VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"AGRUPAMIENTO SEMÁNTICO DE CONTEXTOS DEFINITORIOS"

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS (COMPUTACIÓN)

P R E S E N T A:

ALEJANDRO MOLINA VILLEGAS

DIRECTOR DE TESIS: Dr. Gerardo Sierra Martínez

MÉXICO, D.F., 2009.

Para mi padre, mi gran maestro.

Sin tu « Blues » la vida sería un error.

Agradecimientos

A mi asesor, el Dr. Gerardo Sierra por creer en mi, al Dr. Juan Manuel Torres por el apoyo que me ha brindado para continuar con mis estudios, a los miembros del Grupo en Ingeniería Lingüística por su entusiasmo y compañerismo, a Ariadna, al personal del posgrado en Ciencias e Ingeniería de la Computación por brindarme un hogar, al Dr. Fernando Arámbula, al Dr. Alfonso Medina por sus enseñanzas y a la Dra. Sofía Galicia por haberme introducido en esta interesante área.

A mi familia y a mis amigos que me han llenado de alegría.

Introducción

RESUMEN: En este estudio presentamos un método eficaz de agrupamiento automático de definiciones, el cual refleja relaciones semánticas entre éstas. Parte del método aquí presentado está basado en una técnica recientemente estudiada, conocida como *energía textual*, la cual nunca antes ha sido utilizada para el problema de agrupamiento. En este trabajo de investigación se desarrollan las ideas subyacentes a la implementación de un algoritmo de agrupamiento de definiciones como un módulo que puede ser agregado a una aplicación Web; dejando al descubierto la posibilidad de otras aplicaciones como el agrupamiento de pequeñas unidades de información desplegadas por los motores de búsqueda (*snippets*).

ANTECEDENTES

En la actualidad, la cantidad de información acumulada de manera escrita resulta apabullante, debido a la masificación de información en formato electrónico que ha sido propiciada por el uso intensivo de la Web. La abundancia de definiciones en la red no escapa a esta realidad, ya que las definiciones representan un recurso invaluable para cualquier área del conocimiento pues aportan información útil. Es ahí donde se concentra el conocimiento y la información que se va a transmitir acerca de un término. Por supuesto que no nos referimos a la situación donde un comité de sabios se reúnen para determinar la mejor descripción de todos los objetos del mundo. La necesidad de expresar conceptos a través de las palabras nos obliga, la mayoría de las veces, a convertirnos en lexicógrafos de manera cotidiana y sin percatarnos de ello. Como seres humanos, tenemos la capacidad de acumular conocimiento y, como seres sociales, la de transmitirlo. Ante este panorama, surge la necesidad de contar con herramientas eficaces de extracción y procesamiento de definiciones.

Existen estudios recientes concernientes a la extracción automática de definiciones en texto digital. La idea fundamental de los métodos de extracción radica en el hecho de que se han encontrado patrones que fungen como indicadores de definiciones (Pearson, 1998). Por ejemplo en la frase:

> *Carlos Godino **define** la arquitectura naval **como** la ciencia que trata de los conocimientos necesarios para la construcción de los buques.*

El verbo *define* indica que lo que viene a continuación es el término que se define, en este caso *la arquitectura naval*, y el adverbio *como* introduce la definición.

Partiendo del supuesto de que es posible detectar automáticamente las definiciones en su contexto y luego agruparlas por su contenido, surge la idea de desarrollar el sistema *Describe*[1], un sistema que obtiene definiciones inmersas en la Web y, a su vez, es capaz de formar grupos de estas definiciones, de acuerdo con la acepción a la que aluden. Este sistema se consulta a través de una interfaz Web de tipo buscador; es un buscador de definiciones.

Uno de los componentes principales del *Describe* es el módulo de agrupamiento que se encarga de organizar y presentar la información al usuario. El algoritmo utilizado para generar los grupos semánticos a partir de una colección de definiciones es presentado aquí y su implementación corresponde al módulo de agrupamiento del sistema *Describe*.

La idea de agrupar definiciones por su contenido semántico no es nueva. Sierra y MacNaught (2000) presentan interesantes avances con el alineamiento de palabras inmersas en definiciones. Siguiendo estas ideas, Castillo (2000) expone un algoritmo para extracción de grupos de palabras en el cual utiliza con éxito diversas técnicas propias del área de procesamiento del lenguaje natural, tales como: truncamiento de palabras (*stemming)* y el uso de listas de paro (*Stop list)*. En este estudio sacamos provecho de las investigaciones antes mencionadas para plantear y desarrollar un algoritmo de agrupamiento capaz de detectar y agrupar los diferentes significados o acepciones de un mismo término a partir de su aparición en un sitio Web.

PLANTEAMIENTO DEL PROBLEMA

El problema que resolvemos en esta tesis se puede enunciar de la siguiente manera: dada una colección de definiciones de un término, es necesario estructurar grupos de tal forma que aquellas definiciones que poseen el mismo significado sean asociadas a un mismo grupo y dos grupos distintos hagan referencia a campos semánticos distintos. Por ejemplo, considere las siguientes definiciones del término *virus*:

1. *Programa que tiene la capacidad de infectar a otros programas para modificarlos e incluir una copia de si mismo en ellos.*

[1] www.describe.com.mx

2. *En tecnología de seguridad en computadoras, un virus es un programa auto replicable que se expande insertando copias de sí mismo en códigos ejecutables o documentos.*
3. *Se entiende un programa diseñado para alterar el funcionamiento de los equipos, sin la autorización o conocimiento del usuario.*
4. *Virus es la máxima expresión de la modernidad en el rock nacional.*
5. *Virus es una banda de rock argentina fundamental del new wave de los años '80, liderada por Federico Moura hasta su muerte, en diciembre del 1988 a causa del VIH. Su hermano Marcelo tomó entonces el rol de vocalista principal y la banda continuó funcionando hasta fines de 1989.*
6. *El organismo más pequeño que puede causar una infección.*
7. *Microorganismo más pequeño que puede causar infecciones y para sobrevivir necesita estar dentro de una célula viva.*

Una posible, y aceptable, estructura de grupos generada por el algoritmo de agrupamiento, podría ser la siguiente:

GRUPO 1:

1. *Programa que tiene la capacidad de infectar a otros programas para modificarlos e incluir una copia de si mismo en ellos.*
2. *En tecnología de seguridad en computadoras, un virus es un programa auto replicable que se expande insertando copias de sí mismo en códigos ejecutables o documentos.*
3. *Se entiende un programa diseñado para alterar el funcionamiento de los equipos, sin la autorización o conocimiento del usuario.*

GRUPO 2:

4. *Virus es la máxima expresión de la modernidad en el rock nacional.*
5. *Virus es una banda de rock argentina fundamental del new wave de los años '80, liderada por Federico Moura hasta su muerte, en diciembre del 1988 a*

causa del VIH. Su hermano Marcelo tomó entonces el rol de vocalista principal y la banda continuó funcionando hasta fines de 1989.

GRUPO 3:

6. *El organismo más pequeño que puede causar una infección.*
7. *Microorganismo más pequeño que puede causar infecciones y para sobrevivir necesita estar dentro de una célula viva.*

Otro ejemplo, menos trivial, lo da la siguiente colección para el término *gen*:

1. *El gen es la unidad de la herencia.*
2. *Cada gen es una secuencia de ácido nucleico que lleva la información que determina un polipéptido concreto .*
3. *Un gen es una secuencia de ADN que codifica una proteína, ARNt o ARNr .*

No hay que ser un experto para darse cuenta de que existe cierta relación entre las definiciones 2 y 3, basta con recordar que ADN es la sigla de ácido desoxirribonucleico lo cual podríamos incluso deducir por el contexto de ambas frases (*gen es una secuencia de...*). Tampoco es difícil notar que en 1, se define *gen* con un enfoque distinto que en 2 y 3. Tal vez dividiríamos la colección en dos grupos, uno donde concebimos al gen como *una unidad hereditaria* y otro donde lo concebimos como *una secuencia*. Observe que hemos usado nuestro conocimiento previo del mundo para realizar tales inferencias. Para una computadora, esta tarea resulta completamente imposible. En lugar de eso, utilizaremos una representación de los textos y una serie de técnicas numéricas para formar los grupos.

En el presente estudio, dejamos de lado las aproximaciones de carácter teórico-lingüístico, ya que consideramos que caracterizar semánticamente las frases tomando en cuenta todos los factores lingüísticos involucrados no haría mas que añadir complejidad innecesaria a la metodología. Por un lado, observaríamos que la gramática –una rama formal y bien estructurada de la lingüística– no basta para explicar el significado de las oraciones. Y si quisiéramos saber en que grado la comprensión de las oraciones se explica sin la gramática, rápidamente nos daremos cuenta que la gramática proporciona descripciones estructurales idénticas para oraciones cuyo significados son diferentes: *El perro mordió a aquel hombre* vs *El gato rasguñó a esa mujer*; y, descripciones estructurales diferentes para oraciones con el

mismo significado: *El perro mordió a aquel hombre* vs. *Aquel hombre fue mordido por el perro*.

Lo anterior muestra que una gramática no asegura el significado de sus partes pero si añade complejidad. No obstante, podemos valernos de métodos de análisis numérico que hasta ahora han mostrado ser una buena aproximación en procesamiento automático de lenguaje natural y que han dado buenos resultados en áreas como: minería de textos (*text mining*), recuperación de información (*information retrieval*), procesamiento de lenguaje natural (*natural language processing*), Web semántica (*semantic web*), etcétera. Hemos resuelto el problema en términos de estas áreas de las ciencias de la computación; más concretamente, con el uso de una representación vectorial para las frases y con la aplicación de un método de aprendizaje no supervisado.

OBJETIVOS

Objetivo General: Contribuir al desarrollo de métodos eficientes de agrupamiento de fragmentos textuales de información mediante aproximaciones de aprendizaje no supervisado.

Objetivos específicos:

1. Utilizar métodos estadísticos para el desarrollo de un algoritmo de agrupamiento de contextos definitorios.
2. Reconocer las características de los contextos definitorios que permitan elegir la metodología adecuada.
3. Consolidar un corpus de contextos definitorios adecuado para la experimentación de nuestra investigación.
4. Implementar el algoritmo de agrupamiento en algún lenguaje de programación.
5. Evaluar experimentalmente la eficacia del algoritmo de agrupamiento de contextos definitorios.

JUSTIFICACIÓN

Hoy en día, la información es tan abundante y tan valiosa que el desarrollo de herramientas para el procesamiento inteligente de contenido textual representa una contribución importante a *la sociedad de la información*. En esta investigación, tenemos el propósito de generar un

método eficaz de agrupamiento de definiciones que refleje las relaciones semánticas entre éstas.

La aplicación del agrupamiento de textos es una técnica que ha sido de utilidad en diversas áreas. Por ejemplo, en recuperación de información, la hipótesis del agrupamiento (*cluster hypothesis*) formula que los documentos en el mismo grupo tienden a ser relevantes a los mismos requerimientos de información (Jardine y vanRijsbergen, 1971). Es decir que si una persona tiene la necesidad de obtener información acerca de un tema y encuentra un documento que le resulta relevante, es muy probable que los documentos en el mismo grupo (después de aplicar un método de agrupamiento) también le resulten relevantes.

Creemos que el estudio es conveniente porque los resultados obtenidos son de interés tanto en las ciencias de la computación como en lingüística. Por el lado de la computación, involucra técnicas de inteligencia artificial, procesamiento de lenguaje natural, minería de datos y recuperación de información. Más específicamente, se utilizan los conceptos de: gramáticas y lenguajes formales, redes neuronales, *cluster analysis*, lematización, entre otros. Por el lado de la lingüística, involucra las áreas de lingüística computacional y lexicografía.

ESTRUCTURA DE LA TESIS

En el capítulo 1 introducimos formalmente el concepto de contexto definitorio al que nos referiremos durante todo el trabajo, presentamos sus elementos constitutivos principales y describimos una taxonomía de los mismos. Finalizamos el capítulo describiendo la idea de extracción de definiciones y sus limitantes, de esta manera, hacemos evidente la posibilidad de acoplar el módulo de extracción con el módulo de agrupamiento aquí descrito.

El capítulo 2 versa sobre el agrupamiento, mejor conocido como *clustering* en las ciencias de la computación. En este capítulo describimos los elementos teóricos necesarios para comprender la propuesta de solución al problema de agrupamiento que se nos presenta, así como las características que nos llevaron a la solución aquí presentada.

En el capítulo 3 explicamos detalladamente el algoritmo de agrupamiento de contextos definitorios. Primero, caracterizando la entrada al algoritmo y después describiendo los pasos necesarios para producir los grupos de definiciones que reflejen relaciones semánticas.

El capítulo 4 describe la consolidación de un corpus de prueba para nuestro algoritmo y los resultados de aplicar el algoritmo a dicho corpus. También mostramos gráficamente los resultados obtenidos por medio de árboles jerárquicos, así como tablas donde se muestran algunos coeficientes utilizados para evaluar el algoritmo de agrupamiento en general.

Finalmente damos una síntesis de la interpretación de los resultados y mostramos, a manera de conclusión, algunos aspectos importantes de las evaluaciones. Al final vinculamos el trabajo realizado en esta tesis con futuros desarrollos relacionados con tecnologías del lenguaje.

1. CONTEXTOS DEFINITORIOS

1.1. DELIMITACIÓN DE LOS CDS

Por contexto definitorio (CD) se entenderá todo aquel fragmento textual de un documento especializado donde se define un término (Alarcón *et al.*, 2008). Por ejemplo:

1. *El buque inicialmente lo podemos considerar como un flotador que trata de permanecer en posición vertical frente a perturbaciones exteriores.*

El fragmento anterior es un contexto definitorio debido a que posee un término, en este caso *buque*, con su respectiva definición: *como un flotador que trata de permanecer en posición vertical (…)*. Como vemos en el ejemplo, el término y su definición se relacionan a través de la estructura sintáctica *considerar como;* a este tipo de estructuras que ligan los dos elementos más importantes de un CD se les llama patrones definitorios (PD).

Por lo general, estos contextos son de extensión breve y terminan en el primer punto. Sin embargo, existen casos donde pueden ir más allá del primer punto y casos donde pueden terminar antes de éste (Hernández, 2009). Un ejemplo de CD que termina después del punto es:

2. *Se conoce como transformador de corriente a aquél cuya función principal es cambiar el valor del la corriente de uno más o menos elevado a otro con el cual se puedan alimentar instrumentos de medición control o protección, como amperímetros, instrumentos registradores, reveladores de sobre-corriente, etc. Su construcción es semejante a la de cualquier tipo de transformador, ya que fundamentalmente consiste de un devanado primario y uno secundario.*

En el ejemplo anterior la definición del término *trasformador de corriente* se extiende más allá del punto y del primer párrafo. Por otra parte, también pueden darse casos donde el CD termina antes, por ejemplo:

3. *La acción es entendida como la conducta intencionada proyectada por el agente, en cambio el acto es definido como la acción cumplida.*

En el contexto anterior la definición del término *acción* termina hasta la palabra *agente*, como se puede observar.

Para este trabajo de investigación no consideraremos los casos donde el contexto definitorio termina antes o después del punto. Solamente consideraremos contextos definitorios que terminan, junto con la frase, en el punto, como en el ejemplo 1.

1.2. Elementos Constitutivos de los CDs

Los elementos constitutivos principales de un contexto definitorio son: el término, la definición y el patrón definitorio. El término (T) es la unidad sobre la cual se aporta información relevante y la definición (D) es la información sobre dicho término. Estos dos constituyentes de un CD se encuentran conectados mediante un patrón definitorio (PD), el cual puede ser un signo de puntuación, una marca tipográfica o un verbo. Además de estos existen otros elementos dentro de los contextos definitorios como los patrones pragmáticos (PP), los cuales ya han sido definidos y establecidos en otros trabajos (Alarcón *et al.*, 2003). En el siguiente CD, el término es *gen*, la definición es *la unidad de la herencia* y ambos están conectados mediante el patrón definitorio *se define como*.

> *En términos generales, el gen se define como la unidad de la herencia.*

En la figura 1.1 se observa una representación de la estructura de un contexto definitorio, donde los elementos mínimos constitutivos T y D, unidos mediante el patrón definitorio PD, conforman una unidad que puede estar modificada o no por un elemento optativo, un patrón pragmático (PP). En el ejemplo anterior, *En términos generales* representa un patrón pragmático.

En seguida se describen los elementos constitutivos principales de los contextos definitorios: el término, la definición y el patrón definitorio.

Figura 1.1. Elementos constitutivos de los contextos definitorios.

El **término** es la unidad sobre la cuál se aporta información relevante y puede tener estructuras sintácticas diferentes. El núcleo de un término generalmente será nominal, aunque no se debe descartar que en ocasiones pueda ser de otro tipo, como verbal o adjetival (Alarcón, 2006). Dependiendo del área especializada es común que en ocasiones lo que se defina esté más relacionado con fórmulas o elementos que, si bien no siguen patrones morfosintácticos comunes, sí representan una unidad de conocimiento. En los experimentos desarrollados en este trabajo consideramos únicamente términos de tipo nominal, por tratarse de los más comunes. La tabla 1.1 muestra ejemplos de los tipos de términos que pueden hallarse en los contextos definitorios.

Término	**Tipo**
Célula	Nominal
Gen supresor de tumores	Nominal
Suturar	Verbal
AC-137	no lingüístico

Tabla 1.1. Ejemplos de términos.

Por su parte, la **definición** en un contexto definitorio es donde está contenido el conocimiento y la información que se va a transmitir acerca del término con la finalidad de llegar a su comprensión cabal.

Para la realización de este proyecto utilizamos una serie de definiciones encontradas en la Internet. Como es de esperar, existe una gran diversidad en el uso del español para transmitir conceptos por medio de definiciones.

La tabla 1.2 muestra ejemplos de definiciones encontradas en la Web. El inicio y fin de la definición está delimitado por las marcas <D> y </D> respectivamente. El término es delimitado por las marcas <T> y </T>.

Definición	Fuente
La <T>aguja</T> es <D>la mejor compañera de la mujer</D>	uncajonrevuelto.arte-redes.com/
La <T>barra</T> es <D>el jugador número doce dentro de la cancha, que se involucra en cada una de las jugadas a favor o en contra de su equipo o jugador favorito</D>	www.critica.com.pa/archivo/10042007/dep05
la <T>célula</T> es <D>una comunidad de amor</D>	ucucartagena.ohlog.com/manual-para-el-lider-de-celula.oh31100.html

Tabla 1.2. Ejemplos de definiciones.

Ahora bien, en un CD, los dos elementos constitutivos anteriores están ligados mediante un **patrón definitorio (PD)**. Los patrones definitorios son de dos tipos: patrones tipográficos (PT) o sintácticos (PS). Los primeros (PT) están formados por elementos tipográficos, como mayúsculas, subrayado o signos de puntuación. La marca de dos puntos es un ejemplo de patrón tipográfico en: *liebre***:** *Especie de conejo.* Por su parte, los patrones sintácticos (PS) se dividen en dos: marcadores reformulativos definitorios (MRD), como *es decir, o sea, esto es,* etc.; y patrones verbales definitorios (PVD), los cuales son estructuras sintácticas específicas, como la secuencia *se puede definir como* en: *el índice de aridez se puede definir como el porcentaje de la falta de agua*.

Los PVD pueden ser simples o compuestos. Estos son simples cuando sólo están formados por el verbo sin ninguna otra partícula gramatical y compuestos cuando el patrón presenta una estructura que puede incluir preposiciones, adverbios, pronombres, etc. Por ejemplo:

> <PVD>*Se define como*</PVD> <T>*tensión de contacto*</T> <D>*al valor de la tensión que se presenta, al paso de la corriente a tierra, entre las masas metálicas conectadas a tierra y el terreno circunvecino, que puede eventualmente, en alguna forma, entrar en contacto con una persona*</D>.

En este caso encontramos que el PVD *se define como* es de tipo compuesto, porque está formado no sólo por una conjugación del verbo definir, sino también por el clítico *se* y el adverbio *como*. Observe que los PVD generalmente utilizan verbos metalingüísticos, como

definir, *denominar*, *describir* (Rodríguez, 1999). Aunque también utilizan verbos no metalingüísticos, como *ser*, *conocer*, *identificar*, los cuales pueden funcionar como definitorios según el contexto.

Para este estudio hemos considerado únicamente CDs con patrones verbales definitorios (PVD), dado que Alarcón (2003) muestra una metodología más o menos establecida para su extracción.

Es importante mencionar que los patrones verbales definitorios se encuentran asociados a un tipo de definición específica debido al tipo de relación que establecen entre el término y su definición. Por ejemplo, patrones como *consta de* o *formado por* suelen introducir contextos donde se describen las partes que conforman un término determinado, mientras que *sirve para* suele introducir la funcionalidad o la utilidad del término definido. En la siguiente sección exponemos una clasificación de los contextos definitorios de acuerdo con los tipos de definiciones y patrones verbales que contienen.

Por último, cabe señalar que a pesar de que existen diversos estudios sobre los elementos constitutivos de los CD**,** para el presente estudio basta con conocer los elementos explicados hasta ahora. Aguilar (2009) presenta un estudio muy completo en el tema.

1.3. TIPOS DE DEFINICIONES

Como mencionamos anteriormente, los patrones verbales se encuentran asociados a un tipo de definición específica según la relación que establecen entre el término y su definición. En lo que se refiere al estudio de contextos definitorios en español, se plantean cuatro tipos de definiciones basadas en el modelo de definición aristotélico, el cual se explica por la expresión:

X = genero próximo + diferencia específica

Donde, *X* sustituye al término, el *género próximo* representa la categoría general a la cual pertenece el término y la *diferencia específica* son las características del término que lo distinguen de todos los demás elementos pertenecientes a una misma categoría (Aguilar, 2009).

Los tipos de definiciones identificados son: analítico, funcional, extensional y sinonímico.

Las **definiciones analíticas** son aquellas que presentan un género próximo que expresa la categoría más general a la cual pertenece el término, así como la información (diferencia específica) que permite distinguir el término de otros elementos de su misma clase. Algunos patrones verbales asociados a las definiciones de este tipo son: *ser+un, definir+como, entender+ como, identificar+como*. Por ejemplo:

> *Carlos Godino define la arquitectura naval como la ciencia que trata de los conocimientos necesarios para la construcción de los buques.*

Esta definición, cuyo término es *la arquitectura naval*, tiene como género próximo a *la ciencia* y como diferencia específica a *que trata los conocimientos necesarios* (…).

Las **definiciones extensionales** enumeran las partes que conforman al término visto como un todo. Algunos verbos ligados a este tipo son: *constar, contener, comprender, incluir*. Un ejemplo de definición extensional es:

> *La zona límite incluye planicies, costeras, marismas, áreas de inundación, playas, duna y corales.*

En este contexto definitorio no hay un género próximo sino una partición que muestra varios componentes del término *zona límite*.

En las **definiciones funcionales** no se presenta el género próximo y, en cambio, se introduce una diferencia específica donde se explica la función o el uso particular del término. Algunos patrones verbales relacionados con estas definiciones son: *funcionar, encargarse +de, permitir, servir+para*. El siguiente fragmento es un CD con una definición de tipo funcional en el cual, mediante el PVD *permitir,* se aporta dentro de la definición la funcionalidad del término *la técnica de velocimetría de imágenes de partícula.*

> *La técnica de velocimetría de imágenes de partícula, permite medir la velocidad de un campo de flujo bi o tri dimensional.*

Por último, las **definiciones sinonímicas** establecen una equivalencia entre el término y su definición. Por ejemplo:

> *La tensión de base se le llama también tensión unidad.*

En el contexto definitorio anterior, a través del PVD *llamar también,* se expresa una relación de igualdad entre los términos *tensión de base* y *tensión unidad.*

La tabla 1.3 muestra de manera resumida los tipos de definiciones y algunos de sus patrones verbales asociados.

Patrón Verbal	**Tipo**
ser + determinante + nombre	Analítica
definir como	Analítica
entender como	Analítica
concebir como	Analítica
identificar como	Analítica
constar de	Extensional
formar de	Extensional
contener	Extensional
tener	Extensional
usar en | como | para	Funcional
utilizar en | como | para	Funcional
servir en | como | para	Funcional
ser + igual +a	Sinonímica
equivaler +a	Sinonímica

Tabla 1.3. Tipos de definición y patrones definitorios asociados.

En este trabajo nos concentraremos en las definiciones de tipo analítico, funcional y extensional. Dejamos de lado las de tipo sinonímico, dado que aún no han sido ampliamente exploradas.

En la sección siguiente, mencionamos algunas ideas y técnicas que motivaron la extracción automática de definiciones a partir de patrones verbales. Una temática que consideramos como el antecedente directo y una de las motivaciones del trabajo que aquí presentamos.

1.4. EXTRACCIÓN DE CDS

Recordemos que existen patrones lingüísticos asociados a las definiciones. Jennifer Pearson, en su libro *Terms in Context* (1998), nos explica las características de dichos patrones y su rol

dentro de los contextos definitorios. Mediante la búsqueda de patrones, esta autora encontró que podía extraer información valiosa para la formulación de definiciones con un esfuerzo computacional mínimo. Pearson no propone un sistema complejo para la extracción automática de este tipo de información, pero sí comprueba que mediante la búsqueda de las ocurrencias de los patrones definitorios se pueden encontrar términos y definiciones.

Bajo este panorama surgen diversas propuestas con un enfoque y aplicación distinto**,** pero con la idea general de identificar contextos ricos en conocimiento de manera automática. También se comparte la idea de que deben buscarse patrones recurrentes que ayuden a encontrar información relevante sobre términos.

Partiendo de la hipótesis de que los contextos definitorios pueden ser extraídos automáticamente mediante la búsqueda de patrones, Alarcón (2009) presenta la tesis de doctorado titulada *Extracción automática de contextos definitorios en corpus especializados* desarrollada en el Instituto Universitario de Lingüística Aplicada (UPF) en colaboración con el Grupo de Ingeniería Lingüística (II, UNAM). Como parte de sus resultados, el autor presenta el prototipo de *ecode* (acrónimo de Extractor de Contextos Definitorios): un sistema de reconocimiento y extracción automática de definiciones inmersas en texto libre.

Además de la extracción, *ecode* es capaz de distinguir tres tipos de definiciones de acuerdo con en el patrón verbal que éstas presentan. Esta clasificación corresponde con alguno de los tipos: analítico, funcional o extensional. Uno de los objetivos de este estudio es añadir un módulo de agrupamiento a la salida del sistema *ecode*, con la finalidad de revelar relaciones semánticas más sutiles entre definiciones asociadas a un mismo término y que comparten un mismo tipo de definición.

En este punto resulta importante señalar la diferencia entre clasificación y agrupamiento. En la clasificación, las categorías son previamente definidas mientras que en el agrupamiento las categorías son *descubiertas* durante el proceso. Para ejemplificar, supongamos que el sistema *ecode* extrae una colección de contextos definitorios asociados al término *virus*, los CDs de tipo analítico son separados, por el *ecode*, de los extensionales y de los funcionales a partir del patrón verbal definitorio que presentan.

El siguiente grupo de definiciones simula el resultado de la extracción, de *ecode*, de los CDs del término *virus*. Aparece en negrita el patrón verbal definitorio de cada CD. El sistema distingue que las definiciones 1 a 6 corresponden al tipo analítico, la 7 y la 8 son de tipo extensional y la 9 corresponde al tipo funcional:

1. *En tecnología de seguridad en computadoras, un virus* ***es un*** *programa auto replicable que se expande insertando copias de sí mismo en códigos ejecutables o documentos.*
2. ***Se entiende un*** *programa diseñado para alterar el funcionamiento de los equipos, sin la autorización o conocimiento del usuario.*
3. *Virus* ***es la*** *máxima expresión de la modernidad en el rock nacional.*
4. *Virus* ***es una*** *banda de rock argentina fundamental del new wave de los años '80, liderada por Federico Moura hasta su muerte, en diciembre del 1988 a causa del VIH. Su hermano Marcelo tomó entonces el rol de vocalista principal y la banda continuó funcionando hasta fines de 1989.*
5. *Un virus* ***es el*** *organismo más pequeño que puede causar una infección.*
6. *Un virus* ***es el*** *microorganismo más pequeño que puede causar infecciones y para sobrevivir necesita estar dentro de una célula viva.*
7. *Un virus* ***contiene*** *instrucciones para iniciar una serie de "eventos" que afectan la computadora infectada.*
8. *En esencia un virus* ***consta de*** *ácido nucleico rodeado por una cubierta de. proteína, llamada cápsida.*
9. *Es un virus,* ***sirve para*** *robar contraseñas bancarias y ese tipo de cosas.*

Dentro de las definiciones de tipo analítico (1 a 6) reconocemos que existen distintas acepciones para *virus*. Es decir, a pesar de ser todas definiciones analíticas, el significado varía según el ámbito en el cual se define el término: computación, biología o música contemporánea. Es por esto que en los contextos 1 y 2 se define el término *virus* como *un programa*, mientras que en 3 y 4 como *una agrupación musical* y en 5 y 6 como *un organismo*.

El sistema *ecode* es capaz de reconocer y separar éstas 6 definiciones analíticas del resto de las definiciones, pero es incapaz de indicar y distinguir los diversos significados de un término dentro de un mismo tipo de definición. Es de esperar que este fenómeno se manifieste de manera similar en las definiciones funcionales y en las extensionales y que el problema se acentúe a medida que el sistema recupera mayores cantidades de información. Ante esto, como medida de solución y mejoramiento de resultados se plantea la posibilidad de

entregar la salida del sistema *ecode* a un sistema de agrupamiento que distinga las diferentes acepciones de un término asociadas a un mismo tipo de definición .

En el siguiente capítulo describimos las bases teóricas que sustentan el método de agrupamiento de definiciones propuesto en este trabajo.

2. BASES TEÓRICAS DE AGRUPAMIENTO

2.1. AGRUPAMIENTO

No cabe duda que agrupar objetos por sus características es una habilidad natural en los seres humanos. Los sustantivos comunes son etiquetas que distinguen objetos en alguna categoría: *ave*, *perro*, *casa*. Existen incontables ejemplos del uso de agrupamientos para el beneficio de la ciencia. En biología, la teoría de clasificación de organismos es conocida como taxonomía, la tabla periódica nos ayuda a entender la estructura del átomo, la clasificación de estrellas en astronomía, etcétera.

Agrupar, en el más amplio sentido, es reunir objetos similares pero en este trabajo nos referimos, más específicamente, a la tarea de utilizar un método de aprendizaje no supervisado (*Duda et al.*, 2001) para reunir textos, sin incluir información lingüística adicional ni utilizar un conjunto de ejemplos de entrenamiento previo.

El conjunto de técnicas que se utiliza para agrupar datos en tales condiciones ha tenido diversos nombres dependiendo del área de estudio y de la época. En nuestros días, *cluster analysis* o *clustering* son probablemente los nombres preferidos. En este trabajo nos referimos con el término **agrupamiento** al conjunto de técnicas.

El problema de agrupamiento (*clustering*) se puede enunciar como sigue: Dado un conjunto de n elementos, generar una asignación de los elementos en k grupos de tal manera que la similitud entre los objetos dentro del mismo grupo sea máxima mientras que la similitud entre grupos distintos sea mínima (Everit, 2001). Por ejemplo, considere que los objetos en cuestión son personas y que el criterio de similitud es la edad. Probablemente asignaríamos a cada persona en uno de los siguientes grupos: *niños*, *jóvenes*, *adultos* o *ancianos*. En general, sea $D=\{d_1, d_2, ..., dn\}$ un conjunto con n elementos. Decimos que $S=\{D_1, D_2, ... , D_k\}$ es un agrupamiento de los elementos en D si para $q=1,2,...,k$ se tiene que $D_q \subseteq D$, pero $D_q \neq \varnothing$. Es decir que todos los grupos son subconjuntos de D pero ninguno de ellos es vacío. Note que si S es la solución óptima del problema de agrupamiento para el conjunto D, es porque existe la certeza de que la similitud entre los objetos dentro del mismo grupo es máxima mientras que la similitud entre grupos distintos es mínima o, de manera equivalente, la distancia intra-cluster es mínima, mientras que la distancia inter-cluster es máxima.

Aunque el problema de agrupamiento es fácil de enunciar y de entender, computacionalmente resulta difícil (no polinomial o NP). En teoría de la complejidad, un problema NP pertenece a la clase de problemas para los que no se ha encontrado una solución eficiente (Cook, 1971). Para introducir la noción de la dificultad del problema de agrupamiento, considere un algoritmo de fuerza bruta para el problema de agrupamiento, el cual necesita evaluar al menos tantas soluciones como particiones tenga el conjunto. Para un conjunto con 100 elementos, el número de particiones, o número de Bell (1934), es $1.6x10^{114}$ (¡una cifra con 115 dígitos!). Esto quiere decir que no hay esperanza en resolver el problema de manera óptima y, típicamente, alguna aproximación de tipo voraz (*greedy*) debe ser utilizada (Cormen *et al.*,1990). Dicho de otra manera, para el problema de agrupamiento no existe un método universal y mucho menos eficiente que pueda aplicarse en todas sus instancias posibles. Esto ha conducido a la creación de diversas aproximaciones y técnicas que intentan resolver el problema de agrupamiento para distintas aplicaciones. Halkidi *et al.* (2001) nos brindan un compendio de las diversas aproximaciones al problema de agrupamiento y una muy buena introducción a los conceptos teóricos en los que están fundamentadas.

En general, el problema de agrupamiento incluye los siguientes componentes:

- **Una representación vectorial de los objetos a agrupar:** Consiste en seleccionar y codificar las características de interés de cada objeto para que puedan ser procesadas por un computadora.

- **La definición de un criterio de distancia o similitud entre los objetos:** Una medida para evaluar el grado de relación entre los objetos con base en un criterio predeterminado y en la codificación de sus características.

- **La aplicación de un método sistemático para la generación de grupos:** La elección de un criterio que permita asignar los objetos iterativamente hasta alcanzar una solución aceptable.

A lo largo del capítulo presentamos las ideas fundamentales de los componentes que integran el algoritmo de agrupamiento semántico de contextos definitorios. Describimos las características de la representación vectorial de textos, introducimos el concepto de energía textual, una técnica en la que está basada nuestra medida de distancia entre definiciones y,

finalmente, mencionamos las propiedades de los algoritmos de agrupamiento jerárquico que resultan los más adecuados para nuestros propósitos.

2.2. REPRESENTACIÓN VECTORIAL DE TEXTOS

Una de las formas más comúnmente utilizadas para la representación de textos digitales en minería de textos y recuperación de información es a través de vectores.

Un vector es un arreglo de números, los cuales definen magnitudes en un espacio que es común a otros vectores. En la figura 2.1 se representan gráficamente los vectores $u = (1,1,2)$ y $v = (0,0,1)$ como puntos en un espacio euclidiano. La *caja* representa el espacio común donde *viven* ambos vectores, en tanto los ejes (las flechas en la figura) representan dimensiones ortogonales en dicho espacio. Note que cada vector tiene una entrada con cierta magnitud por cada una de las dimensiones representadas y que dichas magnitudes se ven reflejadas en la posición final del vector.

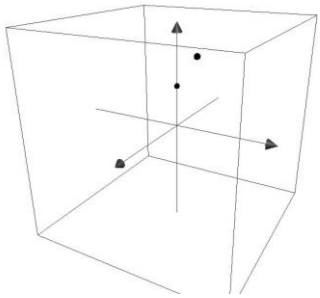

Figura 2.1 Dos vectores representados en un espacio.

Los vectores se utilizan para representar las propiedades de los objetos -reales o abstractos- que nos interesa comparar. Para ello, existe la necesidad de definir un origen en el espacio, la magnitud del vector, la dirección del vector y su sentido. En la figura 2.1 el origen esta determinado por el punto de intersección de los ejes, la magnitud de los vector es la distancia que hay entre el origen y los puntos y tanto la dirección como el sentido están determinados por la magnitud de las entradas de los vectores.

Gerald Salton (1971) propuso un modelo en el cual es posible representar el texto por medio de vectores. En dicho modelo, cada palabra conforma una dimensión en un espacio vectorial común a todos los documentos de una colección. Es decir, cada palabra es una dimensión y cada documento es un vector en el espacio definido por las palabras.

Antes de presentar el modelo, introducimos algunas definiciones necesarias para la comprensión del resto del capítulo.

- Un **documento** es una cadena de longitud arbitraria pero finita de símbolos gráficos denominados términos.

- A lo largo de este capitulo, entendemos como **término** una entidad léxica que puede ser representada mediante un símbolo o la unión de varios de ellos. Por ejemplo, una palabra del español como *manzana* puede representar un término, una frase como *Estados Unidos Mexicanos* también. Asimismo, un término puede ser un símbolo ininteligible como *Viv* o *A4*. Debemos tener cuidado para no confundir este sentido con el del capítulo anterior, donde con término nos referimos a la unidad sobre la cuál se aporta información relevante en una definición. En este capítulo nos referimos siempre a la acepción de entidad léxica.

- Una **colección** es un conjunto de documentos.

- Un **diccionario** es una lista de términos únicos que aparecen en una colección.

Dicho lo anterior, comenzaremos diciendo que en el modelo vectorial se representan formalmente los documentos como vectores y los términos como entradas. Esto es:

$$d_j = (w_{j1}, w_{j2}, ..., w_{jp}).$$

En esta notación, d_j representa el documento j y la entrada w_{ji} en el vector toma algún valor, denominado peso, que representa alguna propiedad del término i en d_j. La forma más extendida de asignar el peso de un término es conocida como *tf-idf* (Spärck, 1972), la cual consiste en una función que asigna un valor a un término de acuerdo con el número de veces que éste aparece en un documento (*term frequency*) y la proporción de dicho término en la colección (*inverse document frequency*). Por otra parte, existe una caracterización más simple, el modelo booleano, en el cual se considera simplemente que los términos están presentes o ausentes en un documento. Como resultado, los pesos de los términos son asumidos todos binarios, esto es, $w_{ji} \in \{ 1, 0 \}$.

Independientemente del esquema de pesos, el modelo vectorial permite representar una colección de *n* documentos como una matriz *M* de *n* x *p*, conocida como matriz documento-término, donde la entrada w_{ji} representa el peso del término *i* en el documento *j*:

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & \cdots & \cdots & w_{np} \end{bmatrix}.$$

El modelo vectorial de Salton fue concebido con el propósito específico de establecer las propiedades teóricas en la tarea de recuperar documentos relevantes de una colección a partir de una expresión de consulta (*query*). Dicha expresión es concebida como un *seudodocumento* que puede ser representado como un vector en el mismo espacio de la colección. La tarea a la que nos referimos, es el objetivo principal del área de recuperación de información (*Information Retrieval*), la cual ha cobrado mucho auge en los últimos años con el uso, cada vez más extendido, de los motores de búsqueda en la Web. Grossman (2004) resulta excelente referencia para profundizar en este tema.

La relevancia del modelo vectorial en esta tesis radica en el hecho de que nos brinda una alternativa para la representación de un conjunto de definiciones. En dicha representación, las definiciones son vectores con tantas entradas como términos distintos en el diccionario, y el valor de las entradas en cada vector toman el valor de 1 o de 0. Considere por ejemplo las definiciones siguientes como una colección:

1. *El gen es la unidad de la herencia.*

2. *El gen es una unidad fundamental.*

Eliminando las palabras de uso común del español (palabras funcionales), un posible diccionario de la colección sería: [*gen*, *unidad*, *herencia*, *fundamental*], y la representación vectorial de las definiciones queda:

1. [1 1 1 0].

2. [1 1 0 1].

Finalmente la matriz documento-término *M* que representa la colección del ejemplo se expresa como:

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

2.3. MEDIDAS DE SIMILITUD ENTRE VECTORES

Recordemos que la solución al problema de agrupamiento para un conjunto dado es una asignación de los objetos en grupos, tal que la similitud entre los objetos dentro del mismo grupo es máxima, mientras que la similitud entre grupos distintos es mínima. Hasta ahora no hemos mencionado cuál es el criterio con el que se determina dicha similitud entre los objetos.

En este contexto, la similitud es una cantidad que expresa el grado de semejanza entre dos objetos (codificados como vectores) de acuerdo con los valores de sus entradas. Ésta se calcula por medio de una función conocida como medida de similitud, la cual toma dos vectores (pertenecientes al mismo espacio) y regresa un valor real comúnmente en el intervalo [0,1], que cuantifica su semejanza. De manera análoga, una medida de distancia es una función que determina la disimilitud entre dos vectores. Existe una relación entre ambas medidas: sean u y v dos vectores, y sean $\delta(u,v)$ y $s(u,v)$ su distancia y su similitud, respectivamente; si ambas funciones están acotadas al intervalo [0,1], entonces, $s(u,v) = 1 - \delta(u,v)$. Cuando la distancia es 1, los objetos son muy distintos, la similitud es 0; cuando la distancia es 0, los objetos son prácticamente iguales y la similitud es 1.

Muchas medidas de distancia se han creado hasta ahora. Una muy conocida, por su utilidad en geometría, es la distancia euclidiana, en la cual se determina qué tan alejados se encuentran dos puntos en el espacio. La distancia euclidiana está relacionada con toda una familia de distancias denominada *familia de Minkowsky* (Cha, 2008) cuya expresión general está dada por:

$$d_m(u,v) = \sqrt[r]{\sum_{k=1}^{p} |u_k - v_k|^r}.$$

Donde p es el número de entradas de los vectores y cuando $r = 2$, obtenemos la distancia euclidiana.

Una medida de similitud muy utilizada en recuperación de información está basada en el ángulo del coseno que definen dos vectores. Ésta se denomina similitud coseno o producto coseno y para los vectores *u* y *v* se define como:

$$s_{\cos}(u,v) = \theta_{uv} = \frac{u \bullet v}{\|u\|\|v\|}.$$

Que no es más que el producto punto de *u* y *v* normalizado; un concepto bien conocido en álgebra vectorial.

El impacto de utilizar una u otra distancia para determinar la similitud entre vectores es evidente en la figura 2.2, en la cual se ilustran los vectores *a*, *b* y *c* compartiendo un mismo espacio vectorial. Según la figura, la distancia euclidiana entre *a* y *b* es menor que entre *a* y *c*; pero, el ángulo (producto coseno) entre *a* y *c* es menor que entre *a* y *b*.

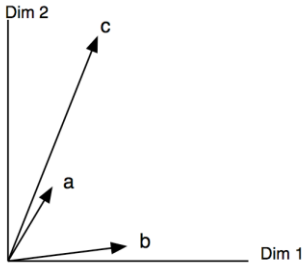

Figura 2.2 Ejemplo del impacto en la elección de una medida de similitud.

El ejemplo anterior nos sirve para introducir la pregunta: ¿Cuál es la medida de similitud o distancia que se debe utilizar? La respuesta es: depende; hasta nuestro conocimiento no hay nada estipulado al respecto. Cada conjunto de datos debe ser analizado por el investigador (el más familiarizado con el problema en cuestión). En esta tesis proponemos una medida de distancia nunca antes utilizada. Dicha medida, está basada en un concepto conocido como la *energía textual*, una aproximación de redes de neuronas artificiales.

2.4. LA ENERGÍA TEXTUAL

Partimos del *modelo de Ising*, uno de los modelos de magnetismo más sencillos en la mecánica estadística (Binder, 2001). En la versión 2D del modelo se consideran espines

atómicos dispuestos en una retícula rectangular en el plano X-Y, como se muestra en la figura 2.3.

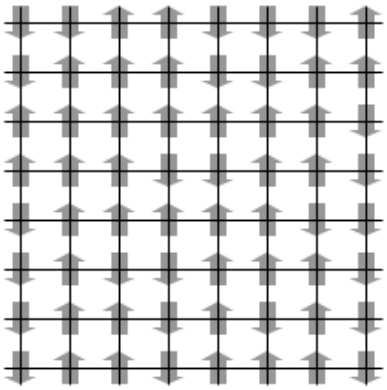

Figura 2.3. Retícula de espines del modelo de Ising.

Inspirado por este modelo, Hopfield (1982) construyó una red neuronal recurrente con capacidad de recuperar patrones a partir de un conjunto de ejemplos dados y la denominó *red de memoria asociativa*. En esta red, las unidades tienen asociados dos posibles valores de activación (0 ó 1). La configuración de los valores de activación en las unidades determina el estado de la red. Cada estado, a su vez, tiene asociado un número conocido como la *energía de la red*.

Uno de los inconvenientes del modelo de Hopfield es que solamente una fracción de patrones puede ser *recordado* correctamente, limitando su uso en aplicaciones prácticas. No obstante, el equipo de investigadores de procesamiento de lenguaje natural del *Laboratoire d'Informatique d'Avignon* (Francia) observó que el comportamiento de la red de memoria asociativa de Hopfield puede ser ampliamente explotado en aplicaciones de procesamiento de lenguaje (Fernandez *et al*., 2007) pues tomando como base el modelo vectorial, se puede interpretar la matriz documento-término como una red neuronal de Hopfield al definir los documentos como cadenas de neuronas, haciendo que la neurona *i* esté activa si el término *i* aparece en la frase o inactiva si el término *i* está ausente.

Consideremos nuevamente la matriz documento-término *X*, donde los valores x_{ji} representan la presencia o ausencia del término *i* en el documento *j*. En una configuración de memoria asociativa, los valores x_{ji} en la matriz equivalen al espín de las unidades en la red:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix}.$$

En este punto surge una consideración importante entre el modelo de Hopfield y la analogía para procesamiento de lenguaje natural: mientras que en las memorias asociativas no se toma en cuenta la interacción de una unidad consigo misma, en el modelo de energía textual esta interacción es importante. Así, la regla de correlación de Hebb (1949) es muy similar a la versión de Hopfield y se expresa por:

$$J = X^T \times X.$$

De donde se obtiene que la energía textual de interacción se calcula según la expresión:

$$E = -\frac{1}{2} X \times J \times X^T.$$

Que en términos de *X* se expresa:

$$E = -\frac{1}{2} X \times (X^T \times X) \times X^T = -\frac{1}{2} (X \times X^T)^2.$$

Fernandez et al. (2007a) centraron su interés en las relaciones entre términos y frases y denominaron dicha interacción: *la energía textual de un documento*, que ha servido, entre otras cosas, para ponderar las frases en un documento y generar resúmenes automáticos, así como para detectar fronteras temáticas a partir de cambios bruscos en las cadenas de texto. Los estudios realizados sobre esta técnica concluyen que se han obtenido buenos resultados con el uso de la energía textual en el sistema de resumen automático (Fernandez *et al.*, 2008), así como en la detección de fronteras temáticas. Los resultados indican que la calidad es independiente del tamaño de los textos, de los temas abordados y de cierta cantidad de ruido semántico inherente a la lengua.

En el capítulo 3, deducimos una medida de distancia entre vectores a partir del modelo de energía textual presentado aquí. Tal medida es utilizada como criterio de semejanza en un algoritmo de agrupamiento de tipo jerárquico.

2.5. ALGORITMOS DE AGRUPAMIENTO

Un algoritmo de agrupamiento es un método de aprendizaje no supervisado (Duda ,2000), que reúne una colección de objetos (documentos) en subgrupos. La clasificación de textos tiene el mismo objetivo, pero a diferencia de los algoritmos de agrupamiento, los métodos de clasificación se consideran métodos de aprendizaje supervisado pues precisan información previa de la colección, así como un conjunto de ejemplos de entrenamiento o el número de categorías a considerar, para su funcionamiento. Existen algoritmos de agrupamiento que no requieren conocimiento del número de grupos *a priori* y que tampoco requieren una etapa de entrenamiento.

Entre los diversos tipos de algoritmos de agrupamiento, dos de ellos resultan los más utilizados:

- **Algoritmos Particionales**: Dividen completamente el conjunto de datos de manera tal que todos los grupos unidos restituyen al conjunto y ningún grupo comparte elementos con algún otro grupo. Los algoritmos más comunes de este tipo son k-means (MacQueen, 1967), k-medoids y CLARA (Clustering Large Applications) (Ng y Han , 1994).

- **Algoritmos Jerárquicos**: Forman (o dividen) grupos sucesivamente hasta generar una estructura de árbol de los datos. Algunos métodos representativos son: el de Zhang Ramakrishnman y Linvy, (1996), y ROCK (Guha *et al.*, 1999).

Existen otros tipos de algoritmos como los basados en modelos de lógica difusa (Dunn,1973), los basados en densidad (Ester *et al.*, 1998) y los mapas de Kohonen (Kohonen, 1995) que son un tipo de red neuronal.

Hasta ahora no se ha demostrado que algún tipo de algoritmo sea sustancialmente superior a otro. En general, es una cuestión de la aplicación que se está considerando y de la naturaleza de los datos. Sin embargo, hay quienes argumentan que los algoritmos jerárquicos generan mejores agrupamientos (Jain y Dubes, 1998).

Lo que sí es evidente es que la idea fundamental de los algoritmos de agrupamiento jerárquico es simple pero elegante. Resulta *natural* y persuasiva para los seres humanos dado que puede ser fácilmente interpretada. Esta facilidad de interpretación radica en el hecho de que los algoritmos jerárquicos proveen una estructura multinivel de conjuntos anidados conocida como dendrograma (*dendro*=árbol) que se va formando en cada paso del algoritmo. La figura 2.4 muestra la representación gráfica de un dendrograma.

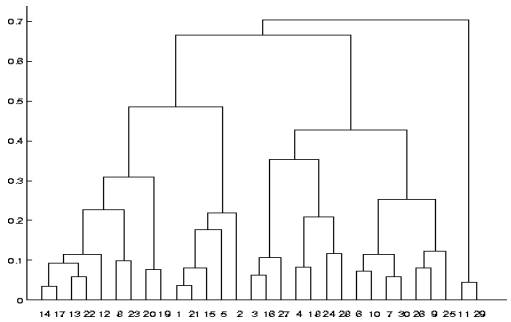

Figura 2.4. Dendrograma.

Existen dos tipos de agrupamiento jerárquico: aglomerativo y divisivo. El algoritmo jerárquico aglomerativo comienza considerando cada objeto como un grupo y en cada iteración fusiona los dos objetos considerando la distancia entre ellos hasta que todos los objetos forman un todo. En el algoritmo de tipo divisivo, se comienza con un solo grupo que unifica a todos los objetos y en cada iteración se divide el grupo de la manera más conveniente. En lo que resta de la sección nos concentraremos en los algoritmos jerárquicos aglomerativos por ser la aproximación que hemos adoptado. Jain *et al.*, (1999) resultan una excelente referencia en el tema de algoritmos jerárquicos.

En el algoritmo jerárquico aglomerativo simple (HAC, por sus siglas en inglés), la entrada es un conjunto de objetos y la salida es un dendrograma. En cada iteración, una función calcula la distancia entre cada par de grupos y con esto se determinan los siguientes grupos a ser combinados para formar un grupo nuevo. El criterio para calcular la distancia entre cada grupo es alguna variante de una familia de métodos denominados *linkage*, ilustrados en la figura 2.5 y explicados a continuación.

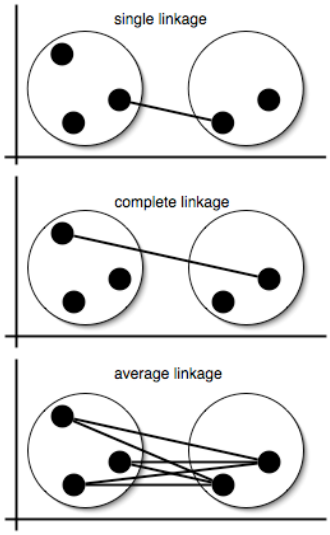

Figura 2.5. Métodos *linkage*.

En el método *single linkage* (Florek *et al.*, 1951), también conocido como *vecino más próximo* o *método de distancia mínima*, la similitud de dos grupos es la similitud de sus miembros más próximos. En este sentido, es un método local, pues solo se toma en cuenta el área donde los dos grupos tienen proximidad máxima. Las partes más distantes de ambos grupos o la estructura de los mismos no se toma en cuenta. Esto causa una tendencia a producir grupos *grandes* y con elementos muy distribuidos pues, como sólo se considera información local, se pueden llegar a fusionar en un solo grupo lo que en realidad podrían ser dos o más. En el método de *vecino más proximo*, si D_i y D_j son dos grupos tales que $d_i \in D_i$ y d_j D_j, la distancia entre D_i y D_j se calcula por medio de la fórmula:

$$Dist(D_i, D_j) = \min_{d_i \in D_i, d_j \in D_j} dist(d_i, d_j).$$

Nótese que *dist* es una función con dominio en el espacio de los objetos (una medida de distancia), a diferencia de *Dist* que tiene dominio en el espacio de los grupos.

En el método del vecino más lejano (*complete linkage*) sugerido por Sorensen (1948), conocido también como *método de distancia máxima*, la distancia entre grupos se representa por la distancia más larga entre un objeto del primer grupo y un objeto del segundo grupo. El método de distancia máxima tiene la ventaja de generar grupos pequeños, cohesivos y bien delimitados y por ello hemos optado utilizarlo en nuestro algoritmo. En el método de *vecino mas lejano*, la distancia entre dos grupos, D_i y D_j, tales que $d_i \in D_i$ y d_j D_j, se define como:

$$Dist(D_i, D_j) = \max_{d_i \in D_i, d_j \in D_j} dist(d_i, d_j).$$

Al igual que en el método *single linkage*, *dist* es una función con dominio en el espacio de los objetos y *Dist* tiene dominio en el espacio de los grupos.

En el método *Average linkage* (Sokal y Michener, 1958) se calcula la distancia entre dos grupos como el promedio de las distancias entre todos los objetos de un grupo y todos los objetos del otro grupo, según la siguiente fórmula:

$$Dist(D_i, D_j) = \frac{\sum_{d_i \in D_i} \sum_{d_j \in D_j} dist(d_i, d_j)}{|D_i| \times |D_j|}.$$

Independientemente del método *linkage* elegido, el agrupamiento jerárquico converge de manera determinista a la integración de un grupo absoluto. Por tanto, la manera de obtener una partición del conjunto de entrada es realizando algún tipo de *corte* en algún punto. Dos criterios de corte comúnmente utilizados son:

- **Corte por umbral de similitud**: Consiste en recuperar como partición del conjunto inicial todos los grupos formados al *podar* el árbol jerárquico (dendrograma) en alguna altura predeterminada.

- **Corte por distancia**: Consiste en considerar solamente los grupos cuya distancia de unión se encuentre por debajo de un umbral predeterminado.

En el capítulo siguiente mostraremos la manera en que utilizamos un algoritmo jerárquico aglomerativo para reunir una colección de CDs en español representados en forma vectorial. En dicho algoritmo, la constitución de los grupos es determinada por medio de un valor de corte por distancia.

3. ALGORITMO DE AGRUPAMIENTO SEMÁNTICO DE CDS

3.1. OBTENCIÓN DE LOS CDS

Recordemos que el sistema *ecode* (Alarcón, 2009) es capaz de extraer una colección de contextos definitorios asociados a un término y a un tipo de definición. De manera adicional, el *ecode* realiza una clasificación de los contextos de acuerdo con el patrón verbal asociado a su tipo: los de tipo analítico, los de tipo extensional y los de tipo funcional.

Esto plantea la posibilidad de *entregar* un archivo de salida del *ecode* a un módulo capaz de formar grupos de definiciones, de acuerdo con la acepción a la que éstas aluden. Por lo tanto, partimos de que la entrada del algoritmo de agrupamiento es una colección de definiciones asociadas a un mismo término y clasificadas en el mismo tipo de definición. De manera adicional, pretendemos simular las condiciones de uso real que tendría la implementación del algoritmo como sistema. Así, consideramos que el primer contacto del texto con el algoritmo debe ser a través de un archivo en texto plano, sin etiquetas, ni ningún tipo de marcaje como XML o HTML, en español, codificado en algún conjunto de caracteres conocido como UTF-8 y con una unidad de texto independiente (una definición) por cada línea del archivo. La figura 3.1 representa una archivo con tales características.

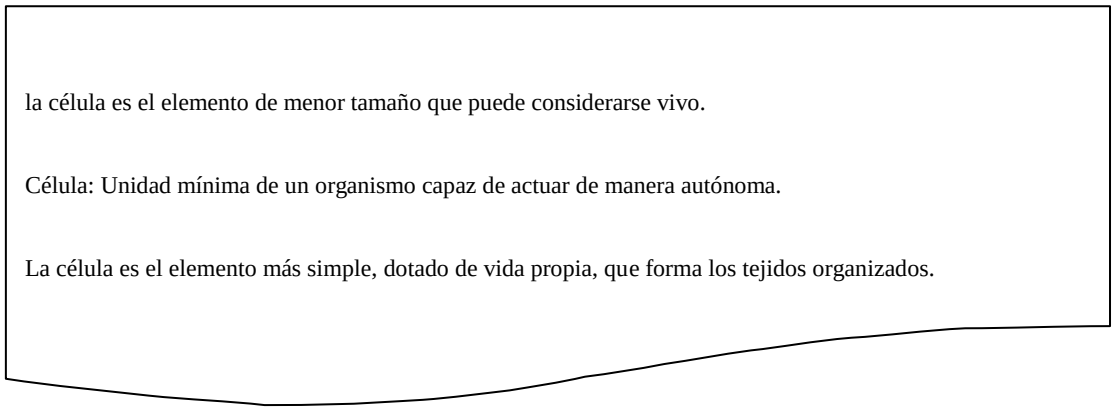

Figura 3.1. Ejemplo de un archivo de colección de definiciones.

Considere que los textos serán extraídos de la Web, y que resulta imposible predecir de cuántas formas encontraremos representado un mismo término, ya que incluso estos pueden presentar faltas de ortografía. Observe en la figura 3.1 que algunas palabras comienzan por mayúsculas, algunas tienen algún signo como coma o punto justo delante de la última letra, y que algunas otras contiene vocales acentuadas. Para una computadora los

símbolos "*Célula"*, "*celula"* y "*célula*," son todos distintos. Para nosotros los humanos es evidente que los tres hacen referencia al mismo término. Para mitigar esta complicación conviene tener un mecanismo que unifique la diversidad de símbolos gráficos.

Cabe mencionar que la procedencia de los contextos definitorios (la fuente original) no tiene relevancia para nuestros propósitos, pero dado que uno de los objetivos es implementar el módulo aquí descrito y ponerlo en producción como aplicación Web, consideramos, según los resultados de diversas pruebas preliminares, que el tamaño de un archivo promedio debe estar entre 12 y 24 KB, es decir, entre 50 y 150 definiciones por archivo.

La siguiente sección describe el tratamiento que debe aplicarse a los archivos para facilitar la generación de un espacio vectorial de la colección que sea representativo pero óptimo para el agrupamiento.

3.2. PREPROCESAMIENTO

La primera etapa del algoritmo de agrupamiento consiste en el preprocesamiento, el cual a su vez esta conformado por tres transformaciones al archivo de entrada: la primera transformación elimina los signos de puntuación y diacríticos, la segunda elimina las palabras con *poca información semántica* y la tercera trunca las palabras para convertirlas en un prefijo.

El objetivo principal del preprocesamiento es reducir el tamaño de la matriz que será generada en la etapa siguiente, aunque cabe mencionar que realizar truncamiento tiene también beneficios en la reducción de dimensiones del espacio vectorial y que la eliminación de signos de puntuación y diacríticos hace más fácil la manipulación de los archivos.

Consideremos nuevamente la figura 3.1 como el texto de entrada al módulo de agrupamiento. Con la intención de reducir la diversidad de símbolos, la primera transformación consiste en un filtro que unifica, por ejemplo, los símbolos 'Ú', 'ú', 'Ü' y 'ü' bajo el común 'u'. Lo mismo hacemos para todas las marcas diacríticas y signos de puntuación. El resultado es un archivo libre de caracteres no-ASCII y mucho más fácil de procesar por la computadora. La figura 3.2 ilustra el resultado de uniformar los textos de la figura 3.1.

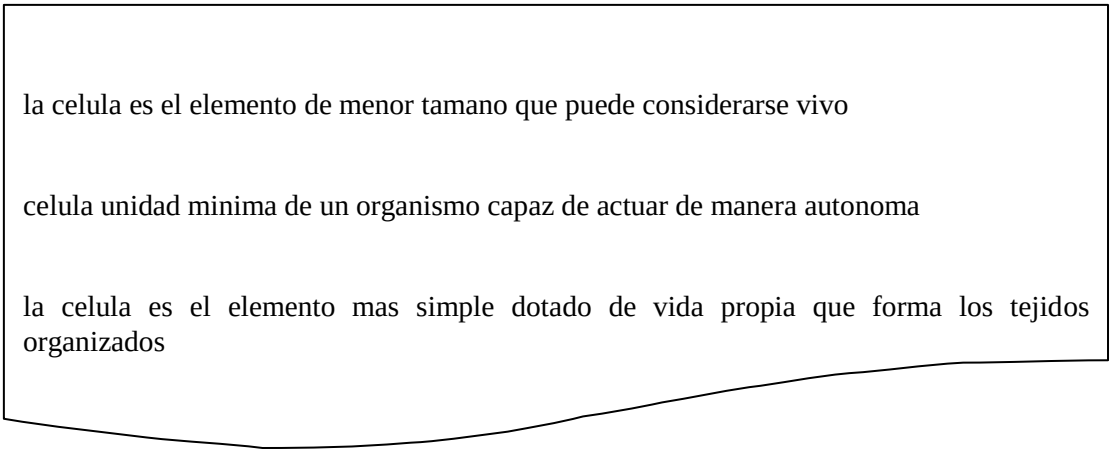

Figura 3.2. Aplicación de filtros para eliminar diacríticos y puntuación.

Note en la figura 3.2 que además, todo el texto queda en minúsculas, los signos de puntuación como: puntos, comas y dos puntos; fueron eliminados, la falta de vocales acentuadas y la sustitución de la letra ‘ñ’ por la letra ‘n’.

En una segunda transformación, los textos son filtrados mediante una *lista de* paro (Manning y Schütze, 1999), que no es más que un archivo que le indica a un programa qué palabras deben ser eliminadas de los textos. El filtrado mediante la lista de paro reduce en gran medida el tamaño del diccionario generado por la colección y representa una práctica muy común en procesamiento de lenguaje natural. La figura 3.3 ilustra parte del contenido de la lista de paro que se aplicó a las definiciones analíticas del término *célula*. Las palabras carecen de acentuación por que la lista de paro es procesada con el mismo filtro para poder empatar los símbolos.

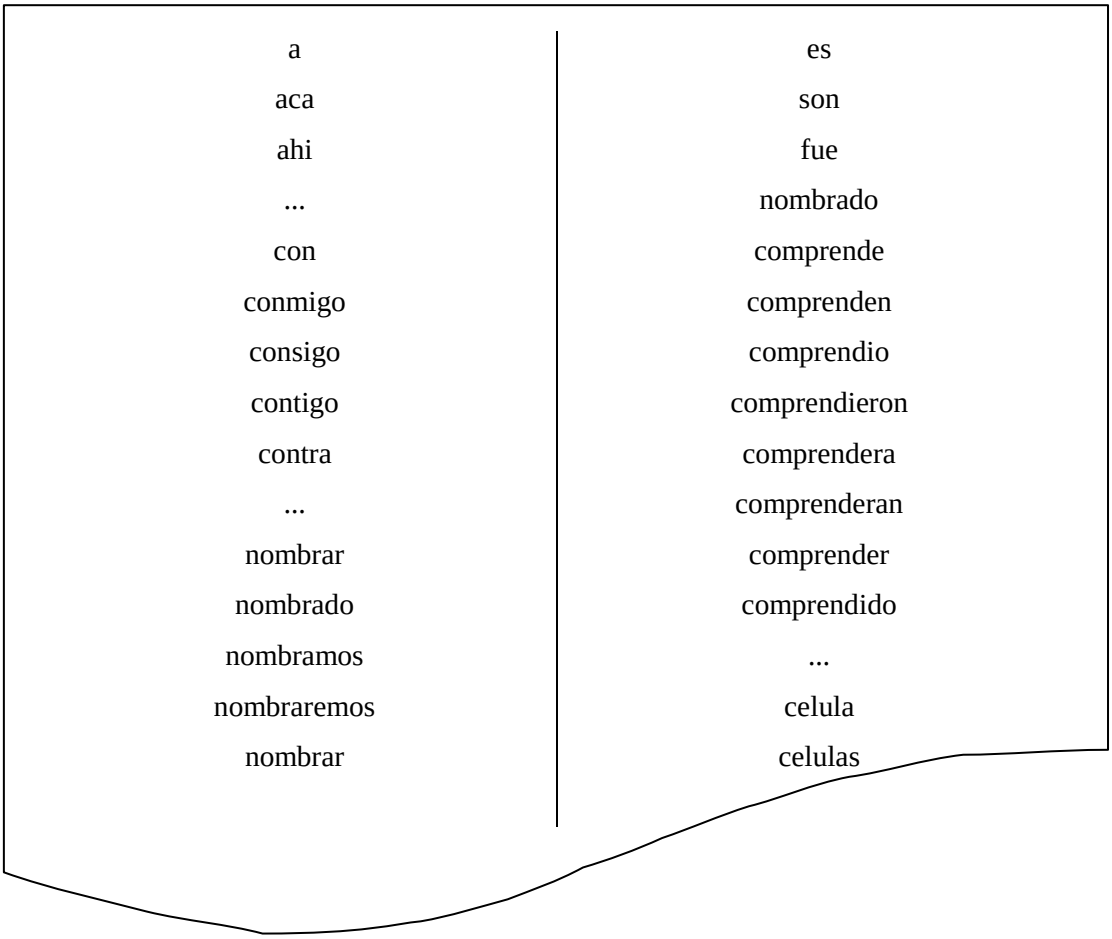

Figura 3.3. Lista de paro para las definiciones analíticas del término *célula*.

Note que las palabras eliminadas son preposiciones, artículos, pronombres y palabras comunes del español. También son eliminados todos los verbos identificados como *definitorios*. Es decir que eliminamos las ocurrencias de los patrones que sirven para definir. Por ejemplo, en el caso de las definiciones analíticas, eliminamos los patrones *es un, se define como, se puede nombrar, se ha comprendido, se puede definir como*, etcétera. También eliminamos el término en su forma singular y plural.

Para nuestro método, tanto los patrones definitorios como el término, no representan elementos discriminatorios dado que aparecen, prácticamente, en todas las definiciones de la colección. De esta manera, damos oportunidad a otros elementos menos comunes de relacionarse de manera más directa. Por ejemplo, en la figura 3.3 observamos las secuencias: *la celula es el elemento de menor tamano* y *la celula es el elemento mas simple*. Después de eliminar palabras de estas secuencias quedan como: *elemento menor tamano* y *elemento*

simple. De esta manera, las secuencias: *menor tamano* y *simple* están más directamente relacionadas con *elemento* por el hecho de estar más próximas.

La figura 3.4 ejemplifica el resultado de la eliminación de las palabras en la lista de paro.

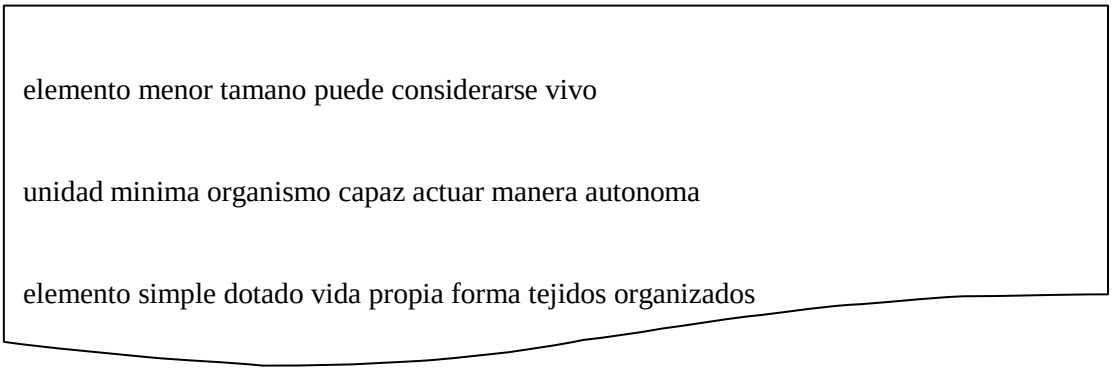

Figura 3.4. Resultado de eliminar las palabras en la lista de paro.

Note que el texto conserva los elementos semánticos más importantes de cada definición e incluso es legible aunque dejó de ser sintácticamente correcto. Paradójicamente, esta transformación lo hace más *amigable* con el procesamiento por computadora.

La última transformación consiste en truncar las palabras mediante el algoritmo de Porter (1980). La intención de esta trasformación es unificar en un solo símbolo aquellas palabras que poseen la misma raíz, y que probablemente están relacionadas semánticamente y su significado varía según los afijos o las flexiones que presentan. Por ejemplo, las palabras *vivo* y *viviente* se unifican en el símbolo *viv* que es la raíz.

Siguiendo con nuestro ejemplo, la figura 3.5 muestra el resultado de aplicar el algoritmo de porter al texto de la figura 3.4.

La aplicación de estas sencillas técnicas reduce enormemente el tamaño del diccionario generado por la colección. La tabla 3.1 muestra la reducción en el tamaño del diccionario aplicando el preprocesamiento a nuestro corpus de prueba.

element menor taman pued consider viv

unid minim organ capaz actu maner autonom

element simpl dot vid propi form tej organiz

Figura 3.5. Resultado de aplicar el algoritmo de Porter.

La reducción de entradas en el diccionario es importante sobre todo por los requerimientos de espacio que representa la generación de la matriz documento-término. En la siguiente sección mostramos la manera en la que se construye dicha matriz, cuyas dimensiones dependen tanto del número de definiciones en la colección como del número de entradas en el diccionario generado por la colección. Posteriormente, al calcular la energía textual asociada a la matriz será evidente la ventaja que representa la reducción del diccionario en tiempo de procesamiento.

		Palabras en el diccionario sin preprocesamiento	**Palabras en el diccionario con preprocesamiento**	**reducción**
Barra	**Analíticas**	2037	579	71%
	Extensionales	1901	604	68%
	Funcionales	1845	567	69%
Célula	**Analíticas**	1440	306	78%
	Extensionales	2717	618	77%
	Funcionales	1319	394	70%
Punto	**Analíticas**	1549	391	74%
	Extensionales	1969	552	72%
	Funcionales	2453	571	76%
Ventana	**Analíticas**	3174	633	80%
	Extensionales	996	251	75%
	Funcionales	1234	343	72%

Tabla 3.1. Reducción en el tamaño del diccionario.

3.3. CONSTRUCCIÓN DEL ESPACIO VECTORIAL

En esta etapa construimos el espacio vectorial generado por las definiciones, es decir, una matriz concebida como un arreglo de vectores que representan documentos. El número de renglones de dicha matriz está determinado por el número de definiciones en la colección; el número de columnas, por el tamaño del diccionario de palabras gráficas (símbolos distintos) de la colección obtenido después del preprocesamiento. La generación de la matriz documento-término se lleva a cabo inicializando una matriz de ceros de *n* x *p* elementos, donde *n* es el número de definiciones y *p* el número de palabras gráficas en el diccionario y posteriormente, el diccionario es recorrido palabra por palabra y un valor de 1 es insertado en la posición correspondiente a la entrada de dicha palabra para todos los documentos que la contienen.

El algoritmo 3.1 muestra el seudocódigo del método *getBoolean* de una clase simbólica que crea la matriz documento-término. Un objeto de la clase *reader* tiene acceso a las propiedades del diccionario de palabras gráficas y a los documentos de la colección.

Método getBoolean

Entrada: lector de la colección con acceso a los documentos y al diccionario
Reader reader

Salida: Matriz documento-termino
Matrix m

```
                    /* apunta al primer termino */

        TermEnum termEnum ← reader.terms()
        TermDocs termDocs  ← reader.termDocs()

                    /* Construye matriz de ceros */

        m ← new Matrix(numDocs,numTerms)
        while ( termEnum.next() )
            Term term ← termEnum.term()
            termDocs.seek(term)

            while ( termDocs.next() )
                int docId ← termDocs.doc()
                int freq  ← termDocs.freq()

                if ( freq > 0 ) then
                    m.set(docId, dic.indexOf(
termEnum.term().text() ) , 1)
                end if

            end while
        end while
```

Algoritmo 3.1. Creación de la matriz del espacio vectorial.

Aunque la matriz generada por el método *getBoolean* es binaria, nada nos impide utilizar un esquema de pesos más sofisticado y llenarla con valores reales. Por ejemplo, podríamos asignar a las entradas la frecuencia de los términos en cada definición. Sin embargo, en el caso de los contextos definitorios, nos enfrentamos con el problema de frecuencias muy bajas en el vocabulario, es decir, las definiciones solamente tendrán una o dos apariciones de cada palabra. Como hemos mencionado antes, existen pocos trabajos que presenten propuestas para asignar pesos en estas condiciones.

La sección siguiente muestra cómo calcular la distancia entre textos a partir de una matriz binaria usando una fórmula basada en la *energía textual* descrita en el capítulo anterior.

3.4. CÁLCULO DE LA ENERGÍA

Una vez que hemos generado una matriz binaria surge la cuestión de comparar dos definiciones a partir de su representación vectorial. Como hemos dicho antes, para hacer agrupamiento necesitamos de algún mecanismo de comparación entre textos que funcione como criterio para determinar cómo están conformados los grupos semánticos, es decir, cuáles textos son similares entre sí y cuáles son diferentes. Hemos elegido derivar una medida de distancia a partir de la matriz de *energía textual* propuesta por Fernandez *et al.* (2008). Esta técnica nos ofrece dos grandes ventajas: funciona para matrices con frecuencias unitarias (presencia-ausencia de términos) y fue concebida desde sus inicios como una aproximación teórica para ponderar las relaciones de significado en textos.

A continuación deducimos la medida de distancia del algoritmo de agrupamiento semántico a partir de la fórmula general de *energía textual*.

Consideremos la matriz documento-término *X*:

$$X = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & \dots & \dots & x_{np} \end{vmatrix}.$$

La matriz de *energía textual*, *E* asociada a *X* se calcula por :

$$E_{textual} = -\frac{1}{2} X \times (X^T \times X) \times X^T = -\frac{1}{2} (X \times X^T)^2$$

Sin pérdida de generalidad, podemos considerar solamente las magnitudes de las entradas de la matriz de energía, pues sabemos que todas son negativas. Esto es:

$$E = -E_{textual}$$

Donde *E* es la matriz de *n* x *n* dada por:

$$E = \begin{vmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ e_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ e_{n1} & \dots & \dots & e_{nn} \end{vmatrix}.$$

En general, puede ocurrir que la máxima energía (la mayor magnitud) no se encuentre en los elementos de la diagonal. Pero, por simplicidad vamos a considerar que sí es así y los vamos a dejar fuera de los cálculos posteriores, asumiendo que un documento tiene la máxima energía consigo mismo y que no hay necesidad de conservar dichos valores para futuras comparaciones.

Aunado a lo anterior, tenemos la propiedad de que los valores de energía son simétricos en las matrices triangulares asociadas. Es decir, $e_{ij} = e_{ji}$.

Esto nos permite vaciar los valores de *E* en un arreglo unidimensional de *distancia energética*.

$$D_{ener} = [e_{12}, e_{13}, e_{14}, ..., e_{1n}, e_{23}, e_{24}, ..., e_{2n}, ..., e_{n-1n}]$$

Sólo resta restringir los valores de las entradas de *Dener* en el rango $[0,1]$. Para esto restamos el valor máximo del vector D_{ener} a todas sus entradas y dividimos por dicho valor. Esto es :

$$DistEner = \frac{máx(D_{ener}) - D_{ener}}{máx(D_{ener})}$$

De esta forma, hemos calculado la distancia entre documentos a partir de la matriz de energía textual. Tenemos ahora la posibilidad de utilizar un algoritmo de agrupamiento para generar una estructura de grupos utilizando esta distancia como criterio.

3.5. AGRUPAMIENTO DE DEFINICIONES

Una vez que la proximidad entre textos es calculada por medio de la distancia energética, generamos los grupos usando un algoritmo jerárquico aglomerativo simple. Un algoritmo de

tipo jerárquico nos ofrece dos grandes ventajas: la primera de ellas es que no requiere que el número de grupos sea especificado previamente, la segunda y tal vez la más importante, es que tenemos la intuición de que enfatiza relaciones encontradas empíricamente entre las definiciones. Un ejemplo de este tipo de relación es el siguiente. Considere las definiciones:

1. *El **gen** es la unidad de la herencia.*
2. *El **gen** es la unidad física, funcional y fundamental de la herencia.*

Se ve claramente como la segunda definición corresponde a una versión *más completa* que la primera. Es de esperar, que en un algoritmo de tipo jerárquico, el grupo al que pertenece la primera definición sea considerado como un subgrupo de aquel que contiene la segunda.

El método utilizado para comparar los grupos en el algoritmo jerárquico es el método de vecino más lejano (*complete linkage*). Preferimos este método porque genera grupos pequeños, cohesivos y bien delimitados, brindándonos la posibilidad de mejorar la precisión de los grupos. Asumimos que esto puede tener repercusiones en el recuerdo (*recall*), dado que se corre el riesgo de generar grupos de un solo elemento que finalmente serán ignorados en la etapa de presentación de los resultados. Aceptamos ese riesgo, pues nos parece que si los datos agrupados muestran buena cohesión y reflejan mejor las relaciones semánticas, las pérdidas en cantidad son aceptables.

El criterio para determinar el número de grupos generados es un valor umbral de *corte por distancia*. Con dicho valor, indicamos el valor máximo en distancia que puede haber entre dos grupos. Por ejemplo, si determinamos que el valor umbral de corte por distancia es 0.1 significa que aquellos grupos cuya distancia es mayor a 0.1 no son unificados. La figura 3.6 representa esquemáticamente esta situación.

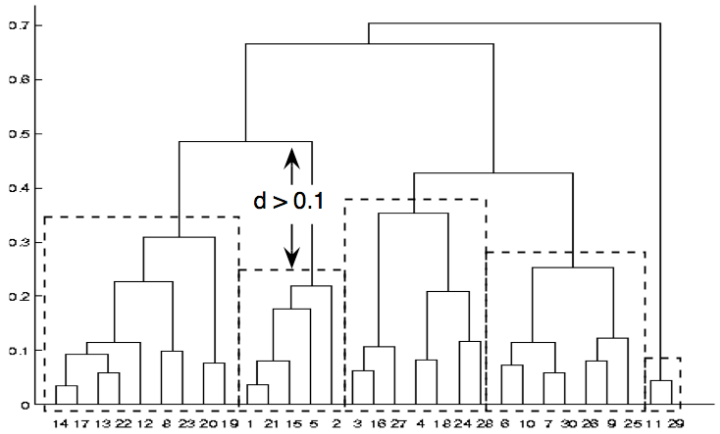

Figura 3.6. Dendrograma con corte de grupos por distancia.

Además, el algoritmo de agrupamiento jerárquico genera un dendrograma. Lo cual tiene dos ventajas: por un lado, nos permite calcular algunos coeficientes de comparación entre agrupamientos, como veremos más adelante; y por el otro, representamos gráficamente los resultados obtenidos en cada ejecución del módulo. De este modo podemos hacer inferencias interesantes sobre los resultados del algoritmo con una simple inspección visual del dendrograma.

3.6. PRESENTACIÓN GENERAL DEL ALGORITMO

Hasta este punto, hemos visto como convertir una colección de textos en una matriz binaria y hemos sugerido la aplicación de algunos métodos numéricos para comparar vectores entre si y para generar una estructura de grupos mediante un algoritmo jerárquico.

Presentamos, en el algoritmo 3.2, el algoritmo de agrupamiento semántico de contextos definitorios.

El algoritmo puede ser concebido en tres grandes etapas:

1. En la primera, el texto es procesado hasta llegar a la representación vectorial usando diversas técnicas de procesamiento de lenguaje natural.
2. En la segunda etapa, se calcula la distancia entre cada vector utilizando la matriz de energía textual.
3. La última etapa consiste en aplicar agrupamiento jerárquico con el método de *vecino más lejano*.

El algoritmo genera una partición del conjunto de definiciones *D*, es decir, que todos los subconjuntos D_i son disjuntos y *D* es la unión de todos los subconjuntos D_i para i=1,2,...,k.

Algoritmo de Agrupamiento semántico de contextos definitorios

Entrada: conjunto de definiciones
$D=\{d_1, d_2, \ldots, d_n\}$

Salida: Partición del conjunto D
$S=\{D_1, D_2, \ldots, D_k\}$

```
                    /*  Prepara los datos  */

D' ← PreProcesa(D)
B  ← GeneraEspacioVectorial(D')
E  ← CalculaEnergiaTextual(B)

for i=1 to n do
      D_i ← {d_i}
end for

k ← n
S ← { D_1,...,D_n}
                    /*  Genera los grupos  */
repeat
      Dist ← DistanciaEntreGrupos(S,E)
      d ← ∞

      for i=1 to k-1 do
          for j=i+1 to k do
              if Dist(i,j) < d then
                  d ← Dist(i,j)
                  u ← i
                  v ← j
              end if
          end for
      end for

      k ← k-1
      D_nuevo ← D_u ∪ D_v
      S ← S ∪ D_nuevo - D_u - D_v

until k=1;
```

Algoritmo 3.2. El algoritmo de agrupamiento semántico de CDs.

El algoritmo realiza de manera general los pasos siguientes:

1. Se procesa el archivo: con la finalidad de reducir el tamaño del espacio vectorial, se aplica eliminación de palabras y truncamiento.
2. Se construye el espacio vectorial: considerando la presencia o ausencia de las palabras gráficas se genera una matriz binaria documento-término.
3. Se calcula la energía textual: se derivan las distancias distancias entre los vectores a partir de la matriz de energía textual.
4. Se aplica HAC: Se genera la estructura de grupos aplicando un algoritmo jerárquico aglomerativo utilizando el método de vecino más lejano y la distancia derivada de la energía textual.
5. Se muestran los resultados: De acuerdo con un criterio preestablecido mostramos los grupos con mayor oportunidad de reflejar relaciones semánticas.

La figura 3.7 presenta de manera esquemática el algoritmo de agrupamiento. En la figura queda ilustrado el efecto de cada etapa en el algoritmo.

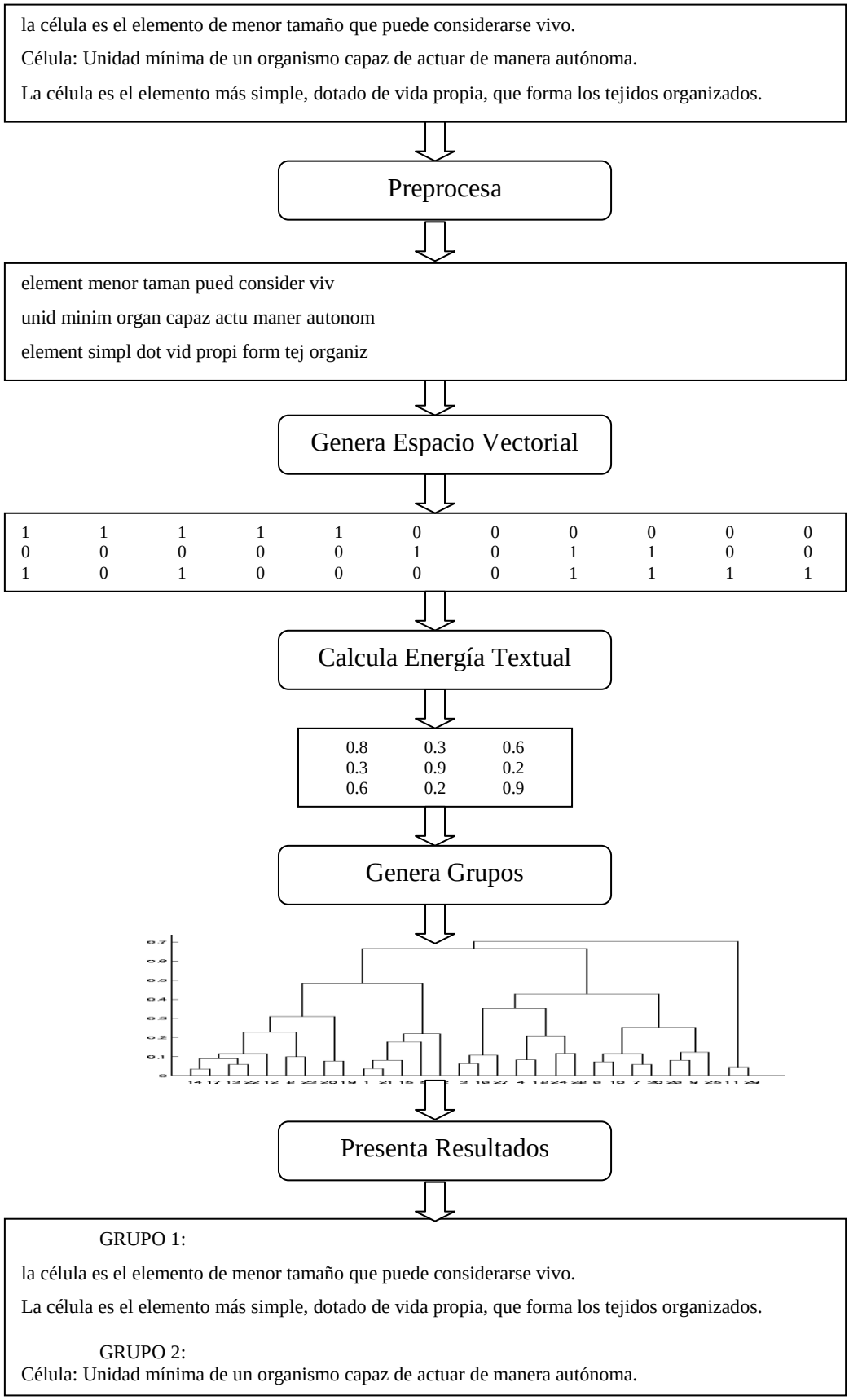

Figura 3.7. Etapas del algoritmo de agrupamiento semántico de CDs

4. APLICACIÓN DEL ALGORITMO AL CTPE

4.1. EL CORPUS DE TÉRMINOS POLISÉMICOS EN ESPAÑOL (CTPE)

La necesidad de crear nuestro propio corpus de experimentación fue inminente. Los CDs son una estructura discursiva recientemente estudiada y todavía son escasos los corpus disponibles para su uso. Además, las características necesarias para la experimentación en este trabajo son tan particulares que resultó más que conveniente la creación de nuestros propios datos. Las características deseables en nuestros corpus se centran, por un lado, en la cantidad de acepciones que un término puede tener según su contexto (polisemia) y, por otro, en la probabilidad de extraer suficientes observaciones de la Web. Lo primero resulta evidente si consideramos que entre más significados puedan identificarse de un término, más claramente se revelarán los grupos semánticos. Lo segundo se refiere a que necesitamos contar con datos suficientes para que nuestros resultados sean representativos. Por estas razones, resultó conveniente seleccionar cuidadosamente los términos a incluir en nuestro corpus: el Corpus de Términos Polisémicos en Español (CTPE).

Con esta idea, elegimos primeramente diez términos que cumplen con dos características: son ambiguos o tan generales que su significado está condicionado por el contexto y es muy probable encontrarlos en diversos contextos de uso común. Basándonos además en las consideraciones reportadas por Alameda (1995), los términos inicialmente elegidos fueron: *aguja*, *barra*, *cabeza*, *casco*, *célula*, *golpe*, *punto*, *serie*, *tabla* y *ventana*. Después de elegidos los términos, surge la cuestión de cómo encontrar en la Web contextos definitorios asociados a estos términos.

Decidimos combinar una lista de los términos elegidos con una lista de patrones verbales definitorios para formar cierto tipo de expresión que denominaremos patrón definitorio. Un ejemplo de patrón definitorio para el término *aguja* y con el patrón verbal definitorio *Ser + Determinante* es: *la aguja* ***es un***.

La figura 4.1 ilustra el contenido de un archivo de patrones definitorios con el símbolo <T> representando un término genérico.

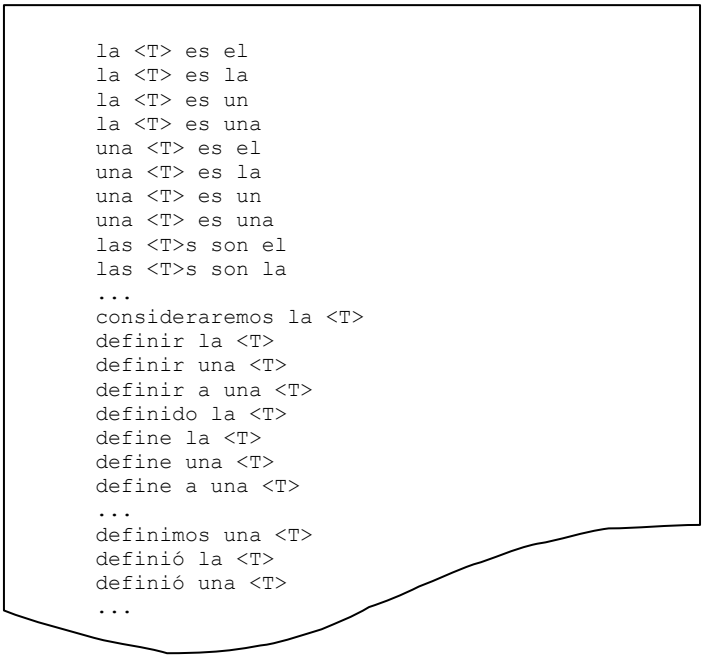

Figura 4.1. Archivo de patrones definitorios.

Lo anterior pone de manifiesto la posibilidad de utilizar dichos patrones definitorios como expresiones de consulta (*query*) en motores de búsqueda de la Web. La API BOSS de Yahoo![1] nos permitió realizar esta minería. Gracias a esta plataforma, logramos extraer las referencias de fragmentos textuales que contienen patrones definitorios en la red. Sin embargo, cabe mencionar que la precisión de nuestra minería se vio ceñida por las limitaciones heredadas del sistema de recuperación. Yahoo! no sabe que estamos intentando localizar contextos definitorios y, por ende, nos responde con un sinnúmero de documentos que contienen un patrón definitorio pero que no son definiciones. Recordemos que un Contexto Definitorio es un fragmento de texto donde se define un término mediante el uso de un patrón verbal definitorio. Por ejemplo, en el fragmento:

> *Una **aguja es un** filamento de metal u otro material duro, de tamaño relativamente pequeño, generalmente recto, afilado en un extremo y con el otro acabado en un ojo o asa para insertar un hilo*

Podemos observar que el patrón *Ser+Determinante*, conecta el término *aguja* con su definición. Sin embargo, en la práctica, es posible – y muy común – encontrar fragmentos

[1] Yahoo! Search BOSS (Build your Own Search Service). http://developer.yahoo.com/search/boss/

textuales donde aparece un término y solo un contexto de uso que no es, propiamente, una definición. Por ejemplo, en el fragmento:

En general, el miedo a la aguja es el más frecuente.

Observamos que, como en el caso anterior, aparece el término aguja y el patrón *ser+Determinante*, pero evidentemente no se trata de una definición. Decimos que un fragmento de texto es un Candidato a Contexto Definitorio (CCD) si contiene un término y alguna instancia de algún patrón verbal definitorio pero no necesariamente es un contexto definitorio.

Una vez reunida la información expedida por Yahoo!, decidimos reducir el número de términos en nuestro estudio a tan sólo cuatro. La razón principal radica en la gran cantidad de información recuperada (cerca de 3700 resultados por término, en promedio), pues uno de los criterios de evaluación del presente estudio implica la lectura de la información por un humano. Consideramos que basta con conservar algunos de ellos para aplicar el algoritmo de agrupamiento y analizar los resultados. Los términos finalmente seleccionados para conformar el corpus fueron: *barra*, *célula*, *punto* y *ventana*. La tabla 4.1 muestra la cantidad de candidatos a contextos definitorios extraídos de la Web para cada término y tipo de definición.

	Analíticos	**Extensionales**	**Funcionales**	
Barra	1863	307	467	**2637**
Célula	5352	649	533	**6534**
Punto	1702	422	750	**2874**
Ventana	1534	587	565	**2686**
	10451	**1965**	**2315**	**14731**

Tabla 4.1. Candidatos a Contextos Definitorios extraídos de la Web.

Después de separar manualmente los CDs de los CCDs, observamos una proporción importante de ruido en la información. La tabla 4.2 muestra la cantidad de CDs obtenidos para el corpus de términos polisémicos en español.

Las secciones siguientes describen los criterios de evaluación del algoritmo, así como los resultados de aplicar el algoritmo de agrupamiento semántico de contextos definitorios al CTPE.

		CCDs	CDs	Ruido
Barra	**Analíticas**	1863	148	92%
	Extensionales	307	117	62%
	Funcionales	467	103	78%
Célula	**Analíticas**	5352	92	98%
	Extensionales	649	135	79%
	Funcionales	533	66	87%
Punto	**Analíticas**	1702	101	94%
	Extensionales	422	96	77%
	Funcionales	750	152	79%
Ventana	**Analíticas**	1534	212	86%
	Extensionales	587	107	81%
	Funcionales	565	93	83%

Tabla 4.2. Precisión de la extracción.

4.2. ACERCA DEL ANÁLISIS CUALITATIVO

El análisis cualitativo del algoritmo consistió en la lectura y la interpretación de los grupos generados. El algoritmo fue ejecutado para todas las colecciones del CTPE, es decir, para cada archivo que asocia el par: (término, tipo de definición). Se implementó un programa que varía el parámetro de distancia de corte por distancia desde 0.10 hasta el valor de 1.00, con una diferencia de 0.01 entre cada ejecución. Sólo fueron reportados los grupos que reúnen al menos dos definiciones, dejando fuera de la presentación de resultados los grupos con tan solo una definición: esto quiere decir que, si una definición no *alcanza* a integrarse en un grupo, queda fuera de la presentación final de los datos. Es por ello que, si el parámetro de corte cambia su valor, puede darse la ilusión de que *aparecen* definiciones. Con el fin de que el lector observe todas las definiciones originalmente incluidas en los experimento, incluimos al final de los archivos de resultados el grupo cuyo valor de parámetro de corte es 1.00.

Los archivos de resultados tienen el siguiente formato:

```
---------- corte = x -------------------------
GRUPO n:
def1
def2
...
defi

GRUPO m:
...
----- i definiciones recuperadas en g grupos ----
---------- corte = y -------------------------
...
----- j definiciones recuperadas en h grupos ----
```

Los separadores indican una ejecución completa del algoritmo, tomando *x/y* como valores de corte por distancia. El resultado de la ejecución con dicho valor de corte son los *g/h* grupos mostrados entre ambos separadores, donde *g/h* es el número de grupos generados por el algoritmo e *i/j* representan el número total de definiciones mostradas.

El formato de los grupos por su parte es el siguiente: cada grupo se identifica anteponiendo la palabra GRUPO seguida de un identificador (un valor entero) que indica el número de la iteración en la que los objetos que de ese grupo fueron reunidos.

Un ejemplo para el término *barra* y el tipo analítico con el valor de corte 0.5 es el siguiente:

```
------------  corte = 0.5 -------------------------

GRUPO 5:

La Barra es el mejor grupo de cuarteto que haya.

La Barra es un grupo de Cuarteto Cordobés, formado en 1994 como desprendimiento de la orquesta Tru-la-lá.

La barra para mi es el mejor grupo de cuarteto qe hai en cordoba.


GRUPO 12:

las barras son la verdadera aficion que apoya incondicional mente a su equipo.

Las barras son una copia de los modelos sudamericano de apoyo a los equipos.


GRUPO 72:

Una barra es un elemento ideal que se caracteriza por tener una directriz recta, una sección constante y estar formada por un solo material.

Las barras son el resultado del depósito de los materiales sólidos transportados por  fuertes lluvia.


GRUPO 98:

La barra es la franja vertical que separa las mitades izquierda y derecha del tablero.

La barra es la franja vertical que separa el tablero por la mitad.


GRUPO 102:

La Barra es el cuarteto "digerible" para los rockeros, los chetos, los os empresarios o los estudiantes que hace poco tiempo miraban con cierto desprecio o recelo la música popular cordobesa.

La Barra es el cuarteto que más iniciativas tomó al margen de la lógica que obliga a grabar dos discos anuales y presentase en el circuito de clubes tradicionales.


GRUPO 104:

una barra es una comunidad de personas que van a un estadio a animar, alentar su equipo.

las barras son las agrupaciones de personas, que adscritas bajo un nombre y afiliadas o no a una institución que las agrupa, asisten al estadio a apoyar a su equipo de su corazón.
```

GRUPO 106:

La Barra es una ciudad balnearia uruguaya ubicada sobre la costa del océano Atlántico.

La Barra es una ciudad balnearia ubicada a la altura de la desembocadura del arroyo Maldonado.

La Barra es una pequeña población balnearia en la costa uruguaya, dentro del departamento de Maldonado, en las cercanías de Punta del Este.

GRUPO 111:

La barra es un resumen de los servicios clave de Yahoo que está incrustada en la parte superior del navegador.

Las barras son unos cilindros huecos que discurren por el interior de las botellas y que están ancladas en su parte superior a la tija.

GRUPO 113:

La Barra es el centro de la movida de pubs, restaurantes bailables, discos y establecimientos similares.

La Barra es el centro de la vida nocturna, con pubs y discotecas que le dan una ritmo particular y propia.

GRUPO 114:

las barras son el alma de los equipos pero también son considerados la oveja negra de la familia del fútbol.

las barras son el accesorio interior perfecto para la familia entretenida y amigos, son también extremadamente prácticas para utilizar el espacio.

GRUPO 117:

la barra es la banda mas pero mas grande de cordoba

Sin duda la barra es la banda mas grande de cordoba.

GRUPO 134:

La Barra es una zona de alto valor comercial así como una buena inversión si se tiene en cuenta el desarrollo vertiginoso en edificaciones y negocios que ha tenido los últimos 10 años.

Una barra es una estructura verticalista y además, buen negocio.

------ 26 definiciones recuperadas en 12 grupos ------

Consideramos que la interpretación es el criterio más importante en esta evaluación y es por ello que hemos publicado en la Web los resultados sobre los cuales se realizó el análisis cualitativo. No pretendemos ser exhaustivos ni formales en el análisis. Sin embargo, en las secciones 4.4, 4.5 y 4.6 se discuten los resultados que nos parecieron más relevantes para cada tipo de definición. Invitamos a que el lector compruebe la calidad de los resultados del algoritmo en la siguiente dirección:

http://saussure.iingen.unam.mx/~amolinav/resultados/

4.3. ACERCA DEL ANÁLISIS CUANTITATIVO

El análisis cuantitativo presenta dos tipos de evaluación: la primera de ellas concierne a las características de los agrupamientos en función del valor de corte por distancia y la segunda proporcionan información acerca del dendrograma *per se*.

Para el análisis de las características de los agrupamientos generados, se presentan las gráficas del comportamiento del algoritmo al variar el valor de corte por distancia desde 0.10 hasta 1.00, con un avance de 0.01 en cada ejecución. Los indicadores utilizados para analizar el algoritmo son: el número de grupos generados, el recuerdo y la precisión.

Definimos el recuerdo como la proporción de definiciones integradas a algún grupo con respecto al total de las definiciones utilizadas como entrada del algoritmo. Es decir, cuántas definiciones logramos integrar en el agrupamiento. El recuerdo es una función que devuelve un valor entre 0 (si no se conformó ningún grupo con al menos dos definiciones) y 1 (si todas las definiciones fueron integradas en algún grupo). La fórmula para calcular el recuerdo es la siguiente:

$$r = \frac{\left|\{total\_definiciones\} \cap \{definiciones\_asignadas\_a\_un\_grupo\}\right|}{\left|\{total\_definiciones\}\right|}.$$

La precisión por su parte es la proporción de intrusos en un agrupamiento generado. Nos indica cuantos errores cometimos al integrar los grupos. La precisión devuelve también un valor entre 0 (si no es posible determinar la acepción de ninguno de los grupos) y 1 (si ningún grupo contiene intrusos). La fórmula para la precisión utilizada es:

$$p = \frac{\left|\{definiciones\_asignadas\_a\_un\_grupo\} - \{Intrusos\}\right|}{\left|\{definiciones\_asignadas\_a\_un\_grupo\}\right|}.$$

Para la evaluación del dendrograma, se utilizaron los coeficientes de inconsistencia (*inconsistency*), sugerido por Zahn (1971), y cofenético (*cophenetic*), sugerido por Sokal (1962). Estos coeficientes están enfocados al agrupamiento jerárquico por sí mismo, dado que ambos están basados en el dendrograma generado por la aplicación de HAC. Como hemos mencionado, en un dendrograma, cualesquiera dos objetos son finalmente enlazados en algún nivel. La altura de dicho enlace representa la distancia entre los dos grupos que contienen

dichos objetos. Esta altura es conocida como la distancia cofenética y al comparar dicha distancia con la distancia entre los objetos –la energía textual en este caso– evaluamos si la estructura generada refleja adecuadamente las distancias entre los objetos. Si la estructura de grupos es válida. El coeficiente de correlación se calculó según la ecuación:

$$c = \frac{\sum_{i<j}(d_{ij} - \bar{d})(D_{ij} - \bar{D})}{\sqrt{\sum_{i<j}(d_{ij} - \bar{d})^2 \sum_{i<j}(D_{ij} - \bar{D})^2}}.$$

Donde d_{ij} es la distancia entre el vector *i* y el vector *j* según la medida utilizada, D_{ij} representa la distancia cofenética, $\bar{d}$ y $\bar{D}$ son el promedio de la distancias entre objetos y el promedio de las distancias coféneticas, respectivamente. Entre más se acerca a uno el valor, mayor es la calidad de la solución propuesta.

Por su parte, el coeficiente de inconsistencia caracteriza cada enlace en el árbol comparando su altura con la altura promedio de otros enlaces en el mismo nivel de jerarquía. Valores mayores del coeficiente indican menos similitud de los objetos unidos por el enlace. Para cada enlace, el coeficiente de inconsistencia se puede calcular según la fórmula:

$$I(e) = h_e - \frac{\mu_e}{\sigma_e}$$

Donde h_e es la altura del enlace *e*, μ_e y $\sigma_e$son, respectivamente, la media aritmética y la desviación estándar de las alturas de los enlaces bajo *e* en la jerarquía.

4.4. RESULTADOS PARA LAS DEFINICIONES ANALÍTICAS

Para ejemplificar los resultados a nivel cualitativo en las definiciones analíticas haremos un análisis de los resultados para el término *punto*. Observaremos que a medida que el valor de corte aumenta, se recuperan más definiciones y en grupos cada vez más específicos.

Comenzamos por el valor de corte 0.59, donde tenemos solamente dos grupos y cuatro definiciones recuperadas. El resultado del algoritmo es:

```
------------  corte = 0.59 -------------------------

GRUPO 72:
```

En su costado más simple, el punto es una manera de anudar de hilos en series paralelas, entrelazando las filas para formar una tela tejida

En su lado más simple, el punto es una manera de entrelazar los hilos en series paralelas, formando filas y finalizar con una tela tejida

GRUPO 73:

El punto es un elemento geométrico adimensional, no es un objeto físico; describe una posición en el espacio, determinada en función de un sistema de coordenadas preestablecido

El punto se considera como elemento geométrico fundamental, sin dimensión, sin propiedades físicas, como un ladrillo de todas las demás construcciones que realizará el software para identificar, localizar o relacionar entre si las características geométricas de la muestra sometida a análisis

------ 4 definiciones recuperadas en 2 grupos ------

Donde observamos que las acepciones están muy bien delimitadas como *una manera de costura* o *una entidad geométrica*, pero existe la desventaja de que solo recuperamos cuatro definiciones de un conjunto de cien. Si el valor de corte es aumentado a 0.66, aparece una nueva acepción, el punto es *una pausa*; se introducen tres nuevas definiciones y aparece un intruso, *una pequeña mancha* en el grupo de *costura*:

---------- corte = 0.66 -------------------------

GRUPO 69:

El punto es un elemento geométrico adimensional, no es un objeto físico; describe una posición en el espacio, determinada en función de un sistema de coordenadas preestablecido

El punto se considera como elemento geométrico fundamental, sin dimensión, sin propiedades físicas, como un ladrillo de todas las demás construcciones que realizará el software para identificar, localizar o relacionar entre si las características geométricas de la muestra sometida a análisis

GRUPO 70:

El punto: es una pausa que indica que ha terminado una oración, un párrafo o un texto

El punto es una pausa que indica que ha terminado una oración

GRUPO 75:

En gráficos de ordenador o computadora e imprenta, un punto es una pequeña 'mancha' combinada con otras en una matriz de filas y columnas para formar un carácter o un elemento gráfico de un dibujo o diseño

En su costado más simple, el punto es una manera de anudar de hilos en series paralelas, entrelazando las filas para formar una tela tejida

En su lado más simple, el punto es una manera de entrelazar los hilos en series paralelas, formando filas y finalizar con una tela tejida

------ 7 definiciones recuperadas en 3 grupos ------

Si aumentamos el corte a 0.75 aumenta el número de definiciones introducidas a 13, hay un intruso en el grupo de *costura* y aparecen dos nuevas acepciones de punto, *una unidad visual* y *un signo de puntuación*. Es motivo de discusión si las acepciones de *pausa* y *un signo de puntuación* deberían ser consideradas como un solo grupo, en cuyo caso diríamos que el algoritmo empieza a separar indebidamente. También podríamos argumentar que al contrario, el algoritmo hace bien en separar porque existen sutilezas en la explicación de ambos significados: en el grupo 74 se entiende que el punto es un símbolo ortográfico que aparece en un medio escrito, mientras que en el grupo 79 nos referimos a un instante de tiempo.

```
------------  corte = 0.75 -------------------------


GRUPO 67:

En gráficos de ordenador o computadora e imprenta, un punto es una pequeña
'mancha' combinada con otras en una matriz de filas y columnas para formar
un carácter o un elemento gráfico de un dibujo o diseño

En su costado más simple, el punto es una manera de anudar de hilos en
series paralelas, entrelazando las filas para formar una tela tejida

En su lado más simple, el punto es una manera de entrelazar los hilos en
series paralelas, formando filas y finalizar con una tela tejida


GRUPO 71:

El punto es la unidad mínima de información visual, y está caracterizado por
su forma, tamaño, color y ubicación

El punto. Es la unidad más simple, irreductiblemente mínima, de comunicación
visual


GRUPO 72:

El punto es un elemento geométrico adimensional, no es un objeto físico;
describe una posición en el espacio, determinada en función de un sistema de
coordenadas preestablecido

Un punto es un elemento fijo con unas coordenadas concretas y si por el
pasan varias líneas que definen el armado del dibujo y en el se apoyan los
vértices del Modelo Digital del Terreno se nos hace necesario tenerlo
perfectamente identificado y accesible en todo momento

El punto se considera como elemento geométrico fundamental, sin dimensión,
sin propiedades físicas, como un ladrillo de todas las demás construcciones
que realizará el software para identificar, localizar o relacionar entre si
las características geométricas de la muestra sometida a análisis


GRUPO 74:

El punto es el signo de puntuación que cierra cualquier enunciado, de modo
que, en caso de que se transcriban textos con comillas, con guiones o con
otros signos, estos deben preceder al punto

El punto (.), es un signo de puntuación que cierra oraciones o frases con
sentido completo


GRUPO 79:

El punto es una pausa que indica que ha terminado una oración
```

```
El punto: es una pausa que indica que ha terminado una oración, un párrafo o
un texto
El punto es una pausa que indica que ha terminado una oración

------  13 definiciones recuperadas en 5 grupos   ------
```

Para comenzar con los resultados del análisis cualitativo presentamos la figura 4.2, que muestra los resultados de variar el valor de corte por distancia en el algoritmo de agrupamiento para las colecciones de definiciones del tipo analítico.

En la tabla 4.3 se muestran algunos resultados estadísticos asociados a los dendrogramas obtenidos al aplicar el algoritmo de agrupamiento en las colecciones de definiciones del tipo analítico extraídas de la Web para los términos: *barra*, *célula*, *punto* y *ventana*. Se muestran los resultados del coeficiente cofenético y de la medida de distancia aplicada para efectos de comparación de los resultados. El lado izquierdo de la tabla reporta la energía máxima, mínima y promedio obtenida directamente de la medida de distancia utilizada para comparar los vectores. Análogamente, el lado derecho muestra la distancia cofenética máxima, mínima y promedio. La columna del extremo izquierdo muestra el coeficiente cofenético de cada dendrograma.

Finalmente, la tabla 4.4 muestra algunos resultados asociados a los dendrogramas obtenidos al aplicar el algoritmo de agrupamiento en las colecciones de definiciones del tipo analítico. Se muestran algunos resultados asociados al coeficiente de inconsistencia: la media y mediana tomando como muestra el conjunto de enlaces de cada dendrograma, el valor máximo de la desviación estándar y, en la columna del extremo derecho, el valor de inconsistencia más alto obtenido por algún enlace.

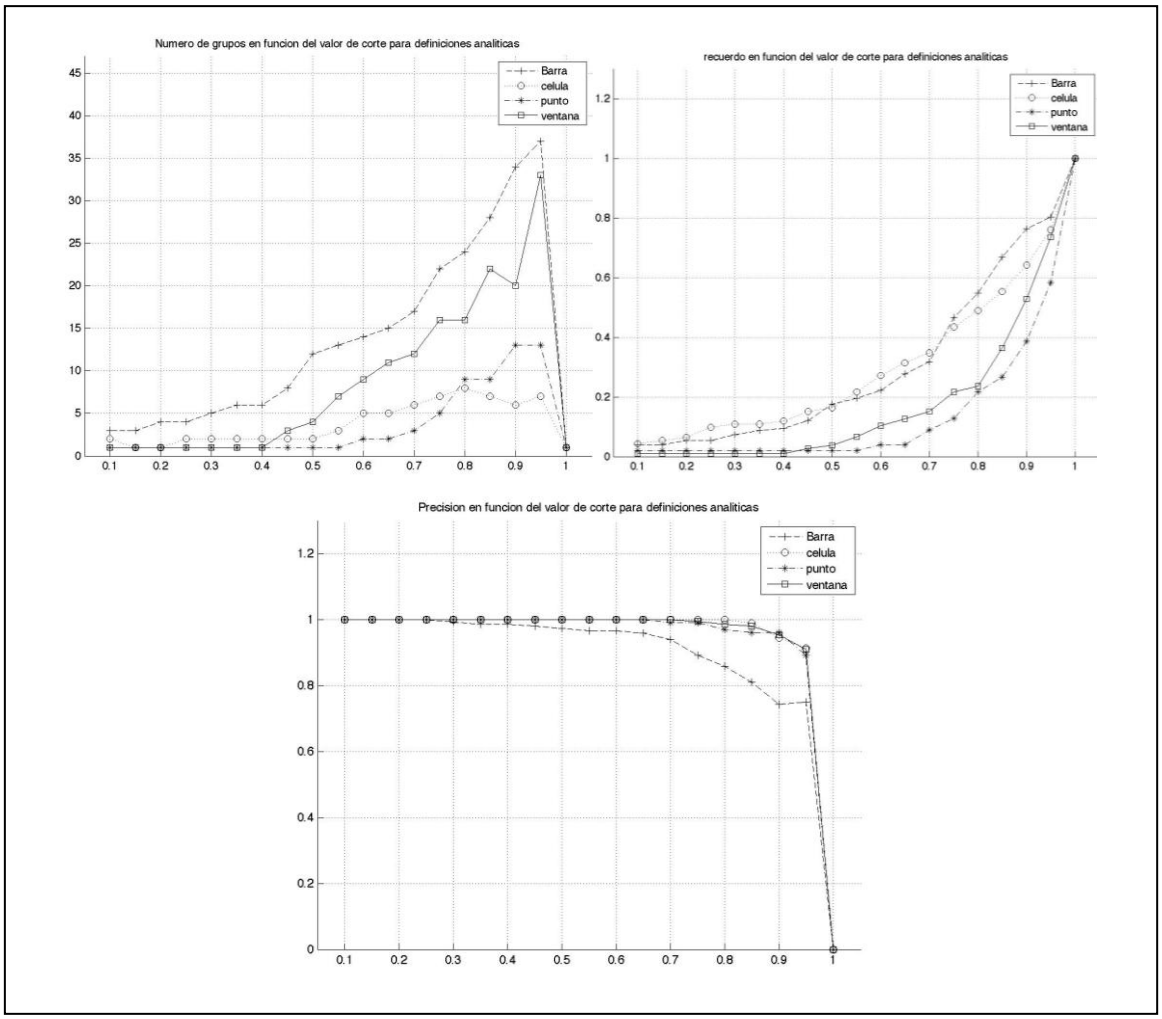

Figura 4.2. Resultados de los agrupamientos para las definiciones analíticas.

	Distancia entre documentos (Energía)			Distancia cofenética			**Coeficiente cofenético**
	máx	**mín**	**prom**	**máx**	**mín**	**avg**	
Barra	1	0	0.9689	1	0	0.9825	0.8817
Célula	1	0	0.8175	1	0	0.8819	0.9333
Punto	0.9689	0	0.6378	0.9689	0	0.7815	0.8279
Ventana	1	0	0.9631	1	0	0.9865	0.7671

Tabla 4.3. Resultados del coeficiente cofenético para las definiciones analíticas.

	Inconsistencia			**Máximo valor de inconsistencia**
	media	**mediana**	**máx(DesvStd)**	
Barra	0.4456	0.5774	0.4265	1.0722
Célula	0.5780	0.7071	0.1895	1.1446
Punto	0.5705	0.7071	0.1149	1.1366
Ventana	0.5209	0.7071	0.4001	1.1547

Tabla 4.4. Resultados del coeficiente de inconsistencia para las definiciones analíticas.

4.5. RESULTADOS PARA LAS DEFINICIONES EXTENSIONALES

Los resultados del análisis cualitativo de las definiciones extensionales serán ejemplificados con el término *célula*. Se observa una relación importante entre el valor de corte y el número de grupos generados. Para el valor de corte de 0.50 y hasta el valor 0.64 se obtienen tan solo dos grupos, mientras que para el valor 0.98 se obtienen 23 grupos.

En los grupos primeramente identificados la célula contiene o bien *información* o bien *habilidad/recursos/caminos para formar moléculas*. Mostramos a continuación el agrupamiento generado para el valor 0.64 y todos los valores de corte anteriores:

```
------------  corte = 0.64 -------------------------


GRUPO 5:

La célula contiene en sus cromosomas toda la información genética necesaria
para el funcionamiento y la reproducción del organismo entero del que ella
forma parte.

La célula contiene toda la información sobre la síntesis de su estructura y
el control de su funcionamiento y es capaz de transmitirla a sus
descendientes, es decir, la célula es la unidad genética autónoma de los
seres vivos

La célula contiene toda la información sobre la síntesis de su estructura y
el control de su funcionamiento y es capaz de transmitirla a sus
descendientes, es decir, la célula es la unidad genética autónoma de los
seres vivos


GRUPO 104:

las células cuentan con caminos para formar moléculas más pequeñas a partir
de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con habilidades para formar moléculas más pequeñas a
partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con recursos para formar moléculas más pequeñas a partir
de moléculas grandes, y a este proceso se le llama catabolismo
```

------ 6 definiciones recuperadas en 2 grupos ------

Posteriormente, cuando el valor de corte es 0.7, aparece un grupo donde la célula cuenta con *sistemas* (grupo 97). De igual manera, aparece un grupo donde la célula consta de *membrana, citoplasama y núcleo* (grupo 93), que es un grupo que esperábamos encontrar pronto. Observamos también que existe un grupo mal formado, el grupo 90, el cual contiene una definición que debería pertenecer al grupo 93 y otra que debió integrarse al grupo 97.

---------- corte = 0.7 ----------

GRUPO 90:

Una célula consta de un núcleo y su citoplasma circundante, con una envoltura altamente especializada denominada membrana celular

la célula cuenta con un sistema especializado en cortar las proteínas del citoplasma, por lo que no es difícil imaginar la existencia de un sistema equivalente en la membrana celular, cuyo primer componente puede ser la nicastrina

GRUPO 92:

La célula contiene en sus cromosomas toda la información genética necesaria para el funcionamiento y la reproducción del organismo entero del que ella forma parte.

las células contienen información hereditaria necesaria para las funciones de regulación de la célula y para transmitir información a la siguiente generación de células

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

Una celula contiene toda la informacion hereditaria necesaria para el control de su propio ciclo y del desarrollo y el funcionamiento de un organismo

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

GRUPO 93:

Una célula consta de tres partes: membrana, citoplasma y núcleo

Todas las células constan de tres partes principales: La membrana citoplasmática, el citoplasma y una región nuclear que alberga el material genético

GRUPO 97:

las células cuentan con dos sistemas de enzimas, que tienen la función de introducir en el substrato un átomo de oxígeno proveniente del oxígeno molecular (oxigenasas de función mixta)

Para ello las células cuentan con sistemas reguladores, como son las vías de los SREBP y de LXR, que, en respuesta a cambios en las concentraciones celulares de esteroles, modifican la expresión de genes implicados en el

metabolismo lipídico, como son el del receptor de LDL, enzimas de la biosíntesis de colesterol y la lipogénesis y moléculas implicadas en la excreción de colesterol

GRUPO 99:

las células cuentan con caminos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con habilidades para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con recursos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

------ 14 definiciones recuperadas en 5 grupos ------

El siguiente valor de corte notable es para el valor 0.74 donde se introduce la acepción de célula en electrónica, donde la célula *se compone de ánodo y cátodo*. Note también que se han introducido varias definiciones nuevas que no se encontraban para valores de corte anteriores. También se detecta que el grupo de *membrana, citoplasama y núcleo* se dividió en dos, el grupo 3 y el grupo 88.

---------- corte = 0.74 -------------------------

GRUPO 3:

La célula consta de: membrana celular, citoplasma y núcleo.

Una célula consta de un núcleo, un citoplasma y la membrana celular

Una célula consta de un núcleo y su citoplasma circundante, con una envoltura altamente especializada denominada membrana celular

la célula cuenta con un sistema especializado en cortar las proteínas del citoplasma, por lo que no es difícil imaginar la existencia de un sistema equivalente en la membrana celular, cuyo primer componente puede ser la nicastrina

GRUPO 86:

La célula contiene en sus cromosomas toda la información genética necesaria para el funcionamiento y la reproducción del organismo entero del que ella forma parte.

Las células contienen la información hereditaria necesaria para la regulación de las funciones celulares y para la transmisión de la información

las células contienen información hereditaria necesaria para las funciones de regulación de la célula y para transmitir información a la siguiente generación de células

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

Una celula contiene toda la informacion hereditaria necesaria para el control de su propio ciclo y del desarrollo y el funcionamiento de un organismo

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

GRUPO 88:

Una célula consta de tres partes: membrana, citoplasma y núcleo

La Célula: Consta de una membrana que es su frontera externa, un citoplasma, unos organelos, un núcleo separado por la membrana nuclear

Todas las células constan de tres partes principales: La membrana citoplasmática, el citoplasma y una región nuclear que alberga el material genético

GRUPO 91:

las células cuentan con dos sistemas de enzimas, que tienen la función de introducir en el substrato un átomo de oxígeno proveniente del oxígeno molecular (oxigenasas de función mixta)

Para ello las células cuentan con sistemas reguladores, como son las vías de los SREBP y de LXR, que, en respuesta a cambios en las concentraciones celulares de esteroles, modifican la expresión de genes implicados en el metabolismo lipídico, como son el del receptor de LDL, enzimas de la biosíntesis de colesterol y la lipogénesis y moléculas implicadas en la excreción de colesterol

GRUPO 93:

las células cuentan con caminos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con habilidades para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con recursos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

GRUPO 109:

La celula consta de - el cuerpo de celula con los dos contactos de conexión hacia la unidad de evaluación - el cátodo de plata - el ánodo de plomo - el electrolito- la membrana de teflon permeable al oxígeno (pero no porosa), que representa la superficie límite entre el interior de la célula y el medio de medición

la célula se compone de dos electrodos (ánodo y cátodo) separados por un electrolito

------ 20 definiciones recuperadas en 6 grupos ------

Un agrupamiento notable es para el valor de corte 0.80, donde además de las agrupaciones ya mencionadas, aparecen dos nuevas acepciones. En una de ellas se considera que la célula *cuenta con genes* (grupo 106) y, aunque es una definición perteneciente también al campo de la biología, proporciona una forma distinta de concebir la composición de la célula, tal vez una más apegada a la genética. En otro grupo de reciente formación (grupo 94), observamos un fenómeno interesante y alentador: a partir de la primera definición este grupo

podemos inferir que la célula se compone de un núcleo que está envuelto en protoplasma que a su vez está envuelto por una membrana; mientras que, en la segunda, la explicación es semánticamente equivalente, pero estructuralmente inversa, parafraseando, la célula consta de una membrana que envuelve al protoplasma en el cual está el núcleo.

---------- corte = 0.8 -------------------------

GRUPO 5:

Las células constan de tres partes: membrana, núcleo y plasma

La célula consta de 3 partes: La membrana plásmatica, el citplásma y el núcleo

GRUPO 12:

La célula consta de: membrana celular, citoplasma y núcleo.

Una célula consta de tres partes: membrana, citoplasma y núcleo

contiene otros elementos en su núcleo, además de un citoplasma y una membrana

Una célula consta de un núcleo, un citoplasma y la membrana celular

La Célula: Consta de una membrana que es su frontera externa, un citoplasma, unos organelos, un núcleo separado por la membrana nuclear

Todas las células constan de tres partes principales: La membrana citoplasmática, el citoplasma y una región nuclear que alberga el material genético

Una célula consta de un núcleo y su citoplasma circundante, con una envoltura altamente especializada denominada membrana celular

la célula cuenta con un sistema especializado en cortar las proteínas del citoplasma, por lo que no es difícil imaginar la existencia de un sistema equivalente en la membrana celular, cuyo primer componente puede ser la nicastrina

GRUPO 82:

las células cuentan con dos sistemas de enzimas, que tienen la función de introducir en el substrato un átomo de oxígeno proveniente del oxígeno molecular (oxigenasas de función mixta)

Para ello las células cuentan con sistemas reguladores, como son las vías de los SREBP y de LXR, que, en respuesta a cambios en las concentraciones celulares de esteroles, modifican la expresión de genes implicados en el metabolismo lipídico, como son el del receptor de LDL, enzimas de la biosíntesis de colesterol y la lipogénesis y moléculas implicadas en la excreción de colesterol

GRUPO 84:

las células cuentan con caminos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con habilidades para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

Las células cuentan con recursos para formar moléculas más pequeñas a partir de moléculas grandes, y a este proceso se le llama catabolismo

GRUPO 90:

La célula contiene en sus cromosomas toda la información genética necesaria para el funcionamiento y la reproducción del organismo entero del que ella forma parte.

Las células contienen la información hereditaria necesaria para la regulación de las funciones celulares y para la transmisión de la información

Las células contienen información hereditaria indispensable para regular sus funciones y ser transmitida a las siguientes generaciones

las células contienen información hereditaria necesaria para las funciones de regulación de la célula y para transmitir información a la siguiente generación de células

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

Una celula contiene toda la informacion hereditaria necesaria para el control de su propio ciclo y del desarrollo y el funcionamiento de un organismo

La célula contiene toda la información sobre la síntesis de su estructura y el control de su funcionamiento y es capaz de transmitirla a sus descendientes, es decir, la célula es la unidad genética autónoma de los seres vivos

GRUPO 94:

La célula se compone de un núcleo envuelto en protoplasma, alrededor del cual hay una membrana que separa la célula de su medio ambiente

La célula consta de una MEMBANA CELULAR que envuelve una masa viscosa y granulosa llamada PROTOPLASMA, en la cual se encuentran todos los ORGANELOS CELULARES, incluido el NÚCLEO

GRUPO 102:

La celula consta de - el cuerpo de celula con los dos contactos de conexión hacia la unidad de evaluación - el cátodo de plata - el ánodo de plomo - el electrolito- la membrana de teflon permeable al oxígeno (pero no porosa), que representa la superficie límite entre el interior de la célula y el medio de medición

la célula se compone de dos electrodos (ánodo y cátodo) separados por un electrolito

GRUPO 106:

una de las funciones que la célula ejerce es la de su propia reproducción, y la célula cuenta con genes específicos para llevarla a cabo

La célula cuenta con genes propios de acuerdo a su especialización y están en disposición de encendido o apagado con una secuencia específica fundamentada en un tiempo intervienen en la actividad celular que se rige por eventos que ocurren desde la vida embrionaria hasta la vida adulta

------ 28 definiciones recuperadas en 8 grupos ------

Para valores posteriores de corte, los grupos se vuelven demasiado específicos y resulta complicado determinar si en realidad es preciso seguir aumentando el valor de corte.

Por otro lado, nuevas acepciones son reveladas, como es el caso del agrupamiento para el valor de corte 0.89, donde se descubre la acepción de célula en el área de redes de computadoras al formarse el siguiente grupo:

```
GRUPO 73:

Las células constan de un campo de información de 48 octetos y una cabecera
de 5 octetos, la cual contiene un conjunto de informaciones de control, como
identificadores, que se utilizan para identificación de las conexiones y
encaminamiento

La célula se compone de. un campo para la información y una cabecera para
su. identificación
```

Como primer resultado del análisis cualitativo, la figura 4.3 muestra el comportamiento del algoritmo al variar el valor de corte por distancia. Se puede observar la relación que existe entre este parámetro y el número de grupos generados, así como en el recuerdo y la precisión de los agrupamientos.

La tabla 4.5 se muestran los resultados obtenidos al aplicar el algoritmo de agrupamiento en las definiciones del tipo. Se muestran los resultados del coeficiente cofenético y de la distancia energética para efectos de comparación de los resultados. El lado izquierdo de la tabla reporta la energía máxima, mínima y promedio obtenida directamente de la medida de distancia utilizada para comparar los vectores. El lado derecho muestra la distancia cofenética máxima, mínima y promedio. La columna del extremo izquierdo muestra el coeficiente cofenético de cada dendrograma.

En la tabla 4.6**,** se muestran la media y mediana asociadas al coeficiente de inconsistencia, tomando como muestra el conjunto de enlaces de cada dendrograma. También se incluyen: el valor máximo de la desviación estándar y el valor de inconsistencia máximo obtenido por algún enlace.

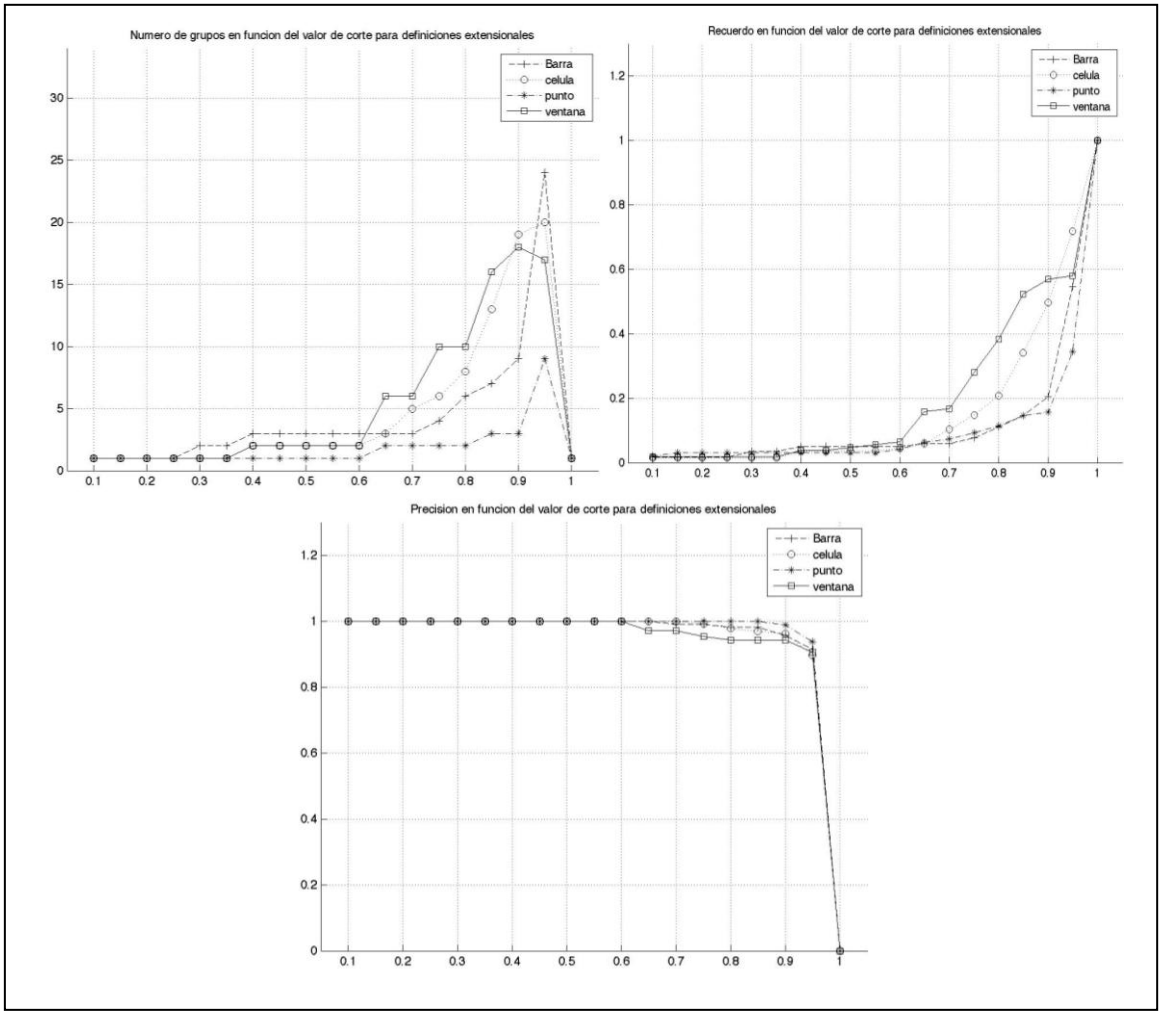

Figura 4.3. Resultados de los agrupamientos para las definiciones extensionales.

	Distancia entre documentos (Energía)			Distancia cofenética			**Coeficiente cofenético**
	máx	**mín**	**prom**	**máx**	**mín**	**avg**	
Barra	1	0	0.9803	1	0	0.9898	0.8076
Célula	1	0	0.9583	1	0	0.9852	0.8099
Punto	1	0	0.9846	1	0	0.9919	0.9401
Ventana	1	0	0.9385	1	0	0.9693	0.7981

Tabla 4.5. Resultados del coeficiente cofenético para las definiciones extensionales.

	Inconsistencia			**Máximo valor de inconsistencia**
	media	**mediana**	**máx(DesvStd)**	
Barra	0.4785	0.6265	0.4602	0.9527
Célula	0.5306	0.7071	0.4200	1.1529
Punto	0.4934	0.7071	0.3489	1.1547
Ventana	0.4427	0.5774	0.4714	0.7071

Tabla 4.6. Resultados del coeficiente de inconsistencia para las definiciones extensionales.

4.6. RESULTADOS PARA LAS DEFINICIONES FUNCIONALES

Para el análisis cualitativo de las definiciones funcionales nos enfocaremos en los resultados del término *punto*. Las primeras dos funciones de *punto* se detectan desde el valor de corte 0.69. En estas dos primeras acepciones, el punto se emplea para *indicar el final de una oración* o permite *regresar al equipo de Schuster*.

```
------------  corte = 0.69 -------------------------


GRUPO 112:

se dice que el punto se emplea al final de una oración para indicar que lo
que la precede forma sentido completo

El punto se emplea para indicar el final de una oración, señala que lo
escrito antes tiene sentido completo


GRUPO 114:

el punto permite al equipo de Schuster, que vio el partido en la grada por
sanción, regresar a la zon UEFA junto con el Atlético y a costa del
Recreativo, que sucumbió hoy en Santander ante el Racing en un partido lleno
de sorpresas

el punto permite al equipo de Schuster, que veía el partido en la grada por
sanción, regresar a la zona UEFA junto con el Atlético y a costa del
Recreativo, que ha sucumbido en Santander ante el Racing en un partido lleno
de sorpresas (4-3)


------  4 definiciones recuperadas en 2 grupos   ------
```

Conforme aumentamos el valor de corte a 0.75 aparece la funcionalidad del punto para *separar* como se muestra en el grupo 113.

```
------------  corte = 0.75 -------------------------


GRUPO 110:
```

el punto permite al equipo de Schuster, que vio el partido en la grada por sanción, regresar a la zon UEFA junto con el Atlético y a costa del Recreativo, que sucumbió hoy en Santander ante el Racing en un partido lleno de sorpresas

el punto permite al equipo de Schuster, que veía el partido en la grada por sanción, regresar a la zona UEFA junto con el Atlético y a costa del Recreativo, que ha sucumbido en Santander ante el Racing en un partido lleno de sorpresas (4-3)

GRUPO 112:

se dice que el punto se emplea al final de una oración para indicar que lo que la precede forma sentido completo

El punto: Se emplea para indicar el final de una oración o cuando se ha terminado de exponer una idea en un escrito (idea completa)

El punto se emplea para indicar el final de una oración, señala que lo escrito antes tiene sentido completo

GRUPO 113:

La coma, según la ISO, se emplea para separar la parte entera de la parte decimal en las expresiones numéricas del sistema decimal (aunque admite que en los textos en inglés se emplee en ese caso un punto), mientras que el punto se usa con una función análoga en sistemas no decimales, como los empleados para las horas, los años, los grados, los puntos tipográficos, etcétera

Los puntos se emplean para separar los cuatro octetos que se describen como valores numéricos decimales independientes en la dirección IP

------ 7 definiciones recuperadas en 3 grupos ------

Para el valor 0.88, aparece una nueva funcionalidad del punto: *reforzar y decorar* (grupo 41), pero se nota que el algoritmo no puede resolver la mezcla de conceptos que hay entre la función del punto en una oración. Por un lado *indica una pausa*, por otro *divide secciones* y por otro *señala una idea*. Esta mezcla natural en los conceptos de punto se refleja en la mezcla de grupos que el algoritmo obtiene.

------------ corte = 0.88 -------------------------

GRUPO 3:

El punto se emplea para separar oraciones, indica una pausa larga en la lectura y debe marcarse con entonación descendente

El punto se utiliza para señalar una pausa larga, que marca el final de una frase u oración.

El punto se utiliza para dar fin a una oración e indicar una pausa entre las distintas ideas que se expresan

GRUPO 41:

los puntos se emplean para reforzar y decorar un orillo, por ejemplo, el punto de lengüeta, el punto de festón y el punto de ojete; para perfilar una forma, como el hilván, el punto de cadeneta o el bordado de hilos tendidos; para rellenar una superficie se utiliza el punto al pasado o el punto de

hoja; y, finalmente, pueden emplearse el punto de armas retorcido o el punto de armas anudado para crear relieve

el punto (.) se usa para crear una cadena formada por otras dos

GRUPO 104:

La coma, según la ISO, se emplea para separar la parte entera de la parte decimal en las expresiones numéricas del sistema decimal (aunque admite que en los textos en inglés se emplee en ese caso un punto), mientras que el punto se usa con una función análoga en sistemas no decimales, como los empleados para las horas, los años, los grados, los puntos tipográficos, etcétera

el punto se utiliza como separador de miles y la coma como símbolo decimal

Los puntos se emplean para separar los cuatro octetos que se describen como valores numéricos decimales independientes en la dirección IP

GRUPO 105:

el punto permite al equipo de Schuster, que vio el partido en la grada por sanción, regresar a la zon UEFA junto con el Atlético y a costa del Recreativo, que sucumbió hoy en Santander ante el Racing en un partido lleno de sorpresas

el punto permite al equipo de Schuster, que veía el partido en la grada por sanción, regresar a la zona UEFA junto con el Atlético y a costa del Recreativo, que ha sucumbido en Santander ante el Racing en un partido lleno de sorpresas (4-3)

El punto permite al equipo de Quique Sánchez Flores situarse tercero a diez del líder, mientras que el rojiblanco lleva ya nueve partidos sin ganar y está sumido en la medianía de la tabla

GRUPO 106:

El punto se emplea para indicar el final de una oración

se dice que el punto se emplea al final de una oración para indicar que lo que la precede forma sentido completo

El punto: Se emplea para indicar el final de una oración o cuando se ha terminado de exponer una idea en un escrito (idea completa)

El Punto (.): Se emplea para indicar el sentido de una oración, simple o compuesta, aunque se siga tratando el mismo asunto

El punto se emplea para indicar el final de una oración, señala que lo escrito antes tiene sentido completo

El punto se utiliza para señalar el final de una oración, de un párrafo o de un texto que tiene sentido en sí mismo

------ 17 definiciones recuperadas en 5 grupos ------

Notemos que las distinciones entre los significados del término *punto* se aclaran a medida que aumentamos el valor de corte. El efecto de tener más grupos es poder discernir mejor la diferencia entre ellos, sus puntos en común y sus diferencias. Por otro lado, tenemos como consecuencia directa la introducción de más definiciones al agrupamiento, lo que puede causar que sea más complicado analizar la presentación de los resultados inmediatamente.

El último agrupamiento que mostramos es para el valor de corte 0.93, donde observamos que el número de grupos es todavía *manejable*, pero los grupos generados resultan demasiado específicos.

```
------------  corte = 0.93 -------------------------

GRUPO 17:

el punto se usa para separar dos oraciones que afirman ideas en cierta forma independientes

los puntos se usan para separar oraciones y no listados


GRUPO 30:

los puntos se emplean para reforzar y decorar un orillo, por ejemplo, el punto de lengüeta, el punto de festón y el punto de ojete; para perfilar una forma, como el hilván, el punto de cadeneta o el bordado de hilos tendidos; para rellenar una superficie se utiliza el punto al pasado o el punto de hoja; y, finalmente, pueden emplearse el punto de armas retorcido o el punto de armas anudado para crear relieve

el punto (.) se usa para crear una cadena formada por otras dos


GRUPO 96:

No olvidar que, en inglés, el punto se usa para separar las fracciones, y los múltiplos no se separan con ningún signo

tradicionalmente el punto se emplea, en las expresiones numéricas, para separar los millares, millones, miles de millones, etc.

La coma, según la ISO, se emplea para separar la parte entera de la parte decimal en las expresiones numéricas del sistema decimal (aunque admite que en los textos en inglés se emplee en ese caso un punto), mientras que el punto se usa con una función análoga en sistemas no decimales, como los empleados para las horas, los años, los grados, los puntos tipográficos, etcétera

el punto se utiliza como separador de miles y la coma como símbolo decimal

en europa el punto se utiliza para separar cifras de tres y la coma para marcar fin de las cifras enteras, al contrario que se utiliza en toda america

Los puntos se emplean para separar los cuatro octetos que se describen como valores numéricos decimales independientes en la dirección IP


GRUPO 97:

El punto se emplea para indicar el final de una oración

se dice que el punto se emplea al final de una oración para indicar que lo que la precede forma sentido completo

El punto: Se emplea para indicar el final de una oración o cuando se ha terminado de exponer una idea en un escrito (idea completa)

El Punto (.): Se emplea para indicar el sentido de una oración, simple o compuesta, aunque se siga tratando el mismo asunto

El punto se emplea para indicar el final de una oración, señala que lo escrito antes tiene sentido completo

El punto se emplea para separar oraciones, indica una pausa larga en la lectura y debe marcarse con entonación descendente
```

El punto se utiliza para señalar una pausa larga, que marca el final de una frase u oración.

El punto se utiliza para señalar el final de una oración, de un párrafo o de un texto que tiene sentido en sí mismo

El punto se utiliza para dar fin a una oración e indicar una pausa entre las distintas ideas que se expresan

GRUPO 98:

El punto permite caminar despacio a ambos equipos, ya que el Pontevedra se aleja de las primeras posiciones y el Dépor B no logra huir todavía de las posiciones de descenso

el punto permite al equipo de Schuster, que vio el partido en la grada por sanción, regresar a la zon UEFA junto con el Atlético y a costa del Recreativo, que sucumbió hoy en Santander ante el Racing en un partido lleno de sorpresas

el punto permite al equipo de Schuster, que veía el partido en la grada por sanción, regresar a la zona UEFA junto con el Atlético y a costa del Recreativo, que ha sucumbido en Santander ante el Racing en un partido lleno de sorpresas (4-3)

El punto permite al equipo de Quique Sánchez Flores situarse tercero a diez del líder, mientras que el rojiblanco lleva ya nueve partidos sin ganar y está sumido en la medianía de la tabla

Los puntos sirven para armar un ranking de atletas de la modalidad de kata (la de Antonio Díaz) y uno de kumite, sin distinción de sexos, y en individual y equipos

GRUPO 102:

El punto permite al líder seguir con holgada ventaja sobre el segundo

El punto permite al líder seguir con holgada ventaja sobre el segundo

GRUPO 112:

el punto permite a los linenses continuar en zona de ascenso pero con solo un punto de ventaja sobre el Sanluqueño

el punto permite a los de Ortuondo continuar en la zona privilegiada de la clasificación después de un encuentro donde los ceutíes pusieron las ganas y las oportunidades mientras que el Extremadura puso el orden y la entrega

GRUPO 114:

el punto se emplea, también, como marcador de la posición de los decimales (en inglés, por ejemplo).

El punto permite también al Atlético mantener la cuarta posición, gracias a la derrota del Racing y al empate del Sevilla

GRUPO 119:

los puntos permiten obtener vuelos y estancias en hoteles de todo el mundo a los mejores precios

Los Puntos permiten a todos en Xbox LIVE -membresías Silver y Gold- juntar y mejorar la experiencia de entretenimiento, abriendo el Mercado a todo un mundo conectado con Xbox LIVE

GRUPO 120:

```
Los puntos permiten a la comunidad hacerse una idea del grado de implicación
de cada usuario y de la ayuda prestada al portal, además de poder presumir
ante tus amigos

Los puntos sirven pa estar arriba de la tabla y poder presumir de ser el que
más trampas haces.


------  34 definiciones recuperadas en 10 grupos   ------
```

Para comenzar con los resultados cualitativos, se muestra la figura 4.4 con los resultados para el número de grupos, el recuerdo y la precisión en función del valor de corte por distancia.

En la tabla 4.7 se muestran algunos resultados asociados al coeficiente cofenético la tabla 4.8 muestra algunos resultados asociados al coeficiente de inconsistencia.

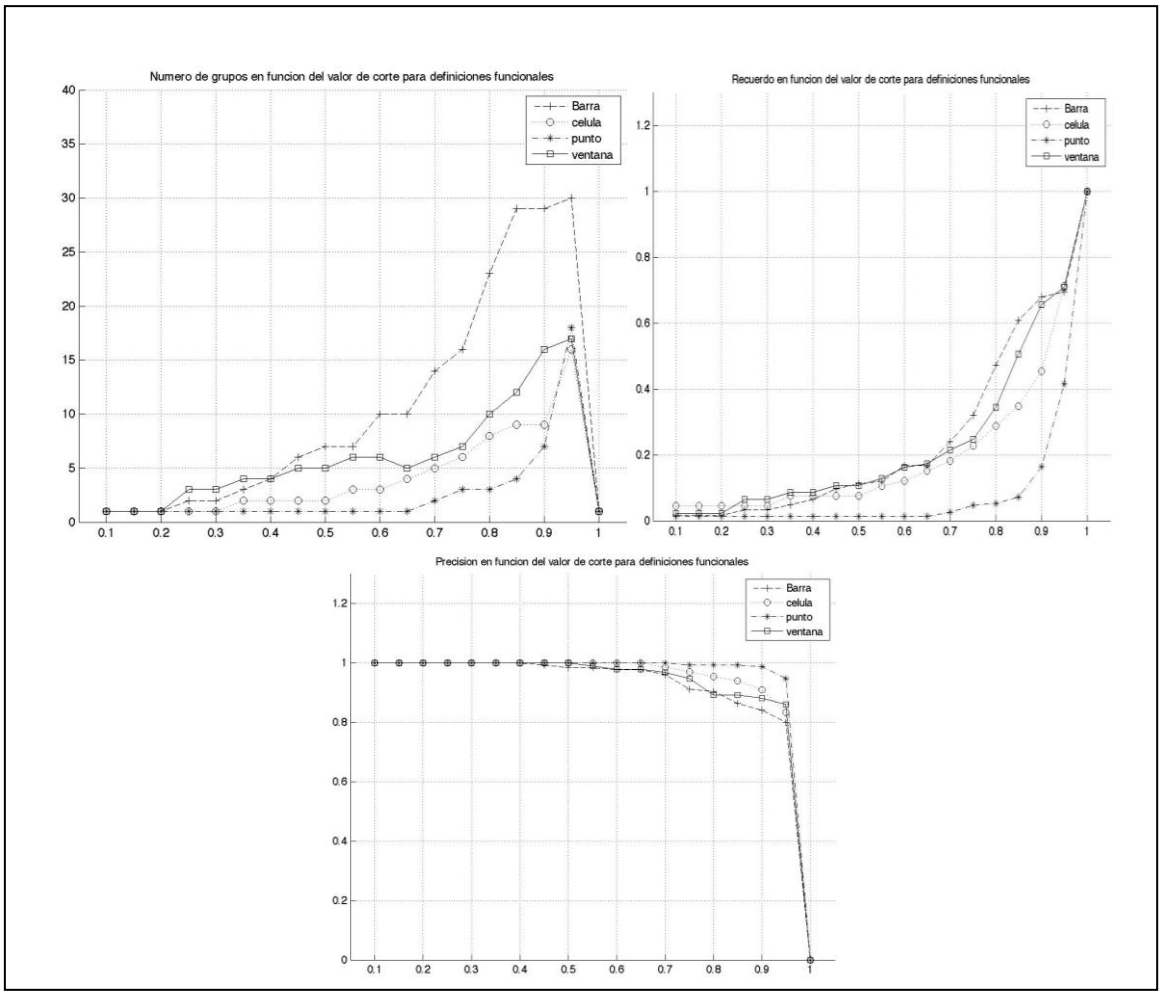

Figura 4.4. Resultados de los agrupamientos para las definiciones funcionales.

	Distancia entre documentos (Energía)			Distancia cofenética			**Coeficiente cofenético**
	máx	**mín**	**prom**	**máx**	**mín**	**avg**	
Barra	0.8646	0	0.6016	0.8646	0	0.6974	0.8570
Célula	1	0	0.9711	1	0	0.9875	0.8496
Punto	1	0	0.9904	1	0	0.9958	0.8646
Ventana	1	0	0.9376	1	0	0.9687	0.7819

Tabla 4.7. Resultados del coeficiente cofenético para las definiciones funcionales.

	Inconsistencia			**Máximo valor de inconsistencia**
	media	**mediana**	**máx(DesvStd)**	
Barra	0.6058	0.7071	0.1625	1.0256
Célula	0.4460	0.5774	0.5427	1.1547
Punto	0.5018	0.7071	0.5667	1.1547
Ventana	0.4767	0.7071	0.5657	1.1209

Tabla 4.8. Resultados del coeficiente de inconsistencia para las definiciones funcionales.

4.7. ANÁLISIS DE RESULTADOS

El análisis cualitativo resulta alentador. En general, se generan agrupamientos que reflejan las distintas acepciones que tiene un término, pero también se distinguen sutilezas existentes en las definiciones del mismo dentro de una misma acepción. Podemos afirmar que el algoritmo agrupa definiciones con precisión, es decir, que logra reunir definiciones estructuralmente distintas pero equivalentes en significado. Considere como ejemplo de este fenómeno las siguientes dos definiciones agrupadas para el término *célula* de tipo extensional con valor de corte 0.8. En la primera definición, abajo mostrada, podemos inferir que la célula se compone de un núcleo que está envuelto en protoplasma y que a su vez el protoplasma está envuelto por una membrana; mientras que la segunda es estructuralmente inversa pero semánticamente equivalente, esto es, la célula consta de una membrana que envuelve al protoplasma en el que está el núcleo.

1. La célula se compone de un núcleo envuelto en protoplasma, alrededor del cual hay una membrana que separa la célula de su medio ambiente
2. La célula consta de una MEMBANA CELULAR que envuelve una masa viscosa y granulosa llamada PROTOPLASMA, en la cual se encuentran todos los ORGANELOS CELULARES, incluido el NÚCLEO

Otro resultado importante es que, a medida que el valor de corte tiende a uno, los grupos se vuelven más específicos y a veces indeseablemente explícitos en el significado de las palabras que los conforman. Tomemos como ejemplo los resultados para el término *punto* de tipo funcional con valor de corte 0.99. Note que la diferencia principal entre los grupos 1 y 2, abajo mostrados, es que en el grupo 1 la función principal del punto es *indicar/marcar/señalar*, mientras que en el grupo 2 la función principal es *terminar/finalizar*. Sería deseable que el algoritmo unificara ambos grupos en uno solo.

GRUPO 1:

El punto se emplea para indicar el final de una oración

se dice que el punto se emplea al final de una oración para indicar que lo que la precede forma sentido completo

El punto: Se emplea para indicar el final de una oración o cuando se ha terminado de exponer una idea en un escrito (idea completa)

El Punto (.): Se emplea para indicar el sentido de una oración, simple o compuesta, aunque se siga tratando el mismo asunto

El punto se emplea para indicar el final de una oración, señala que lo escrito antes tiene sentido completo

El punto se emplea para separar oraciones, indica una pausa larga en la lectura y debe marcarse con entonación descendente

El punto se utiliza para señalar una pausa larga, que marca el final de una frase u oración.

El punto se utiliza para señalar el final de una oración, de un párrafo o de un texto que tiene sentido en sí mismo

El punto se utiliza para dar fin a una oración e indicar una pausa entre las distintas ideas que se expresan

El punto se utiliza para indicar versículos o capítulos que no forman un rango

GRUPO 2:

El punto sirve para terminar una frase con sentido completo

El punto (.) .- Se emplea al final de cada frase.

El punto se usa para terminar la oración y las abreviaturas

los puntos se usan para terminar una oración

También se observó que, a medida que se incluyen más definiciones en el agrupamiento, se introduce más *ruido* en los grupos generados. Nos referimos a los intrusos que aparecen en los grupos generados y que son incongruentes con la mayoría de las definiciones de un grupo. Este fenómeno podemos atribuirlo directamente al uso de un algoritmo de tipo jerárquico. Tomemos como ejemplo los resultados para el término *punto* de tipo analítico. Cuando el valor de corte es 0.66, el algoritmo constituye el siguiente grupo:

GRUPO 72:

En su costado más simple, el punto es una manera de anudar de hilos en series paralelas, entrelazando las filas para formar una tela tejida

En su lado más simple, el punto es una manera de entrelazar los hilos en series paralelas, formando filas y finalizar con una tela tejida

Al aumentar el valor de corte a 0.67, se agrega a este grupo una definición más que parece incongruente con la acepción de *una manera de anudar/entrelazar hilos*, la definición mencionada es:

En gráficos de ordenador o computadora e imprenta, un punto es una pequeña 'mancha' combinada con otras en una matriz de filas y columnas para formar un carácter o un elemento gráfico de un dibujo o diseño

Se puede verificar que para cualquier valor de corte mayor a 0.67, esta última definición nunca vuelve a ser separada de las del grupo 72 mostradas anteriormente. Atribuimos esto al uso de un algoritmo de tipo jerárquico aglomerativo dado que en este tipo de algoritmos, una vez que un objeto es integrado a un grupo ya no puede ser reasignado a otro grupo ni ser eliminado del agrupamiento. Sin embargo, cabe mencionar que en general, la precisión comienza a disminuir solamente hasta que el valor de corte es superior a 0.65.

También se puede verificar que para obtener como recuerdo al menos la mitad de la colección es necesario que el valor del corte sea superior a 0.85, lo cual conllevará a una disminución en la precisión.

Con respecto a los resultados en la precisión, podemos decir que en general el algoritmo mantiene un valor muy alto hasta que el valor de corte llega a 0.90, valor en el cual la precisión se empieza a ver rápidamente afectada.

Teniendo en cuenta los resultados, se puede considerar que independientemente del tipo de definición, el comportamiento del algoritmo puede ser dividido en tres zonas según el valor de corte:

1. Zona 1. Valores de corte entre 0.0 y 0.70: En esta zona se obtiene muy alta precisión (más de 90% aprox.) y muy bajo recuerdo (menos de 40% aprox.). Se debe utilizar un valor de corte en este intervalo cuando se quieran obtener acepciones comunes, pocos grupos (5 grupos aprox.) con pocas definiciones y sin la presencia de intrusos en los grupos generados.

2. Zona 2. Valores de corte entre 0.75 y 0.85: Esta zona se caracteriza por tener precisión alta (80% aprox.) y recuerdo intermedio (50% aprox.). Este intervalo representa un buen compromiso entre la precisión y el número de grupos generados (10 grupos aprox.).

3. Zona 3. Valores de corte entre 0.95 y 0.99: Esta zona proporciona precisión media (50% aprox.) y recuerdo muy alto (80% aprox.). El número de grupos generados es generalmente excesivo pero cada grupo es muy preciso en el significado (20 grupos aprox.).

5. CONCLUSIONES Y TRABAJO FUTURO

SÍNTESIS: Con base en los resultados obtenidos en la aplicación del algoritmo de agrupamiento semántico al corpus de términos polisémicos en español podemos concluir que hemos logrado los objetivos específicos planteados al inicio de este trabajo:

- Se definió la metodología con base en las características de los contextos definitorios.
- Se utilizaron diversos métodos estadísticos para la resolución del problema planteado.
- Se consolidó un corpus apropiado para la experimentación.
- Se implementó el algoritmo en diversos módulos programados en lenguajes de programación de alto nivel.
- Se analizaron los resultados de la aplicación del algoritmo al corpus tanto a nivel cualitativo como cuantitativo.

De igual manera, observamos que el algoritmo desarrollado agrupa, con muy buena precisión, un conjunto de definiciones para un mismo término de acuerdo con su acepción.

El valor de corte por distancia recomendado para el algoritmo, según las evaluaciones, debe estar entre 0.75 y 0.85.

5.1. APORTACIONES TEÓRICAS

Dentro de las aportaciones de carácter teórico, está la descripción del algoritmo en notación de seudocódigo (Algoritmo 3.2) ya que esto permite que sea fácilmente implementado en cualquier lenguaje de programación de alto nivel.

También es de destacar una primera aproximación para la generación de una medida de distancia entre textos basada en energía textual. Creemos que es conveniente un análisis más exhaustivo al respecto.

5.2. APORTACIONES PRÁCTICAS

La aportación práctica más importante del presente estudio es la implementación del algoritmo de agrupamiento. El algoritmo aquí descrito puede ser utilizado directamente para agrupamiento de resultados de motores de búsqueda (*snippets*), ya que es independiente del idioma y no requiere de ningún tipo de anotación lingüística, como serían las etiquetas POS (*parts of speech*). Tampoco requiere un conjunto de entrenamiento previo ni es necesario indicar el número de grupos a generar y, a diferencia de otros algoritmos similares como *lingo*

(Osinski *et al.*, 2004), nuestro algoritmo es fácilmente configurable, pues depende únicamente de un parámetro: el valor de corte por distancia.

Otra aportación es la creación del corpus de términos polisémicos. Como mencionamos anteriormente, los contextos definitorios son una estructura recientemente estudiada y los corpus que de ellos pueden encontrarse son escasos. El corpus desarrollado para la experimentación de este estudio representa por si mismo un aporte para aquellos que deseen consultarlo y analizarlo.

5.3. TRABAJO FUTURO

El vínculo más inmediato del estudio aquí presentado con el trabajo futuro será la adaptación del módulo de agrupamiento al sistema *Describe*. Con ello pretendemos poner en producción tanto el módulo de agrupamiento como algunos otros desarrollos pertenecientes al Grupo de Ingeniería Lingüística de la UNAM (GIL-UNAM), entre los cuales destacan: el extractor de contextos definitorios (*ecode*), el extractor de candidatos a CDs mediante la API de Yahoo! y un etiquetador POS basado en el de Brill y adaptado al español de México (Méndez, 2009).

Creemos que los resultados obtenidos tanto a nivel teórico como práctico son alentadores y que valdría la pena darle seguimiento al trabajo aquí presentado. Por supuesto que existen diversas mejoras posibles al algoritmo; una de ellas sería el desarrollo de un esquema de pesos especializado para CDs, uno que incluya consideraciones de tipo lingüístico. Otra mejora posible es el desarrollo de un algoritmo para la selección de frases representativas de los grupos, o mejor aún, para la generación automática de etiquetas de los grupos (*automatic labeling*, *topic detection*). Esta última aplicación conlleva un estudio amplio en la temática de compresión de frases, una técnica de reciente estudio y que puede ser explotada para diversas tareas de procesamiento de lenguaje natural como la generación de resumen automático.

Con éste panorama, el Grupo de Ingeniería Lingüística (UNAM-México) en colaboración con el *Laboratoire Informatique d'Avignon* (UAPV-Francia) han sometido a consideración ante la convocatoria Conacyt-Ecos 2009 el proyecto titulado: *compresión automática de frases*, en el cual el autor de esta tesis, presentará el doctorado bajo la dirección del Dr. Juan Manuel Torres-Moreno y el Dr. Gerardo Sierra Martínez y en colaboración con otros investigadores como el Dr. Florian Boudin, el Dr. Rodrigo Alarcón y el Dr. César Aguilar. Estamos seguros que este vínculo fortalecerá de manera importante el desarrollo de diversas áreas especializadas en tecnologías del lenguaje en México.

REFERENCIAS:

1. Alarcón, R. (2006). *Primeras aproximaciones a la extracción automática de contextos definitorios*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

2. Aguilar, C. (2009). *Análisis lingüístico de definiciones en contextos definitorios*. México: Universidad Nacional Autónoma de México.

3. Alameda, J. R. & Cuetos, F. (1995). *Diccionario de Frecuencia de las unidades lingüísticas del catellano,* (Vols. I y II). Servicio de Publicaciones. Universidad de Oviedo.

4. Alarcón, R. & Sierra, G. (2003). "El rol de las predicaciones verbales en la extracción automática de conceptos". *Estudios de Lingüística Aplicada*, 38, 129 - 144. México: Centro de Enseñanza en Lenguas Extranjeras, Universidad Nacional Autónoma de México.

5. Alarcón, R. & Sierra, G. (2006). "Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios". *Avances en la Ciencia de la Computación*, VII Encuentro Nacional de Ciencias de la Computación. Hernández, A., Zechinelli, J.L. (eds). San Luís Potosí: MSCC.

6. Alarcón, R. (2003). *Análisis lingüístico de contextos definitorios en textos de especialidad*. Tesis de licenciatura. México: Universidad Nacional Autónoma de México.

7. Alarcón, R. (2009). *Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECCODE (extractor de candidatos a contextos definitorios)*. Tesis de Doctorado. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

8. Alarcón, R., Bach, C. & Sierra, G. (2008). "Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica". *Revista de la Sociedad Española de Lingüística*, 37, 247 - 278. Madrid: Sociedad Española de Lingüística.

9. Bell, E. (1934). "Exponential Numbers". *Amer. Math. Monthly*, 41, 411-419.

10. Benítez, V. (2008). *Anáforas en la expansión de contextos definitorios: una propuesta de etiquetado*. Tesis de licenciatura. México: Universidad Nacional Autónoma de México.

11. Binder, K. (2001), "Ising model", *Hazewinkel*, Michiel, Encyclopaedia of Mathematics, Kluwer Academic Publishers, ISBN 978-1556080104.

12. Cha, S. (2008). "Taxonomy of Nominal Type Histogram Distance Measures". *American Conference on Applied Mathematics*. Massachusetts.

13. Cook, S., A. (1971). "The complexity of theorem proving procedures". *Proceedings, Third Annual ACM Symposium on the Theory of Computing*, ACM, New York.

14. Cormen, Leiserson, & Rivest. (1990) *Introduction to Algorithms*, Greedy Algorithms.

15. Duda, R. O., Hart, P. E. & David G. S. (2000). *Pattern Classification*. (2nd Edition). Wiley-Interscience. 274, 357, 503, 505, 511.

16. Duda, R. O., Hart, P. E. & Stork, D. G. (2001). "Unsupervised Learning and Clustering", *in Pattern classification (2nd edition)*, p. 571, Wiley, New York, ISBN 0-471-05669-3.

17. Dunn, J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *Journal of Cybernetics* , 32-57.

18. Ester, M., Kriegel, H., Sander, J., & Wimmer, M. (1998). “Incremental Clustering for Mining in a Data Warehousing Environment”. *Proceedings of 24th VLDB conference*.

19. Estopà, R. (2001). “Elementos lingüísticos de las unidades terminológicas para su extracción automática”. *La terminología científico–técnica: reconocimiento, análisis y extracción de información formal y semántica*, 67 – 80, Cabré, M. T. y Feliú, J. (eds.). Barcelona: Institut Universitari de Lingüística Aplicadac, Universitat Pompeu Fabra.

20. Everit, B., Landau, S. & Leese, M., (2001). *Cluster Analysis*, London: Arnold, 4 edición. the oxford university press.

21. Fernandez, s., SanJuan, E. & Torres-Moreno, J. (2007). “Textual Energy of Associative Memories: performants applications of ENERTEX algorithm”. *Text summarization and topic segmentation*, 861-871. Aguascalientes: MICAI 2007.

22. Fernandez, S., SanJuan, E. & Torres-Moreno, J. (2008). *Enertex : un système basé sur l'énergie textuelle*. Avignon: TALN 2008.

23. Fernandez, S., SanJuan, E. & Torres-Moreno, J.M. (2007a). "Energie textuelle de mémoires associatives". *Conference TALN 2007*. Toulouse (France), 2007 5-8 june. Pages 25-34.

24. Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus H. & Zubrzycki S. (1951). “Sur la liason et la division des points d’un ensemble fini”. *Colloquium Mathematicae* 2: 282-285.

25. Gower, J. & Legendre, P. (1986), “Metric and ecuclidean properties of dissimilarity coefficients”. *Journal of classification*, 5, 5-48.

26. Grossman, D. & Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer. ISBN 1402030037, 978140203003.

27. Guha, S., Rastogi, R., & Shim, K. (1998). An efificient Clustering Algorithm for Large Databases. *Proceedings of the ACM SIGMOD Conference*.

28. Halkidi, M., Batistakis, Y., & Vzirgiannis, M. (2001). “On Clustering Validation Techniques”. *Journal of Intelligent Information System* , 107-145.

29. Hebb, D.O. (1949), *The organization of behavior*, New York: Wiley.

30. Hernández, A. (2009). *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. Tesis de licenciatura. México: Universidad Nacional Autónoma de México.

31. Hopfield, J . (1982). “Neural networks and physical systems with emergent collective computational abilities”. *Proceedings of the National Academy of Sciences of the USA*, 9, 2554–2558.

32. Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Englewood, Cliffs, NJ: PrenticeHall. 384, 503, 506

33. Jain, A.K., Murty, M. N. & Flynn, P. J. (1999). “Data clustering: a review”. *ACMComput. Surv*, 31:264–323. 384, 504, 506, 508.

34. Jardine, N., & vanRijsbergen, C. J. (1971). “The use of hierarchic clustering in information retrieval”. *Information Storage and Retrieval*, 7:217–240. 357, 506, 510.

35. Kohonen, T. (1995). “Self-Organizing Maps”. *Series in Information Sciences* ,Vol. 30. Springer .

36. MacQueen, J. (1967). “Some Methods for Classification and Analysis of Multivariate

Observations". *Proceedings of 5th Berkley Sysmposium on Mathematical Statiscs and Probability*, 281-297.

37. Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : The MIT Press.

38. Méndez, C. (2009). *Adaptación del método Brill para el etiquetado morfosintáctico del Español del Siglo XVI*. Tesis de maestría, UNAM.

39. Ng, R., & Han, J. (1994). *Effecient and Effective Clustering Methods for Spatial Data Mining*. Santiago, Chile.

40. Osinski, S., Stefano, J. W. & Weiss, D. (2004). "Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition". *Intelligent Information Systems 2004*. 359-368

41. Pearson, J. (1998). *Terms in context*. Ámsterdam: John Benjamin's.

42. Porter, M. (1980). "An algorithm for suffix stripping". *Readings in information retrieval*. San Francisco CA: Morgan Kaufmann Publishers Inc.

43. Rijsbergen, C. J. (1979). Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow.

44. Rodríguez, C. (1999). *Operaciones Metalingüísticas Explícitas en Textos de especialidad*. Trabajo de investigación. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

45. Salton, G. (1971). *The SMART Retrieval System: Experiments in automatic document procesing*. NJ : Prentice Hall, Englewood Cliffs.

46. Sierra, G. & McNaught J. (2000). "Extracting semantic clusters from MRDs for an onomasiological search dictionary". *International Journal of Lexicography*, Vol. 13 (4).

47. Sokal, R. & Michener, C.D. (1958). *A statistical method for evaluating systematic relitionships*. University of Kansas Science Bulletin, 38, pp. 1409-1438.

48. Sokal, R. R. & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. Taxon, 11:33-40

49. Sorensen, T. (1948), "A method of estimating groups of equal amplitude . *plant sociology based on similarity of species content*, Biologiske Skrifter , 5, 1-34.

50. Spärck, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation*, 28 (1), 11–21.

51. Zahn, C. T. (1971). "Graph-theoretical methods for detecting and describing Gestalt clusters." IEEE Transactions on Computers, Vol. C-20, Issue 1, , pp. 68–86.

52. Zhang, T., Ramakrishnman, R., & Linvy, M. (1996). "BIRCH: An Efficient Method for Very Large Databases". *ACM SIGMOD*. Montreal, Canada.

53. Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge Mass.