor siempre ofreció el mismo resultado. Esto comprueba la claridad y precisión en la eliminación de las ideas secundarias y la permanencia de la idea principal de los textos a resumir.

[19] Daniel Signorelli

Mi conclusión general es que el resumidor (al menos con el texto utilizado en mi caso) no funciona. Esto a partir del hecho de que en ninguno de los tres niveles de resumen, se mostró cual es el sujeto del tema, es decir, "el cacao". Aunque, cabe destacar que el resto del texto, conserva la idea central del párrafo. Con esto quiero decir, que si al principio de los tres resúmenes apareciera el sujeto, el resumen, seria entonces muy bueno.

Por otra parte, considero que el resumidor tiene algún problema cuando encuentra paréntesis en el texto original. Esto lo indico ya que como se puede observar al principio de los tres resúmenes, lo primero que aparece es),. Tal vez por ello el resumidor haya omitido el sujeto (el cacao).

C.C.Reconocimiento

[21] Carmen Rodriguez

Según mi criterio un buen resumen debería:
1. Extraer los aspectos más importantes del tema,
2. Resumir las opiniones o comentarios de cada autor y
3. Explicar brevemente todo el tema.

Entonces, para realizar la evaluación del resumidor se considerará si cumple con los tres criterios descritos anteriormente. Aunque siendo una salida de factor 1, debería comenzar a eliminar algunos comentarios que amplían la idea fundamental, como por ejemplo: " ... especialmente para el Reino Unido, en muchos sentidos".

[22] Jose Brito

Normalmente, el resumen de un texto implica una reducciòn del mismo, considerando sòlo aspectos esenciales del tema abordado. Esto comprende la identificaciòn de lo relevante, y con frecuencia, la reescritura o creaciòn de nuevas oraciones (acciòn que el resumidor no desarrollò) para expresarse concisamente.

las palabras resaltadas demuestran un acercamiento aceptable a los tòpicos del texto. Si a un humano se le muestra la lista de tòpicos: La epidemia, cambios, la mayorìa de los consumidores, del consumo de carne de bovino al de carne de pollo, un poco favorecida la agricultura ecològica, como tendencia, una poblaciòn proporcionalmente màs envejecida que requiere una menor, y demanda. Si se le solicita una redacciòn coherente relacionàndolos, muy probablemente "capturarà" el tema de la crisis planteada por la EEB (aunque sin necesariamente circunscribirla a Europa).

Adicionalmente, los resultados reflejan un problema para manejar ciertos tòpicos y sustantivos separados por conectores. En particular, en "del consumo de carne de bovino al de carne de pollo", la preferencia por "pollo" en lugar de las otras carnes no posee una justificaciòn clara (un resumen más apropiado habría presentado "del consumo de carne de bovino al de otras carnes"). Igual ocurre con "la generalizaciòn de las dietas para adelgazar o de mantenimiento" que se encuentra al mismo nivel de la otra causa sì resaltada, y de la cual la separa el conector "y".

En definitiva, con el texto dado el resumidor se desempeña adecuadamente con los factores 0 y 1. En Màximo, sin embargo, el resultado es insatisfactorio por cuanto el mensaje se desvirtùa mucho: esa oraciòn no puede transmitir la idea principal del texto.

[26] Alexy Sanchez

Un resumen debe dar cuenta de la información más relevante y presentarla de forma clara. Entonces, un resumidor debe identificar las partes más importantes y relacionarlas de forma lógica para presentarlas de forma clara. A mi juicio las partes más importantes son los argumentos, es decir, lo que se declara como consecuencia, o lo que es lo mismo, lo que se infiere a partir de premisas. Entonces, el resumidor debe ser capaz de identificar las unidades de relación que anuncian los argumentos, y además, tener la capacidad (no creo que podamos hablar de habilidad) de reexpresarlos de manera clara, reutilizando las unidades de relación.

Declaremos entonces nuestras premisas para evaluar el resumidor de textos:
- Un buen resumidor debe identificar las unidades de relación.
- Un buen resumidor debe identificar el argumento principal del párrafo.
- Un buen resumidor debe relacionar las ideas de forma clara

www.bdigital.ula.ve
C.C.Reconocimiento

Glosario

A

Actuación lingüística
Uso observable del conocimiento del hablante sobre la lengua (Moreno, 1998).

Agente
Fuerza que posee una entidad que realiza una acción y que, de esa manera, modifica una situación (Beaugrande y Dressler, 1997).

Anáfora
Referencia a una expresión previa en un discurso en lenguaje natural. Generalmente usa un pronombre para referirse a personas, lugares o cosas previamente mencionadas (Moreno, 1998).

C

Cadena de tópicos
Secuencia de tópicos de un párrafo correspondientes a la ideas de cada oración. Permiten al lector moverse en contextos conocidos o familiares a lo largo del texto (Williams, 1990).

Claridad
Manera precisa de presentar los actores y acciones de un relato, los cuales corresponden respectivamente a los sujetos y verbos de las oraciones (Williams, 1990).

Coherencia
Conectividad de contenido subyacente de un texto. La coherencia regula la posibilidad de que sean accesibles entre sí e interactúan de un modo relevante los componentes del mundo textual, es decir, la configuración de los conceptos y de las relaciones que subyacen bajo la superficie del texto (Beaugrande y Dressler, 1997).

Cohesión
Conectividad superficial de un texto. La cohesión representa la función comunicativa de la sintaxis que dirigen y mediatiza la operación de acceso a elementos lingüísticos. (Beaugrande y Dressler, 1997).

Competencia lingüística
Conocimiento que cada hablante tiene de su lengua materna (Moreno, 1998).

Complemento
Constituyente de una cláusula, tal como una frase sustantiva o adjetiva, que expresa algo sobre el ***sujeto*** de la cláusula.

Concepto
Estructura de conocimiento (o contenido cognitivo) que el hablante puede activar o recuperar en su mente con mayor o menor unidad o congruencia (Beaugrande y Dressler, 1997).

Contexto
Elemento o conjunto de elementos adyacentes a una unidad lingüística que son pertinentes para su interpretación. Los elementos contextuales pueden localizarse antes y/o después de la unidad seleccionada (Tuson, 2000).

Corpus
Conjunto de textos en lenguaje natural que incluye información extra tales como etiquetas para cada palabra indicando el constituyente gramatical. Esta gran cantidad de textos en lenguaje natural es usada para acumular estadísticas de datos lingüísticos (Moreno, 1998).

www.bdigital.ula.ve

D

Descriptores
Términos característicos y representativos que describen el contenido de un documento (Arntz y Picht, 1995).

E

Elipsis
Situaciones cuyas oraciones son abreviadas o eliminan un constituyente, dejando parte de ellas para ser entendidas por el contexto (Moreno, 1998).

Estilo
Resultado de una determinada elección entre opciones diversas que se ha realizado durante el proceso de producción de un texto o de una serie de textos (Beaugrande y Dressler, 1997).

C.C.Reconocimiento

F

Frase nominal
Unidad formada por al menos un sustantivo. Una frase sustantiva incluye generalmente uno o más palabras, pero puede estar formada como mínimo por un nombre o pronombre.

G

Género (gramatical)
Categoría gramatical propia del sustantivo que refleja la clasificación de los nombres en masculino, femenino y neutro en algunas lenguas indoeuropeas (Tuson, 2000).

Gramática
Definición abstracta de un conjunto de elementos estructurados y bien formados. Una gramática formal es una especificación rigurosa y explícita de la estructura de una lengua. Esta es escrita con un formalismo gramatical, es decir, una lengua artificial creada para describir lenguas naturales. Su uso se debe a que es un lenguaje bien definido, riguroso, facilita la evaluación de hipótesis, y permite desarrollar predicciones. (Moreno, 1998).

www.bdigital.ula.ve

I

Infinitivo
Forma básica del verbo. No presenta marca de categoría inflexional en: tiempo, modo, aspecto, voz, persona y nombre.

L

Lengua común
Núcleo de la lengua del que participan todos los miembros de una comunidad lingüística.

Lenguaje especializado
Área de la lengua que aspira a una comunicación unívoca y libre de contradicciones en un área especializada determinada y cuyo funcionamiento encuentra un soporte decisivo en la terminología establecida (Arntz y Picht, 1995).

Lingüística
Ciencia que estudia el lenguaje humano en todas sus formas.

Lingüística computacional
Disciplina aplicada que estudia los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales (Winograd, 1983).

M

Marcador discursivo
Información interactiva que guía la trayectoria interpretativa de los usuarios textuales (Beaugrande y Dressler, 1997).

Morfología
Disciplina lingüística que estudia la estructura interna de la palabra tanto desde el punto de vista de la flexión (desinencias verbales y nominales), como desde el punto de vista de la formación del verbo (derivación o composición) (Tuson, 2000).

N

Número (gramatical)
Categoría gramatical de nombres, pronombres y verbos que expresa la cantidad.

O

Oración
Unidad sintáctica que forman sujeto y predicado (Tuson, 2000).

P

Parser
Algoritmo o conjunto de instrucciones que relacionan cadenas de símbolos con el conocimiento lingüístico almacenado. El parser es un mecanismo computacional que utiliza una gramática y un diccionario para establecer si una oración es gramatical o no. El problema que tiene que resolver el parser es puramente sintáctico: reconocer las oraciones gramaticales y asignarles una estructura. Otros componentes se encargan de la interpretación (Winograd) (Moreno, 1998).

C.C.Reconocimiento

Participio

Unidad léxica derivada de un verbo que tiene algunas de las características y funciones de verbos y adjetivos.

Pragmática

Disciplina que enfoca el lenguaje desde el punto de vista del uso y de los usuarios del sistema en las circunstancias concreta en las que se realiza el intercambio lingüístico.

PLN Procesamiento del Lenguaje Natural

Área de conocimiento que investiga los modelos computacionales del lenguaje para desarrollar programas informáticos que simulan la capacidad lingüística humana.

Pronombre demostrativo

Palabras que funciona como un nombre al sustituir una frase nominal. Los pronombres demostrativos son determinantes deícticos que indican una referencia espacial, temporal dentro del discurso. Por ejemplo: estos, esos, este, etc.

R

Relaciones

Vínculos que se establecen entre los conceptos que aparecen reunidos en un mundo textual determinado, cada vínculo recibe una denominación según los conceptos que conecte (Beaugrande y Dressler, 1997).

S

Semántica

Dominio de la lingüística cuya finalidad es el estudio del significado (Tuson, 2000).

Sintaxis

Conjunto de relaciones superficiales que interconectan gramaticalmente componentes textuales (Beaugrande y Dressler, 1997).

Superficie textual

Palabras que realmente se escuchan o se leen. Esta compuesto por expresiones lingüísticas presentadas por alguien en la interacción e identificadas por los receptores. Estos componentes dependen unos de otros conformes a sus convenciones y a unas formalidades gramaticales.

Sustantivación

Proceso derivativo mediante el cual se obtiene un vocablo nominal a partir de un verbo o de un adjetivo (Tuson, 2000).

T

Terminología

Conjunto completo de conceptos de un área especializada y sus denominaciones (Arntz y Picht, 1995). Según la Teoría Comunicativa de la Terminología, disciplina de la lingüística aplicada que estudia el uso de las unidades de significación especializada dentro del discurso especializado (Tuson, 2000).

Tesauro

Vocabulario controlado y dinámico, de términos que mantienen entre sí relaciones semánticas y genéricas y que resulta aplicable a un área particular del conocimiento. Es un instrumento lingüístico para indizar documentos (Lerat, 1997).

Texto

Unidad superior a la oración, formada por oraciones con distribución discursiva. Es una actividad comunicativa humana prototipicamente cultural e intencionada, que cumple con siete normas de textualidad: cohesión, coherencia, intencionalidad, aceptabilidad, situacionalidad, intertextualidad e informatividad (Beaugrande y Dressler, 1997). Un texto es un espacio semiótico (Zunzunegui, 1990).

Texto especializado

Unidad del discurso que trata aspectos representativos de un área temática. También llamado texto expositivo porque ofrece información de un área temática especializada.

Token

Elemento básico del lenguaje, que se expresan como un grupo de símbolos en un documento. El token es una representación especial de piezas de texto (Thomas, 1999).

Tokenización

Proceso de división de la entrada en distintos tokens.

Tópico

Frase nominal caracterizada o comentada por el resto de la oración. Para Williams, el tópico es el sujeto psicológico de la oración dado a conocer en las primeras palabras (Williams, 1990).

www.bdigital.ula.ve
C.C.Reconocimiento

V

Verbo

Núcleo del sintagma predicativo cuyas categorías son tiempo, modo, aspecto, voz, persona y nombre (Tuson, 2000).

Verbo auxiliar

Tipo de verbo que se emplea para formar tiempos perifrásticos. Un verbo auxiliar es un verbo que acompaña al verbo léxico (primario) en una frase verbal, expresando las distinciones gramaticales no llevadas por el verbo léxico, tal como: persona, número, aspecto, tiempo y voz.

Voz activa

Categoría relacional del verbo que indica la relación entre el sujeto agente y el sujeto paciente. En esta caso el agente aparece con caso nominativo en aquellas lenguas que tienen morfemas de caso; el sujeto paciente aparece como complemento directo del verbo (caso acusativo) y el verbo aparece en forma no marcada (activa) (Tuson, 2000).

Voz pasiva

Categoría relacional del verbo que indica la relación entre el sujeto agente y el sujeto paciente. En este caso el sujeto paciente se convierte en sujeto (caso nominativo), el agente aparece representado por un sintagma preposicional (en algunos casos) y el verbo aparece en forma marcada (pasiva) (Tuson, 2000).

www.bdigital.ula.ve

C.C.Reconocimiento

Indice

www.bdigital.ula.ve
C.C.Reconocimiento

H

I

www.bdigital.ula.ve

J

K

L

M

N

P

Q

R

C.C.Reconocimiento

www.bdigital.ula.ve

C.C.Reconocimiento

Bibliografía

Acero, I., Alcojor, M., Díaz A. y Gómez, J.M. 2001. Generación automática de resúmenes personalizados. Procesamiento del Lenguaje Natural, 27.

Adam, J. M. 1992. Les textes: types et prototypes. París, Nathan.

Arntz, R. y Picht, H. 1995. Introducción a la Terminología. Madrid: Fundación Germán Sánchez Ruipérez. Traducido por Irazazábal A et al.

Bach, E. 1974. Syntactic Theory. Nueva York, Holt, Rinehart & Winston.

Barwise, J. y Perry, J. 1983. Situations and Attitudes. Cambrigde, M.I.T. Press.

Beaugrande, R.A. 1980. Text, Discourse, and Process. Norwood, N.J., Ablex; Londres, Longman.

Beaugrande, R.A. y Dressler, W. U. 1997. Introducción a la Lingüística del texto. 1ª edición en español. Editorial Ariel, S.A. Barcelona España. ISBN: 84-344-8215-0.

Behrens, L. y Rosen, L. J. 1982. Writing and Reading across the curriculum. Little, Brown & Company (Canada) Limited, Boston USA. ISBN: 0-316-39132-4.

Bosque, I. y Demonte, Violeta. 1999. Gramática Descriptiva de la Lengua Española. Real Academia Española. Colección Nebrija y Bello.

Bowen, D.L. and L.M. Byrd. 1983. A portable Prolog compiler. In L.M. Pereira, ed., Proceedings of the Logic Programming Workshop 1983, Lisbon, Portugal. Universidade Nova de Lisboa.

Cabré, M. T.; Gonzalo, Consuelo y García Yerba. 2000. Documentación, terminología y traducción. Madrid: Síntesis.

www.bdigital.ula.ve

Cabré Castellví, M. T.; Estopà Bagot, R., Vivaldi Palatresi, J. Automatic term detection: A review of current systems. (falta datos de publicación).

Charniak, E. 1975. Organization and Inference in a Frame-Like System of Commom-Sense Knowledge. Castagnola, Institute for Semantic and Cognitive Studies.

Charniak, E. 1993. Statistical Language Learning. Cambridge, The MIT Press.

Chomsky, N. 1957. Syntactic structures. La Haya, Mouton.

Contreras, H. y Dávila, J. 2001. Procesamiento del lenguaje natural basado en una "gramática de estilos" para el idioma español. CLEI'2001. Mérida, Venezuela.

Cooper, W.S. 1971. A definition of relevance for information retrieval. Information Storage and Retrieval, 7, 19-37.

Covington, Michael A. 1994. Natural Language Processing for Prolog Programmers. Artificial Intelligence Programs. The University of Georgia Athens, Georgia. Prentice Hall, Englewood Cliffs.

Cowie, J. y Lehnert, W. 1996. Information Extraction, Communications of the ACM. Enero 1996, Vol. 39, No. 1.

Dávila, J.; Astorga, L.; Márquez, M., Contreras, H., Myerston, J. y Parra, M.. "Introducción a la lingüística computacional con una perspectiva interdisciplinaria". Terminómetro. Ed. Unión Latina. Número. 6. Paris. 2002.

Dávila, J. y Contreras, H. 2002. Una gramática de estilos para resumir textos en español. XVIII Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN), Septiembre 2002. Valladolid, España.

C.C.Reconocimiento

Deransart, Ed-Dbali, y Cervoni. 1996. Prolog: The Standard. Springer-Verlag, New York.

Dijk, T.A. van. 1977. Semantic Macro-Structures and Knowledge Frames in Discourse Comprehension. Cognitive Processes in Comprehension, M.A. Just and P.A. Carpenter, Eds., Lawrence Erlbaum, Hillsdale, N.J., pp. 3-32.

DiMarco, C. y Hirst, G. 1993. A computational theory of goal-directed style in syntax. Computational Linguistics. 19, 3 septiembre 451-499.

Grice, H., P. 1975. Logic and conversation. Cole y Morgan (Eds.), pp. 41-58.

Hahn, U. y Mani, I. 2000. The Challenges of Automatic Summarization. Computer IEEE. Noviembre.

Hand, T. F. 1997. A Proposal for a Task-based Evaluation of text Summarization Systems. En Proceddings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

Hahn, U. y Reimer, U. 1999. Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge-Based Abstraction. Avanced in Automatic Text Summarization. I. Mani y M. Maybury, Eds., MIT Press, Cambridge, Mass., pp. 215-232.

Hoffmann, L. 1985. Kommunikationsmittel Fachsprache: Eine Einführung. 2° reedición revisada. Tübingen: Gunter Narr Verlag.

Kamp, H. y Reyle, U. 1993. From Discourse to Logic. Dordrecht, Kluwer.

King, M. 1996. Evaluating Natural Language Processing Systems. Communications of the ACM. Enero, Vol. 39, No. 1.

Kupiec, J.; Pedersen, J. y Chen, F. 1995. A Trainable Documment Summarizer. Proc. 18th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, E.A. Fox, P. Ingwersen, and R. Fiel, eds., ACM Press, New York. pp. 68-73.

www.bdigital.ula.ve

Lerat, Pierre. 1997. Las lenguas especializadas. Ariel Lingüística. Barcelona, España.

Locke W.N. y Booth A.D. 1955. Machine Translation of Languages, Technology Press of MIT and Wiley, Cambridge, Mass.

Lyons, J. 1968. Introduction to Theorical Linguistics. Cambridge University Press.

Manaris, B. Z. y Slator, B. M. 1996. Interactive Natural Language Processing: Building on Success. Computer, IEEE.

Mandler, J. y Johnson, N. 1997. Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, pp. 111-51.

Mani, I., E. Bloedorn, y B. Gates. 1998. Using Cohesion and Coherence Models for Text Summarization. In Working Notes of the AAAI'98 Spring Symposium on Intelligent Text Summarization, 69 76. Stanford, CA.

Manganaris, Stefanos. 2001. Reading between the lines. Refining your text mining techniques. Data Minner. DB2 Magazine, quarter 1.

Maña, M., Buenaga, M. y Gómez J.M. 1998. Diseño y evaluación de un generador de resúmenes de texto con modelado de usuario en un entorno de recuperación de información. En XIV Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN'98), 23-25 septiembre, Alicante (España). Publicado en Procesamiento del Lenguaje Natural, nº 23, septiembre, 32-39.

Miller, G. 1956. The magical number seven, plus or minus two. Psychological Review, 63, pp. 81-97.

Miller, G. y Johnson-Laird, P. N. 1976. Language and Perception. Cambridge, Cambridge University Press.

C.C.Reconocimiento

Moreno Sandoval, A. 1998. Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos. Madrid: Sintesis.

Munakata, Toshinori. 1999. Knowledge Discovery. Communications of the ACM. Noviembre. Vol. 42 No. 11.

Ochoa A., J. I., Tutor: Davila, J. 1999. Formulación de Normas Lógicas para el discurso escrito de noticias. Tesis de Grado no publicada de Ingeniería de Sistemas, Universidad de Los Andes. Mérida-Venezuela.

Manuel Palomar; Lidia Moreno; Jesus Peral; Rafael Munoz; Antonio Ferrandez; Patricio Martinez-Barco; Maximiliano Saiz-Noeda. 2001. An Algorithm for Anaphora Resolution in Spanish Texts. Computational Linguistics, Volume 27, Number 4, December 2001.

Pereira, F.C.N. y Warren, D.H.D. 1986. Definite Clause Grammars for Language Analysis-a Survey of the Formalism and a Comparison with Augmented Transition Networks. Artificial Intelligence. 13:231-278. pag. 101-138.

Pereira, F. 1996. Sentence Modeling and Parsing. En The State of the Art of Human Language Technology, capítulo 3.6.

Quesada M., José F. y Amores C. José G. 2000. Diseño e Implementación de sistemas de traducción automática. Universidad de Sevilla. Secretariado de publicaciones.

Russell, S. y Norving, P. 1995. Artificial Intelligence: A modern approach. Prentice Hall Series in Artificial Intelligence.

Salton, G. et al. 1997. Automatic Text Structuring and Summarization. Information Processing & Management. Vol. 33, No. 2. pp. 193-207.

www.bdigital.ula.ve

Schank, R. C. 1975. Conceptual Information Processing. Amsterdam: North-Holland.

Schank, R. y Abelson, R. 1977. Script, Plans, Goals ans Understanding. Hillsdale, N. J., Erlbaum.

Schreiber, A. Th. and B. J. Wielinga. SWI: Making Knowledge Technology Work. IEEE Expert, 11(2):74-76, April 1996.

Sommerville, I. 1992. Software Engineering. 4ª Edición. Addison-Wesley.

Thomas, Bernd. 1999. Token-Templates and Logic Programs for Intelligent Web Search. University of Koblenz-Landau, Abteilung Landau, Institut für Informatik.

Tuson, Jesús. 2000. Diccionari de Lingüística. Barcelona, España. Biblograf s.a.

Warren, D.H.D. 1983. The Runtime Environment for a Prolog Compiler Using a Copy Algorithm. Technical Report 83/052, SUNY and Stone Brook, New York, 1983. Major revision, March 1984.

Wiebe, J., Hirst, G. y Horton, D. 1996. Language use in Context. Communications of the ACM, Enero 1996, Vol. 39, No. 1.

Williams, J. 1990. Style: Toward Clarity and Grace. The University of Chicago Press. Chicago and London.

Winograd, T. 1983. Language as a Cognitive Process: Syntax. Reading, Addison-Wesley.

C.C.Reconocimiento