INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA
UNIVERSITAT POMPEU FABRA

Programa de doctorado: Lingüística Aplicada
Bienio 2001-2003

Tesis doctoral

Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente

John Jairo Giraldo Ortiz

Tesis doctoral
Para optar al título de doctor por la Universitat Pompeu Fabra

Dirigida por: Maria Teresa Cabré Castellví

UNIVERSITAT POMPEU FABRA

Barcelona, 2008

Dipòsit legal: B.37271-2008
ISBN: 978-84-692-0972-1

A ti.

Índice

Agradecimientos

Ningún proyecto que se emprenda en la vida es fácil y menos aún su conclusión. Hoy llego al final de una etapa en mi vida profesional que seguramente dará paso a nuevos y mayores retos. Es tiempo pues de hacer un alto y recordar múltiples pasajes de mi vida personal y profesional reciente. En todos y cada uno de ellos han intervenido personas que, como si de una obra de teatro se tratase, han entrado y salido de escena. Gracias a todas ellas he podido interpretar mi papel y ha llegado la hora de hacer público mi agradecimiento a todas ellas porque sin su concurso mi papel habría sido mucho más difícil de hacer.

Deseo expresar mi profundo agradecimiento a todas aquellas personas e instituciones que me han acompañado y apoyado a lo largo de los años que ha durado esta investigación. Gracias a todas ellas por demostrarme su amistad sincera, por ayudarme a sortear momentos difíciles y por brindarme siempre todo el apoyo que he necesitado. Agradezco en primer lugar a mi directora, Maria Teresa Cabré, por haber confiado en mí y por haberme sabido guiar a lo largo de todo este proceso. Así mismo, agradezco a todos mis colegas y amigos del *Institut Universitari de Lingüística Aplicada* por toda su ayuda, constancia y generosidad, en especial a Judit Feliu, Vanesa Vidal, Mariona Barrera, Carme Bach, Jordi Vivaldi, Gabriel Quiroz, Jaume Llopis, Mercè Lorente y Araceli Alonso. Igualmente merecen mención Anne Condamines, Aurélie Picton y Marianne Vergez de l'*Université Toulouse-Le Mirail* quienes me acogieron amablemente en su centro y compartieron conmigo conocimiento y experiencias durante mi pasantía de investigación. También deseo hacer público mi agradecimiento a mis buenos amigos Luisa Zapata y Jaime Peón por su hospitalidad. Gracias a la *Agència de Gestió d'Ajuts Universitaris i de Recerca* de la *Generalitat de Catalunya* por su apoyo económico, necesario para llevar a feliz término este trabajo. Y, por último, pero no menos importante, mi

gratitud a mi familia por todo el apoyo moral y por el amor que siempre me ha irradiado desde la distancia.

Abreviaturas y siglas empleadas

AF	Acronym Finder
CGH	Corpus de genoma humano
CGPP	Corpus general de la prueba piloto
CMA	Corpus de medio ambiente
CT-IULA	Corpus Tècnic del Institut Universitari de Lingüística Aplicada
ECON	Economía
EN	Inglés
FD	Forma desarrollada
FR	Francés
GH	Genoma humano
INF	Informática
LCS	Longest Common Subsequence
LSP	Lenguajes para propósitos específicos
MA	Medio ambiente
NC	Nombre común
NP	Nombre propio
POS	Part of Speech
SCGH1	Subcorpus de genoma humano de nivel de especialidad bajo
SCGH2	Subcorpus de genoma humano de nivel de especialidad alto
SCGH2	Subcorpus de genoma humano de nivel de especialidad medio
SCMA1	Subcorpus de medio ambiente de nivel de especialidad bajo
SCMA2	Subcorpus de medio ambiente de nivel de especialidad medio
SCMA3	Subcorpus de medio ambiente de nivel de especialidad alto
SP	Español
TCT	Teoría Comunicativa de la Terminología
UCE	Unidad de conocimiento especializado
UF	Unidad fraseológica
UMLS	Unified Medical Language System
UO	Unidad oracional
URL	Unidad de reducción léxica
UT	Unidad terminológica

Resumen

Las siglas se consideran un fenómeno de reducción léxica. En la actualidad es fácil encontrar textos especializados repletos de siglas, lo que evidencia su función como variantes denominativas. Por un lado, pueden ser variantes léxicas y semánticas, puesto que se usan como sinónimos de sus formas desarrolladas. Por otro lado, se consideran variantes pragmáticas, ya que su finalidad es facilitar la lectura de los textos por parte de los expertos.

Las siglas han llamado la atención de diferentes colectivos profesionales como lexicógrafos, traductores, bibliotecólogos, lingüistas, informáticos y expertos (principalmente del campo de la biomedicina). Sin embargo, a pesar de que estas representan la forma reducida de términos de origen sintagmático, no se ha prestado suficiente atención a su descripción y análisis desde el punto de vista de la terminología

El estudio de las siglas es relevante en terminología ya que aparecen con frecuencia en los textos especializados. Además, aparte de ser importantes para todos los ámbitos de especialidad mencionados anteriormente, las siglas son especialmente importantes en las tareas de recuperación de información y desambiguación.

En la actualidad, la investigación sobre siglas se desarrolla en dos líneas, a saber: teórico-descriptiva y aplicada. Se destacan algunos trabajos teórico-descriptivos previos como son los de Calvet, 1980; Rodríguez, 1981; 1985; Algeo, 1991 y Fijo, 2003. En el plano aplicado sobresalen los trabajos llevados a cabo por Taghva & Gilbreth, 1999; Pustejovsky, 2001; Larkey, 2002; Dannélls, 2005 y Zahariev, 2004.

Aunque muchos autores han estudiado antes el fenómeno de la siglación, sus investigaciones se han centrado en la lengua inglesa en la gran mayoría de las veces. Por tanto, existe una falta de trabajos similares en otras lenguas como el español. En realidad, sólo se conoce un trabajo, Fijo (2003), que ha abordado este tema bajo la perspectiva de la terminología en español. Además, toda la investigación precedente ha presentado problemas en lo que se refiere a la falta de consenso en el establecimiento del concepto y tipología de las siglas.

El propósito de esta tesis es el análisis y la descripción de las siglas de los ámbitos de genoma humano y medio ambiente. Este trabajo consta de cuatro partes principales, a saber: 1) Definición y tipología de siglas; 2) Análisis siglométrico (estadístico); 3) Análisis lingüístico y 4) Criterios para la identificación semiautomática de siglas en textos en español.

En primer lugar, el capítulo de definición y tipología se ha realizado a partir de una revisión amplia de la bibiliografía existente. La selección de la definición y tipología ha servido posteriormente para la selección de las unidades del corpus. En segundo lugar, el análisis estadístico se ha realizado para determinar la frecuencia de uso de estas unidades y su impacto en la terminología de cada ámbito de especialidad. Los resultados obtenidos en genoma humano y medio ambiente se han comparado con los dominios de informática y economía. En tercer lugar, el análisis lingüístico se ha llevado a cabo para conocer las propiedades lingüísticas de las siglas. Y en cuarto y último lugar, se han establecido los patrones de detección de siglas en español como base para la posterior creación de un sistema de recuperación de estas unidades.

Finalmente, este trabajo propone tres elementos a tener en cuenta cuando se trabaja con siglas. Por una parte, llama la atención sobre la calidad de los bancos de siglas que existen en Internet y propone que dichos recursos se ajusten a los estándares de tratamiento de datos como los que se aplican en recursos similares como los bancos de datos terminológicos. Por otra parte, este trabajo propone varios criterios para el reconocimiento de candidatos a par sigla-forma desarrollada, los cuales deben tenerse en cuenta en el momento de diseñar una herramienta para tal propósito. Por

último, se proponen algunos elementos que aportan las siglas a la caracterización del discurso especializado.

Palabras clave: siglas, terminología, español, corpus, genoma humano, medio ambiente

Abstract

Initialisms are considered a lexical reduction phenomenon. Today it is easy to find specialized texts full of them, a clear evidence of terminological variation. On the one hand, initialisms can be lexical and semantic variants since they are used as synonyms of their expansions. On the other hand, initialisms are seen as pragmatic variants since they are supposed to facilitate the expert's reading.

Initialisms have attracted wide interest of lexicographers, translators, librarians, linguists, programmers and experts (mainly from biomedicine). However, despite initialisms represent the shortening of equivalent terms, no attention has been paid to their description and analysis from the point of view of terminology.

The study of initialisms is relevant in terminology since they appear frequently in specialized texts. Additionally, apart from being important for all the abovementioned subject fields, initialisms are specially important for information extraction and disambiguation.

Currently, research on initialisms has two trends: theoretical description and automatic identification and disambiguation from texts. Previous research on the theoretical description has been carried out by Calvet, 1980; Rodríguez, 1981; Quirk *et. al.*, 1985; Algeo, 1991; Fijo, 2003, among others. Likewise, Taghva & Gilbreth, 1999; Pustejovsky, 2001; Ananiadou, 2002; Larkey, 2002; Dannélls, 2005; and Zahariev, 2004, stand out among the authors that have carried out researches on automatic identification and disambiguation.

Although many authors have studied the initialisms before, their research has been focused on English language in the vast majority of the cases. Therefore, there is a lack of description and analysis in other languages such as Spanish. In fact, there is

only one work, Fijo (2003), that has dealt with this subject under a terminological perspective. In addition, all the previous research has shown inconsistencies regarding initialisms' concept and typology.

The aim of this thesis is the analysis and description of the initialisms in the Genomics and Environment subject fields. It is divided into four main parts: 1) Definition and typology; 2) Corpus quantitative analysis; 3) Corpus qualitative analysis, and 4) Initialisms recognition patterns.

Firstly, definition and typology have been set from a comprehensive checking of the bibliography. It is the starting point to select the initialisms from a corpus of Spanish texts. Secondly, corpus quantitative analysis has been done in order to establish the frequency of use of these units and their impact on each subject field's terminology. The results have been compared with other domains such as Computer Science and Economics. This has shown similar trends between Genomics-Computer Science and Environment-Economics, respectively. Thirdly, corpus qualitative analysis has been carried out to establish the linguistic properties of initialisms. Fourthly, initialisms recognition patterns have been established as a basis for the creation of a further initialisms acquisition system.

Finally, this work proposes three different elements to take into account while dealing with initialisms. Firstly, it calls attention on the quality of current on line abbreviation databanks. It proposes a series of criteria to better their quality. Secondly, it proposes several criteria to recognize initialism-expansion pair candidates in Spanish, and thirdly, it proposes different elements observed in the initialisms that can be useful to add in the characterization of the specialized discourse.

Key words: Initialisms, Terminology, Spanish, Corpus, Genomics, Environment

Introducción

Introducción

> "De tout temps, l'esprit humain, se sentant empêtré dans les longueurs de l'écriture, a cherché à abréger par tous les moyens" (Losson, 1990: 8).

El fenómeno de la reducción de las palabras ha existido desde tiempos inmemoriales. Ya en nuestra era, los romanos hicieron uso frecuente de esta técnica, especialmente en sus manuscritos sobre actividades políticas, económicas y jurídicas. A medida que se desarrolló su cultura, las palabras aumentaron y se volvieron más complejas. El imperio pronto descubrió que, cuando el tiempo y el espacio eran limitados, las inscripciones largas se podían acomodar con mayor facilidad en los pergaminos y en las piedras si se abreviaban las palabras comunes. La cultura y la lengua latinas sirvieron de herramienta básica para la ciencia y la tecnología europeas, de ahí que el hábito de abreviar las palabras haya llegado hasta nuestros días. Dentro de su vasta producción de abreviaciones encontramos: i.e. (*id est*), *e.g.* (*exempli gratia*), S.P.Q.R. (*Senatus Populusque Romanus*), INRI (*Iesus Nazarenus Rex Iudaeorum*), e IMP (*Imperatori*), entre otras.

No obstante, el verdadero auge de las abreviaciones se ha dado desde mediados del siglo XX gracias al rápido avance de la ciencia y de la técnica y a la consecuente generación de información que debe escribirse, almacenarse, recuperarse y transferirse cada vez a mayor velocidad. A menudo, se requiere transmitir la mayor cantidad de conocimiento con el menor número de palabras posible, lo cual se logra, en primer lugar, con las estrategias de emisión de un mensaje. En muchos casos los autores de textos crean unidades de reducción léxica (normalmente siglas) como alternativa para referirse periódicamente a unidades sintagmáticas preexistentes. De esta manera, se reduce el tamaño de los textos (exigencia frecuente de los editores de publicaciones) y se evita la repetición frecuente de este tipo de unidades que hacen pesada la lectura. En segundo lugar, con el desarrollo de la informática y las

telecomunicaciones, gracias a las cuales se ha facilitado el almacenamiento, procesamiento y transmisión rápida de datos. En definitiva, la aparición constante de siglas da fe de que la reducción léxica es un fenómeno vigente, acrecentado por el avance de las ciencias y de la técnica y propio de una lengua influenciada a menudo por factores como la economía lingüística, la mnemotecnia, la estilística o los criterios editoriales.

Estudios como los llevados a cabo por Bloom (2000) confirman el uso creciente de las siglas en el discurso especializado. En el artículo *Acronyms, Abbreviations and Initialisms* este autor ofrece un interesante estudio diacrónico sobre la aparición de las siglas en las publicaciones más prestigiosas del campo de la urología. En él concluye que la presencia de las siglas en los manuales y revistas especializadas ha ido en contínuo aumento a lo largo del siglo XX.

En su estudio, Bloom encontró, por ejemplo, que entre 1917 y 1935 el *Journal of Urology* y el *British Journal of Urology* no incluían ninguna abreviación más allá de las esperadas como cc (centímetros cúbicos), mgm (miligramos), o Fig. (figura). Incluso, en algunas ocasiones, hasta las abreviaciones más habituales se omitían en favor de sus formas desarrolladas, por ejemplo *bacillus* y *milligrams*. En 1935, el manual *The Principles and Practice of Urology* de Hinman sólo presentaba una abreviación, KUB, la forma reducida de *Kidneys, ureters and bladder*. La edición de 1944 del manual *Urological Surgery* de Dodson empleaba algunas abreviaciones estandarizadas como Hb, cc, No., y Fig. En la década siguiente, en 1950, aparecía el volumen 63 del *Journal of Urology*, dedicado casi exclusivamente a *Transurethral Resection of the Prostate* (TURP); aunque sólo se registró el uso de la sigla TUR dentro de una tabla. Una década más tarde, en 1960, esa misma obra volvió a editar un volumen monográfico sobre TURP. En esta ocasión, la sigla TUR apareció cuatro veces en un artículo mientras que en los demás aparecía sólo su forma desarrollada (FD). En 1970 se revierte el escaso empleo de las siglas en las publicaciones del área. En efecto, en este año el *Journal of Urology* publicó cinco artículos sobre TURP, en dos de los culaes se registró una amplia frecuencia de uso de la sigla TUR llegando hasta las 34 ocurrencias. En 1973 se editó el primer número de una nueva

publicación del área titulada *Urology*. El primer artículo mostraba 21 ocurrencias de seis siglas diferentes; situación que se asoció a la aparición por aquel entonces de la biología molecular y su impacto en la medicina. Algunas de las siglas recogidas fueron DNA, RNA y ACTH. Para 1990, las siglas TUR y TURP ya se usaban ampliamente en las publicaciones escritas. El volumen 66 del *British Journal of Urology* publicó dos artículos dedicados específicamente a *Transurethral Resection of the Prostate*. En el primero, se definía TURP y posteriormente se empleaba la sigla nueve veces en el texto y seis veces en tablas. Adicionalmente, se contaron otras 18 abreviaciones como CVP (*Central venous pressure*), RISA (*Radioiodinated serum albumin*), NS (*Not significant*), etc. En el segundo, se registró el uso de la sigla TURP cuatro veces en el texto y en una tabla. Se encontraron además cinco siglas especializadas (BHP, TURP, TO, EDTA, UICC), así como 16 abreviaciones de carácter general. En 1996, el *British Journal of Urology* introdujo su versión de la "Piedra Roseta"; es decir, una tabla de abreviaciones en cada número con el ánimo de evitar la coaparición de la sigla y la forma desarrollada dentro de los textos, a pesar de que las convenciones internacionales señalaban que cada abreviación debía ir "definida" en el *abstract*. Por último, Bloom afirma que, a finales del siglo XX, las siglas y los demás tipos de abreviación estaban totalmente arraigados en toda la literatura sobre urología. Una buena muestra de este cambio la encontró en el primer número de *Journal of Urology* del año 2000, el cual salió repleto de abreviaciones. Sólo en los *abstracts* aparecieron 47 abreviaciones diferentes que representaban 285 ocurrencias, siendo las más frecuentes PSA (87 ocurrencias) y BCG (30 ocurrencias).

Si bien es cierto que muchas abreviaciones son efímeras porque obedecen a las necesidades puntuales de un momento dado, también es cierto que muchas otras están destinadas a la perdurabilidad. De todas formas, el uso de estrategias de reducción léxica conlleva el riesgo de que el lector encuentre difícil la comprensión de un texto si no tiene la competencia suficiente sobre el tema. Aunque muchos legos puedan reconocer siglas como ADN, ARN o VIH, para el gran público la mayoría de las siglas pueden resultar indescifrables.

Muchas voces se han declarado abiertamente en contra de la creación y uso indiscriminado de las abreviaciones (*cf*. Morgan, 1985; Green, 1990; Cheng, 1997, 2002, 2005; Walling, 2001; Farber, 2002; Fallowfield, 2002; de Granda, 2003; Guardiola, 2003; Lader, 2002; Fred, 2003; Rowe, 2003; Jack, 2003; Bradley, 2004; Isaacs, 2007). Un buen ejemplo de esta posición lo constituye Garner cuando afirma en el *Oxford Dictionary of American Usage and Style* que "*One of the most irritating types of pedantry in modern writing is the overuse of abbreviations, especially abbreviated names… many writers — especially technical writers… allow abbreviated terms to proliferate, and their prose quickly becomes a hybrid-English system of hieroglyphs requiring the reader to refer constantly to the original uses of terms to grasp the meaning. This kind of writing might be thought more scholarly than ordinary straightforward prose. It isn't. Rather, it's tiresome and inconsiderate writing; it betrays the writer's thoughtlessness toward the reader and a puerile fascination with the insubstantial trappings of scholarship*" (Garner, 2000: 2).

A pesar de la existencia de algunas voces disidentes, la realidad es que las abreviaciones, particularmente las siglas, han ganado terreno en todos los ámbitos. Bloom ha demostrado cómo el uso de estas unidades ha logrado consolidarse en las publicaciones del campo de la urología. En la actualidad, todo tipo de texto, desde una carta hasta un artículo científico es suceptible de contener siglas. Las siglas suelen pasar inadvertidas salvo hasta que se desconoce su significado ya que entorpecen la comprensión del texto. Cada ámbito de especialidad posee siglas específicas en función de su terminología. De ahí que, en un intento por recoger y documentar este tipo de unidades, surjan constantemente recursos electrónicos y en papel tales como diccionarios, glosarios o bancos de datos. Prueba de ello son las bases de datos creadas para almacenar el caudal de siglas provenientes de todos los ámbitos de conocimiento, entre las que se sobresalen *Acronym finder*, *Acrophile*, *Acronym server*, *Wiley InterScience*, *Abbreviations.com*, *Acronym search* y *Acromed*.[1]

1 *Acronym Finder* es un banco de abreviaciones sobre informática, tecnología, telecomunicaciones e industria militar.
Acrophile es un banco de abreviaciones extraídas de casi 1 millón de páginas web de organismos gubernamentales de los EUA.

En definitiva, la formación y uso de unidades de reducción léxica como las siglas se ha convertido en un objeto de estudio interesante por motivos como la proliferación constante (hecho que demuestra la influencia que los hablantes ejercen sobre su lengua por medio de sus conocimientos, necesidades y costumbres), la controversia que generar sobre la conveniencia de su uso, y la falta de consenso alrededor de su concepto y clasificación.

Los fenómenos de reducción léxica, y especialmente la siglación, interesan a campos como: lingüística, discurso, neología, traducción, lexicología, terminología, redacción técnica, enseñanza de lenguas con fines específicos o lingüística computacional. Dentro del discurso especializado, y concretamente de la terminología en español, existen pocos trabajos dedicados al estudio de las siglas, una de las formas de reducción léxica más frecuentes.[2] Por consiguiente, esta tesis aborda el estudio de la siglación desde la óptica del discurso especializado de genómica y medio ambiente con un propósito doble. Por una parte, avanzar en el estudio de este tipo de unidades en el discurso especializado en español y, por otra parte, establecer los criterios para la detección y extracción de estas unidades en español.

Esta investigación se divide en nueve capítulos, a saber:

El primer capítulo presenta los resultados de la fase de exploración y delimitación del tema objeto de estudio. El segundo capítulo se ha dedicado al establecimiento de la definición y tipología de las siglas. El tercer capítulo muestra los aspectos teóricos en

---

Acronym server es un banco de abreviaciones de diversos ámbitos.
Wiley InterScience es un banco de abreviaciones pertenecientes a 17 ámbitos de especialidad diferentes como aeronáutica, arquitectura, ingeniería civil, etc.
Abbreviations.com es un banco de abreviaciones pertenecientes a 10 categorías que contienen 132 subcategorías diferentes.
Acronym search es un banco de abreviaciones procedentes de ámbitos como informática, industria militar, finanzas, contabilidad, aeronáutica, deportes, etc.
Acromed es un banco de abreviaciones del área de la biomedicina, recogidas a partir de los resúmenes de las publicaciones existentes en *Medline*.

2 Entendemos por unidad de reducción léxica (o abreviación) todo fenómeno de acortamiento léxico en general.

los que se ha enmarcado esta tesis. El cuarto capítulo presenta un análisis de las siglas desde el punto de vista de los diccionarios en línea. El quinto capítulo detalla la metodología aplicada en esta investigación. El sexto capítulo muestra el análisis estadístico a partir del estudio del corpus de siglas constituido para este estudio. El séptimo capítulo presenta el análisis lingüístico de las siglas con mayor frecuencia dentro del corpus de trabajo. El capítulo octavo se dedica a la evaluación de los sistemas de extracción de siglas existentes y a los criterios de diseño de un modelo de extracción de siglas para el español. Finalmente, el capítulo noveno contiene las conclusiones generales y las líneas de trabajo futuro.

Capítulo 1

Capítulo 1

Antecedentes de esta investigación

> "El grado de especialización de la comunicación condiciona no solo la densidad terminológica de un texto, sino también la cantidad de variación expresiva para hacer referencia al mismo concepto. Un texto altamente especializado suele ser preciso, conciso y sistemático; la terminología que utiliza tiende a la monosemia y a la univocidad. A medida que disminuye el grado de especialización, el discurso va adquiriendo características que lo acercan al discurso no especializado: en el plano semántico, variación conceptual, redundancia, ambigüedad, falta de precisión estricta; en el plano de la expresión, un alto índice de sinonimia, pero sobre todo un uso muy alto de fórmulas parafrásticas que explican analíticamente el mismo concepto que en un nivel especializado sería designado inequívocamente por un término" (Cabré, 1999: 171).

1. Fase de exploración del tema

La realización de esta tesis ha estado precedida por un trabajo de exploración del tema titulado "Siglas y variación vertical en textos sobre genoma humano y medio ambiente", que se expone a continuación.

Dentro de la comunicación especializada se produce gran variedad de textos orales y escritos que corresponden a distintos niveles de especialización en función de los receptores (expertos, aprendices de experto o gran público) y de las funciones que persigue este discurso (discutir, enseñar o divulgar el tema). De acuerdo con estos dos parámetros, que están en la base de la variación vertical del discurso especializado, se distinguen tres tipos de textos:

1) De especialización alta (producidos por expertos y destinados a expertos de la misma área);

2) De especialización media, a veces denominados didácticos (producidos por expertos y destinados a formar a futuros expertos), y
3) De especialización baja, comúnmente denominados de divulgación (producidos por expertos o mediadores lingüísticos y destinados al gran público interesado en una materia o tema).

En el gráfico 1 se ilustra cada uno de los niveles de especialización de los textos y su correspondencia con el tipo de destinatario.

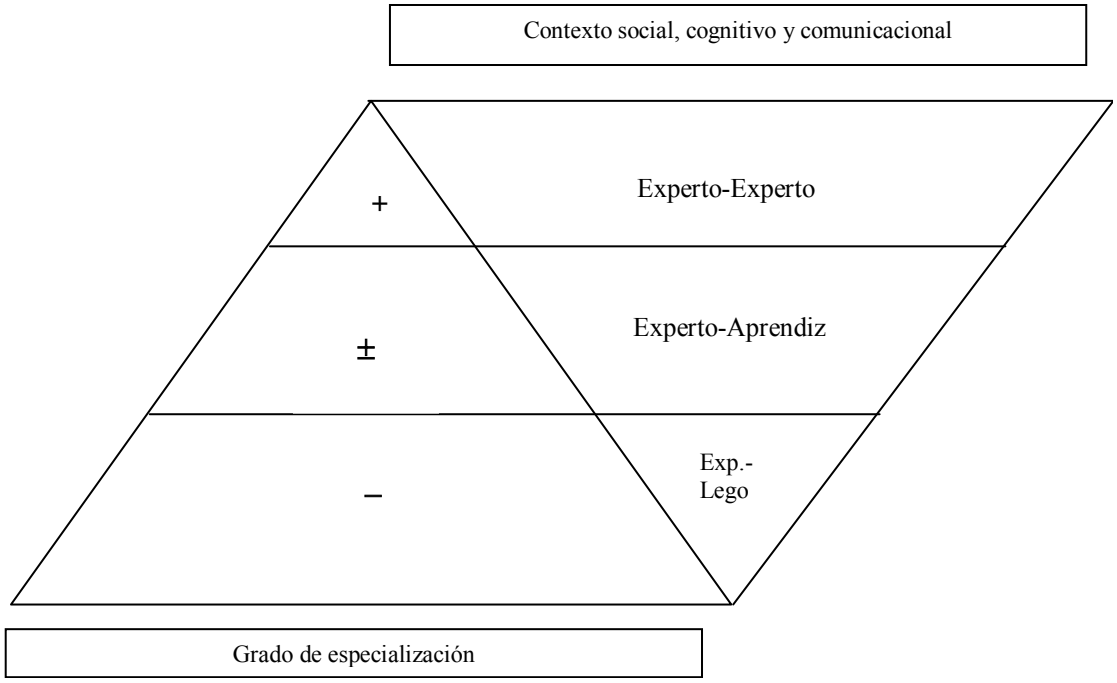


Gráfico 1. Niveles de especialización de los textos

Partimos de la hipótesis de que la frecuencia, el tipo de siglas y la presencia de sus formas desarrolladas en el texto eran factores que permitían distinguir los textos de uno u otro nivel de especialización. Es decir, las siglas podrían ser un factor discriminante del nivel de especialización de los textos dentro de la variación vertical del discurso.

Las siglas de cada ámbito de conocimiento resultan de difícil comprensión para el lector, en especial para el lego. Generalmente, esta limitación se da por dos motivos.

En primer lugar, porque la sigla encapsula un sintagma pleno, lo que genera opacidad cuando se desconoce la relación de equivalencia entre dicho sintagma y la sigla. En segundo lugar, porque la sigla puede generar ambigüedad cuando se desconoce el verdadero significado dentro del contexto en el que se encuentra. A pesar de que este problema se puede evitar con la inclusión de la forma desarrollada, ésta puede originar otro tipo de fenómenos como la variación denominativa y la redundancia dentro del texto.

Empleamos el término genérico "variación denominativa" en el sentido de la alternancia o coaparición de un par sigla-forma desarrollada dentro de un texto. Este tipo de variación puede clasificarse en:

1) Variación formal. Se trata propiamente de la alternancia de la sigla y su forma desarrollada; *e.g.*: PCR/*Polymersase Chain Reaction*.
2) Variación por traducción. Pueden darse dos casos: a) cuando se traduce la sigla; *e.g.*: USA por EUA; b) cuando se traduce la forma desarrollada; *e.g.*: PCR (*Polymersase Chain Reaction*) por PCR (reacción en cadena de la polimerasa).
3) Variación por grafemas. Se trata del uso indistinto de los caracteres en mayúscula o minúscula que conforman la sigla; *e.g.*: YACS/YACs; RNAsa/Rnasa.

Los objetivos que se trazaron para la fase de exploración fueron:

1) Analizar el comportamiento de las siglas y de sus formas desarrolladas en textos sobre genoma humano (GH) y medio ambiente (MA) de distinto nivel de especialización.

2) Analizar comparativamente los datos en cada uno de los niveles de especialización de los corpus GH y MA.

Con estos objetivos se pretendía, por un lado, validar y afinar la hipótesis sobre las siglas como factor discriminante de la variación vertical del discurso especializado y, por otro lado, formular nuevas hipótesis a la luz de los datos experimentales.

Para cumplir con los objetivos marcados se diseñó una prueba piloto. Con ella se buscaba determinar si la incidencia de las siglas y sus formas desarrolladas permitía diferenciar los niveles de especialización de los textos de un ámbito específico. En concreto, esta prueba se limitó al análisis de la frecuencia y de la variación denominativa en un corpus de textos escritos de diferente nivel de especialización.

1.1 Metodología

La metodología para llevar a cabo la prueba piloto comprendió cinco fases principales, a saber:

1) Constitución del corpus. A partir del Corpus Técnico del Institut Universitari de Lingüística Aplicada (CT-IULA) se constituyó el corpus general de la prueba piloto (CGPP), que se dividió en corpus de genoma humano (CGH) y corpus de medio ambiente (CMA). Cada corpus estaba conformado por tres subcorpus (SC) pertenecientes a cada uno de los niveles de especialización: (SCGH1), (SCGH2), (SCGH3), (SCMA1), (SCMA2) y (SCMA3).
2) Obtención de las unidades de análisis. A cada subcorpus de texto se le aplicó una expresión regular para obtener los candidatos a sigla y a sigla-forma desarrollada.
3) Análisis y depuración de las unidades. De acuerdo con la definición y tipología de siglas establecidas previamente, se analizaron de forma manual los listados de candidatos para seleccionar las siglas y eliminar el ruido existente. En este paso también se observó el comportamiento de las siglas y de sus formas desarrolladas en cada ámbito y nivel de especialización.
4) Vaciado de las unidades seleccionadas. Las siglas resultantes se almacenaron en una base de datos.

5) Análisis y cotejo de los datos. Se comparó la información por niveles y ámbitos de especialización.

El gráfico 2 muestra los pasos metodológicos que se han seguido.

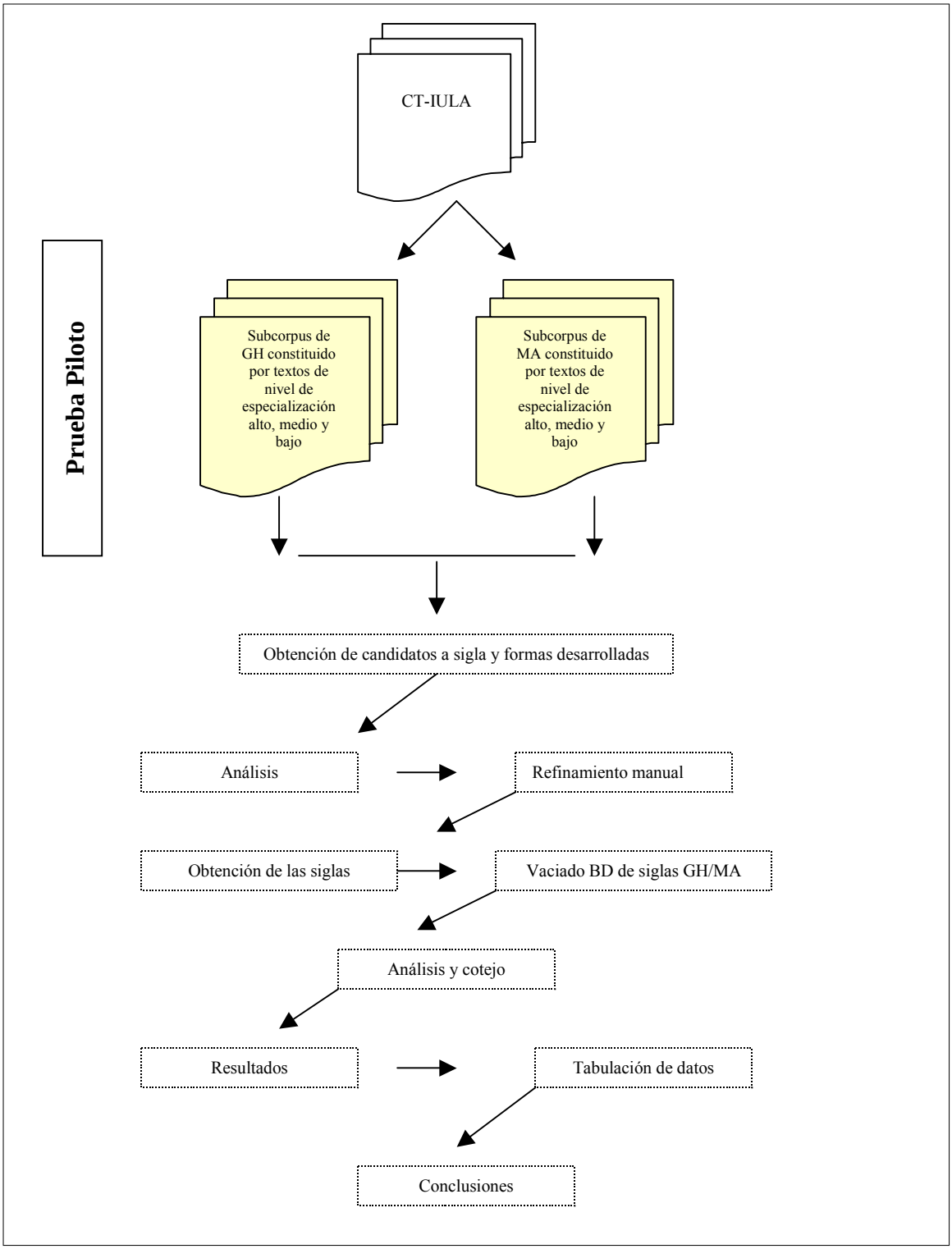


Gráfico 2. Metodología de la prueba piloto

1.1.1 Constitución del corpus

Para constituir el corpus de la prueba piloto se utilizó el Corpus Técnico del IULA (CT-IULA), el cual posee documentos marcados, lematizados y analizados morfológicamente.

El CT-IULA cuenta con una base de datos que describe las características de cada documento incorporado. Se partió de esta base de datos para la selección previa de los textos a analizar. La información básica que presenta cada documento es:

1) Formato (papel o electrónico)
2) Lengua
3) Título
4) Número de muestras
5) Número de palabras
6) Subárea temática, etc.

Los textos que conforman los corpus de GH y MA pertenecen a diferentes subcampos tal y como se muestra en la siguiente tabla.

Campo de especialidad	Subcampos de especialidad	Tipo de documento
GH	Estructura interna Ingeniería genética Biotecnología Enfermedades Inmunología Farmacogenoma Neurociencia Filogenia Eugenesia Diferenciación Investigación genética	Artículo científico Tesis/tesina Manual Artículo de divulgación
MA	Medio natural Asentamientos humanos Impacto ambiental Política y derecho ambiental Ciencia y tecnología ambiental y energética Medio social y organizaciones	

Tabla 1. Selección de áreas temáticas y tipos de documentos

Los textos y, en consecuencia, los subcorpus seleccionados se clasificaron por niveles de especialidad alto, medio y bajo. Para ello se empleó la siguiente información:

1) Código del documento en la BD
2) Código *.sgm* del documento
3) Número de palabras
4) Título del documento
5) Tipo de publicación (revista, manual, tesis, actas, artículo de prensa, etc.)
6) Grado de especialización del texto.

Como se ha mencionado anteriormente, el corpus general de la prueba piloto (CGPP) se formó a partir de un corpus de textos sobre genoma humano (CGH) y otro de textos sobre medio ambiente (CMA). Estos dos corpus estaban formados a su vez por tres subcorpus, correspondientes a los tres niveles de especialidad. Los corpus de genoma humano (CGH) y medio ambiente (CMA) sumaban 129.772 y 126.374 palabras, respectivamente.

La siguiente tabla presenta las características del corpus empleado.

Corpus general de la prueba piloto (CGPP)		
Corpus de genoma humano (CGH)	**Nº de palabras en cada subcorpus**	**Nº total de palabras**
Subcorpus de genoma humano nivel bajo (SCGH1)	46.352	
Subcorpus de genoma humano nivel medio (SCGH2)	44.509	
Subcorpus de genoma humano nivel alto (SCGH3)	38.911	129.772
Corpus de medio ambiente (CMA)		
Subcorpus de medio ambiente nivel bajo (SCMA1)	43.469	
Subcorpus de medio ambiente nivel medio (SCMA2)	44.293	
Subcorpus de medio ambiente nivel alto (SCMA3)	38.612	126.374
Nº total de los corpus de GH y MA		256.146

Tabla 2. Corpus general de la prueba piloto

1.1.2 Obtención de las unidades de análisis

Para la obtención de los datos se utilizó BwanaNet, una herramienta de interrogación que permite, entre otras cosas, hacer búsquedas en línea sobre el CT-IULA y efectuar búsquedas por patrones lingüísticos.[3,4]

BwanaNet es un sistema mayoritariamente lingüístico. Vivaldi & Bach (2003: 1) afirman que "BwanaNet permite consultas no sólo sobre formas, sino también sobre lemas y categorías morfológicas, que pueden combinarse ya que el corpus sobre el cual se hacen las consultas está etiquetado morfosintácticamente". En este trabajo se ha utilizado esta herramienta, ya que en conjunto con las demás herramientas del CT-IULA, permiten un trabajo de extracción y análisis más rápido, cómodo y eficiente.

En nuestro caso la estrategia de búsqueda de las siglas consistió en interrogar a los corpus de GH y MA mediante la formulación de expresiones regulares que permitieran obtener una lista de candidatos a sigla.[5] Se utilizaron dos expresiones regulares para detectar, por un lado, candidatos a sigla y, por otro lado, pares de candidatos sigla-forma desarrollada.

Para la búsqueda de las siglas se establecieron intuitivamente los siguientes patrones:

1) Palabras que contengan varias letras mayúsculas consecutivas o que se combinen con una o más letras minúsculas;
2) Palabras que contengan una mayúscula inicial combinada siempre con uno o más caracteres numéricos y, en algunos casos, con una o más letras minúsculas.

Para obtener los candidatos a sigla se interrogó a BwanaNet mediante la siguiente expresión regular:

3 http://brangaene.upf.es/bwananet/index.htm
4 Véase Vivaldi & Bach (2003: 1).
5 Una expresión regular se usa para indicar que se busca un patrón determinado.

Expresión regular	[pos="N4.*" & word="([0-9a-záéíïóúüñ]*[0-9A-ZÁÉÍÓÚ]+[a-z0-9áéíïóúüñ]*[0-9A-ZÁÉÍÓÚ]+[0-9a-záéíïóúüñ]*)+"]
Ámbito de la consulta	All
Acciones	group Last match word;

Esta expresión permite hallar, bien en todo el CT-IULA o bien en un documento específico, todas las concordancias de los nombres (N4.*) que presenten: a) palabras con más de una letra mayúscula, sean consecutivas o no y que se combinen o no con una o más letras minúsculas, y b) palabras con una mayúscula inicial combinada siempre con una o más cifras y, en algunos casos, con una o más letras minúsculas.[6]

Para detectar las formas desarrolladas de las siglas se siguieron los mismos patrones establecidos en el caso anterior pero seguidos de un paréntesis.[7] Las formas desarrolladas se obtuvieron mediante la siguiente expresión regular:

Expresión regular	[pos="N4.*" & word="([0-9a-záéíïóúüñ]*[0-9A-ZÁÉÍÓÚ]+[a-z0-9áéíïóúüñ]*[0-9A-ZÁÉÍÓÚ]+[0-9a-záéíïóúüñ]*)+"]
Ámbito de la consulta	All
Acciones	set Context 1 s; cat Last;

Esta expresión regular incluye aquellos candidatos a sigla que poseen más de dos mayúsculas consecutivas y que aparecen después de un punto y seguido o de un punto y aparte. Con ello se facilita el reconocimiento de ciertas siglas como es el caso de HGS, que puede llegar emplearse sin artículo; *e.g.*:

> <m00190> <s>HGS y sus asociados han analizado ARNm de muchas muestras de tejidos sanos de próstata y de tumores de próstata benignos y malignos.</s>

6 Todas las etiquetas que se utilizan para las concordancias complejas están descritas en el etiquetario del IULA que se encuentra en la ayuda de la página de "concordancia compleja".

7 Para extraer las siglas con la ayuda de BwanaNet es necesario partir de la opción de búsqueda experta. A continuación se detallan los tres pasos a seguir en la interfaz del programa:
-Ventana superior. Aquí se indica una expresión regular en particular.
-Ventana intermedia. Permite seleccionar el ámbito de la consulta en el que se quiere aplicar la expresión regular. Las diferentes posibilidades son: cualquier parte del texto (*all*), títulos (*head*), celdas de tablas (*cell*), ítems de listas (*ítem*) y el resto del texto (s).
-Ventana inferior. En esa ventana se definen todas las acciones posteriores a la acción de búsqueda en el patrón representado mediante la expresión regular.

1.1.3 Análisis y depuración de las unidades

Para eliminar el ruido y depurar los datos era necesario usar una definición y una tipología de unidades de reducción léxica (URL). En tal sentido, partimos de la base de que existen tres formas de reducción léxica bien diferenciadas:

1) Abreviaturas. Unidades de reducción léxica formadas mediante la eliminación de algunos de sus grafemas; *e.g.*: megavatio (MW/Mw), caballo de fuerza (HP/Hp), kilómetro (Km/km).

2) Siglas. Unidades de reducción léxica de una estructura sintagmática, formadas con las letras, cifras y/o los símbolos iniciales de sus elementos. Una sigla forma una secuencia cuya pronunciación es alfabética, silábica o ambas; *e.g.*: PCR, TS, TEP, Grb2.

A este grupo pertenecen los acrónimos, un tipo de sigla formada por varios grupos de letras, cuya pronunciación es exclusivamente silábica; *e.g.*: HUGO, ICONA, SASE. Además de los acrónimos, se incluyen en este grupo los cruces (*blends*), o sea, aquellas unidades formadas mediante la combinación de las partes de dos unidades léxicas; *e.g.*: CalTech (California+Technology).

3) Formas truncadas (*clippings*). Unidades de reducción resultantes, por lo general, de la eliminación del final de una unidad léxica; *e.g.*: micrófono (micro); demostración (demo), televisión (tele).

En la fase inicial de la depuración se obtuvo un listado de candidatos a siglas así como algunas unidades que por su estructura se asemejan a las siglas; por ejemplo, ciertas palabras cortas escritas en mayúsculas. Posteriormente, con el refinamiento manual, se obtuvieron las siglas. El refinamiento consistió en el proceso mediante el cual se escogieron manualmente los candidatos que se ajustaban a la definición y tipología de siglas establecidas en este trabajo.

El ruido encontrado en la lista de candidatos a sigla estaba representado por unidades del siguiente tipo:

1) Apellidos: McLeod, DeLisi, McCArty, DiGeorge, etc.
2) Fórmulas químicas: NaCl
3) Secuencias de bases químicas: ATGC, AGC, GAT
4) Artículos: EL, LA, LOS, LAS
5) Pronombres: SUS
6) Palabras en mayúscula: CIENCIA, PROYECTO, GENOMA, etc.

Además, surgió otro tipo de reducción léxica, las abreviaturas (MW, HP, Kmh, etc.), el cual se descartó puesto que no pertenece al tipo de unidad objeto de estudio.

El procesamiento manual también sirvió para establecer otros patrones de detección posibles, que podrían mejorar la expresión regular y, por consiguiente, las listas de candidatos a sigla.

Entre los nuevos patrones se encontró que el rango de letras mayúsculas consecutivas de una sigla parecía situarse entre 2 y 4. Este criterio podría servir para eliminar parte del ruido generado por falsos candidatos como GENOMA, PROYECTO, URALITA, etc.

En cuanto a la expresión regular empleada para la obtención de formas desarrolladas de la siglas se observó que arrojaba el mismo tipo de ruido que la anterior; es decir, apellidos, fórmulas químicas, secuencias de bases, etc. Además, no alcanzaba a cubrir todas las posibilidades de aparición de una forma desarrollada, pues se limitaba a buscar sólo aquellas unidades que presentaban el patrón “sigla (forma desarrollada)”. Esta limitación llevó a considerar nuevos patrones de búsqueda de formas desarrolladas como:

1) sigla **o** forma desarrollada

2) sigla**,** forma desarrollada**,**

El proceso de depuración de cada listado de candidatos siguió los siguientes pasos:

Para aquellos casos en los que no se detectó la forma desarrollada de la sigla, fue necesario confirmar su condición de sigla. Para ello, en primera instancia, se procedió a buscarla en el resto del corpus del IULA mediante el módulo de "búsqueda estándar" de BwanaNet. En el caso de no obtener un resultado positivo, se buscaba en otras fuentes, empezando por las más específicas y terminando por las más generales, así:

1) Búsqueda en fuentes específicas
    (a) En todo el corpus del campo de especialidad tratado (GH o MA)
    (b) En bases de datos de siglas, diccionarios y glosarios electrónicos en Internet
        (i) Sobre genoma humano
            1. AcroMed
            2. Medical Dictionary on-line
            3. Diccionari Enciclopèdic de Medicina
            4. Human Genome Acronym List
            5. Glosarios de Biotecnología
            6. Glosarios de Genética
            7. Genetics home reference
            8. Merck Source
        (ii) Sobre medio ambiente
            1. Compendium of Environmental and professional Acronyms
            2. U.S. Environmental Protection Agency
            3. U.S. Global Change Research Information Office
            4. Diccionario de la contaminación
            5. EEA Multilingual Environmental Glossary
    (c) En páginas web
        (i) List of Acronyms on the Literature on Genome Research
        (ii) Abreviaturas de genes
        (iii) Nombres de proteínas y genes
        (iv) Índice de acrónimos y siglas comunes en Bioquímica y Biología molecular
2) Búsqueda en fuentes generales
    (a) En buscadores de siglas
        (i) Acronym finder

(ii) Acrophile
(iii)Acronym server
(iv)Abbreviations.com
(v) Acronym search

(b) En buscadores generales
(i) Google
(ii) Scirus
(iii)Vivísimo

1.1.4 Presentación de los datos

Los datos que se consideraron para cuantificar las siglas fueron los siguientes:

1) Total de ocurrencias de candidatos a sigla en cada corpus;
2) Total de ocurrencias de siglas en todo el corpus, en cada subcorpus y en subcorpus del mismo nivel;
3) Total de variantes formales, traductivas y grafémicas de una sigla en cada subcorpus.

1.1.5 Análisis y cotejo de los datos

1.1.5.1 Las siglas en el corpus de genoma humano (CGH), subcorpus de nivel de especialidad bajo, medio y alto

1) Nº total de ocurrencias de candidatos a sigla: 1977
2) Nº total de unidades de reducción léxica (incluidas siglas): 214
3) Nº total de siglas: 205
4) Nº de siglas que presentan variación formal (pares de sigla-forma desarrollada): 75
5) Nº de siglas que presentan variación por traducción: 4

- DNA/ADN
- RNA/ARN
- USA/EEUU
- HIV/VIH

1.1.5.2 Las siglas en el corpus de medio ambiente (CMA), subcorpus de texto de nivel de especialidad bajo, medio y alto

1) Nº total de ocurrencias de candidatos a sigla: 164
2) Nº total de unidades de reducción léxica (incluidas siglas): 36
3) Nº total de siglas: 30
4) Nº de siglas que presentan variación formal (pares sigla-forma desarrollada): 19
5) Nº de siglas que presentan variación grafémica: 1
6) EUA / EEUU
7) Nº de siglas que presentan variación por traducción: 1

    - BOD / DBO

1.1.5.3 Análisis contrastivo de las siglas en ambos campos de especialidad

1) Porcentaje de siglas con respecto a cada total de unidades de reducción léxica (URL)
    a) GH: 95,79% de URL son siglas
    b) MA: 83,33% de URL son siglas
2) Porcentaje de variación formal (pares sigla-formas desarrolladas)
    a) GH: 36,58% de las siglas presentan forma desarrollada
    b) MA: 63,33% de las siglas presentan forma desarrollada
3) Porcentaje de variación por traducción
    a) GH: 1,95% de las siglas presenta variación por traducción
    b) MA: 3,33% de las siglas presenta variación por traducción

La siguiente tabla muestra en conjunto todos los datos presentados hasta el momento en este apartado.

GH				MA				
Corpus	**Nº siglas**	**Nº FD**	**Nº var. trad.**	**Corpus**	**Nº siglas**	**Nº FD**	**Nº var. trad.**	**Nº var. graf.**
CGH	205 (100%)	75 (36,58%)	4 (1,95%)	**CMA**	30 (100%)	19 (63,3%)	1 (3,33%)	1 (3,33%)

Tabla 3. Análisis del corpus de GH y MA empleado en la prueba piloto

1.1.5.4 Las siglas en relación con la variación vertical

1.1.5.4.1 Análisis de las siglas en relación con la variación vertical en el corpus de genoma humano

El corpus de genoma humano (CGH), conformado por 129.772 palabras, arrojó un total de 214 URL a partir de las cuales se observó que 205 eran siglas. A su vez, el 36,58% del total de siglas, es decir 75, iban acompañadas de su forma desarrollada (variación formal) y sólo un 1,95% de ellas presentaba variación por traducción.

Como puede apreciarse en la tabla 4, el corpus de genoma humano constaba de 1 subcorpus de texto por cada nivel de especialización. La progresión esperada, de acuerdo con la variación vertical, era que a menor nivel de especialización habría mayor grado de variación. Al observar los datos de este corpus en conjunto vemos que, en efecto, el subcorpus de nivel bajo (SCGH1) es el que presenta mayor variación, representada en la aparición de pares sigla-forma desarrollada y variantes por traducción. Sin embargo, existe una diferencia notoria entre el nivel medio y el alto (SCGH2 y SCGH3). En otras palabras, aquí se invierte la tendencia en el grado de variación y es el subcorpus de nivel alto el que supera al de nivel medio (cuando se esperaba que fuera al contrario). Expresamos los resultados anteriores mediante la siguiente formalización:

$$[\text{Nivel bajo}]_{\text{variación alta}};\ [\text{Nivel medio}]_{\text{variación baja}};\ [\text{Nivel alto}]_{\text{variación media}}.$$

1.1.5.4.2 Análisis de las siglas en relación con la variación vertical en el corpus de medio ambiente

El corpus de medio ambiente, conformado por 126.374 palabras, mostró un total de 36 URL, de las cuales 30 correspondían a siglas. Las formas desarrolladas (variación formal) que acompañaban a las siglas dentro de los textos representan el 63,33% del total de siglas, es decir, 19 siglas. La variación por traducción llegó al 3,33% del total, o sea, 1 sigla; además, se registró un caso de variación grafémica, que representa el 3,33% del total.

Al igual que en el caso anterior, la tabla 4 indica los datos para el corpus de medio ambiente. Como señalamos en el apartado anterior, se presupone que el nivel de especialización bajo es el que produce el mayor grado de variación. Sin embargo, al analizar los datos correspondientes se observa que esta regla no se cumple. En efecto, el subcorpus de nivel alto (SCMA3) es el que mayor variación muestra, seguido por los de nivel bajo (SCMA2) y medio (SCMA1). Los resultados anteriores se formalizan así:

[Nivel bajo] $_{\text{variación media}}$; [Nivel medio] $_{\text{variación media}}$; [Nivel alto] $_{\text{variación alta}}$

Por tanto, aquí se nota que la progresión de la variación se ha invertido. Se esperaba que el orden de los textos según su grado de variación fuera:

[Nivel bajo] $_{\text{variación alta}}$; [Nivel medio] $_{\text{variación media}}$; [Nivel alto] $_{\text{variación baja}}$.

1.1.5.5 Análisis contrastivo de las siglas en relación con la variación vertical en campos de especialidad

Se esperaba que la progresión de la variación disminuyera a medida que aumentara el nivel de especialización de los textos; así, se tendría textos de nivel bajo con mayor variación, textos de nivel medio con más o menos variación y textos de nivel alto con

menor variación. No obstante, para el caso de genoma humano analizado aquí se dio la siguiente situación:

[Nivel bajo] variación alta; [Nivel medio] variación baja; [Nivel alto] variación media

Y en lo que respecta a medio ambiente la situación que se observó fue la siguiente:

[Nivel bajo] variación media; [Nivel medio] variación media; [Nivel alto] variación alta

A pesar de que se procuró que el tamaño del CGH y CMA fuera lo más cercano posible, se encontró entre ambos una diferencia notoria en cuanto a la presencia de siglas. El número de siglas es claramente superior en el CGH donde se hallaron 205 siglas contra las 30 que se encontraron en el CMA. Estos datos parecen coincidir con la siguiente afirmación de Baudet (2001: 34): "*D'un domaine à l'autre (informatique, droit, pisciculture, politique internationale...) et d'une langue à l'autre (français, anglais...), la fréquence des sigles rencontrés dans le discours peut varier fortement*".

1.1.6 Las siglas con respecto a cada nivel de especialización

1.1.6.1 Análisis de las siglas en los textos de nivel de especialización bajo (SCGH1 y SCMA1)

1) Con relación a las URL
   a) GH: 79 siglas de 80 URL
   b) MA: 7 siglas de 10 URL
2) Con relación a la variación formal (pares sigla-forma desarrollada)
   a) GH: 42 siglas con forma desarrollada (53,16% del total de siglas)
   b) MA: 2 siglas con forma desarrollada (28,57% del total de siglas)
3) Con relación a las variantes por traducción
   a) GH: 2 siglas con variante por traducción (2,53% del total de siglas)
   b) MA: 1 sigla con variante por traducción (14,28% del total de siglas).

1.1.6.2 Análisis de las siglas en los textos de nivel de especialización medio (SCGH2 y SCMA2)

1) Con relación a las URL
   a) GH: 47 siglas de 55 URL
   b) MA: 8 siglas de 8 URL
2) Con relación a la variación formal (pares sigla-forma desarrollada)
   a) GH: 8 siglas con forma desarrollada (17% del total de siglas)
   b) MA: 2 siglas con forma desarrollada (25% del total de siglas)
3) Con relación a las variantes por traducción
   a) GH: ninguna sigla con variante por traducción
   b) MA: 1 sigla con variante por traducción (12,5% del total de siglas).

1.1.6.3 Análisis de las siglas en los textos de nivel de especialización alto (SCGH3 y SCMA3)

1) Con relación a las URL
   a) GH: 79 siglas de 79 URL
   b) MA: 15 siglas de 18 URL
2) Con relación a la variación formal (formas desarrolladas)
   a) GH: 25 siglas con forma desarrollada (31,64% del total de siglas)
   b) MA: 15 siglas con forma desarrollada (100% del total de siglas)
3) Con relación a las variantes por traducción
   a) GH: 2 siglas con variante por traducción (2,53% del total de siglas)
   b) MA: ninguna sigla con variante por traducción.

Los datos presentados arriba se recogen en la siguiente tabla.

GH				MA				
Corpus	**Nº siglas**	**Nº FD**	**Nº var. trad.**	**Corpus**	**Nº siglas**	**Nº FD**	**Nº var. trad.**	**Nº var. graf.**
CGH	205 (100%)	75 (36,58%)	4 (1,95%)	**CMA**	30 (100%)	19 (63,3%)	1 (3,33%)	1 (3,33%)
SCGH1	79 (100%)	42 (53,16%)	2 (2,53%)	**SCMA1**	7 (100%)	2 (28,57%)	1 (14,28%)	-
SCGH2	47 (100%)	8 (17%)	-	**SCMA2**	8 (100%)	2 (25%)	1 (12,5%)	-
SCGH3	79 (100%)	25 (31,64%)	2 (2,53%)	**SCMA3**	15 (100%)	15 (100%)	-	-

Tabla 4. Comparación de los subcorpus de nivel de especialización alto, medio y bajo

Mediante el cotejo de la información anterior puede observarse que el subcorpus de nivel de especialización bajo (SCGH1) fue el que más número de siglas presentó, 79 en total, de las cuales 42 iban acompañadas de sus respectivas formas desarrolladas y 2 con variantes por traducción. Le sigue en cantidad el nivel alto (SCGH3), el cual mostró, al igual que en el caso anterior, 79 siglas, 25 de ellas con sus formas desarrolladas y 2 con variante por traducción. Finalmente, el nivel de especialización medio (SCGH2) fue el que menor número de siglas reflejó, 47 en total, de las cuales 8 poseían formas desarrolladas dentro del texto pero ninguna variante por traducción.

En lo que respecta al ámbito de medio ambiente, el subcorpus que más siglas presentó fue el nivel alto (SCMA3) con un total de 15 unidades e igual número de formas desarrolladas y ninguna variante formal. Le sigue el nivel bajo (SCMA1) con 7 siglas, 2 con sus respectivas formas desarrolladas y tan sólo una con variante por traducción. Por último, el nivel medio (SCMA2) con 8 siglas, 2 formas desarrolladas y una variante por traducción.

Se deduce entonces que, la incidencia de las siglas en ambos campos de especialidad varía según el nivel de especialización aunque no en la proyección esperada. En GH el nivel de especialización más rico en siglas es el bajo mientras que en MA es el alto.

1.1.7 Conclusiones

Al comienzo del trabajo de exploración se formulaba como hipótesis general de partida que "la frecuencia y el tipo de siglas en el discurso especializado, así como la presencia o ausencia de la forma desarrollada en el texto, eran factores que permitían diferenciar el nivel de especialización de los textos. Se planteaba, por tanto, que las siglas podrían ser un factor discriminante del nivel de especialización de los textos dentro de la variación vertical del discurso".

Esta hipótesis se corroboró parcialmente. No se logró comprobar que el número de siglas y variantes formales fuera proporcional al nivel de especialización del texto en que aparecen. Esto es, a mayor nivel de especialización menor nivel de variación. Sin embargo, sí se logró confirmar que las siglas inciden en el discurso especializado puesto que introducen variación denominativa, reflejada tanto por la expresión de sus formas desarrolladas (variantes formales) como por sus variantes por traducción y por grafemas. Hecho que está en consonancia con un principio de la Teoría Comunicativa de la Terminología (TCT) que establece que: "Todo proceso de comunicación comporta inherentemente variación, explicitada en formas alternativas de denominación del mismo concepto (sinonimia) o en apertura significativa de una misma forma (polisemia). Este principio es universal para las unidades terminológicas, si bien admite diferentes grados según las condiciones de cada tipo de situación comunicativa..." (Cabré, 1999: 85).

2. Fase del proyecto de tesis

A partir de las conclusiones del trabajo de exploración titulado "Siglas y variación vertical en textos sobre genoma humano y medio ambiente", se concretó el enfoque de la tesis mediante la segunda fase de la investigación, denominada Proyecto de tesis.

Este ha aportado a la tesis tres elementos básicos como son: 1) supuestos de partida; 2) objetivos de la tesis, y 3) metodología.

El proyecto de tesis ha partido de las siguientes consideraciones:

1) Que la hipótesis de partida formulada sólo pudo confirmarse parcialmente, al no lograr comprobar que el número de siglas, formas desarrolladas y variantes formales fuera proporcional al nivel de especialización del texto en que aparecían; *i.e.,* a mayor nivel de especialización menor nivel de variación;
2) Que existía una dificultad real para conformar un corpus representativo en español, que incluyera suficientes textos de alto nivel de especialidad en MA y de bajo nivel de especialidad en GH[8], y
3) Que en el IULA se llevaban a cabo estudios paralelos sobre el grado de especialización de los textos mediante técnicas de medición de densidad terminológica. Puesto que las siglas se consideran formas reducidas de una unidad terminológica, dichos estudios incluían *per se* el análisis e influencia de las siglas en los diferentes niveles de especialidad.

Por estas razones se desestimó la continuación del estudio por la vía de la incidenca de las siglas en el grado de especialización de los textos y se decidió enfocar la investigación hacia una vía más general, es decir, al estudio de las siglas desde la perspectiva global de los ámbitos de genoma humano y medio ambiente.

8 Actualmente el CT-IULA no procesa ningún documento nuevo por lo que asumir esta tarea individualmente implicaba una enorme inversión de tiempo, lo que retrasaría considerablemente el cronograma de trabajo.

2.1 Supuestos de partida

A la luz de los datos obtenidos en la prueba piloto, realizada en la fase experimental, se exponen los supuestos de partida que serán validados o rechazados a lo largo de la investigación.

1) Desde el punto de vista descriptivo

   a) El mecanismo de reducción léxica "siglación" es ampliamente utilizado en el discurso especializado;
   b) La frecuencia de las siglas varía de acuerdo con el ámbito temático;
   c) Dentro del discurso especializado, las siglas corresponden mayoritariamente a unidades de conocimiento especializado, concretamente a unidades terminológicas (UT);
   d) Las siglas especializadas son elementos nominales;
   e) Dependiendo del grado de fijación en el discurso, la sigla puede aparecer junto a su forma desarrollada;
   f) Las siglas, aunque tengan categoría N, no presentan las mismas posibilidades que los N ni en relación a su flexión ni a su combinación;
   g) Las siglas, al ser un fenómeno de reducción léxica y, por tanto, de economía lingüística, son más frecuentes que su respectiva forma desarrollada en un texto dado, funcionando como variantes;
   h) Las preposiciones y conjunciones se omiten normalmente dentro de las siglas;
   i) Las siglas, cuando proceden del inglés, pasan mayoritariamente al español sin traducirse.

2) Desde el punto de vista de la detección

   a) La mayoría de las siglas están compuestas por tres o cuatro caracteres alfanuméricos;

b) Las siglas que proceden del inglés no se corresponden exactamente con las iniciales de su forma desarrollada en español;
c) El paréntesis es la forma más común de aparición de las formas desarrolladas de las siglas. Sin embargo, existen otras estrategias de introducción bien de la sigla o bien de la forma desarrollada; por ejemplo: la palabra clasificadora "abreviatura". Este hecho hace pensar que muchas de las formas desarrolladas aparecen "alejadas" de su sigla correspondiente dentro del texto, lo cual puede tener repercusiones en los sistemas de detección y extracción. También se harán búsquedas con otras palabras clasificadoras como "sigla" y "acrónimo";
d) Los patrones para hallar las siglas y sus respectivas formas desarrolladas son diversos; de ellos se han identificado cuatro *a priori*:

i) Forma desarrollada (sigla). Este patrón indica que las siglas aparecen a la derecha de su correspondiente forma desarrollada, la cual va entre paréntesis. En efecto, estudios llevados a cabo en corpus sobre biología molecular en inglés, han mostrado que las formas desarrolladas que se ubican a la izquierda de la sigla en la línea de concordancias representan más del 90% de los casos (*cf*. Nenadić, 2002). En el listado obtenido a partir de nuestro corpus también se constata este hecho; por ejemplo:

<doc_codi m00190> <s>La distrofia muscular de Duchenne (DMD), llamada así por el neurólogo francés Guillaume Benjamin Amand Duchenne (1806-1875), que la describió, es la distrofia más frecuente entre los varones (1 de cada 3500 niños varones).</s>

ii) Sigla (forma desarrollada). Las siglas aparecen a la izquierda de su forma desarrollada que, como en el caso anterior, también va entre paréntesis; por ejemplo:

<doc_codi a00007> <s>Francia tiene un plan energético oficial que proyecta obtener para 1.990 alrededor de 8.000 millones de TEP (toneladas equivalentes de petróleo) entre la explotación forestal y la de residuos agrícolas y se están ensayando cultivos energéticos forestales con chopos, con un rendimiento de 8 a 12 toneladas por hectárea y año.</s>

iii) Forma desarrollada_o_sigla. Con este patrón se representa la ocurrencia de las siglas y sus formas desarrolladas enlazadas mediante la letra ‘o’, lo cual evidencia, al igual que los paréntesis en los casos anteriores, el carácter de variación denominativa presente en estas unidades; por ejemplo:

<doc_codi a00207> <s>En los nefelómetros la turbiedad se lee como "Unidades de turbidez nefelométrica" o NTU (ver referencia 3).</s>

iv) Sigla_o_forma desarrollada. Finalmente, este último patrón representa que primero aparece la sigla y luego su correspondiente forma desarrollada, enlazadas mediante la letra ‘o’; por ejemplo:

<doc_codi m00369> <s>El GMP cíclico está emparentado estructuralmente con otro segundo mensajero, bastante más conocido, el AMPc o adenosina monofosfato cíclico.</s>

2.2 Objetivos de la tesis

La presente tesis se propone un objetivo general doble. Por un lado, analizar los aspectos teórico-descriptivos del fenómeno de la siglación; por el otro, proponer los criterios para el diseño de una aplicación para el reconocimiento de las siglas en español.

1) En el plano teórico, pretendemos analizar el concepto y la tipología de las siglas y revisar su tratamiento en la bibliografía. Concretamente nos proponemos:

   a) Establecer el estado de la cuestión para determinar el tratamiento que se le ha dado a estas unidades en disciplinas como la terminología y la lexicología, y
   b) Delimitar el objeto sigla y su tipología de modo que puedan hacerse generalizaciones lingüísticas de éste como unidad del léxico en el discurso especializado.

En el plano descriptivo, nos proponemos observar las siglas y sus características lingüísticas y estadísticas. Concretamente nos proponemos:

c) Analizar las características lingüísticas de las siglas del discurso especializado a partir del corpus unificado de siglas de GH y de MA;
d) Analizar desde el punto de vista de la estadística descriptiva la incidencia de las siglas en el discurso de GH y MA, respectivamente, y
e) Probar que la frecuencia de aparición de las siglas varía de un ámbito de especialidad a otro y determinar las consecuencias que este factor tiene en el discurso.

2) En el plano aplicado, nos proponemos determinar los criterios para el diseño de un sistema de detección y extracción semiautomática de siglas. Concretamente nos proponemos:

a) Establecer el estado de la cuestión para determinar qué aplicaciones similares existen en la actualidad y con qué estrategias operan;
b) Establecer los patrones comunes para la detección de siglas. Para ello partiremos del corpus unificado de siglas de GH y MA;
c) Establecer los criterios a tener en cuenta para el diseño de un prototipo de detector/extractor de siglas para el español.

2.3 Metodología

Finalmente, en el Proyecto de tesis también se establecieron los criterios para la conformación del corpus y la metodología de trabajo de la tesis, los cuales se detallan en el capítulo 5.

Capítulo 2

Capítulo 2

Las siglas: definición y tipología

1. Antecedentes

> "En el acelerado mundo que nos ha tocado vivir, la sigla cumple una función importante, necesaria, incluso, a veces, imprescindible. Tan viva y casi frecuente como las voces comunes, aparece en todo tipo de textos ocupando su lugar por propio derecho, tan en su sitio como cualquiera de los artilugios que el hombre ha inventado para hacer más fácil y cómodo su vivir diario" (Martínez de Sousa, 1984: 13).

Muchos han sido los autores que han tratado el tema de las siglas. Han descrito algunos rasgos propios de estas unidades e igualmente han sugerido emprender estudios más profundos y detallados. Baudet (2002: 93) por ejemplo, afirma que la siglación constituye un procedimiento de creación léxica que, sin ser exclusivo de la ciencia y de la técnica, se utiliza bastante en áreas que van desde la matemática hasta la medicina, pasando por todas las ramas de la técnica. Considera que la siglación es un procedimiento muy eficaz para acelerar la comunicación, aunque presenta el inconveniente de generar abundante homonimia. Por todas estas razones, Baudet sostiene que la siglación amerita un estudio cuidadoso por parte de lexicógrafos y, sobre todo, de terminólogos.

Aunque el fenómeno de la reducción léxica está presente prácticamente en todas las lenguas, en español el tema ha sido tratado de manera tangencial, en especial si se mira desde la óptica del discurso especializado. La mayoría de los estudios se han llevado a cabo desde la lexicología donde se cuenta con trabajos como los de

Rodríguez, 1981; Alvar & Miró, 1983; Martínez de Sousa, 1984 y Casado Velarde, 1985. Y desde la terminología donde destacan los trabajos de Cardero, 2002 y Fijo 2003. De todos ellos resaltamos por su profundidad los siguientes:

1) Rodríguez (1981) presenta un estudio descriptivo, fundamentalmente lexicológico y gramatical, titulado "Análisis lingüístico de las siglas: especial referencia al español e inglés". En él se coteja el comportamiento lingüístico de las siglas con el de los vocablos normales de la lengua. Es un estudio que se basa en textos escritos provenientes de diarios, revistas y glosarios pertenecientes al discurso general. Como complemento recurre a una encuesta con diversos informantes, para dar así un fundamento más empírico a los postulados propuestos.

Todo el material, escrito y oral, constituye el corpus para establecer por vía de inducción las regularidades en el uso de las siglas. El corpus alberga las formas más generalizadas, pero también da cuenta razonada de otras variantes más ocasionales. Por ejemplo, el registro de OTAN como lexema más corriente de la *Organización del tratado del Atlántico Norte*, y el uso esporádico de NATO.

2) Fijo (2003) presenta un estudio titulado "Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas". En él trata de describir todos los aspectos relacionados con la creación, uso y traducción de las siglas como términos pertenecientes al ámbito de la enfermería. Se trata de un estudio contrastivo inglés-español que parte de un modelo integrador de las perspectivas lingüística y terminológica. La metodología consiste en la recopilación del corpus paralelo (50 textos) para el análisis, aplicación del método terminográfico y el formato electrónico (Multiterm) en que se ha almacenado la información. Los datos, correspondientes a 65 siglas en inglés y 58 en español, se analizan desde el punto de vista de la estadística y, a partir de ellos, se llega a los resultados y se formulan las conclusiones.

A pesar de los trabajos realizados hasta el momento, no se cuenta con estudios contrastivos sobre siglas en el discurso especializado en español y menos aún con criterios para sistemas de detección y extracción a partir de textos en esta lengua.

2. Concepción y clasificación

2.1 Estado de la cuestión

> "Tanto las denominaciones como las definiciones relacionadas con las formas abreviativas están sometidas actualmente a análisis crítico por los especialistas, y no hay al respecto unanimidad en los criterios de esquematización, taxonomía y aplicación" (Martínez de Sousa, 1993: 23).

Una de las quejas constantes por parte de los estudiosos de los fenómenos de reducción léxica ha sido la falta de consenso a la hora de delimitar los conceptos de cada uno de estos fenómenos. (*cf.* Calvet, 1980: 5; Rodríguez, 1981: 15; Martínez de Sousa, 1984: 17; Gehénot, 1990: 106; Zolondek, 1991: 1; Bauer, 1999: 172; López Rúa, 2000: 366; Fijo, 2003: 76). En efecto, Rodríguez (1981:15) sostiene que "una de las necesidades más acuciantes en el estudio de los métodos abreviatorios es el establecimiento de una terminología más estandarizada. La dificultad estriba en la propia naturaleza de la abreviación, que alberga un abigarrado haz de tipos y subtipos difíciles de categorizar y delimitar. Como resultado, a menudo aparecen entremezclados al ser tratados por los distintos autores". Esta dificultad no es exclusiva de la lengua española sino que también está presente en lenguas como el inglés y el francés. En su estudio sobre las siglas en inglés López Rúa (2000: 366) sostiene que "*If one goes back to the fundamental question –what is an acronym? –, and looks for an answer in the large amount of literature available, the result could not be more discouraging. The lack of agreement and explicitness on the part of scholars concerning terminology, definitions and classifications has been a constant*

before and even after the specific term ***acronym*** *was coined. […] confusion, overlapping and inconsistency concerning definitional criteria are generally acknowledged and sometimes regretted, but seldom confronted and only exceptionally challenged […]*".

Como se acaba de señalar, la problemática por la falta de consenso y la gran cantidad de definiciones y tipologías sobre las formas de reducción léxica ha sido objeto de preocupación constante por parte de los estudiosos de la lengua.[9] A modo de ejemplo, tomamos el estudio de Bauer (1990: 172), quien hace una radiografía del asunto. Para su investigación toma el campo de la informática, por tratarse justamente de uno de los campos donde se ha expandido con mayor celeridad el fenómeno de la reducción léxica. Bauer ha encontrado que, tanto en artículos especializados como en diccionarios, a las abreviaturas, acrónimos y siglas se les denomina genéricamente fenómenos de reducción léxica; aunque también se encuentran términos compuestos como "siglas acronímicas" y "abreviaturas acronímicas". Además, ha encontrado que el acrónimo a veces se muestra como un tipo de sigla y que ésta a su vez puede constituir una abreviatura.

A raíz de esta situación, este autor se dedicó a rastrear una treintena de diccionarios (monolingües y bilingües en varias lenguas), además de algunos artículos científicos. Su análisis arrojó como resultado que:

1) En la mitad de las obras consultadas en francés, alemán, inglés e italiano los términos acrónimo y sigla se consideran sinónimos puros. Y cita en orden cronológico los siguientes ejemplos:

> 'DEAK 1973, 5: amér. *acronym* = fr. *sigle;*
> BATTAGLIA 1980, 141: ital. *acronimo* = ital. *sigla;*
> FERRANTE/CASSIANI 1983, 20: ital. *acronimo* = fr. *sigle;*
> DBI, 46: angl. *acronym* = fr. *sigle;*
> GABRIELLI 1989, 62: ital. *acronimo* = ital. *sigla*)

9 El punto más crítico está entre los tipos sigla y acrónimo.

O bien una forma de reducción léxica define a otra; por ejemplo:

GEHENOT 1973, 135: fr. *acronyme* = "mot formé de *sigles* correspondant aux premières lettres d'une expression composée";
GARZANTI 1982, 27: ital. *acronimo* = "nome formato dalle lettere iniziali di altre parole: *sigla*";
RDG, 6: amér. *acronym* (1943) -) fr. *acronyme* (1970)= "*sigle* prononcé comme un mot ordinaire";
ROBERT 1986, 96: fr. *acronyme* = "*sigle* prononçable comme un mot ordinaire";
CARDONA 1988, 25: ital. *acronimo* = "una *sigla* composta (...) dalle sole iniziali";
HACHETTE 1988, 14: fr. *acronyme* = "*sigle* que l'on prononce comme un mot ordinaire, sans l'épeler";
PAVEL 1988, 1: angl. *acronym*/ fr. *acronyme*= "*sigle* prononcé en un mot")".

2) En aproximadamente el 40% de las obras de referencia consultadas, los acrónimos se definen con la ayuda del término *mot* (*word, Wort, parola*); por ejemplo:

WEBSTER 1972, 9: angl. *acronym* = "a *word* formed from the initial letter or letters of each of the successive parts or major parts of a compound term";
LEXIS 1975, 22: fr. *acronyme*= "*mot* constitué par les premières lettres des mots composant une expression complexe";
RANDOM, 13: angl. acronym= "a *word* formed from the initial letters or groups of letters of words in a set phrase";
DTN 7: angl. *acronym*—fr. *acronyme* = esp. *acrónimo*= "mot prononçable constitué de séquences d'initiales d'autres mots";
CONRAD 1985, 19: allm. *Akronym*="aus den Anfangsbuchstaben mehuerer Wörter gebildetes Kurz *wort*";
DUDEN, 104: allm= *Akronym*="aus den Anfangbuchstaben mehuerer Wörter gebildetes *Wort*".

3) En cuanto al resto de obras consultadas, Bauer encontró definiciones que recurrían a términos diferentes a los de *mot* (*word, Wort, parola*) para significar el objeto; por ejemplo:

CALVET 1973, 31: angl. *initials* = fr. *sigle*= "phénomène consistant à prendre la première lettre de chaque mot d'un groupe"- fr. *acronyme* = "phénomène consistant à prendre la première syllabe de chaque mot";
LEXIS 1975, 1654: fr. *sigle* = "groupe de lettres initiales constituant l'abréviation de termes fréquemment employés" fr. *abréviation* = "réduction d'un mot à une suite plus courte d'éléments (première syllabe), ou réduction d'un composé à ses initiales";
GOOSSE, 1975, 60: fr. *Sigle* = "où les lettres sont prononcées d'après leur nom" —"où elles ont leur valeur phonétique normale";
GEORGE 1977, 33: fr. *sigle* = "unité lexicale résultant de la juxtaposition des lettres initiales des mots, ou de quelques-uns des mots, qui forment un composé ou une locution";
KNOBLOCH 1986, 6: angl.. *acronym* = *allm. Abkürzung* = "(...) Initialkürzung (...) Suspension (...) Kontraktion (...)").

De todas formas, la comparación de los ejemplos dados por los autores citados no deja las cosas en claro. En la siguiente tabla, Bauer indica la fluctuación definitoria relativa a "acrónimo", "sigla" y "abreviatura".

Lema (=lema abreviado)	Definido como "acrónimo" en:	Definido como "sigla" en:	Definido como "abreviatura" en:
ALGOL **algol**	LEXIS 75 DBI 85 DI 81 ROBERT 86	HUMBLEY 88 _____ _____ _____	DFAI 88 _____ _____ _____
CAO-TAO	PAVEL 88 _____	BOISSY 88 DI 88	_____ DFAI 88
NATO-OTAN	KNOBLOCH 86	GEORGE 77	_____
S.M.I.C	_____	LEXIS 75	LEXIS 75
S-N-C-F.	_____	CALVET 73	LEXIS 75
Unesco **U.N.E.S.C.O**	HACHETTE 88 CONRAD 85 _____	CALVET 73 LEXIS 75 IBAIA 85	_____ _____ _____
U.R.S.S.	GEHÉNOT 73 CARDONA 88	CALVET 73 _____	_____ _____

Tabla 5. Fluctuación definitoria relativa a "acrónimo", "sigla" y "abreviatura" (Bauer, 1990)

El panorama dibujado por los autores antes citados conduce a la necesidad de una revisión de las concepciones sobre el objeto "sigla" con el ánimo de adoptar para nuestro estudio la definición más integradora y flexible del fenómeno. En efecto, este paso ha constituido la condición previa para el inicio de nuestra investigación. Se ha recopilado el mayor número posible de bibliografía referente a las definiciones dadas por diversos autores respecto de los fenómenos de reducción léxica en general y de las siglas en particular. En primer lugar, se han clasificado todas las definiciones según su procedencia. De este modo, han resultado dos grandes grupos, a saber:

1) Bibliografía especializada: artículos, libros y tesis;
2) Otras fuentes: diccionarios (generales y especializados), gramáticas, normas técnicas, manuales de estilo y criterios de política lingüística.

En segundo lugar, los dos grupos antes mencionados se han organizado cronológicamente y de acuerdo con la lengua en que se han publicado (*cf.* cuadro 1)

Cuadro 1. Organización cronológica y lingüística de la bibliografía rastreada para el concepto y tipología de siglas

I. Bibliografía especializada (artículos, libros y tesis)
A. Inglés
Algeo (1991)
Larkey *et al.* (2000)
López (2000)
B. Francés
Calvet (1980)
Mitterand (1986)
Losson (1990)
Nakos (1990)
Zolondek (1991)
Percebois (2001)
Vandaele & Pageau (2006)
C. Español
Mejía (1980)
Rodríguez (1981)
Alvar & Miró (1983)
Martínez de Sousa (1984)
Casado Velarde (1985)
Gehénot (1990)
Cabré (1993)
Abreu (1997)
Estopà (2000)
Cardero (2002)
Fernández (2002)
Fijo (2003)
Alcaraz (2003)
Bezos (2007)
D. Catalán
Mestres (1985)
Mestres i Serra (1996)

II. Diccionarios (generales y especializados), gramáticas, normas técnicas, manuales de estilo y criterios de política lingüística
A. Inglés
Quirk *et al.* (1985)
Brusaw *et al.* (1987)
Mossman (1992)
Burnett (1994)
Bussmann (1996)
Matthews (1997)
Huddleston *et al* (2002)
McArthur (2003)
Merriam-Webster Dictionary (2003)
Crystal (2003)
B. Francés
Dubois *et al.* (1994)
Dictionnaire Le Nouveau Petit Robert (2001)
(Gouvernement de Québec, 2002)
(http://grammaire.reverso.net)
Le trésor de la langue française informatisé (2002)
C. Español
Mounin (1982)
Cerdà *et al.*, (1986)
Lázaro Carreter (1990)
Cardona (1991)
Diccionario de Lingüística, Larousse-Terrncat (1992)
Colás (1994)
Alcaraz & Martínez (1997)
Bastons & Font (2001)
DRAE (2001)
Maldonado (2002)
D. Catalán
Mestres *et al.* (1995)
Capó & Veiga (1997)
Pérez Saldanya (1998)
Mestres i Serra & Guillén (2001)
Gran diccionari de la llengua catalana (2003)

En tercer lugar, el análisis sobre la definición de sigla adoptada por cada uno de los autores estudiados aquí ha permitido detectar tres modos diferentes de concebir el fenómeno de la siglación:

1) Autores que no hacen o no explicitan la distinción entre sigla y acrónimo;
2) Autores que establecen la distinción entre sigla y acrónimo;
3) Autores que consideran al acrónimo como un tipo de sigla (*cf.* cuadro 2).

Centramos nuestra mirada en el segundo y tercer grupos puesto, que son los que aportan verdaderos elementos para la elección de la definición de las siglas.

El segundo grupo incluye aquellos especialistas que consideran el acrónimo como una "unidad de reducción léxica independiente de la sigla". Este grupo comprende, a su vez, dos concepciones diferentes del acrónimo, a saber:

1) El acrónimo se distingue por su pronunciación silábica**.** En este subgrupo se inscriben la mayoría de los especialistas estudiados. Sin embargo, por cuestión de espacio nos limitamos a reseñar los autores más representativos, veamos:

Losson (1990: 22) sostiene que el acrónimo se compone a partir de sílabas iniciales y se pronuncia como una palabra. Afirma, además, que la diferencia real entre las siglas y los acrónimos no es más que una cuestión de forma.

Algeo (1991: 9) concibe el acrónimo como una unidad formada por las letras iniciales de las palabras de una expresión, al igual que los alfabetismos, pero pronunciado según las reglas de la ortografía; *e.g.*: scuba (*self-contained underwater breathing apparatus*).[10]

10 Algeo considera los alfabetismos o inicialismos como abreviaciones que usan las letras iniciales de las palabras de una expresión, pronunciadas mediante su deletreo; *e.g.*: TV, AA, ABL, ACS, CATV, etc.

Zolondek (1991: 1) afirma que una unidad se considera un acrónimo cuando se pronuncia silábicamente. Tanto la acronimia como la siglación dependen de la morfosintaxis; se trata en ambos casos de una unidad "monoreferencial" de elementos normalmente separados "referencialmente".

2) El acrónimo es el resultante de la unión de dos extremos opuestos de dos palabras.

Rodríguez (1981: 21) afirma que la etimología de "acrónimo" predispone este término al ensanchamiento de su significado y de este modo a cierta ambigüedad. En efecto, *akros* significa "extremidad", por lo que puede aludir tanto al extremo —létrico o silábico, inicial o final— como a un segmento morfemático. Entre los autores que comparten este enfoque se encuentran:

Alvar & Miró (1983: 1) sostienen que el acrónimo es la unión de los extremos opuestos de dos palabras: el principio de la primera y el final de la segunda, o el final de la primera y el comienzo de la última; *e.g.*: autobús (*auto*móvil ómni*bus*), *bit* (*bi*nary digi*t*), etc.

Martínez de Sousa (1984: 17) afirma que el acrónimo es un tipo de "abreviación por contaminación". Esta unidad resulta de la fusión en una sola de truncamientos iniciales o finales (cualquiera sea la sucesión) de las voces que forman un término compuesto o sintagma. Tales truncamientos están constituidos normalmente por sílabas, pero puede darse también por mezcla de sílabas y letras, generalmente iniciales; *e.g.*: buna (de *bu*tadieno + *na*trum), motel (*mo*tor + ho*tel*), etc.

Cabré (1993: 177) considera que los acrónimos son palabras formadas por la combinación de segmentos (normalmente dos) de un sintagma desarrollado, que pueden adoptar diferentes formas según los segmentos que lo integran; *e.g.*: Termesp (*term*inología *esp*añola), informática (*infor*mación auto*mática*), etc.

Fijo (2003: 70) equipara el concepto de acrónimo en español con el de *blend* en inglés. Considera que es un procedimiento morfológico consistente en la formación de una palabra a partir de dos o, muy raramente, tres unidades léxicas estando representadas, al menos una de ellas, por un fragmento (una o más sílabas) de su significante. Distingue dos tipos de acrónimos aplicables a esta concepción:

a) Acrónimos formados por el fragmento inicial de una palabra y el fragmento final de otra; *e.g.*: Eurovisión (*European*+*television*);

b) Acrónimos en los que una de las palabras de la base se conserva completa, fragmentándose sólo la(s) otra(s); *e.g.*: cantautor (cantante+autor).

Finalmente, el tercer grupo está constituido por aquellos expertos que consideran el acrónimo como un "tipo de sigla" que se pronuncia silábicamente, veamos:

Alcaraz (2003: 44) afirma que cuando las siglas tienen una estructura silabeable, no se suelen pronunciar sus constituyentes por separado, sino que se articulan como un todo. Son los llamados "acrónimos".

Alcaraz & Martínez (1997), basándose en Rodríguez (1993: 275), sostienen que el acrónimo es una variedad de sigla formada por las iniciales de los componentes de una unidad léxica, que se lexicaliza y se adapta por completo a las formas fonotácticas, gráficas, etc., de la llamada forma canónica del lenguaje. Por ejemplo, sida y Renfe son "acrónimos" formados respectivamente a partir de "síndrome de inmunodeficiencia adquirida" y de "Red de Ferrocarriles Españoles", y se leen como palabras corrientes.

Maldonado (2002) considera que los acrónimos pertenecen al tipo de siglas denominadas mixtas, es decir, cuando se componen con letras no sólo iniciales, o con

letras de palabras de significado gramatical (nexos y artículos).[11] Se escriben con inicial mayúscula y el resto minúscula. Ejemplo: Renfe (***Re****d* ***N****acional de los* ***F****errocarriles* ***E****spañoles*), Banesto (***Ban****co* ***E****spañol de Crédi****to***).

Nakos (1990) señala que las siglas que pueden pronunciarse como una palabra se llaman acrónimos, por ejemplo, UNESCO. También las denomina siglas acrónimas.

Rodríguez (1981: 26) sostiene que las siglas pueden derivar en dos procedimientos diferentes: literación (deletreo) y acronimia (pronunciación silábica).

En el presente trabajo consideramos la sigla como una "unidad de reducción formada por caracteres alfanuméricos procedentes de una unidad léxica de estructura sintagmática. Una sigla forma una secuencia cuya pronunciación puede ser alfabética, silábica o ambas; *e.g.*: PCR, TS, TEP, Grb2". Esta definición es el producto de integrar los dos puntos de vista anteriores. Por un lado, consideramos que el acrónimo es un tipo de sigla que se pronuncia silábicamente y, por otro lado, que el acrónimo se puede formar mediante la combinación de segmentos de un sintagma desarrollado. En el cuadro 2 se presenta la clasificación de los autores según su concepción del objeto acrónimo. De esta manera, creemos que se despeja uno de los aspectos más problemáticos de la definición de sigla hallado a lo largo de la revisión bibliográfica. Por tanto, a partir de aquí se establecen la definición y tipología de sigla, necesarias para la realización de las siguientes fases de esta investigación.

11 Algunos autores como Martínez de Sousa (1984: 34), Mestres (1996: 15) y Casado Velarde (1985: 20) también denominan a las siglas mixtas "impropias" o "sigloides".

Grupo 1
Autores que no hacen o no explicitan la distinción entre sigla y acrónimo:

1. Burnett (1994)
2. Colás (1994)
3. Capó & Veiga (1997)
4. Estopà (2000)
5. Larkey (2000)
6. Lázaro Carreter (1990)
7. Matthews (1997)
8. Mejía (1980)
9. Mestres i Serra (1985, 1995)
10. Mitterand (1986)
11. Mounin (1982)

Grupo 2
Autores que sostienen que el acrónimo se diferencia de la sigla por el modo de pronunciación o de formación:

Concepción 1: Pronunciación silábica

1. Algeo (1991; 2003)
2. Bastons & Font (2001)
3. Brusaw (1987)
4. Bussmann (1996)
5. Calvet (1980)
6. Crystal (2003)
7. Diccionario Le trésor de la langue fraçaise informatisé (2002)
8. Diccionario Merriam-Webster (2003)
9. Gran diccionari de la llengua catalana (2003)
10. Gehénot (1990)
11. Huddleston *et al.*(2002)
12. López Rúa (2000)
13. Losson (1990)
14. McArthur (2003)
15. Mossman (1992)
16. Norma ISO 1087-1 (2000)/Norma ISO 12620 (1999)
17. Percebois (2001)
18. Pérez Saldanya (1998)
19. Vandaele & Pageau (2006)

Concepción 2: Unión de extremos opuestos de dos palabras

1. Alvar & Miró (1983)
2. Cabré (1993)
3. Cardero (2002)
4. Casado Velarde (1985)
5. Diccionario Larousse/Termcat (1992)
6. Fijo (2003)
7. Martínez de Sousa (1984)

Grupo 3
Autores que consideran los acrónimos como un tipo de sigla:

1. Abreu (1997)
2. Alcaraz (2003)
3. Alcaraz & Martínez (1997)
4. Arntz & Picht (1995)
5. Bezos (2007)
6. Cardona (1991)
7. Cerdà (1986)
8. Diccionario DRAE (2001)
9. Diccionario Le Nouveau Petit Robert (2001)
10. Dubois (1994)
11. Fernández (2002)
12. Gouvernement du Québec (2002)
13. Grammaire reverso.net (2000)
14. Maldonado (2002)
15. Mestres & Guillén (2001)
16. Nakos (1990)
17. Quirk (1985)
18. Rodríguez González (1981)

Cuadro 2. Enfoques sobre la conceptualización de sigla y acrónimo según la bibliografía revisada

2.2 Concepto y clasificación de las siglas en este trabajo

Como se ha dicho antes, entendemos por sigla toda "unidad de reducción formada por caracteres alfanuméricos procedentes de una unidad léxica de estructura sintagmática. Una sigla forma una secuencia cuya pronunciación puede ser alfabética, silábica o ambas; *e.g.*: PCR, TS, TEP, Grb2".

Distinguimos dos tipos de siglas, a saber:

1) Siglas propias. Unidades de reducción formadas exclusivamente a partir de las iniciales de unidades léxicas de estructura sintagmática; *e.g.*:

   SSCP (***S**ingle-**s**trand **c**onformational **p**olymorfism*)
   PCR (***P**olymerase **c**hain **r**eaction*)
   DMD (***D**istrofia **m**uscular de **D**uchenne*)

2) Siglas mixtas. Unidades de reducción en las que se han utilizado caracteres secundarios (letras que no son iniciales de la unidad léxica, cifras, símbolos) u omitido partes fundamentales de la forma desarrollada. También se les denomina siglas impropias o sigloides.

   Las siglas mixtas se clasifican en tres subclases, a saber: siglas mixtas típicas, acrónimos y cruces (*blends*).

   En primer lugar, las siglas mixtas típicas son aquellas unidades que emplean u omiten partes fundamentales de su forma desarrollada y cuya pronunciación puede ser alfabética, silábica o ambas; *e.g.*:

   Grb2 (***G**rowth factor **r**eceptor-**b**ound protein **2***)
   SRY (***S**ex determining **r**egion **Y***)
   SEF (***S**uperficie **ef**icaz*), etc.

   En segundo lugar, los acrónimos son unidades formadas por varios grupos de letras de los elementos de la forma desarrollada, cuya pronunciación es

exclusivamente silábica; es decir, aquellas formas de reducción léxica donde no se ha respetado el principio primario de tomar de las unidades léxicas sólo la letra inicial; *e.g.*:

LINE (***L****ong* ***in****terspersed* ***e****lements*)
HUGO (***Hu****man* ***G****enome* ***O****rganization*)
ICONA (***I****nstituto para la* ***co****nservación de la* ***na****turaleza*), etc.

En tercer lugar, los cruces (también denominados formas aglutinadas o *blends*) son unidades similares al acrónimo pero formadas mediante la combinación de dos segmentos de una unidad léxica de estructura sintagmática y de pronunciación silábica. Según Cabré (1993: 179), los cruces pueden adoptar formas diferentes de acuerdo con los segmentos que los integran:

a) Pueden combinar los segmentos iniciales del primer y segundo elemento del sintagma; *e.g.*:

- GeneBio (***Gene****va* ***Bio****informatics*)
- PubMed (***Pub****lic access to* ***MED****LINE*)
- Agrimed (***Agri****cultura* ***med****iterránea*)

b) Pueden combinar el segmento inicial de la primera unidad y el segmento final de la segunda; *e.g.*:

- Informática (***infor****mación auto****mática***)
- Ofimática (***ofi****cina auto****mática***)

c) Pueden combinar el segmento final de la primera palabra y el segmento inicial de la segunda (o muy raramente los segmentos finales de las dos unidades); *e.g.*:

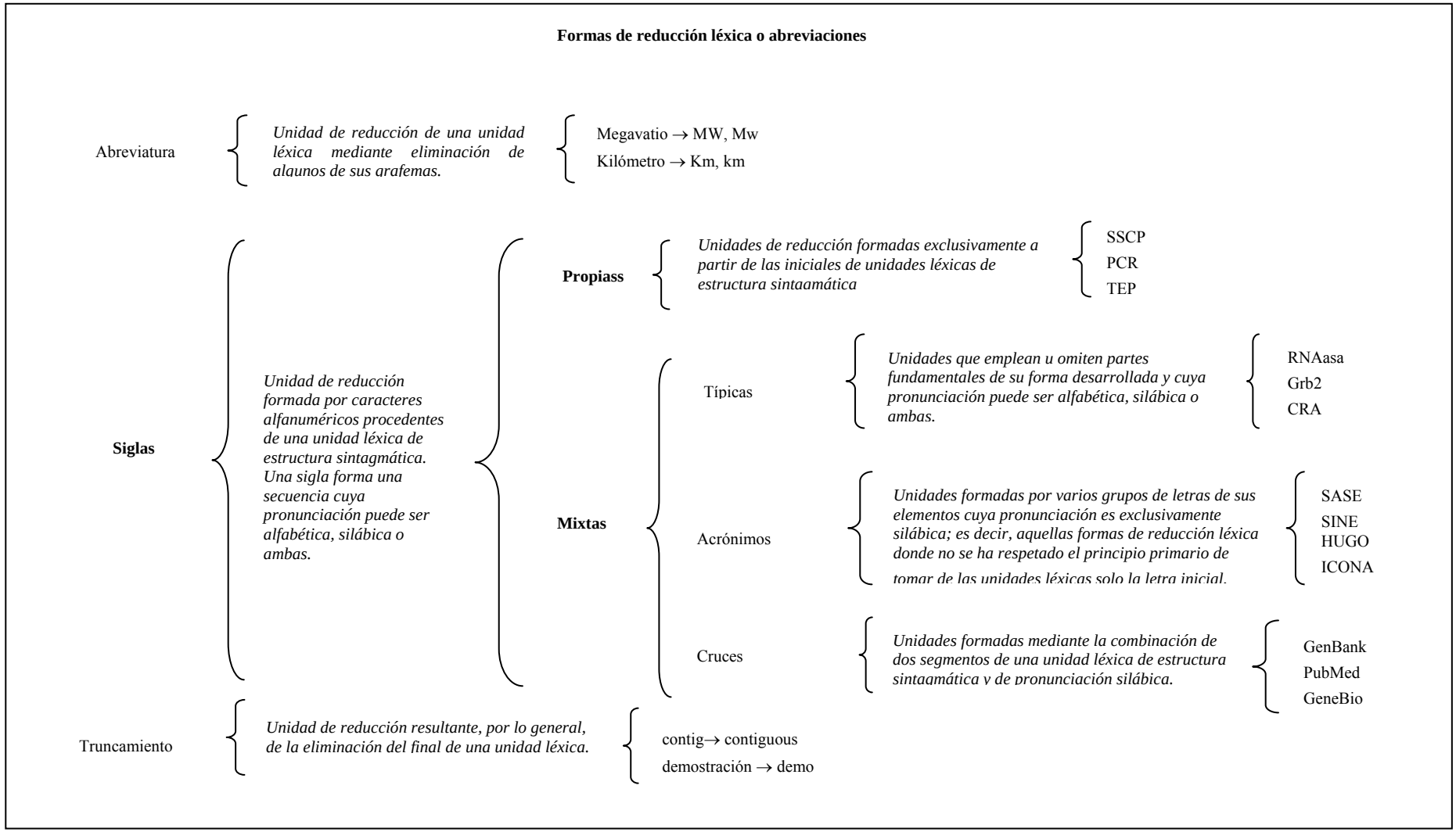
- **Tergal** (*poliés****ter*** ***gal****o*)

Adicionalmente, existe un tipo de unidades denominadas siglónimos. Se trata de aquellas siglas que se han lexicalizado; es decir, que se han incorporado a la lengua general como una palabra y se han sometido a las reglas de ésta. En una primera fase las siglas se escriben en mayúsculas, recurso gráfico que las caracteriza, sin embargo, el resultado final de la lexicalización es la pérdida de las mayúsculas; *e.g.*: *Síndrome de inmunodeficiencia adquirida* → **S.I.D.A** → **SIDA** → **sida.**

Una vez lexicalizadas pueden emplear procedimientos morfológicos como la derivación, por ejemplo: Sida → sídico/sidoso, etc.

El cuadro 3 recoge los conceptos sobre unidades de reducción léxica, tratados a lo largo de este apartado.

Cuadro 3. Clasificación de las siglas y demás formas de reducción léxica



Capítulo 3

Capítulo 3

Marco teórico

1. Los lenguajes de especialidad

Antes de seleccionar una teoría en la cual pueda enmarcarse este trabajo, conviene comentar algunos aspectos relevantes sobre el concepto de discurso especializado. Ciertamente, una revisión bibliográfica sobre el tema deja entrever una variedad de denominaciones para este concepto. Ciapuscio (2003: 25; 2002: 37) advierte sobre este fenómeno cuando sostiene que se debe plantear con toda claridad el problema denominativo, que frecuentemente responde a puntos de vista diferentes sobre el objeto: lenguajes de especialidad, lenguas para propósitos específicos (LSP), textos de especialidad, comunicación especializada, etc. Esta autora opta por la denominación "textos especializados" y los define como "productos predominantemente verbales de registros comunicativos específicos, definidos por los usuarios, las finalidades y las temáticas de los textos. Los textos especializados se refieren a temáticas propias de un domino de especialidad y responden a convenciones y tradiciones retóricas específicas. Los factores funcionales, situacionales y temáticos tienen correlato en el nivel de la forma lingüística, tanto en la sintaxis como en el léxico".

Domènech (1998: 31) coincide con Ciapuscio en la elección de la misma denominación; es decir, texto especializado, y lo define como "una producción lingüística que sirve para expresar y transmitir *conocimiento especializado*, que tiene una serie de características lingüísticas que le dan una especificidad dentro del conjunto de textos producidos en una lengua, y que presenta un conjunto de características pragmáticas que determinan los elementos específicos de su proceso de comunicación (el tema, los usuarios y la situación comunicativa)".

Hoffmann (1998: 51), por el contrario, usa el término "lenguaje de especialidad" y lo define como "el conjunto de todos los recursos que se utilizan en un ámbito comunicativo (delimitable en lo que respecta a la especialidad) para garantizar la comprensión entre las personas que trabajan en dicho ámbito". Para Hoffmann los lenguajes de especialidad son sublenguajes; es decir, sistemas del lenguaje que se actualizan en los textos de ámbitos comunicativos especializados.[12] De este modo, un lenguaje de especialidad es una selección de elementos lingüísticos y de las relaciones que mantienen en textos con una temática restringida. La especificidad de estos lenguajes frente al lenguaje común y los demás sublenguajes se expresa claramente en el vocabulario especializado (terminología), pero también en el uso de determinadas categorías gramaticales, de construcciones sintácticas y estructuras textuales.[13] Así mismo, hay algunas particularidades en la forma de las palabras, en la escritura, en la pronunciación y en el número de signos gráficos.[14]

Arntz y Picht (1995: 28), usan el término "lenguaje especializado" y lo definen como "el área de la lengua que aspira a una comunicación unívoca y libre de contradicciones en un área especializada determinada y cuyo funcionamiento encuentra un soporte decisivo en la terminología establecida".

12 *Op. Cit*. 71 y ss.

13 No todos los lingüistas aceptan el uso del término "lenguaje" para referirse a la comunicación especializada; *e.g*.: Quemada, Cabré, Ciapuscio, entre otros; prefiriéndose términos como "discurso de especialidad" o "texto especializado" (*cf.* Cabré, 1999: 151; Ciapuscio & Kugel, 2002: 41).

14 En este mismo sentido, Rondeau (1983) afirma que "los lenguajes especializados se caracterizan fundamentalmente por su léxico y por las características semánticas generales de los textos que producen: *"La terminologie a pour object la dénomination des notions; ce n'est donc que de façon accesoire que ses preoccupations rencontrent celles de la phonologie, de la morphologie et de la syntaxe"* (Cabré, 1999: 164).

Cabré, al igual que Ciapuscio, reflexiona sobre el problema de la existencia de diversas denominaciones y conceptualizaciones para discurso especializado tanto en su libro *La terminología: teoría, metodología, aplicaciones* como en el artículo "Textos especializados y unidades de conocimiento" (*cf.* Cabré, 1993: 126; 2002: 15). Ya en la primera de las obras citadas Cabré sugiere una definición consensuada para lenguajes de especialidad: "se trata de conjuntos 'especializados', ya sea por la temática, la experiencia, el ámbito de uso o los usuarios; se presentan como un conjunto con características interrelacionadas, no como fenómenos aislados y mantienen la función comunicativa como predominante, por encima de otras funciones complementarias".

En la segunda obra citada, esta autora usa las etiquetas de "texto" y "discurso" como sinónimos. Asimismo, hace referencia a condiciones discursivas cuando habla de las características del proceso discursivo que explican la generación de un texto y que interactúan con él. Cuando se refiere al texto, concebido como estructura, emplea el término "estructura textual". Es de notar que en la bibliografía más reciente Cabré opta por el uso del término "discurso especializado" (*cf.* Cabré, 1999: 151).

A la hora de analizar un texto, Cabré distingue entre las condiciones discursivas (generación, transmisión y recepción del texto) y la estructura textual (organización del texto).

Por un lado, las condiciones discursivas vienen determinadas por las condiciones de producción, transmisión y recepción de los textos, incidiendo, por tanto, en su adecuación. Establecer las condiciones de generación de un texto supone determinar las características de su emisor, receptor, canal, temática y las funciones que se le atribuyen o desean atribuir, etc.

Para Cabré (1999: 160) las características específicas de la comunicación especializada se centran en el emisor, el contenido del mensaje y el receptor; pero

afecta directamente al código y a la estructura lingüística del mensaje.[15] El emisor de un texto especializado no puede ser cualquier hablante de la lengua, sino que ha de ser necesariamente un especialista en la materia, porque sólo desde el conocimiento del tema se puede transmitir su contenido sin que se afecte su carácter de especializado. Esta transmisión puede hacerse de forma directa o indirecta, por medio de los traductores o intérpretes.

Por otro lado, el análisis de la estructura textual comprende tres grandes elementos o estructuras, a saber:

1) Estructura formal: todo texto presenta un formato específico de acuerdo con el género y tipo.

2) Estructura informativa o cognitiva: la información o conocimiento presente en los textos especializados está contenida en buena parte en los términos. Estos términos se organizan a través de relaciones conceptuales que se expresan en los textos, ya sea a través de unidades lingüísticas de carácter relacional u otros tipos de caracteres relacionales. Los textos especializados, en contraste con los generales, se caracterizan además desde el punto de vista

15 Pearson considera que la comunicación experto-experto se da cuando *"writer and reader, or speaker and hearer are assumed to have the same or very similar level of expertise. This expert-expert communicative setting applies to publications in learned journals, academic books, research reports, legal documents such as laws and contracts and any other written documents where the author is writing about his/her area of expertise and addressing readers who are understood to have a similar level of expertise"*. Por otra parte, acerca de la comunicación experto-aprendiz Pearson manifiesta que, *"frequently, experts working within a subject domain are called upon to communicate with others in their field who, while they have some knowledge of the field, do not have the same level of expertise. They may be students of a particular discipline, as in the case of advanced students in third level institutions. They may be people working within the same area but with a different training background, e.g., engineers and technicians, medical specialists and general practitioners...What distinguishes this type of communicative setting from the expert-expert context is the difference in the level of expertise of the writer and reader"* (Pearson, 1998: 36-37). En este mismo sentido Cabré sostiene que "... la consideración de los receptores del discurso especializado nos hace entrar también a distinguir entre discurso altamente especializado o medianamente especializado destinado a especialistas; el discurso didáctico o de aprendizaje, destinado a aprendices de una materia; y el discurso divulgativo*, dirigido al gran público" (1999: 170).
*Para profundizar más sobre el discurso divulgativo véase "Aspectos lingüísticos del discurso". En *La Ciencia empieza en la palabra: análisis e historia del lenguaje científico*. Gutiérrez Rodilla (1998: 326-332).

cognitivo por otras propiedades: precisión, concisión, sistematicidad, etc.[16] La precisión se controla mediante el uso de unidades léxicas no ambiguas, al menos en textos de cierto grado de especialidad; además, sus oraciones en conjunto tienden a ser unívocas. La concisión se manifiesta por la tendencia a describir una idea con el menor número de unidades posible y el uso de términos de estructura poco expandida. La sistematicidad se alcanza cuando hay control de la variación, especialmente la denominativa.

3) Estructura lingüística: todo texto es una red de unidades lingüísticas interrelacionadas entre sí. La estructura lingüística comprende varios niveles:

   a) Nivel morfológico
   b) Nivel sintáctico
   c) Nivel léxico-semántico
   d) Nivel textual

Desde el punto de vista léxico-semántico, el léxico utilizado es preferentemente denotativo, tiende a la monosemia, a la abundante presencia de términos, a los préstamos del latín y del griego, al uso de símbolos y de formas de reducción léxica como abreviaturas, siglas, etc.

En definitiva, la comunicación especializada, que se emplea en el discurso especializado, está influenciada por dos factores: la producción y la ordenación interna del texto, que comportarán siempre un grado de adecuación conforme a la clase de receptor (experto, aprendiz o lego).

Tras este breve repaso se constata que existen diversas denominaciones para el concepto de "discurso especializado", todas ellas condicionadas por las diferentes corrientes o puntos de vista teóricos. A pesar de ello, queda claro que la finalidad de este tipo de discurso radica en la comunicación de un conocimiento especializado

16 En este mismo sentido, Beaugrande & Dressler (1981) proponen 7 características estándar a saber: cohesión, coherencia, intencionalidad, aceptabilidad, informatividad, situacionalidad e intertextualidad. Véase Hoffmann (1998: 77).

dado. Este conocimiento está contenido esencialmente en los términos, de ahí que la terminología sea el principal rasgo diferenciador entre los discursos general y especializado. En este trabajo se ha adoptado el término discurso especializado.

En conclusión, el discurso especializado es la materialización en textos escritos y orales de la comunicación de conocimiento especializado (reflejado morfosintácticamente en los términos y semánticamente en los conceptos). Esta clase de discurso se caracteriza fundamentalmente por dos aspectos; por un lado, por la terminología, tal y como se acaba de mencionar. Por otro lado, por su grado de complejidad, dado en función del tipo de receptor al que se dirige el texto.

2. Las siglas: unidades terminológicas

> "[...] y a poco que un lector se interese por alguna rama del saber aún no frecuentada por él, tropezará con siglas que le resultarán extrañas o inabordables. Es el resultado lógico de la compartimentación y especialización científicas. Y en cada una de ellas, la tendencia general a la abreviación se utiliza dentro de su campo específico. Naturalmente, pueden darse concomitancias, parecidos entre unas siglas y otras" (Izquierdo, 1984: 9)[17]

La terminología y su vertiente aplicada, la terminografía, son uno de los subcampos de la lingüística aplicada con más dinamismo en la actualidad; basta con observar la gran cantidad de bases de datos, glosarios, vocabularios y diccionarios en formato electrónico y en formato papel publicados constantemente.

Partimos de la clasificación macro de las unidades de conocimiento especializado (UCE) propuestas por la Teoría Comunicativa de la Terminología (TCT), a saber: términos o unidades terminológicas (UT), unidades fraseológicas (UF) y unidades oracionales (UO).[18]

17 En: Martínez de Sousa, J. (1984). *Diccionario internacional de siglas y acrónimos*.
18 La TCT, formulada por María Teresa Cabré (1999), aboga por un enfoque menos prescriptivista para explicar la complejidad de las unidades terminológicas (UT).

En cuanto a la forma, las UT suelen clasificarse en diferentes grupos, así:

1) Por número de morfemas: los términos pueden ser simples o complejos; *e.g.*: ácido/acidificación;
2) Por tipos de morfemas que intervienen en la formación de un término complejo: los términos pueden ser derivados o compuestos; *e.g.*: ulceroso, balonmano;
3) Por combinación de palabras que siguen una determinada estructura sintáctica; *e.g.*: impuesto sobre la renta, pantalla líquida, y
4) Por truncaciones de términos de origen complejo como siglas y abreviaturas; *e.g.*: DNA (ácido desoxirribonucleico), etc.[19]

Consideramos que una sigla es una variante reducida de su propia forma desarrollada. Por consiguiente, la sigla posee el mismo estatus que su forma desarrollada, esto es, adquiere la condición de UT. Cabré (2003: 163), afirma, de acuerdo con el modelo de las puertas, que las UT deben cumplir las siguientes condiciones:

1) Desde la óptica del componente cognitivo:
   a) Depender de un campo temático;
   b) Ocupar un lugar preciso en el sistema de conceptos;
   c) Tener un significado específico, determinado según el lugar que ocupen en dicho sistema;
   d) Considerar su significado como una propiedad de la unidad, y
   e) Ser unidades fijas, reconocidas y difundidas con la ayuda de la comunidad de expertos.

2) Desde la óptica del componente lingüístico:
   a) Ser UT bien por su origen léxico o por un proceso de lexicalización;
   b) Poder tener estructura léxica o sintáctica;

19 Véase "Tipos de términos" (Cabré, 1993: 176-218).

c) Explotar todos los mecanismos de formación y los procesos de creación de nuevas unidades;
d) Poder coincidir formalmente con unidades que pertenecen al discurso general;
e) Poder presentarse ya sea como nombres, verbos, adjetivos y adverbios o como estructuras nominales, verbales, adjetivales o adverbiales;
f) Pertenecer a una de las categorías semánticas mayores: entidades, eventos, propiedades o relaciones;
g) Diferenciar su significado dentro de un campo especializado;
h) Extractar su significado a partir de un conjunto de información de una unidad léxica, y
i) Restringir su combinatoria sintáctica sobre la base de los principios combinatorios de todas las unidades léxicas de una lengua.

3) Desde la óptica de su componente comunicativo:
   a) Ocurrir en el discurso especializado;
   b) Adaptarse formalmente a este tipo de discurso de acuerdo con sus características temáticas y funcionales;
   c) Compartir el discurso especializado con unidades que pertenecen a otros sistemas simbólicos o icónicos;
   d) Adquirirse a través del proceso de aprendizaje y, por tanto, ser manejadas por los especialistas del área, y
   e) Ser básicamente denotativas (lo cual no excluye las connotaciones).

Hoffmann (1998, 74) señala que en el vocabulario especializado dominan los sustantivos y los adjetivos por encima de los verbos y otras categorías, porque han de designar la multiplicidad de objetos y manifestaciones hacia las cuales está orientada la actividad especializada. Los sustantivos y adjetivos forman un 60% del léxico de un texto especializado. El estudio de las siglas especializadas cobra relevancia puesto que, como se podrá constatar en los capítulos siguientes, son elementos de carácter nominal (*cf.* Fijo, 2003: 259; Gómez de Enterría, 1992: 267;

Nakos, 1990: 413; Martínez de Sousa, 1984: 39; Rodríguez, 1981: 82; Santoyo, 1980: 19).

3. Una teoría terminológica adecuada

Cabré (2003: 180) sostiene que una teoría puede tener varios grados de adecuación. Así, una teoría es adecuada para la observación si permite la descripción de los datos observados; es adecuada para la descripción si, además de permitir la descripción de los datos observados, permite la descripción de los inobservados que puedan surgir, lo cual le daría el carácter de predictiva y, por último, una teoría es adecuada para la explicación si logra explicar cómo y porqué se producen los datos y de qué manera se obtienen.

La Teoría Comunicativa de la Terminología, propuesta por Cabré (*cf.* 1999, 2002, 2003), es una teoría reciente que estudia las unidades del discurso especializado a partir de la lingüística. Desde 1996 esta autora lleva a cabo investigaciones que apuntan hacia una concepción teórica más amplia, dando cabida a las diferentes opiniones sobre los términos.

Los supuestos de partida de Cabré son:

1) La terminología es al mismo tiempo un conjunto de necesidades, un conjunto de prácticas para solucionar tales necesidades y un campo de conocimiento unificado; y
2) Los elementos centrales de la terminología son las unidades terminológicas (UT).

En primer lugar, se supone que la terminología es una necesidad para todas las áreas relacionadas, por un lado con la generación, y por otro, con la transferencia de

conocimiento especializado (traducción técnica, enseñanza de lenguas con propósitos específicos, documentación, planificación lingüística, redacción técnica, normalización, etc.). Los términos, en su sentido más amplio, son las unidades que mejor vehiculan el conocimiento dentro de un campo de conocimiento específico. En segundo lugar, una aplicación terminológica debe orientarse a la solución de necesidades específicas; de ahí que se deban tener en cuenta los destinatarios y las actividades que ellos desempeñan. De esta manera, variarán en función del destinatario tanto las diferentes aplicaciones (glosarios, lexicones, diccionarios, normas, programas informáticos, etc.) como la información que deben contener (terminología, fraseología, definiciones, variantes, contextos, equivalentes en otras lenguas, ilustraciones, etc.). En tercer lugar, la terminología es una disciplina y como tal es un conjunto formado por principios sobre su objeto de conocimiento. Por consiguiente, una teoría de la terminología debe proveer un marco metodológico lo suficientemente amplio como para satisfacer las posibles necesidades.

El objeto central de la terminología son las unidades terminológicas (UT) de carácter multifacético; es decir, que son al mismo tiempo unidades de conocimiento, de lenguaje y de comunicación. En palabras de Cabré “los términos no forman parte de un sistema independiente de las palabras, de otros sistemas de expresión y comunicación, sino que se solapan con ellos”.

La TCT se basa en estudios empíricos sustentables gracias a corpus lingüísticos. Reconoce la polisemia y la sinonimia como fenómenos reales y existentes en la terminología y los considera como naturales. La metodología de esta teoría tiene como propósito central la adecuación, o sea, la adaptación a las circunstancias propias de un trabajo terminológico dado sin contravenir los principios.

La terminología tradicional ha demostrado ser eficiente en tareas de estandarización en los campos teórico-conceptualmente más consolidados; es decir, en las ciencias puras y aplicadas (ingeniería, matemáticas, física, etc.); mas no en otros campos del conocimiento como las ciencias sociales y humanas, donde hay un mayor grado de variación conceptual y denominativa, lo que las hace áreas difíciles de estandarizar.

Por ello, las ciencias contemporáneas precisan de una teoría terminológica que permita dar cuenta de la diversidad denominativo-conceptual y que describa los términos en todas sus facetas. Tal vacío teórico-metodológico es el que pretende llenar la TCT.

En el artículo titulado "Theories of terminology: their description, prescription and explanation" (2003: 163), Cabré trata de resolver algunas cuestiones clave dentro de la terminología. Así, en lo referente a la adquisición de las UT la autora plantea que, si el discurso es el soporte natural de las UT, será a través de éste que el experto en formación adquirirá dichas unidades.

Respecto de dónde observar las UT queda claro que es necesario acudir al discurso oral y escrito producido por los expertos y dirigido a diferentes tipos de destinatarios, bien por medio de la lengua original o por medio de una traducción o interpretación. En otras palabras, la observación de las unidades terminológicas sólo podrá hacerse en el discurso producido en situaciones de comunicación especializada.

Una vez despejadas las incógnitas sobre la adquisición y el lugar de observación de las UT puede establecerse cuál es el marco idóneo para el estudio de tales unidades. En efecto, Cabré considera que se deben estudiar en el marco de la comunicación especializada. Este tipo de comunicación se caracteriza por las condiciones específicas del emisor, del receptor y del canal; por las condiciones del tratamiento de la información como la categorización precisa (determinada externamente por la estructura conceptual); la fijación y validación de la comunidad de expertos, y por las condiciones sobre la función y objetivos de dicha comunicación.

Una de las inquietudes frecuentes de los investigadores en terminología radica en los criterios de reconocimiento de las UT. A este respecto, Cabré sostiene que este tipo de unidades son de carácter léxico, de estructura morfológica o sintáctica, que ocupan un nodo dentro de la estructura conceptual de un campo de conocimiento dado. Como se ha indicado anteriormente, para Cabré además de las UT existen otras

unidades de conocimiento especializado, a saber: unidades morfológicas, unidades fraseológicas y unidades oracionales especializadas.

Aparte de reconocer las unidades terminológicas es necesario saber cómo se perciben dentro de la teoría lingüística. Ciertamente, desde una teoría del lenguaje natural se perciben como valores especializados de las unidades léxicas contenidas en el lexicón del hablante. De hecho, si se analizan las características fonológicas, morfológicas y sintácticas de las UT no encontraremos propiedades que las diferencien de las unidades del léxico general; en cambio, se sabe que presentan especificidades en su vertiente semántica y pragmática. Por consiguiente, Cabré postula que una unidad léxica no es en sí terminológica o general sino que, por defecto, es una unidad general y adquiere valor especializado cuando por las características pragmáticas del discurso se activa su significado especializado. Adicionalmente, cabe preguntarse cuál teoría del lenguaje se necesita para conservar la multidimensionalidad del objeto. En este sentido, la autora afirma que sólo una teoría lingüística de base cognitiva, funcional y social es capaz de describir las UT en su especificidad, pero también en lo que comparten con las unidades léxicas no especializadas. La gramática desarrollada en esta teoría, que ha de explicar la estructura y el uso de las unidades, necesita contener semántica y pragmática, además de gramática estricta. La semántica es imprescindible para explicar la especificidad del valor terminológico de las unidades léxicas; la pragmática es necesaria para explicar la activación del valor terminológico de estas unidades léxicas.

Parafraseando lo expuesto por Hoffmann, se puede decir que algunos de los rasgos diferenciadores de un discurso de especialidad, como el de los ámbitos de genoma humano o medio ambiente, frente al discurso general se expresan claramente en la terminología, pero también en el uso predominante de determinadas categorías gramaticales, de construcciones sintácticas y estructuras textuales. Así mismo, hay algunas particularidades en la forma de las palabras, en la escritura, en la pronunciación y en el número de signos gráficos.

Finalmente, se ha expresado antes que el instrumento más eficaz para observar, describir y explicar un objeto de estudio es una teoría. En este sentido, el presente trabajo parte de la TCT, dado que las siglas cumplen las condiciones propias de las unidades terminológicas, para dar cuenta de los fenómenos que les atañen al interior del discurso especializado del genoma humano y del medio ambiente.

Capítulo 4

Capítulo 4

Los diccionarios de abreviaciones en línea

1. Antecedentes

> *"Whenever you open a scientific, technical, or economic publication, or even a daily newspaper, you are immediately struck by the number of apparently meaningless letter or syllable combinations which the most knowledgeable reader cannot decipher without the aid of a dictionary or a keen sense of divination"* (Sliosberg).[20]

Un diccionario se concibe como un producto lingüístico, que recoge un tipo determinado de unidades de la lengua, surgidas de la influencia que los hablantes tienen sobre la lengua gracias a sus costumbres, conocimientos y necesidades.

El hombre ha hecho uso de las abreviaciones, en mayor o menor medida, desde hace bastante tiempo. Según Calvet (1980: 13) las siglas han predominado desde la antigüedad en los manuscritos sobre economía y religión. A partir del siglo XIX cobraron fuerza y comenzaron a usarse en los diferentes campos del saber; pero, no es hasta el siglo XX, justo después de la II Guerra Mundial, cuando se da su verdadero auge.

La historia moderna de los diccionarios de abreviaciones se divide en dos grandes etapas. Por un lado, el periodo comprendido entre 1950 y 1980, que da cuenta de la

20 En Mossman, J. (1992). *Acronyms, Intialisms and Abbreviations Dictionary*.

gran cantidad de diccionarios publicados en formato papel. Por otro lado, el periodo comprendido desde 1980 hasta la actualidad, caracterizado por el surgimiento de los diccionarios en línea.

En el presente capítulo se analizan los principales diccionarios de abreviaciones existentes en la red actualmente. En vista de que estos recursos no se dedican a recoger exclusivamente siglas sino todo tipo de abreviaciones, hemos titulado este capítulo los diccionarios de abreviaciones en línea.

1.1 Los diccionarios de abreviaciones en formato papel

El auge de las abreviaciones, en especial de las siglas, ha llevado a la necesidad de recogerlas y documentarlas en diversos recursos a lo largo de los últimos siglos. Dan testimonio de ello obras como el *Tractatus de Siglis Veterum* de Nicolai Johannis (1703),[21] *Abréviations de sociétés, conventionnelles et usuelles*, publicado por el Lloyd Anversois en Amberes (1926),[22] o el *Dictionnaire d'abréviations françaises et étrangères, techniques et usuelles, anciennes et nouvelles*, publicado por Ediciones de Montligeon (1951) *cf.* Gehénot (1990: 105). [23]

Casos recientes como la sigla SARS (*Severe Acute Respiratory Syndrome*) son prueba de que las abreviaciones, también llamadas reducciones léxicas, son un fenómeno vigente, revigorizado por el avance del conocimiento y propio de la lengua que tiende a dejarse influenciar, cada vez más, por factores como la economía lingüística, la mnemotecnia, la estilística y los criterios editoriales. En este sentido, es fácil encontrar hoy en día obras dedicadas a documentar la gran cantidad de

21 Esta obra escrita en latín comprende 314 páginas y 49 capítulos que estudian en detalle el uso de las siglas en un tema particular, apoyado por numerosos ejemplos: derecho, medicina, aritmética, gramática, música, numismática, etc.
22 Abreviaturas marítimas, bursátiles, comerciales (francesas, inglesas, alemanas y españolas), abreviaturas de bancos, gremios, abreviaturas convencionales en uso en los servicios de registro y de propiedades.
23 Este diccionario reúne 8.000 abreviaturas de artes, automoción, aviación, banca, cartografía, química, comercio, ferrocarriles, derecho, electricidad, finanzas, impuestos, industria, jurisprudencia, marina, matemáticas, mecánica, medicina, etc.

abreviaciones que genera la actividad humana; *e.g.*: *Dictionnaire international d'abréviations scientifiques et techniques* (1978), Diccionario internacional de siglas y acrónimos (1984), *Dictionnaire des abréviations et acronymes scientifiques, techniques, medicaux, économiques et juridiques* (1992), *Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols* (1997), *Acronyms, Initialisms and Abbreviations Dictionary* (2003), etc.

1.2 Los diccionarios de abreviaciones en línea

Como se ha mencionado anteriormente, este tipo de diccionarios surge a finales de la década de 1980. Su objetivo era responder a la necesidad de conocer las formas desarrolladas de la gran cantidad de abreviaciones que se venían produciendo en todos los campos del conocimiento.

Los diccionarios en línea superan ampliamente la capacidad de almacenamiento de información de los diccionarios en papel. En general, el formato en línea ofrece, aparte de una interfaz de consulta, la integración de diferentes formatos como pueden ser texto, imagen, sonido o vídeo.

En la actualidad existen diversos diccionarios de abreviaciones disponibles en Internet, entre los que sobresalen *Acronym Server* (1988), *Acronym Finder* (1996), *Wiley InterScience* (1999), *Abbreviations.com* (2001) y *Acronyma* (2004). Esta clase de diccionarios de abreviaciones (principalmente siglas) funciona mediante una interfaz de consulta donde generalmente se pueden buscar dos tipos de información: la sigla y la forma desarrollada.

Se han tomado los tres principales diccionarios de abreviaciones existentes en la web, *i.e.*, *Acronym Finder*, *Abbreviations.com* y *Acronyma*.[24] A cada uno de ellos se les realiza un análisis general y otro específico.

2. Análisis y resultados

Se han realizado dos tipos de análisis, uno general y otro especializado. El análisis general consiste en aplicar a cada diccionario los criterios de calidad para los recursos de la web. El análisis específico consiste en observar los cinco niveles de estructura lexicográfica de cada diccionario para determinar su grado de complejidad y compleción y, por consiguiente, la calidad del tratamiento de las siglas.

Los diccionarios impresos presentan una estructura de la información diferente de los diccionarios electrónicos; organizan la información en una lista de entradas donde cada entrada se crea por medio de un modelo de microestructura, que varía dependiendo del diccionario; *e.g.*:

BP *Autom.* Bechuanalandia (desaparecida). ‖ *Corp.* Deustche Bundespost (Correo Federal Alemán), RFA. ‖ *Econ.* Banco Pastor (España). ‖ *Econ.* Banco Peninsular (España). ‖ *Ind.* British Petroleum Company Ltd. (Compañía Británica de Petróleos), Fund. 1909. ‖ *Pol.* Bayernpartei (Partido Bávaro). Fund. 1946 (RFA).

Diccionario internacional de siglas y acrónimos, Martínez de Sousa (1984).

CTM. Comité de Transportes Marítimos. *Corp.* ● Conférence Technique Mondiale (Conferencia técnica mundial). *Ind.*

Diccionario de siglas y abreviaturas, Alvar (1983).

24 En esta selección se ha tenido en cuenta su representatividad medida en: cantidad de entradas y temáticas tratadas.

2.1 Análisis general: calidad de los diccionarios en línea como recursos de la web

A raíz de la facilidad de acceso que supone Internet, es necesario analizar la calidad de la información que ofrece. En este sentido, y de acuerdo con Sánchez-Gijón (2004: 33), hemos tomado los diez criterios básicos para analizar la calidad de los diccionarios objeto de este estudio:

1) Autoría. Un recurso es más fiable en la medida en que quien lo publica es un experto en el tema.
2) Actualidad. La información actualizada es tan importante como la presencia de las fechas de creación y actualización del recurso.
3) Precisión. La información debe ser precisa tanto en el contenido como en la forma; debe estar libre de errores de coherencia, de ortografía, etc.
4) Tratamiento del contenido. La objetividad, extensión y compleción de la información son imprescindibles para evaluar si se ha dado un buen tratamiento al contenido de un recurso.
5) Originalidad. Cada recurso debe demostrar que su contenido es original. De lo contrario debe citar la fuente de referencia.
6) Propósito. El autor debe explicitar la intención de los contenidos, de manera que se facilite su interpretación.
7) Enlaces a otros recursos. Para determinar la validez de un recurso es necesario que existan enlaces y comentarios sobre recursos similares en contenido y calidad.
8) Ergonomía. El diseño de los recursos debe facilitar la consulta y la navegación interna por medio de un mapa del sitio o menú.
9) Citación. El hecho de que el recurso sea citado en otros sitios es un índice de calidad del mismo.
10) Receptor. Debe tenerse en cuenta el tipo de público al que se dirige el recurso.

La tabla 6 contiene el análisis comparativo correspondiente a los tres diccionarios seleccionados para el estudio.

Criterio	Acronym Finder	Cumple criterio	Acronyma	Cumple Criterio	Abbreviations.com	Cumple Criterio
Autoría	Cualquier usuario (experto o lego) puede remitir una abreviación. Cada abreviación es revisada, verificada en múltiples fuentes, clasificada según las categorías establecidas y editada.	Sí	Cualquier usuario (experto o lego) puede remitir una abreviación. Cada abreviación es revisada y editada antes de su incorporación definitiva al diccionario.	Sí	Cualquier usuario (experto o lego) puede remitir una abreviación. Cada abreviación es revisada y editada antes de su incorporación definitiva al diccionario.	Sí
Actualidad	Permanentemente actualizado. Incorpora 5.000 abreviaciones mensuales en promedio.	Sí	No proporciona información sobre la frecuencia de actualización de la información; pero se ha detectado que el período de revisión es superior a un mes.	+/-	Revisión diaria de las entradas.	Sí
Precisión	Información generalmente precisa, como se ha constatado durante la verificación de las formas desarrolladas de un corpus de 800 siglas del ámbito de Genoma humano. Sin embargo, se han observado errores de acentuación en español y francés. *e.g.*: *petroleo* por *petróleo*. *A* por *À*, etc.	+/-	Al igual que en AF se han detectado algunos errores en español del tipo *acido* por *ácido*. Esto puede deberse a error humano o simplemente porque la entrada se publicó sin editar.	+/-	No se han observado errores de coherencia ni de ortografía.	Sí
Tratamiento del contenido	Cuenta con un editor jefe y varios editores.	Sí	Cuenta con un editor que se encarga de verificar el contenido de la información que se incorpora.	Sí	Cuenta con 310 editores.	Sí
Originalidad	No cita la fuente de las abreviaciones ni de sus formas desarrolladas. Sólo presenta una lista de las personas que han remitido más de 50 abreviaciones.	No	No cita la fuente de las abreviaciones. Se limita a informar que la gran mayoría son extraídas de la web.	No	No cita la fuente de las abreviaciones.	No
Propósito	El vínculo «*About*» presenta toda la explicación sobre los contenidos, propósito y forma de consultar el diccionario.	Sí	Usa la web como corpus para buscar y documentar las abreviaciones.	Sí	El vínculo «*About*» incluye una presentación del recurso así como la definición de sigla, abreviatura y acrónimo.	Sí
Enlaces a otros recursos	Presenta enlaces a otros recursos similares como *Freedictionary.com* y *Acronym Atic*, una base de datos de siglas complementaria.	Sí	No presenta enlaces a otros recursos similares.	No	El vínculo «*Links*» lleva a otros sitios web que contienen siglas en diversas áreas como aviación, biodiversidad, química, ciencias de la tierra, etc.	Sí
Ergonomía	Posee mapa del sitio; además, es de muy fácil manejo para cualquier usuario con conocimientos básicos de internet. Dedica demasiado espacio a la publicidad, lo cual entorpece la navegación por el recurso.	Sí	No tiene mapa del sitio; no obstante, es de muy fácil utilización.	No	No cuenta con mapa del sitio; sin embargo, es un recurso de fácil utilización. La gran cantidad de espacio cedido a la publicidad entorpece la navegación por el recurso.	No
Citación	El sitio es frecuentemente citado por otros recursos en la red, aunque no por los demás diccionarios de abreviaciones analizados aquí.	+/-	No se han encontrado citaciones de este recurso en sitios similares.	No	No se han encontrado citaciones de este recurso en sitios similares.	No
Receptor	Se ofrece a un público muy amplio.	Sí	Se ofrece a un público muy amplio.	Sí	Se ofrece a un público muy amplio.	Sí

Tabla 6. Análisis comparativo de los diccionarios desde el punto de vista de la calidad de los recursos de la web

Este análisis ha permitido observar que:

En general, los diccionarios considerados aquí cumplen cabalmente con 4 de los 10 criterios establecidos; *i.e.*, autoría, tratamiento del contenido, propósito y receptor. Por el contrario, no cumplen con los requisitos de originalidad y ergonomía.[25]

En particular, *Acronym Finder* cumple 7 criterios; *i.e.*, autoría, actualidad, tratamiento del contenido, propósito, enlace a otros recursos, ergonomía y receptor; incumple un criterio (originalidad) y cumple parcialmente 2 criterios (precisión y citación).

Acronyma cumple 4 criterios, a saber: autoría, tratamiento del contenido, propósito y receptor. En cambio, incumple los criterios de originalidad, enlace a otros recursos, ergonomía y citación. Por último, cumple parcialmente los criterios actualidad y precisión.

Abbreviations.com cumple 7 criterios; *i.e.*, autoría, actualidad, precisión, tratamiento del contenido, propósito, enlace a otros recursos y receptor. Sin embargo, no cumple con los criterios de originalidad, ergonomía y citación.

De lo anterior, se deduce que los diccionarios que mejor se adaptan a los criterios de calidad de lo recursos de la web son *Abbreviations.com* y *Acronym Finder*.

25 Datos recogidos y analizados en junio de 2007

2.2 Análisis específico

2.2.1 Estructura

El análisis específico consta de dos partes. En la primera, se observa la estructura lexicográfica de cada diccionario y, en la segunda, se analizan los resultados tras una búsqueda específica.

De acuerdo con Gelpí (2000), partimos de los cinco niveles de la estructura lexicográfica; *i.e.*, hiperestructura, macroestructura, microestructura, iconoestructura y estructura de acceso. Aunque dichos criterios han sido establecidos para los diccionarios en papel no dejan de ser extrapolables a los diccionarios en línea.

La tabla 7 presenta el análisis específico de cada uno de los diccionarios estudiados aquí.

Estructura lexicográfica	Elementos de la estructura	Diccionario		
		AF	**Acronyma**	**Abbreviations.com**
Hiperestructura	Parte inicial.			
	-Título	Sí	Sí	Sí
	-Índice	No	No	No
	-Guía de uso	Sí. Se encuentra bajo el vínculo de *«Help»*	No	Sí
	-Datos metalingüísticos	No	No	No
	-Prólogo/introducción	No	No	No
	Cuerpo del diccionario	3.050.000 abreviaciones, aprox.	480.000 abreviaciones, aprox.	410.000 abreviaciones, aprox.
	Parte final (opcional)			
	-Inf. Fonética	No	No	No
	-Inf. Ortográfica	No	No	No
	-Inf. Gramatical	No	No	No
Macroestructura	Selección de la nomenclatura	Por criterios de frecuencia y adecuación	Por criterios de frecuencia y adecuación	Por criterios de frecuencia y adecuación
	Forma de representación de las entradas	Abreviaciones (entendidas como lemas)	Abreviaciones	Abreviaciones
	Ordenación de las entradas	Alfabética y sistemática[(1)]	Alfabética	Alfabética y sistemática
Microestructura	Lema	Sí	Sí	Sí
	Inf. Gramatical	No	No	No
	Marcas	Ocasionalmente presenta la marca del ámbito al que pertenece la abreviación	No	No
	Contextos	No	No	No
	Equivalencia	Ocasionalmente	No	No
	Definición (forma desarrollada de la abreviación)	Sí	Sí	Sí
	Ejemplos	No	No	No
	Acepciones	No	No	No
	Subentradas	No	No	No
	Inf. Fonética	No	No	No
Iconoestructura	Ilustraciones	No	No	No
Estructura de acceso	Relaciones horizontales	No	No	No
	Relaciones verticales	No	No	No
	Relaciones transversales	No	No	No
	Relaciones externas	Sí	Ocasionalmente	Sí

[26]Tabla 7. Análisis específico

26 [(1)] Las entradas de los diccionarios pueden organizarse de dos formas, a saber: alfabética o sistemática. En concreto, la ordenación sistemática consiste en organizar las entradas del diccionario de acuerdo con un número pre-establecido de descriptores (normalmente temáticos). Se da por descontado que un diccionario pueda usar ambos sistemas de ordenación.

De los datos de la tabla anterior se desprende que existen carencias en la información que ofrecen los recursos bajo estudio. Tanto la hiperestructura como la macroestructura de los recursos son aceptables; sin embargo, no sucede lo mismo con el nivel de la microestructura. En efecto, se ha detectado que de las 10 categorías que conforman este nivel, sólo se cumple con el lema (abreviación) y la definición (forma desarrollada). Sólo *Acronym Finder* incluye, ocasionalmente, las categorías de "marca" y "equivalencia". Ninguno de los diccionarios tiene en cuenta la iconoestructura, a pesar de las facilidades que tienen estos recursos para incorporar imágenes. Finalmente, en lo que concierne a la estructura de acceso, no se han encontrado ni relaciones horizontales, ni verticales ni tampoco transversales. La única relación evidenciada es la externa, la cual se ha encontrado en *Acronym Finder*, *Abbreviations.com* y, ocasionalmente, en *Acronyma*.[27]

2.2.2 Análisis de resultados

Para este análisis se ha escogido la sigla ADN (ácido desoxirribonucleico). Se ha interrogado cada una de las interfaces de los diccionarios escogidos para obtener su forma desarrollada en español. Los resultados son los siguientes:

27 De acuerdo con Gelpí (2000: 16) las relaciones de la estructura de acceso pueden ser: a) horizontales: cuando se da una indicación a otra indicación en el límite del artículo del diccionario, como puede ser la relación que se establece entre la marca gramatical y el lema hacia el cual orienta; b) verticales: cuando se da una indicación entre dos artículos diferentes como puede ser el nexo que vincula un artículo a otro por medio de la marca "véase"; c) transversales: cuando se da una indicación a un componente del diccionario que no forma parte de la nomenclatura principal, como puede ser una marca de asignación temática a la lista de marcas temáticas que ofrece el diccionario en la parte introductoria. Así mismo, Gelpí considera que las relaciones externas de las estructuras de acceso implican la manera como el usuario accede y recupera la información que el diccionario le ofrece. De ahí que sea habitual que un diccionario especializado, por ejemplo, permita la consulta de información enciclopédica por medio de un vínculo. El usuario desea saber, por ejemplo, no sólo qué significa el lema "ley orgánica", sino el tipo de leyes de esta naturaleza que un parlamento ha aprobado en un periodo de tiempo concreto. En nuestro caso, consideramos como relaciones externas la presencia en los diccionarios analizados de vínculos a otros recursos en línea, bien sea buscadores o diccionarios.

2.2.2.1 Acronym Finder (AF)

Acronym Finder presenta los resultados en forma de tabla. Para el caso de la sigla ADN se tiene que la primera columna corresponde al icono "*i*", que representa "*More information & Searches*". La segunda columna muestra el rango (*rank*) o relevancia de la sigla, indicado mediante asteriscos.[28] La tercera y cuarta columnas presentan la sigla y su forma desarrollada, respectivamente. La quinta columna muestra el icono de "libro", que representa que el usuario puede buscar más información en "*Freedictionary.com*" y, por último, la sexta columna indica el enlace de "*Go*" que conduce a la librería "*Amazon*", para buscar libros relacionados con la abreviación consultada.

AF también permite buscar la sigla mediante el "*Acronym Database Surfer*"; es decir, la lista de todas las abreviaciones del recurso organizadas en orden alfabético.

El resultado esperado; *i.e.*, la forma desarrollada "ácido desoxirribonucleico" aparece efectivamente en la quinta entrada. El resultado de la consulta correspondiente a la sigla ADN es el siguiente:

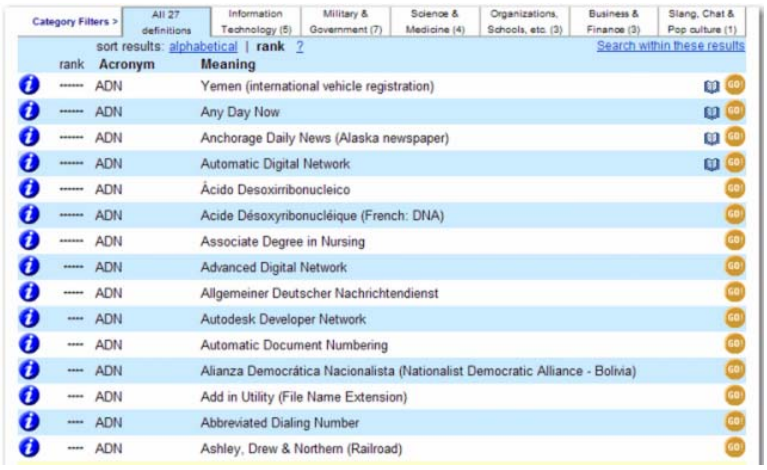


Fig. 1. Resultado en Acronym Finder

28 *Rank* equivale a "importancia". Esto quiere decir que, al igual que en *Acronyma*, las abreviaciones pueden organizarse alfabéticamente y por grado de importancia.

2.2.2.2 Acronyma

Acronyma organiza los resultados de la búsqueda de la sigla ADN en forma de tabla, donde la primera columna corresponde a la sigla y la segunda a la forma desarrollada. Además, en la parte superior de la tabla, se presenta el número de resultados obtenidos en cada lengua, en este caso, 6 significados de la sigla ADN en inglés, 1 en español, 1 en francés y 1 en alemán. No hay resultados en holandés, italiano ni en portugués.

Acronyma permite que los resultados se organicen de dos formas: alfabéticamente o por rango de importancia y, al igual que AF, permite búsquedas de abreviaciones por medio de una lista organizada alfabéticamente disponible en la casilla de interrogación. El siguiente es el resultado de la consulta correspondiente a la sigla ADN:

Fig. 2. Resultado en Acronyma

2.2.2.3 Abbreviations.com

Al igual que los dos diccionarios precedentes, *Abbreviations.com* presenta los resultados en forma de tabla de cinco columnas, donde la primera indica el rango de importancia de la sigla, la segunda la forma desarrollada de la sigla, la tercera la categoría a la que pertenece la sigla (*Academic & Science*, *Computing*,

Miscelaneous, etc.), la cuarta el enlace a la librería *Amazon* y la quinta el enlace a *Google*.

Al contrario de los dos recursos anteriores, el resultado esperado "*ácido desoxirribonucleico*" no aparece.

El siguiente es el resultado de la consulta correspondiente a la sigla ADN:

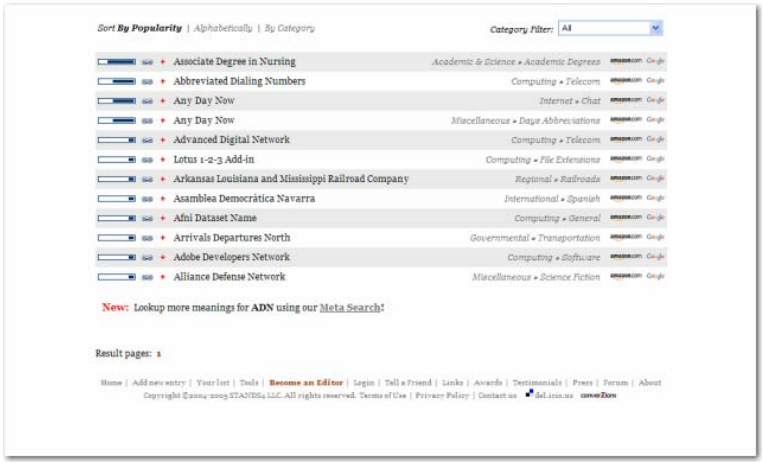


Fig. 3. Resultado en Abbreviations.com

El análisis anterior también ha permitido observar que:

1) Ninguno de los diccionarios distingue el tipo de discurso del que provienen las siglas; es decir, si pertenecen al discurso general o especializado aunque, de cierto modo, *Acronym Finder* y *Abbreviations.com* intentan hacerlo por medio de los *Category filters*.
2) Sólo *Abbreviations.com* intenta clasificar las abreviaciones de acuerdo al campo temático del que proceden. Para ello ha establecido 10 categorías mayores y 132 subcategorías.[29]
3) A pesar del potencial que tienen los diccionarios en línea para almacenar información e incluso incorporar imágenes, no deja de sorprender que la

29 Las 10 categorías son: *Computing, Internet, Academic & Science, Miscelaneous, Medical, Business, Governmental, Community, Regional, International.*

información que proporcionan sobre las abreviaciones sea más reducida que la de la mayoría de diccionarios en papel mencionados anteriormente.

4) A diferencia de *Acronyma*, *Acronym Finder* y *Abbreviations.com* no dan la opción de consultar las abreviaciones por lenguas (SP, EN, FR, etc.).
5) Ninguno de los recursos bajo análisis ofrece la posibilidad de realizar referencias cruzadas mediante enlaces, de modo que puedan consultarse los equivalentes de una abreviación en otras lenguas; *e.g.*:

> **PCR** (Polymerase chain reaction) → (SP)= **RCP** (Reacción en cadena de la polimerasa), o **RCP** (Reacción en cadena de la polimerasa → (EN)= **PCR** (Polymerase chain reaction).

6) Ninguno de los diccionarios brinda información de tipo gramatical, fonológico, grafemático o semántico, valiosa para colectivos profesionales como los traductores, intérpretes y terminólogos. Si se incorporaran contextos de aparición de estas unidades y sus formas desarrolladas quedarían resueltos aspectos como el género de las siglas.
7) Ninguno de los recursos etiqueta las abreviaciones, es decir, no establecen la distinción entre sigla, abreviatura o truncamiento. Una distinción que, sin duda, ayudaría a evitar la confusión terminológica que existe alrededor de la tipología de abreviaciones.

3 Conclusiones

Los fenómenos de abreviación, y especialmente la siglación, interesan a áreas como la traducción, la lexicología, la terminología, la redacción técnica, los LSP o la lingüística computacional.

Cada ámbito de especialidad genera sus propias abreviaciones. De ahí que, en un intento por recoger y documentar este tipo de unidades, surjan constantemente recursos electrónicos y en papel tales como diccionarios, glosarios o bases de datos.

En general, vistos como recursos de la web, los diccionarios de abreviaciones tienen una calidad aceptable. Sin embargo, presentan deficiencias en lo que respecta a criterios como la originalidad y la ergonomía.

Acronym Finder es el recurso de abreviaciones más grande que existe en la actualidad en la web. No obstante, al igual que los demás recursos mencionados en este trabajo, se preocupa por recoger el mayor número de unidades, pero deja de lado gran cantidad de información complementaria sobre las abreviaciones que podría ser de gran utilidad para el perfil de usuarios que a diario consultan estos recursos, entre los que se cuentan traductores, intérpretes, terminólogos, profesores de LSP y redactores técnicos.

De todo lo anterior se deduce que estos recursos, aunque cumplen la función de diccionarios en línea, no se crean con criterios verdaderamente lexicográficos o terminográficos (cosa que si suele suceder con la mayoría de los diccionarios de siglas en papel). Por tanto, hace falta proponer modificaciones o nuevos recursos que efectivamente describan en detalle las características de las abreviaciones para garantizar una información de mayor calidad al mayor número de perfiles de usuario.

En definitiva, para compilar el mayor número de datos posible sobre una abreviación se debería incluir en los diccionarios de abreviaciones, tanto electrónicos como en papel, al menos las siguientes categorías de datos:

1) Abreviación
2) Área o campo temático
3) Lengua en que aparece la abreviación
4) Tipo de abreviación (sigla, abreviatura, truncamiento, etc.)
5) Forma desarrollada (FD) o expansión
6) Fuente de la forma desarrollada
7) Contexto(s) donde aparece(n) la abreviación y su forma desarrollada
8) Pronunciación (silábica, deletreada)

9) Aspectos grafémicos (uso de mayúsculas, minúsculas, etc.)
10) Aspectos sintácticos (género, número)
11) Equivalente en otras lenguas
12) Información sobre el origen de la abreviación como año y lugar de aparición (opcional).

Capítulo 5

Capítulo 5

Metodología

Como hemos indicado en el capítulo 1, a partir de las conclusiones del trabajo de exploración titulado "Siglas y variación vertical en textos sobre genoma humano y medio ambiente", concretamos el enfoque de la tesis mediante la segunda fase de la investigación, denominada Proyecto de tesis, la cual partió de las siguientes consideraciones:

1) Que la hipótesis inicial sólo pudo confirmarse parcialmente, al no comprobarse que el número de siglas y variantes formales fuera proporcional al nivel de especialización del texto en que aparecían; *i.e.*, a mayor nivel de especialización menor nivel de variación.
2) Que existía una dificultad real para conformar un corpus en español, que incluyera suficientes textos de nivel de especialidad alto en MA y de nivel de especialidad bajo en GH[30] y
3) Que en el IULA se llevaban a cabo estudios paralelos sobre el grado de especialización de los textos mediante técnicas de medición de densidad terminológica. Puesto que las siglas se consideran formas reducidas de una unidad terminológica, dichos estudios incluían *per se* el análisis e influencia de las siglas en los diferentes niveles de especialidad.

30 Actualmente el CT-IULA no procesa ningún documento nuevo por lo que asumir esta tarea individualmente implicaba una enorme inversión de tiempo, lo que retrasaría considerablemente el cronograma de trabajo.

Como consecuencia, desestimamos continuar el trabajo focalizado en la incidencia de las siglas en el grado de especialidad de los textos y decidimos orientar la investigación al estudio de las siglas desde una perspectiva general de los ámbitos de genoma humano y medio ambiente.

A partir de aquí se establecieron los objetivos de la tesis, los cuales han quedado expuestos en el capítulo 1 y que, de manera general, resumimos así:

1) Analizar el concepto y la tipología de las siglas y revisar su tratamiento en la bibliografía;
2) Observar las siglas en textos de especialidad de genoma humano y medio ambiente y establecer sus características lingüísticas y estadísticas, y
3) Determinar los criterios para el diseño de un sistema de detección y extracción semiautomática de siglas.

En consonancia con dichos objetivos hemos empleado la siguiente metodología:

1. Constitución del corpus

Para cada ámbito de especialidad, GH y MA, se ha constituido un corpus textual y, a partir de cada uno de ellos, el corpus de siglas. Su proceso de obtención ha seguido fases idénticas a las llevadas a cabo durante la fase de exploración del tema (*cf*. capítulo 1).

1.1 Corpus textual

A partir del CT-IULA se han recogido 158 textos del área de genoma humano, que contienen 999.950 palabras, y 47 textos del área de medio ambiente, que contienen 999.876 palabras. Tal y como se aprecia en la siguiente tabla:[31]

Campo de especialidad	Nº total doc. en español	Nº total de palabras en español
Genoma humano (GH)	158	999.950
Medio ambiente (MA)	47	999.876

Tabla 8. Valores totales de los documentos por campo de especialidad y lengua

1.2 Corpus de siglas

A partir de los corpus textuales se ha conformado el corpus de siglas respectivo. Dicho corpus cuenta con 800 siglas en GH y 317 en MA. Como explicaremos más adelante, hemos fusionado ambos corpus para el análisis lingüístico mientras que para el análisis estadístico los hemos tratado tanto juntos como por separado.

2. Recolección de los datos

1) Depuración manual.[32] Ambos corpus han presentado un nivel de ruido bajo representado, por un lado, en la presencia de otras unidades de reducción léxica como son las abreviaturas y, por otro lado, en unidades que, por su estructura, se asemejan a las siglas, por ejemplo, algunas palabras cortas escritas en

31 BwanaNet es la interfaz web al CQP para interrogar al CT-IULA. Tiene dos modos de interrogación:
-Concordancias (KWIC), que pueden ser: simples (forma/lema), estándar (secuencias forma/lema/categoría) o complejas (lenguaje del CQP restringido);
-Frecuencias, que pueden ser: para un grupo de documentos o para todo el CT.

32 Entendemos por depuración manual el proceso mediante el cual se escogen aquellas unidades de reducción léxica que se ajustan a la definición y tipología de siglas establecidas en este trabajo.

mayúsculas. Se ha procedido a eliminar manualmente del corpus todas aquellas unidades que no fueran estrictamente siglas. El criterio de eliminación ha sido la definición y la tipología de siglas establecidas en esta investigación. Algunos ejemplos del ruido eliminado en el corpus corresponden a elementos como:

1) Apellidos: McLeod, DeLisi, McCArty, DiGeorge;
2) Fórmulas químicas: NaCl;
3) Secuencias de bases químicas: ATGC, AGC, GAT;
4) Artículos (sólo aparecidos en títulos): EL, LA, LOS, LAS;
5) Pronombres (sólo aparecidos en títulos): SUS;
6) Palabras en mayúscula sostenida: CIENCIA, PROYECTO, GENOMA;
7) Otras URL, especialmente abreviaturas (MW, HP, Kmh, etc.), que no constituyen ruido propiamente dicho, pero que no son el foco de nuestro estudio.

Una vez depurado el corpus, surgió la dificultad de que la mayoría de las siglas carecía de su forma desarrollada.[33] Por consiguiente, fue necesario buscar las formas desarrolladas en otras fuentes para, de este modo, comprobar que cada candidato fuera efectivamente una sigla. El procedimiento para buscar las formas desarrolladas faltantes ha sido el siguiente:

En primer lugar, se ha procedido a buscar la sigla y su forma desarrollada en el resto del CT-IULA mediante el módulo de "búsqueda estándar" de BwanaNet. En aquellos casos en los que no se ha logrado confirmar por este medio, se ha continuado la búsqueda en otras fuentes, así:

1) Búsqueda en fuentes específicas
    (a) En todo el corpus del campo de especialidad tratado (GH o MA)
    (b) En bases de datos de siglas, diccionarios y glosarios electrónicos en Internet

33 Es de notar que, dependiendo del grado de fijación de la sigla en el discurso y del estilo editorial, ésta tenderá a aparecer o no junto a su forma desarrollada.

(i) Sobre genoma humano
1. AcroMed
2. Medical Dictionary on-line
3. Diccionari Enciclopèdic de Medicina
4. Human Genome Acronym List
5. Glosarios de Biotecnología
6. Glosarios de Genética
7. Genetics home reference
8. Merck Source

(ii) Sobre medio ambiente
1. Compendium of Environmental and professional Acronyms
2. U.S. Environmental Protection Agency
3. U.S. Global Change Research Information Office
4. Diccionario de la contaminación
5. EEA Multilingual Environmental Glossary

(c) En páginas web
(i) List of Acronyms on the Literature on Genome Research
(ii) Abreviaturas de genes
(iii) Nombres de proteínas y genes
(iv) Índice de acrónimos y siglas comunes en Bioquímica y Biología molecular

2) Búsqueda en fuentes generales
(a) En buscadores de siglas
(i) Acronym finder
(ii) Acrophile
(iii) Acronym server
(iv) Abbreviations.com
(v) Acronym search
(b) En buscadores generales
(i) Google
(ii) Scirus
(iii) Vivísimo

2) Confección del corpus definitivo. Después de la depuración manual y de completar las formas desarrolladas de las siglas, se ha procedido a la creación de una BD para el almacenamiento del corpus definitivo (siglas-formas desarrolladas). Dicho corpus quedó constituido por 800 siglas de GH y 317 de MA, para un total de 1.117 unidades.

3. Análisis de los datos

En primera instancia se han analizado las siglas del corpus de GH. Posteriormente, los resultados derivados de dicho análisis se han cotejado con las siglas del corpus de MA, para buscar las similitudes o disimilitudes y así determinar si se pueden hacer las mismas generalizaciones con miras a la descripción lingüística y al establecimiento de los criterios para la detección semiautomática (reglas de formación, patrones de identificación de pares de sigla-forma desarrollada, etc.).

Para cumplir con el objetivo 1c, se ha efectuado una descripción lingüística de las siglas en el corpus unificado de GH y MA desde los siguientes puntos de vista:

1) Fonético
2) Morfológico
3) Sintáctico
4) Semántico

Para alcanzar el objetivo 1d, se ha efectuado un análisis cuantitativo general de la incidencia de las siglas en el corpus de GH y de MA, que ha comprendido los siguientes aspectos:

1) Análisis estadístico descriptivo
   a) Porcentaje de siglas que presentan su forma desarrollada en el corpus de GH y MA;
   b) Porcentaje de siglas que presentan variantes formales en el corpus de GH y MA;
   c) Tipo de siglas más frecuente en el corpus de GH y MA;
   d) Porcentaje de siglas que representan términos (siglas especializadas) en el corpus de GH y MA;
   e) Porcentaje de siglas creadas en inglés en el corpus de GH y MA;
   f) Porcentaje de siglas creadas en español en el corpus de GH y MA.

2) Análisis estadístico inferencial

Con este análisis se ha buscado conocer cuáles son los rasgos que establecen las diferencias más significativas en los dos ámbitos de especialidad.

Para cumplir con el objetivo 1e se ha llevado a cabo el siguiente análisis:

Contraste de los resultados en los dos ámbitos temáticos: GH y MA. Adicionalmente, se ha conformado otro corpus de textos y siglas de los ámbitos de economía e informática con el ánimo de cotejarlos con los de GH y MA y ver si se comportaban de la misma manera. Las especificaciones de estos dos corpus nuevos se indican al final de este capítulo.

Para alcanzar los objetivos 2a, 2b y 2c se han efectuado las siguientes tareas:

1) Revisión del estado de la cuestión sobre los sistemas de detección de siglas;
2) Evaluación desde el punto de vista del rendimiento de los principales sistemas de detección hallados;
3) Determinación de los patrones de aparición de las siglas en el corpus unificado de siglas de GH y MA;
4) Definición de las estrategias de detección de siglas en textos en español.

4. Presentación de los datos

Se ha diseñado una base de datos para el almacenamiento de las unidades que conforman el corpus de siglas de GH y MA. Los parámetros de descripción de cada campo son los siguientes:

1) Nombre del campo. Indica el nombre con que se representa el campo en la base de datos;

2) Estatus. Indica si el campo es de carácter obligatorio u opcional;
3) Contenido. Presenta una breve descripción del contenido del campo;
4) Abierto/cerrado. Indica si las posibilidades de representación de las informaciones del campo son limitadas o ilimitadas;
5) Formato. Indica el tipo de formato con que se representa el campo; *i.e.*, índice, texto, atributo.

La estructura de la base y los criterios para la entrada sistemática de datos en cada registro se detallan a continuación.

4.1 Campos del encabezado del registro

1) Nombre del campo: Nombre de la base de datos
   Estatus: Obligatorio
   Contenido: Nombre completo de la base de datos
   Abierto/cerrado: Abierto, repetitivo
   Formato: Textual

2) Nombre del campo: Autor
   Estatus: Obligatorio
   Contenido: Nombre abreviado del autor de la base de datos
   Abierto/cerrado: Abierto, repetitivo
   Formato: Textual

3) Nombre del campo: Área temática
   Estatus: Obligatorio
   Contenido: Área temática general a la que pertenece la sigla que se describe en el registro
   Abierto/cerrado: Abierto, no repetitivo
   Formato: Textual

4.2 Campos dependientes del objeto de estudio

1) Nombre del campo: Sigla
   Estatus: Obligatorio
   Contenido: Unidad conformada con caracteres alfanuméricos que pueden estar en mayúsculas y/o minúsculas
   Abierto/cerrado: Abierto, no repetitivo

Formato: Textual

2) Nombre del campo: Lengua de procedencia
Estatus: Obligatorio
Contenido: Nombre abreviado de la lengua de procedencia de la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

3) Nombre del campo: Regla de formación de la sigla
Estatus: Obligatorio
Contenido: Código alfanumérico que representa el tipo y la cantidad de caracteres con que se ha formado la sigla
Abierto/cerrado: Abierto
Formato: Textual

4) Nombre del campo: Forma desarrollada
Estatus: Obligatorio
Contenido: Representación de la forma desarrollada o expansión de la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

5) Nombre del campo: Contexto 1
Estatus: Obligatorio
Contenido: Representación del contexto lingüístico de uso de la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

6) Nombre del campo: Fuente del contexto 1
Estatus: Obligatorio
Contenido: Identificación del texto de donde se extrajo la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

7) Nombre del campo: Contexto 2
Estatus: Opcional
Contenido: Representación del contexto lingüístico de uso de la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

8) Nombre del campo: Fuente del contexto 2
Estatus: Opcional
Contenido: Identificación del texto de donde se extrajo la sigla
Abierto/cerrado: Abierto, no repetitivo
Formato: Textual

9) Nombre del campo: Contexto 3
   Estatus: Opcional
   Contenido: Representación del contexto lingüístico de uso de la sigla
   Abierto/cerrado: Abierto, no repetitivo
   Formato: Textual

10) Nombre del campo: Fuente del contexto 3
   Estatus: Obligatorio
   Contenido: Identificación del texto de donde se extrajo la sigla
   Abierto/cerrado: Abierto, no repetitivo
   Formato: Textual

11) Nombre del campo: Patrones de identificación de sigla-forma desarrollada
   Estatus: Obligatorio
   Contenido: Formalización del patrón de aparición de un par sigla-forma desarrollada dentro de un contexto
   Abierto/cerrado: Abierto, no repetitivo
   Formato: Textual

12) Nombre del campo: Correspondencia de los elementos del par sigla-forma desarrollada
   Estatus: Obligatorio
   Contenido: Grado de correspondencia entre cada carácter de la sigla y su equivalente en la forma desarrollada
   Abierto/cerrado: Cerrado
   Formato: Textual

13) Nombre del campo: Nº de ocurrencias
   Estatus: Obligatorio
   Contenido: Frecuencia de aparición de la sigla dentro del corpus
   Abierto/cerrado: Abierto, no repetitivo
   Formato: Textual

14) Nombre del campo: Tipo de sigla
   Estatus: Obligatorio
   Contenido: Clasificación de la sigla según el tipo al que pertenece
   Abierto/cerrado: Cerrado
   Formato: Textual

15) Nombre del campo: Pronunciación
   Estatus: Obligatorio
   Contenido: Tipo de pronunciación de la sigla (alfabética, silábica)
   Abierto/cerrado: Cerrado
   Formato: Textual

16) Nombre del campo: Grafía

Estatus: Obligatorio
Contenido: Tipo de grafía empleado en la formación de la sigla
Abierto/cerrado: Cerrado
Formato: Textual

17) Nombre del campo: Morfología
Estatus: Obligatorio
Contenido: Aspectos morfológicos presentes en la sigla (derivación, composición, categoría gramatical, formas flexivas)
Abierto/cerrado: Cerrado
Formato: Textual

18) Nombre del campo: Sintaxis
Estatus: Obligatorio
Contenido: Aspectos sintácticos presentes en la sigla (tipo de sintagma, coincidencia de los elementos del par sigla-forma desarrollada, papel de la sigla como sujeto u objeto del verbo, combinatoria de las siglas)
Abierto/cerrado: Cerrado
Formato: Textual

19) Nombre del campo: Semántica
Estatus: Obligatorio
Contenido: Aspectos semánticos presentes en la sigla (sinonimia, homonimia, polisemia)
Abierto/cerrado: Cerrado
Formato: Textual

4.3 Campos de gestión del registro

1) Nombre del campo: Número de la entrada
Estatus: Obligatorio
Contenido: Número consecutivo del registro de la base de datos

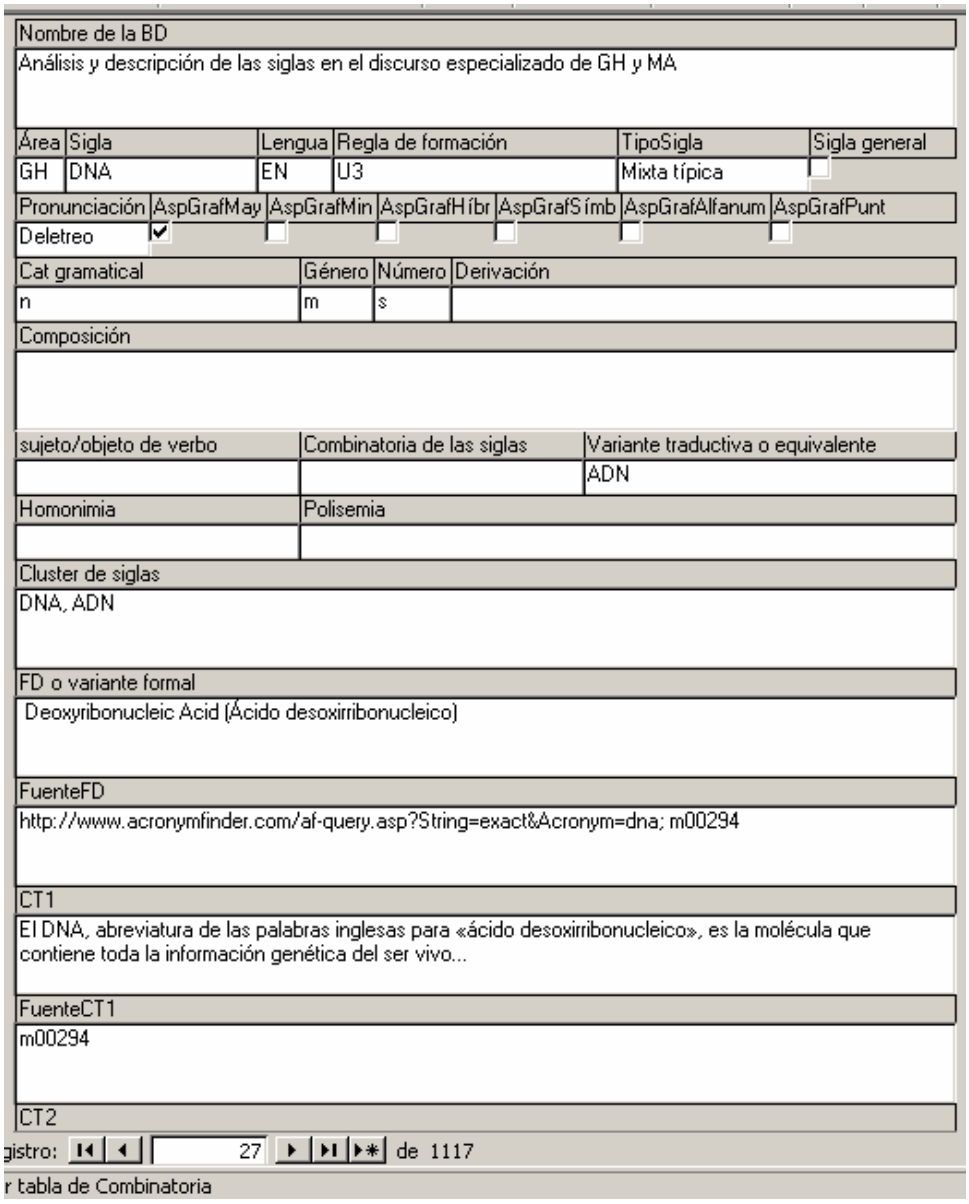
2) Nombre del campo: Fecha de creación
Estatus: Obligatorio
Contenido: Fecha de creación del registro

3) Nombre del campo: Creado por
Estatus: Obligatorio
Contenido: Fecha de la última modificación del registro

4) Nombre del campo: Fecha de modificación
Estatus: Obligatorio
Contenido: Fecha de modificación del registro

5) Nombre del campo: Modificado por
   Estatus: Obligatorio
   Contenido: Nombre del autor de la última modificación.

A continuación se indica el modelo de ficha empleado para la recopilación, análisis y presentación de algunos de los datos obtenidos.



5. Características del corpus de contraste

Como se ha señalado anteriormente, se ha conformado otro corpus de textos y siglas de los ámbitos de economía e informática con el fin de cotejar el comportamiento de estas siglas respecto de las siglas de GH y MA.

5.1 Corpus de economía (ECON)

El corpus de economía está formado por 45 textos pertenecientes a las subáreas de: economía financiera, derecho y economía, microeconomía, métodos matemáticos y cuantitativos, metodología e historia del pensamiento económico, sistemas económicos, economía general y enseñanza, historia económica, economía de la agricultura y de los recursos naturales, economía pública, salud, educación y bienestar, economía internacional, organización industrial, desarrollo económico y cambio tecnológico, administración y economía de empresa, marketing y contabilidad y economía urbana, rural y regional.

Este corpus posee 1.027.995 palabras de las que se han extraído, mediante la misma técnica empleada en los corpus anteriores, 244 siglas diferentes.

5.2 Corpus de informática (INF)

El corpus de informática consta de 50 textos pertenecientes a las subáreas de: hardware, comunicación hombre-máquina, aplicaciones, organización de los ordenadores, software, teoría de la computación y metodología de la computación. Dicho corpus contiene 1.029.172 palabras a partir de las cuales se extrajeron 876 siglas diferentes.

En total el corpus unificado de economía e informática suma 1.120 unidades.

Por último, resumimos en el siguiente esquema todo el diseño metodológico establecido para cumplir con los objetivos trazados.

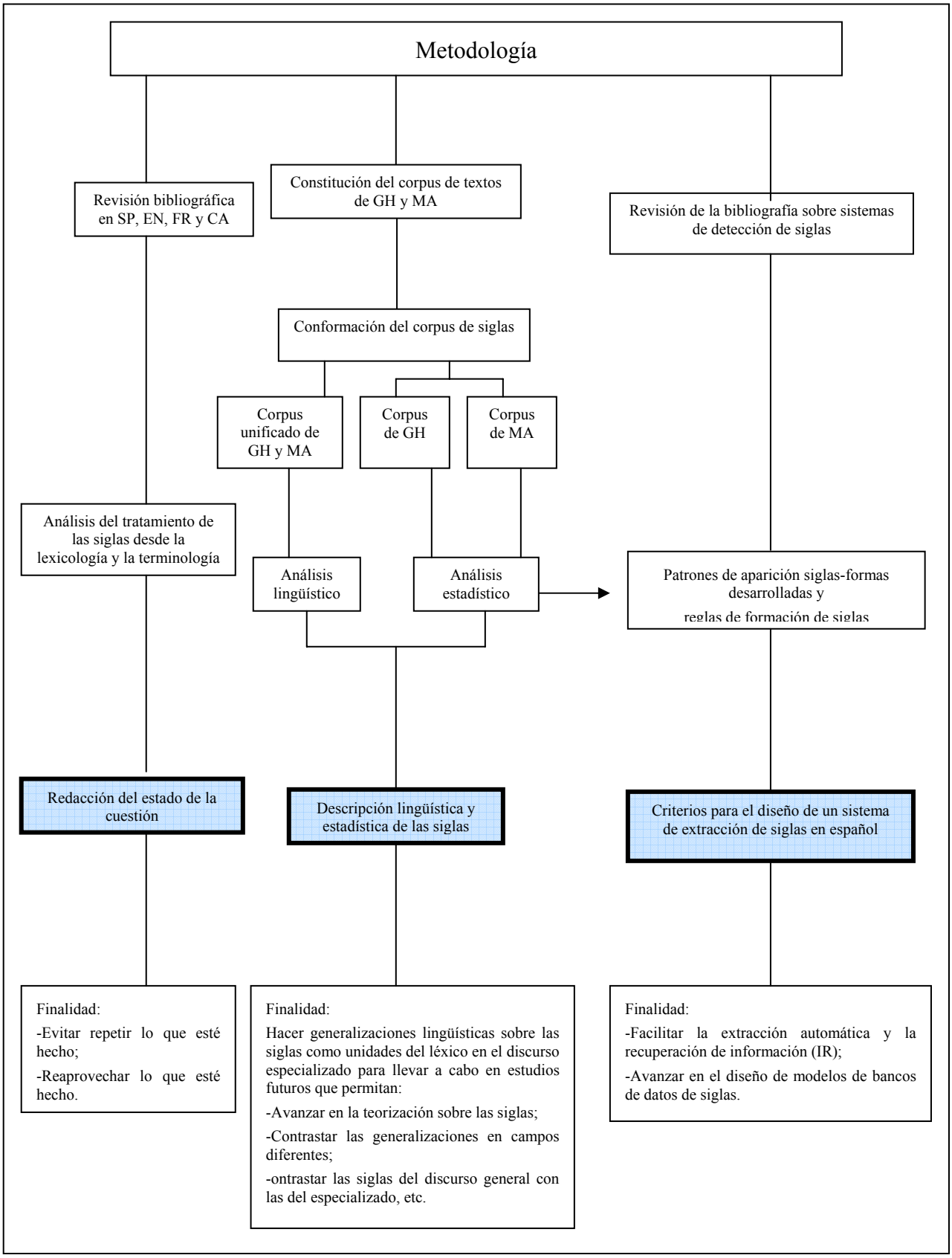


Gráfico 3. Metodología

Capítulo 6

Capítulo 6

Análisis estadístico de las siglas en los ámbitos de genoma humano y medio ambiente

Introducción

En primer lugar, en este capítulo nos proponemos analizar la incidencia de las siglas en el discurso de GH y MA desde el punto de vista estadístico. Y, en segundo lugar, probar que la frecuencia de aparición de las siglas varía de un ámbito de especialidad a otro y determinar las consecuencias que este factor tiene en el discurso.

La "siglometría" (Baudet, 2001: 34) es el término que se emplea para denominar la parte lingüística que estudia las siglas desde el punto de vista de los datos cuantitativos. Por ejemplo, la siglometría es útil para cuantificar el número de caracteres que conforman una sigla, *e.g.*: PCR (*Polymersase chain reaction*) consta de tres caracteres (letras) o BRCA1 (*Breast cancer susceptibility gene 1*), el cual contiene cinco caracteres alfanuméricos. El análisis siglométrico de un corpus sirve, entre otras cosas, para la descripción de las siglas con miras al establecimiento de patrones para su detección semiautomática.

En el presente capítulo se presentan los resultados del análisis siglométrico de nuestro corpus. En primer lugar, se han tomado todos los documentos (individuos)

que conforman los corpus (las poblaciones) de GH y MA. De esta forma, las poblaciones de GH y MA están formadas por 158 y 47 individuos, respectivamente. En segundo lugar, se han aplicado los dos tipos de análisis estadísticos, a saber: descriptivo e inferencial. A partir de los resultados obtenidos se presentan las conclusiones generales y específicas para cada ámbito de especialidad.

1. Análisis estadístico descriptivo

1.1 Análisis estadístico descriptivo de las siglas en el ámbito de GH

El corpus de GH está compuesto por 999.950 palabras, de las cuales 11.026 (1,10%) son siglas. Para este análisis hemos seleccionado el conjunto de variables que consideramos que ofrecen un panorama general de las siglas dentro de este corpus en concreto. Las variables y sus valores correspondientes se detallan en la tabla que aparece a continuación.

	Variable	**Valor**
1	Porcentaje de siglas que originalmente presentan su forma desarrollada en el corpus (variante formal)	36%
2	Porcentaje de siglas que no aparecen originalmente con su forma desarrollada en el corpus	64%
3	Tipo de siglas más frecuente	55% (mixtas)
4	Porcentaje de siglas del discurso especializado	91%
5	Porcentaje de siglas creadas en inglés	71%
6	Porcentaje de siglas creadas en español	29%

Tabla 9. Análisis estadístico descriptivo de las siglas en GH

De los datos de la tabla anterior se infieren varios factores determinantes para la caracterización de estas unidades dentro del ámbito de GH. En primer lugar, se observa un alto porcentaje de siglas que aparecen sin su forma desarrollada (64%). Este hecho no implica dificultad para la detección de candidatos a sigla, pero sí para los candidatos par sigla-forma desarrollada. En segundo lugar, se observa que las siglas mixtas predominan (55%). En otras palabras, en GH las siglas se forman

mayoritariamente sin tener en cuenta el modelo canónico de tomar el carácter inicial de cada uno de los componentes de la forma desarrollada. En tercer lugar, el análisis ha permitido mostrar que el 91% de las siglas de este corpus representa a un término. Finalmente, se ha encontrado que la mayoría de las siglas (71%) se han prestado del inglés, mientras que el 29% restante se ha traducido o creado directamente en español. En síntesis, podemos decir que la mayoría de las siglas de GH se caracterizan por ser especializadas, mixtas y creadas en inglés.

1.2 Análisis estadístico descriptivo de las siglas en el ámbito de MA

El corpus de MA está compuesto por 999.876 palabras, de las cuales 1.583 (0,15%) corresponden a siglas. Con el propósito de establecer un contraste con el ámbito de GH, hemos tenido en cuenta las mismas variables, las cuales se recogen en la siguiente tabla.

	Variable	**Valor**
1	Porcentaje de siglas que originalmente presentan su forma desarrollada en el corpus (variante formal)	47%
2	Porcentaje de siglas que no aparecen originalmente con su forma desarrollada en el corpus	53%
3	Tipo de siglas más frecuente	70% (propias)
4	Porcentaje de siglas del discurso especializado	42%
5	Porcentaje de siglas creadas en inglés	26%
6	Porcentaje de siglas creadas en español	69%

Tabla 10. Análisis estadístico descriptivo de las siglas en MA

La observación de las variables establecidas muestra que, para el caso del corpus de MA, un alto porcentaje de siglas aparece sin su forma desarrollada (53%), lo que supone, como indicábamos anteriormente, una dificultad para los sistemas de detección a la hora de establecer los pares de sigla-forma desarrollada. Así mismo, este análisis ha permitido establecer que las siglas propias son el tipo predominante, llegando a representar el 70% del total de unidades. Las siglas de este ámbito se caracterizan además por representar mayoritariamente nombres de organismos nacionales e internacionales; sólo el 42% representa a algún término. Por último, cabe destacar que la mayoría de las siglas (69%) provienen del español. En resumen,

puede decirse que la mayoría de las siglas del ámbito de MA se caracterizan por ser propias, generales y creadas en español.

El contraste de los análisis descriptivos correspondientes a los dos ámbitos estudiados muestra diferencias notorias. La primera de ellas es justamente el número de siglas en uno y otro dominio; mientras en GH el número de siglas alcanza 11.026 unidades, en MA la cifra tan sólo llega a 1.583 unidades. Por otra parte, en cada ámbito predomina un tipo de sigla diferente. En GH son más numerosas las siglas mixtas en tanto que en MA lo son las siglas propias. Otro factor distintivo tiene que ver con el tipo de unidades que representan las siglas. Las siglas de GH representan mayoritariamente a términos de la especialidad a diferencia de las de MA que representan a nombres de instituciones y entidades, que se enmarcan dentro del discurso general. Finalmente, llama la atención la lengua de procedencia de las siglas. En GH la mayoría han sido creadas en inglés, pero en MA sobresalen las siglas creadas en español. Este último hecho, también puede tener cierta repercusión en los sistemas de detección de siglas, dado que debe tenerse en cuenta la manera en que aparecen las diferentes combinaciones de los pares de sigla-forma desarrollada. Es decir, sigla en inglés-forma desarrollada en español; sigla en inglés-forma desarrollada en inglés; o sigla en español-forma desarrollada en español.

2. Análisis estadístico inferencial de las siglas

Este análisis se compone de dos partes. La primera parte corresponde al análisis de la varianza (ANOVA) de las poblaciones de GH y MA mediante el uso del programa estadístico *Statgraphics*. Para la realización del ANOVA se han tomado los datos provenientes de la medición de un grupo de 23 variables, a saber: número de siglas, siglas diferentes, siglas con forma desarrollada, siglas sin forma desarrollada, siglas con 2, 3, 4, 5, 6 y más de seis caracteres, siglas propias, siglas mixtas, siglas generales (nombres de organizaciones), siglas especializadas, siglas creadas en inglés, siglas creadas en español, siglas creadas en otras lenguas, siglas en

mayúsculas, siglas en minúsculas, siglas híbridas (may+min), siglas con caracteres alfanuméricos, siglas con variación traductiva y marca de plural.

La segunda parte corresponde al análisis de componentes principales (ACP). Esta técnica permite minimizar el número de variables sin que por ello se pierda demasiada información. Con ella se pueden observar los puntos en los que las variables de las dos poblaciones se asemejan o se diferencian.

A continuación se detallan las variables que arrojan los datos más significativos con miras al establecimiento de los rasgos distintivos de cada ámbito de especialidad desde el punto de vista de las siglas.

1) Total de siglas

   Las siglas se crean básicamente como un mecanismo de economía lingüística, estilo editorial y mnemotécnico tendiente a evitar el uso de expresiones excesivamente largas. Tal y como se ha indicado antes en el análisis descriptivo, la diferencia total de siglas en GH y MA es significativa, es muy poco probable que sea fruto del azar.[34] El ANOVA muestra que el número de siglas es superior en GH. El P-valor es 0,0322.

34 En la lectura de los gráficos A representa los valores para medio ambiente y G para genoma humano.

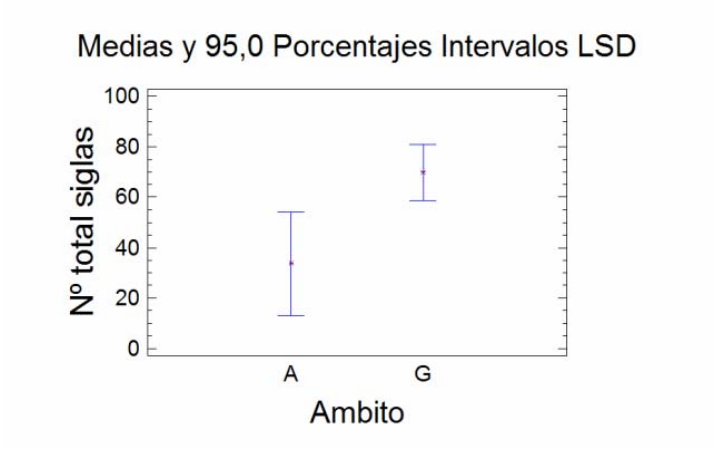


Gráfico 4. Total de siglas

2) Número de caracteres de las siglas

De forma empírica sabemos que las siglas que encontramos a diario presentan una extensión variada. Pero, si abordamos este tema en términos estadísticos ¿cuál sería la extensión promedio de las siglas? Un estudio de Calvet (1980: 21) concluye que la mayoría de las siglas se crean con tres y cuatro caracteres y da como explicación a este fenómeno las siguientes razones:

a) Insuficiencia. Las letras del abecedario se agotarían inmediatamente y se caería en la homonimia. Lo mismo sucedería con las siglas de dos caracteres. La combinación de 28 letras sería muy poca para la creación de nuevas siglas. A esto habría que sumar que las combinaciones de letras altamente usadas como A y C llevarían nuevamente a una gran cantidad de homonimia. Por el contrario, las siglas combinadas con tres letras del abecedario permitirían varios miles de combinaciones sin homónimos (15.600 para el francés). La cifra aumenta considerablemente si las siglas emplean cuatro caracteres (358.800 posibilidades para el francés). A todo lo anterior, se debe añadir que las siglas de uno y dos caracteres tendrían muy poca capacidad de aportar información, lo cual dificultaría en cierto sentido la comunicación.

b) Memorización. A diferencia de una sigla de tres o cuatro caracteres, una sigla de siete u ocho caracteres se considera difícil de retener.

El análisis ANOVA indica una diferencia significativa entre los dos campos de especialidad. Los valores más llamativos se encuentran en la presencia de siglas de 2, 6 y más de 6 caracteres. En los tres casos, MA destaca por tener el mayor número de este tipo de siglas. El P-valor para el número de siglas de 2 caracteres se sitúa en 0,012 mientras que para el número de siglas de 6 y más de 6 caracteres se sitúa en 0,0458 y 0,0162, respectivamente.

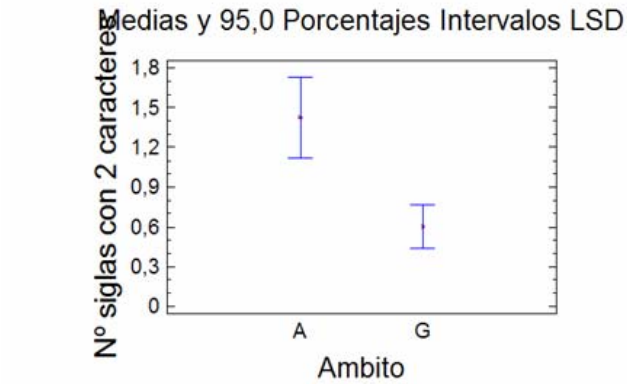


Gráfico 5. Número de siglas de 2 caracteres

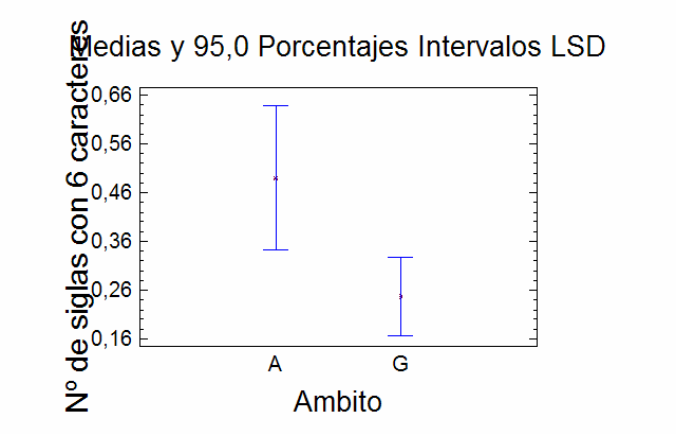


Gráfico 6. Número de siglas con 6 caracteres

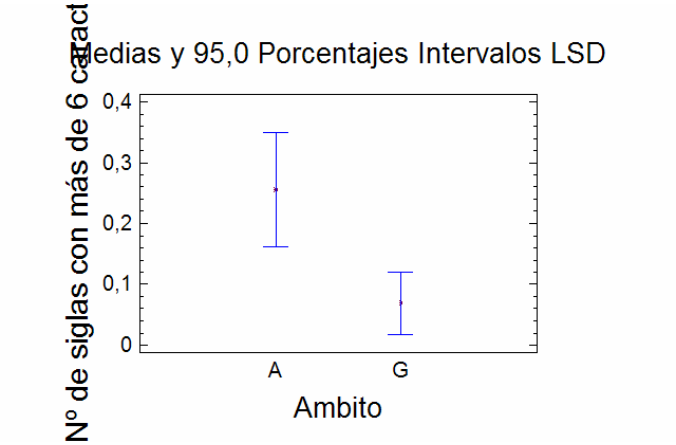


Gráfico 7. Número de siglas con más de 6 caracteres

3) Clases de siglas

Tal y como se ha visto antes en el análisis descriptivo, en cada ámbito de especialidad predomina una clase de siglas diferente. En MA se emplean con mayor frecuencia las siglas propias (*i.e.*, aquellas siglas formadas exclusivamente a partir de las iniciales de los elementos de sus formas desarrolladas). Por el contrario, en GH se emplean con mayor frecuencia las siglas mixtas (*i.e.*, aquellas siglas que bien han sido formadas por caracteres internos de los componentes de la forma desarrollada, cifras, o bien omiten

partes fundamentales de la forma desarrollada).[35] Los P-valores para las siglas propias y para las siglas mixtas son 0,0082 y 0,096, respectivamente.

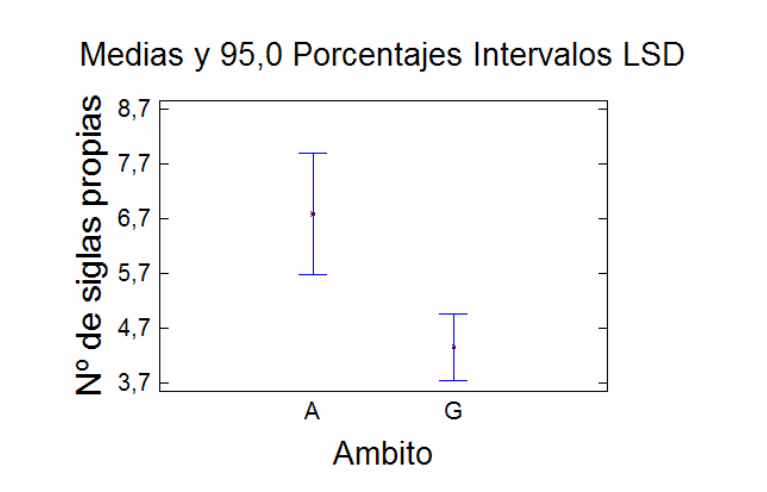


Gráfico 8. Número de siglas propias

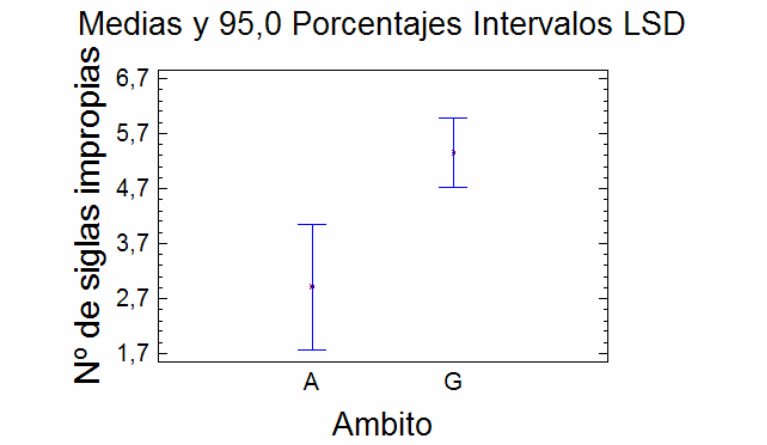


Gráfico 9. Número de siglas mixtas o impropias

35 Algunos autores también denominan a las siglas mixtas siglas impropias o sigloides (*cf*. Martínez de Sousa (1984: 34), Mestres (1996: 15) y Casado Velarde (1985: 20)).

4) Siglas generales y siglas especializadas

El análisis ANOVA muestra que las siglas generales (siglas que corresponden a nombres de instituciones y en general a todas aquellas unidades que no son términos) tienen mayor peso en MA. Sin embargo, en GH ocurre lo contrario, ya que son las siglas especializadas (aquellas que corresponden a términos) las que tienen la frecuencia más alta. El P-valor para las siglas generales es de 0,0000 y para las siglas especializadas 0,0012.

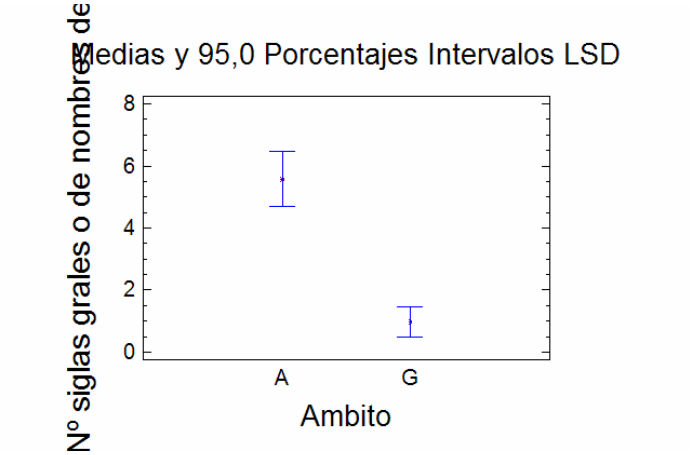


Gráfico 10. Número de siglas generales (nombres de instituciones)

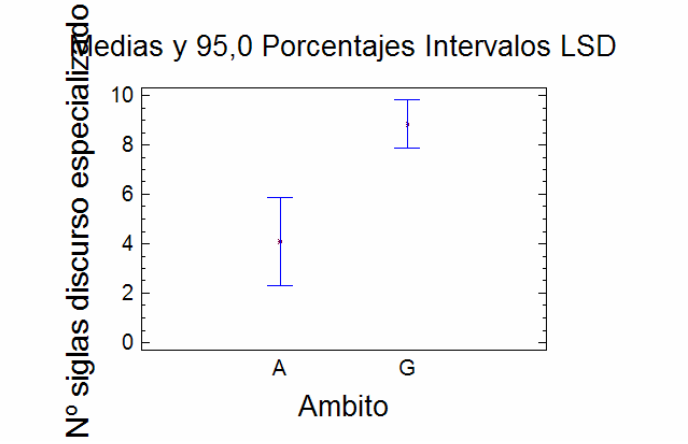


Gráfico 11. Número de siglas del discurso especializado

5) Siglas creadas en inglés, español y en otras lenguas

En lo que concierne a la lengua de procedencia de las siglas, también existe una diferencia marcada entre ambos campos de especialidad. Por una parte, GH tiene una alta concentración de siglas creadas en inglés. Por otra parte, MA presenta una ocurrencia mayor de siglas provenientes del español. Adicionalmente, MA presenta la mayor frecuencia de siglas provenientes de otras lenguas como alemán, francés, italiano, catalán, entre otras. Los P-valor correspondientes para el número de siglas en inglés, español y otras lenguas es 0,0013; 0,0000 y 0,0000, respectivamente.

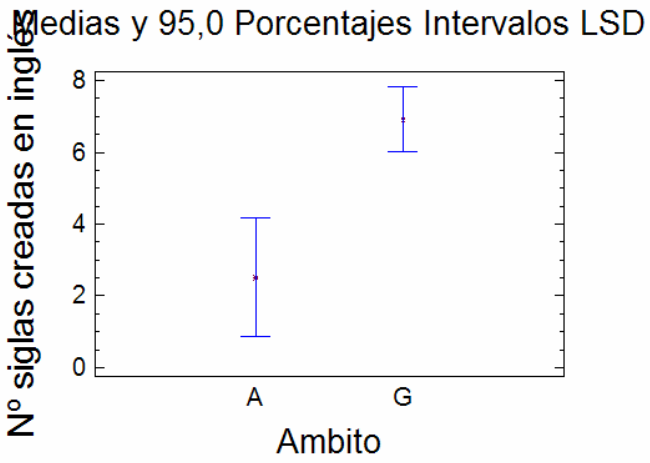


Gráfico 12. Número de siglas creadas en inglés

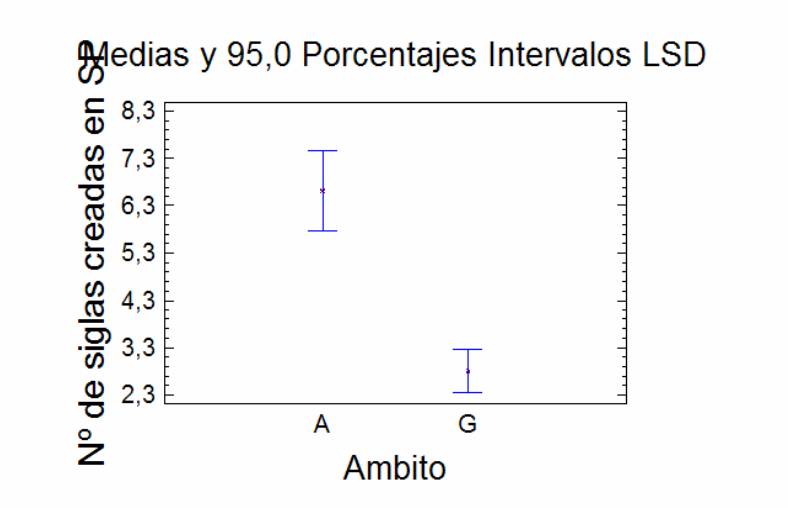


Gráfico 13. Número de siglas creadas en español

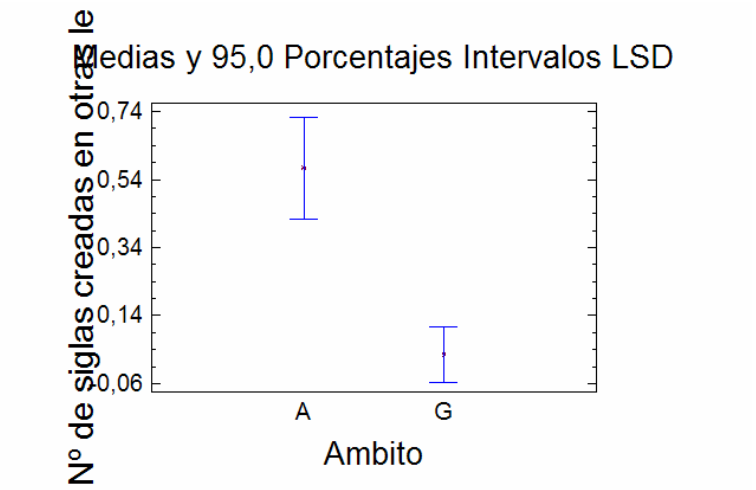


Gráfico 14. Número de siglas creadas en otras lenguas

6) Siglas en mayúsculas, minúsculas e híbridos (may+min)

En este caso, la variable con resultados más significativos ha sido la de siglas formadas con caracteres en mayúsculas y minúsculas (caracteres híbridos), por ejemplo: AcP, ADNmt, AINEs, ARNtAla, ATPásica, etc. Este tipo de siglas predomina en GH. El P-valor 0,0012.

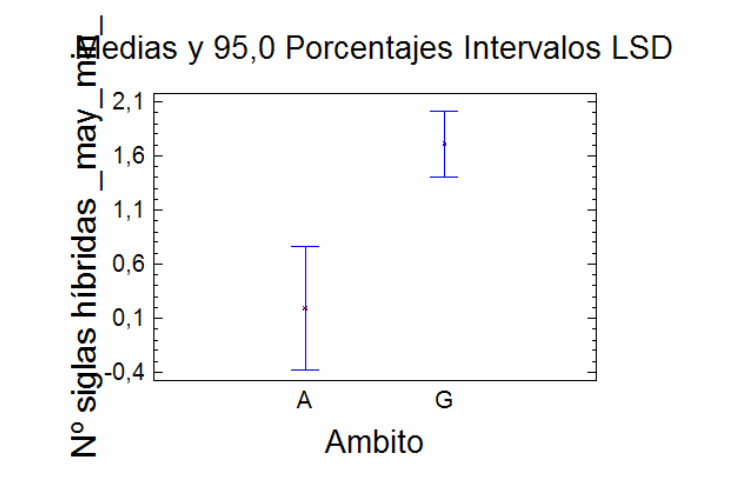


Gráfico 15. Número de siglas híbridas (caracteres en mayúsculas y minúsculas)

7) Siglas con caracteres alfanuméricos

El análisis de esta variable ha revelado que la principal diferencia entre las siglas de GH y MA radica en el empleo de siglas con caracteres alfanuméricos (*i.e.*, ADR2, ATP7B, CD28RC, Cdk5, FGFR1, etc.). En este sentido, al igual que en el caso de las siglas con caracteres híbridos mencionado anteriormente, GH es el ámbito que recurre con mayor frecuencia al uso de este tipo de unidades. El P-valor correspondiente se sitúa en 0,0012.

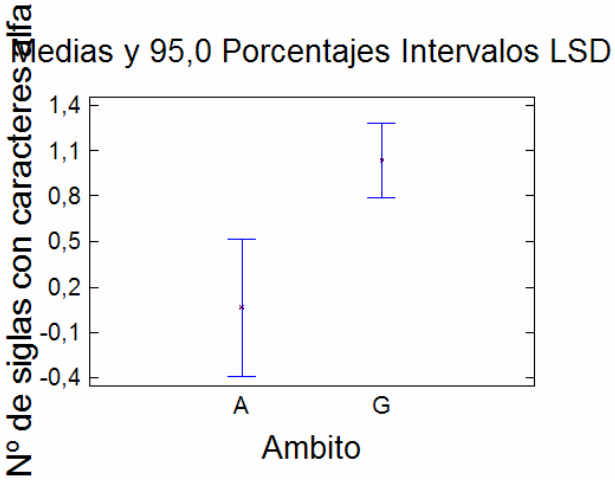


Gráfico 16. Número de siglas con caracteres alfanuméricos

8) Siglas con variación traductiva

Como se ha expuesto antes, entendemos por variación traductiva el fenómeno que se da cuando las siglas son prestadas de una lengua extranjera (L1) y posteriormente traducidas en la lengua de llegada (L2). Este fenómeno es bastante común en GH donde la mayoría de las siglas proceden del inglés, lengua a la que pertenece el 71% del total de siglas, tal y como ha revelado el análisis descriptivo presentado al comienzo de este capítulo. En algunos casos pueden coexistir las dos formas de la sigla, es decir, la forma inglesa su equivalente en español. Entre los casos más comunes se encuentran DNA/ADN, RNA/ARN y USA/EUA. El P-valor de esta variable se sitúa en 0,0094.

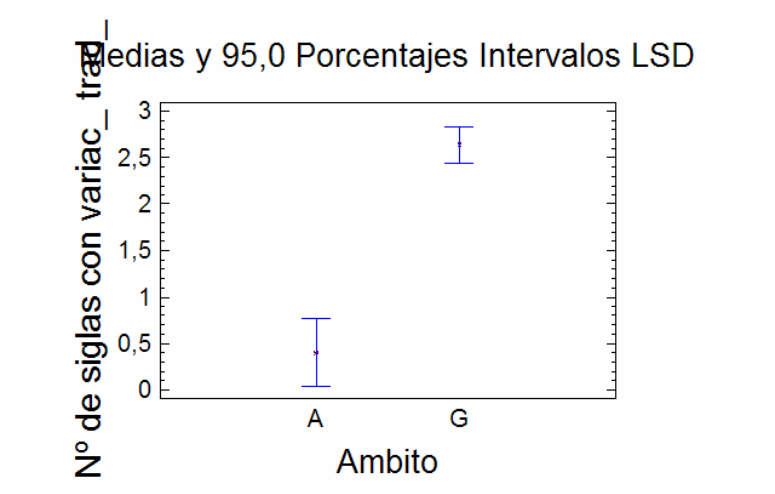


Gráfico 17. Siglas con variación traductiva

9) Siglas con marca de plural

Por último, la variable "marca de plural" ha mostrado la tendencia de las siglas de GH a adoptar la marca de plural anglosajona "s", en parte justamente, porque la mayoría de las siglas de este ámbito provienen del inglés. Así pues, GH es el ámbito de especialidad con mayor presencia de siglas pluralizadas. El P-valor correspondiente es 0,0841.

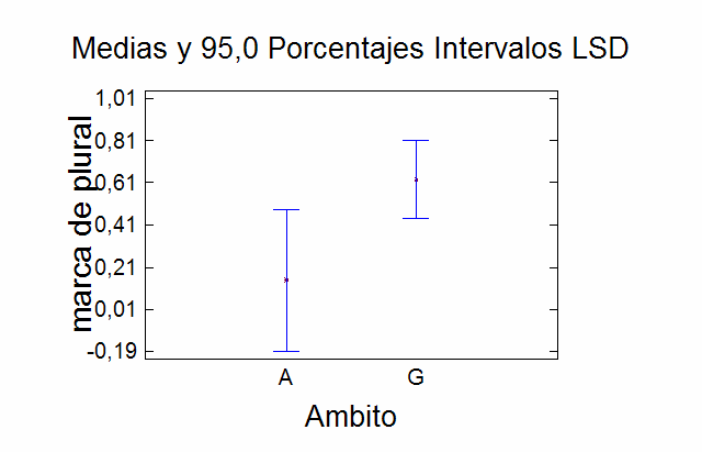


Gráfico 18. Marca de plural

3. Análisis comparativo de los resultados en los ámbitos de GH y MA

La siglometría, reflejada en los dos análisis efectuados anteriormente, ha permitido determinar que el discurso especializado de GH se caracteriza por tener mayormente siglas impropias, especializadas, inglesas, híbridas, alfanuméricas, traducidas y pluralizadas.

En lo que respecta al discurso especializado de MA, puede decirse que se caracteriza por presentar siglas cortas (2 caracteres) y extensas (6 ó más caracteres), propias, generales y españolas.

Como hemos explicado al principio, se ha efectuado un análisis de componentes principales (ACP). Se trata de una técnica estadística de síntesis de la información o reducción del número de variables. Es decir, ante una base de datos con muchas variables, el objetivo será reducirlas perdiendo la menor cantidad de información posible. La elección de los componentes se hace automáticamente por parte del programa, en este caso *Statgraphics*. Los nuevos componentes principales serán una

combinación lineal de las variables originales y, además, serán independientes entre sí.

En este caso, se tienen 23 variables, las cuales han sido reducidas a dos componentes principales, que abarcan la mayor cantidad de información de las 23 variables. Los dos componentes principales obtenidos representan aproximadamente el 65% de la información.

A continuación se presenta la tabla de pesos de los componentes a partir de la cual se ha seleccionado el ACP.

Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	11,7982	51,297	51,297
2	3,03121	13,179	64,476
3	1,50018	6,523	70,998
4	1,25154	5,441	76,440
5	0,959717	4,173	80,613
6	0,805517	3,502	84,115
7	0,693023	3,013	87,128
8	0,534021	2,322	89,450
9	0,494904	2,152	91,602
10	0,451176	1,962	93,563
11	0,35591	1,547	95,111
12	0,286383	1,245	96,356
13	0,231518	1,007	97,362
14	0,203197	0,883	98,246
15	0,173563	0,755	99,000
16	0,113235	0,492	99,493
17	0,0918703	0,399	99,892
18	0,0184807	0,080	99,973
19	0,00294564	0,013	99,985
20	0,00174364	0,008	99,993
21	0,00114946	0,005	99,998
22	0,000372928	0,002	100,000
23	0,000106234	0,000	100,000

Tabla 11. Análisis de componentes principales

Las 23 variables reducidas mediante la técnica de ACP son las que se presentan en la siguiente tabla.

Tabla de Pesos de los Componentes

	Componentes 1	Componentes 2
N° de siglas con 4 caracteres	0,268391	-0,0346993
N° de siglas con 5 caracteres	0,217124	-0,0692105
N° de siglas con 6 caracteres	0,132764	0,23949
N° de siglas con caracteres alfa	0,182331	-0,216197
N° de siglas con FD	0,239047	0,129948
N° de siglas con más de 6 caract	0,127509	0,262571
N° de siglas con variac_ trad_	0,159938	-0,272587
N° de siglas creadas en otras le	0,0506675	0,313577
N° de siglas creadas en SP	0,161089	0,370423
N° de siglas en mayúscula	0,273095	0,0880737
N° de siglas en minúscula	0,0442749	0,013354
N° de siglas impropias	0,265558	-0,136077
N° de siglas propias	0,252992	0,16733
N° de siglas sin FD	0,257628	-0,0820783
N° palabras	0,121038	0,252536
N° siglas con 2 caracteres	0,137143	0,136748
N° siglas con 3 caracteres	0,258178	-0,0288955
N° siglas creadas en inglés	0,259963	-0,208773
N° siglas diferentes	0,289815	0,0144751
N° siglas discurso especializado	0,26202	-0,217631
N° siglas grales o de nombres de	0,128574	0,446056
N° siglas híbridas _may_min_	0,209533	-0,227382
N° total siglas	0,210323	-0,0889206

Tabla 12. Pesos de los componentes principales 1 y 2

La interpretación de los componentes muestra que el componente principal 1 (COMPRIN1) presenta una alta correlación positiva con todas las variables, hecho que permite asimilarlas al ámbito de GH.

El componente principal 2 (COMPRIN2) muestra una alta correlación negativa con las siguientes variables: siglas con 3, 4 y 5 caracteres, con caracteres alfanuméricos, con variación traductiva, sin forma desarrollada, especializadas y número total de siglas, hecho que permite asociarlas al ámbito de MA.

En cuanto al gráfico en dos dimensiones de COMPRIN1 y COMPRIN2, observamos que la mayoría de las variables se sitúan progresivamente hacia los valores extremos de ambos componentes.

El gráfico generado representa al ámbito de MA con color azul y a GH con rojo. Su interpretación revela que los valores negativos se han dado cuando el ACP ha recogido poco de todas las variables. Eso ha implicado que los valores tiendan a concentrarse y, por consiguiente, a que los dos ámbitos tiendan a parecerse. Por el contrario, los valores positivos indican que el ACP ha recogido mucho de todas las

variables. Eso ha supuesto la dispersión de los valores y, por tanto, la tendencia de los dos ámbitos a diferenciarse. El gráfico en dos dimensiones de COMPRIN1 y COMPRIN2 se presenta a continuación.

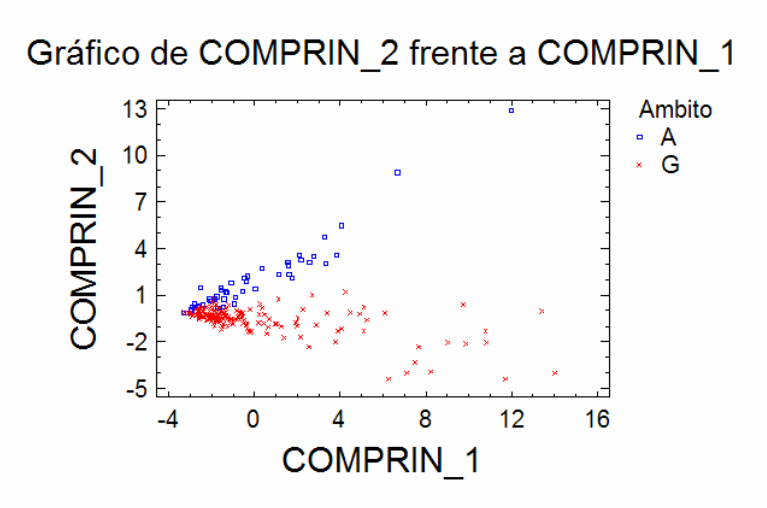


Gráfico 19. Comparación de los componentes principales 1 y 2

4. Genoma humano y medio ambiente *versus* informática y economía

En este apartado contrastamos los ámbitos de genoma humano (GH) y medio ambiente (MA) con los de informática (INF) y economía (ECON). El propósito de esta comparación es determinar si las siglas se comportan de manera similar en otros ámbitos de especialidad. Para dicha comparación realizaremos un análisis estadístico descriptivo a partir de las siguientes variables: número total de siglas, porcentaje de siglas respecto del total de palabras en el corpus, porcentaje de siglas que originalmente presentan su forma desarrollada en el corpus, cantidad de siglas especializadas y porcentaje de siglas creadas en inglés y español.

Para este análisis se han confeccionado los corpus de INF y ECON de tamaño similar al de los corpus de GH y MA. El corpus de informática está formado por 1.027.995 palabras de las cuales 6.130 (0,6%) corresponden a ocurrencias de siglas. Así mismo, el corpus de economía posee 1.029.172 palabras, de las cuales 1.708 (0,02%) son siglas.

Los valores correspondientes a las variables en cada ámbito son los siguientes:

	Variable	Valor			
		GH	MA	INF	ECON
1	Porcentaje de siglas respecto del total de palabras en el corpus	1,10%	0,15%	0,60%º	0,02%
2	Porcentaje de siglas que originalmente presentan su forma desarrollada en el corpus (variante formal)	36%	47%	65%	57%
3	Porcentaje de siglas del discurso especializado	91%	42%	93%	44%
4	Porcentaje de siglas creadas en inglés	71%	26%	84%	24%
5	Porcentaje de siglas creadas en español	29%	69%	15%	71%

Tabla 13. Análisis estadístico descriptivo de genoma humano, medio ambiente, informática y economía

La comparación de los cuatro corpus ha arrojado resultados interesantes. Por una parte, GH e INF son los ámbitos que presentan los mayores porcentajes de siglas en el corpus, siglas especializadas y siglas en inglés. Por otra parte, la variable "forma desarrollada" aparece con mayor frecuencia en los corpus de INF y ECON, mientras que la variable "siglas en español" prepondera en los corpus de MA y ECON.

Así mismo, puede decirse que los ámbitos de ECON y MA son los que presentan los menores porcentajes en las variables: siglas en el corpus, siglas especializadas y siglas inglesas. En cuanto a la variable "forma desarrollada", los valores más bajos se presentan en los campos de GH y MA. Finalmente, INF y GH son los ámbitos con el menor porcentaje de siglas en español.

En conclusión, los cuatro corpus pueden dividirse en dos grupos de acuerdo con la similitud de sus rasgos. Por consiguiente, se observa que los ámbitos de GH e INF por un lado, y de MA y ECON, por otro, tienden a parecerse entre sí. El siguiente

gráfico muestra los valores obtenidos para cada una de las variables consideradas anteriormente.

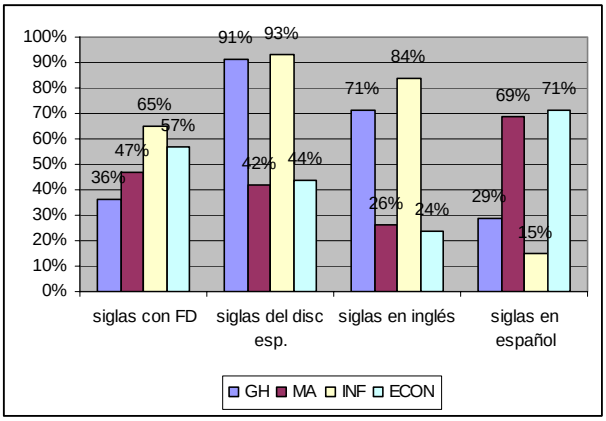


Gráfico 20. Porcentajes de cada una de las variables medidas en los cuatro ámbitos

5. Conclusiones

Al comienzo de este capítulo establecíamos como primer objetivo determinar la incidencia de las siglas en el discurso de GH y MA desde el punto de vista estadístico. Como segundo objetivo se buscaba probar que la frecuencia de aparición de las siglas varía de un ámbito de especialidad a otro y determinar las consecuencias que este factor tiene en el discurso. En tal sentido, puede decirse que las siglas tienen un peso relativamente bajo en el corpus de estudio; este tipo de unidades sólo representa el 1,1% y el 0,15% de palabras en el corpus GH y MA, respectivamente. Sin embargo, es evidente que su creación y uso van en rápido aumento. Dan testimonio de ello estudios diacrónicos como el de Bloom (2000: 4) y el crecimiento constante de las bases de datos de siglas en Internet.

A partir de todos los datos recogidos e interpretados anteriormente se ha concluido que las siglas de los campos de GH y MA tienen rasgos y tendencias que los diferencian entre sí. De manera general, los dos ámbitos se distinguen por: a) la cantidad de siglas, b) tipo de siglas predominante, y c) lengua de procedencia.

De manera particular, la siglometría ha permitido conocer que el discurso especializado de GH tiende básicamente al uso de siglas impropias, especializadas, inglesas, híbridas, alfanuméricas, traducidas y pluralizadas. En cambio, en lo que respecta al discurso especializado de MA, se ha visto que se tiende al uso de siglas cortas (2 caracteres) y extensas (6 o más caracteres), propias, generales y españolas.

Puede decirse que, comparado con MA, GH es un ámbito más rico y diverso desde el punto de vista síglico. Esta condición puede deberse a que se trata de un campo de especialidad de muy rápida evolución, y por esta vía mayor generador de denominaciones nuevas, muchas de las cuales terminan siendo abreviadas por diversos motivos como se ha señalado en otros capítulos.

Por último, la comparación de los resultados de la estadística descriptiva de GH y MA con los de INF y ECON ha permitido conocer los rasgos de las siglas interdominio. De esta forma, GH tiene más afinidad con INF mientras que MA se parece más a ECON, hecho que seguramente tiene relación con el carácter más tecnológico, aplicado y vanguardista de unas áreas respecto de otras.

Capítulo 7

Capítulo 7

Descripción lingüística de las siglas en los ámbitos de genoma humano y medio ambiente

Introducción

El presente capítulo tiene por objetivo el análisis de las características de las siglas desde el punto de vista lingüístico. Para ello hemos tomado de los corpus de GH y MA una muestra de las siglas más frecuentes hasta un máximo de diez por cada tipo de sigla. Los tipos que hemos tenido en cuenta son los que presentamos en el capítulo 2. No es nuestra pretensión con este capítulo realizar un análisis lingüístico profundo y detallado de nuestro objeto de estudio, dado que ya ha sido suficientemente tratado en trabajos previos. Nuestro propósito se orienta más bien hacia la utilización de cada uno de los elementos del análisis lingüístico como medio para la búsqueda de pistas que enriquezcan el conjunto de reglas de formación y patrones de detección de las siglas a tener en cuenta en un sistema de identificación semiautomático de siglas. Partiendo de esta precisión inicial, se ha procedido al estudio de los rasgos de las siglas desde el punto de vista fonético, morfológico, sintáctico y semántico. En concreto, consideramos que los aspectos morfológicos y sintácticos, al ser formales, pueden aportar pistas para el establecimiento de patrones de detección semiautomática de siglas.

En primer lugar, el análisis fonético muestra la manera como se pronuncia cada una de las unidades de análisis, esto es, alfabética (deletreo) o silábica. En segundo lugar,

el análisis morfológico trata los casos de derivación y composición así como la categoría de la palabra que forma el núcleo de la sigla y los aspectos flexivos. En tercer lugar, el análisis sintáctico trata el tipo de sintagma, el papel de la sigla como sujeto u objeto de verbo, la combinatoria de las siglas y la coincidencia entre los elementos de la sigla con los de la forma desarrollada. Por último, con el análisis semántico se buscan los casos de sinonimia y homonimia dentro de las siglas.

La muestra de siglas para este análisis se ha obtenido de la siguiente manera:

1) Unión de los corpus de GH y MA
2) Extracción de las siglas por tipo (propias, mixtas)
3) Selección de las siglas más frecuentes de cada tipo (hasta un máximo de 10)[36]

La siguiente tabla contiene las siglas que conforman la muestra para el presente análisis.

Ámbito	Tipo de sigla	Sigla	Lengua de procedencia	FD
GH	**Propias**	PCR	EN	Polymerase Chain Reaction
		PGH	ES	Proyecto Genoma humano
		RFLP	EN	Restriction Fragment Length Polimorphisms
		VIH	ES	Virus de la inmunodeficiencia humana
		HLA	EN	Human Leukocyte Antigen
		MHC	EN	Major Histocompatibility complex
		VNTR	EN	Variable Number of Tandem Repeats
		ES	EN	Embryonic Stem
		FQ	ES	Fibrosis quística
		LET	EN	Linear Energy Transfer
	Mixtas típicas	ADN	ES	Ácido desoxirribonucleico
		DNA	EN	Deoxyribonucleic Acid
		ARN	ES	Ácido ribonucleico
		RNA	EN	Ribonucleic Acid
		ARNm	ES	ARN mensajero
		Acs	ES	Aberraciones cromosómicas
		BRCA1	EN	Breast Cancer Susceptibility Gene 1
		CFTR	EN	Cystic Fibrosis Transmembrane Conductance Regulator
		ATP	EN	Adenosine Triphosphate
		BRCA2	EN	Breast Cancer Susceptibility Gene 2

36 Algunos tipos de siglas como los cruces son escasos en el corpus por lo que su frecuencia ha resultado ser inferior a 10.

Ámbito	Tipo de sigla	Sigla	Lengua de procedencia	FD
		PCR	EN	Polymerase Chain Reaction
		PGH	ES	Proyecto Genoma humano
	Mixtas acrónimos	ADA	EN	Adenosin Desaminase
		LINE/LINEs	EN	Long Interspersed Nuclear Element
		RANTES	EN	Regulated on Activation, normal T Cells expressed and secreted
		EDTA	EN	Ácido etilendiaminotetracético
		SINE/SINEs	EN	Small interspersed repetitive elements
		HUGO	EN	Human Genome Organization
		YACS	EN	Yeast artificial Chromosomes
		ITIM	EN	Immunoreceptor Tyrosine-based Inhibition Motifs
		UNESCO	EN	Organización de las Naciones Unidas para la Educación, la Cultura y la Ciencia
		JAK2	EN	Janus Kinase 2
	Mixtas cruces	MegaYACs	EN	Mega Yeast artificial Chromosomes
		CalTech	EN	California Institute of Technology
		PiGMap	EN	Pig Gene Mapping Project
		PHRAP	EN	Phragment Assembly Program
		Genpept	EN	GenBank Gene Products
		PubMed	EN	Public MEDLINE
	Siglónimos	Sida	ES	Síndrome de inmunodeficiencia adquirida
MA		CE	ES	Comisión Europea
		DBO	ES	Demanda bioquímica de oxígeno
		UE	ES	Unión Europea
		OCDE	ES	Organización de Cooperación y Desarrollo Económicos
		OD	ES	Oxígeno disuelto
		CP	ES	Código penal
		UCP	ES	Unidad de Carbón Piedra
		DBO5	ES	Demanda bioquímica de oxígeno a los 5 días
		NEI	ES	Nuevos Estados independientes
		ACP	ES	Análisis en Componentes Principales
		EDTA	EN	Ethylenediamine Tetraacetic Acid
	Mixtas típicas	PVC	EN	Polyvinyl Chloride
		DPD	EN	Diethyl-p-phenylene diamine
		PNPP	EN	Para-nitrophenylphosphate
		HCFC	EN	Hydrochlorofluorocarbon
		CPOM	EN	Coarse and fine particulate organic matter
		PCB	EN	Polychlorinated Biphenyls
		ADN	ES	Ácido desoxirribonucléico
		TLm	EN	Threshold Limit
		FCSE	ES	Ley de fuerzas y cuerpos de seguridad
		MEDOC	ES	Mediterráneo occidental
	Mixtas acrónimos	MEDOR	ES	Mediterráneo oriental
		ICONA	ES	Instituto para la Conservación de la Naturaleza
		MARPOL	EN	Convention for the prevention of pollution from ships
		MINER	ES	Ministerio de Industria y Energía
		AEDENAT	ES	Asociación Ecologista de Defensa de la Naturaleza
		PNUMA	ES	Programa de Naciones Unidas para el Medio Ambiente
		NASA	EN	National Aeronautics and Space Administration

Ámbito	Tipo de sigla	Sigla	Lengua de procedencia	FD
		PCR	EN	Polymerase Chain Reaction
		PGH	ES	Proyecto Genoma humano
		TRAGSA	ES	Transformación agraria s.a.
		FAPAS	ES	Fondo para la Protección de los Animales Salvajes

Tabla 14. Muestra de siglas seleccionada para el análisis lingüístico

1. Aspectos fonéticos

En las siglas la pronunciación puede darse de dos formas: alfabética y silábica. La pronunciación alfabética (o deletreo) se da cuando se leen uno a uno los grafemas que componen la sigla. Por el contrario, la pronunciación silábica se da cuando la sigla se lee como una palabra.

Estudios anteriores demuestran que lenguas como el español tienden a la creación de siglas de pronunciación silábica mientras que el inglés tiende a la creación de siglas de pronunciación alfabética (Fijo, 2003: 321). Así mismo, estudios como los realizados por Calvet (1980: 37), sostienen que una sigla de tres caracteres puede estar formada únicamente por consonantes sin mayores problemas para su pronunciación, ya que se pueden deletrear; pero, si se trata de una sigla de ocho caracteres consonánticos, la cuestión sería más difícil. Por consiguiente, cuanto más extensa sea una sigla, mayores posibilidades habrá de que se pronuncie silábicamente; es decir, como una palabra. De esta forma, PCR se pronunciaría por deletreo mientras que UNESCO tendría que pronunciarse como una palabra. Nuestro estudio no pretende comparar la pronunciación de las siglas por lenguas, puesto que nuestro objetivo es el contraste de dos ámbitos de especialidad. Por tanto, centramos esta parte del análisis en la búsqueda del tipo de pronunciación que destaca, de acuerdo con el número de caracteres de la sigla.

La tabla que aparece a continuación muestra las siglas bajo análisis, la lengua de procedencia, la formalización de sus componentes; *i.e.*, consonantes (C) y vocales (V) y el modo de pronunciación.

Tabla 15. Clasificación de las siglas por su modo de pronunciación

	Ámbito	Sigla	Lengua de Procedencia	Formalización de la sigla Consonante (C) Vocal (V)	Pronunciación
1	GH	PCR	EN	CCC	Alfabética
2	GH	PGH	ES	CCC	Alfabética
3	GH	RFLP	EN	CCCC	Alfabética
4	GH	VIH	ES	CVC	Alfabética
5	GH	HLA	EN	CCV	Alfabética
6	GH	MHC	EN	CCC	Alfabética
7	GH	VNTR	EN	CCCC	Alfabética
8	GH	ES	EN	VC	Alfabética
9	GH	FQ	ES	CC	Alfabética
10	GH	LET	EN	CVC	Silábica
11	GH	ADN	ES	VCC	Alfabética
12	GH	DNA	EN	CCV	Alfabética
13	GH	ARN	ES	VCC	Alfabética
14	GH	RNA	EN	CCV	Alfabética
15	GH	ARNm	ES	VCCC	Alfabética
16	GH	ACs	ES	VCC	Alfabética
17	GH	BRCA1	EN	CCCV	Alfabética
18	GH	CFTR	EN	CCCC	Alfabética
19	GH	ATP	EN	VCC	Alfabética
20	GH	BRCA2	EN	CCCV	Alfabética
21	GH	ADA	EN	VCV	Alfabética
22	GH	LINE	EN	CVCV	Silábica
23	GH	RANTES	EN	CVCCVC	Silábica
24	GH	EDTA	EN	VCCV	Silábica
25	GH	SINE	EN	CVCV	Silábica
26	GH	HUGO	EN	CVCV	Silábica
27	GH	YACS	EN	CVCC	Silábica
28	GH	ITIM	EN	VCVC	Silábica
29	GH	UNESCO	EN	VCVCCV	Silábica

30	GH	JAK2	EN	CVC	Silábica
31	GH	MegaYACs	EN	CVCVCVCC	Silábica
32	GH	CalTech	EN	CVCCVCC	Silábica
33	GH	PiGMap	EN	CVCCVC	Silábica
34	GH	PHRAP	EN	CCCVC	Silábica
35	GH	Genpept	EN	CVCCVCC	Silábica
36	GH	PubMed	EN	CVCCVC	Silábica
37	GH	Sida	ES	CVCV	Silábica
38	MA	CE	ES	CV	Silábica (como abrev.)
39	MA	DBO	ES	CCV	Alfabética
40	MA	UE	ES	VV	Silábica (como abrev.)
41	MA	OCDE	ES	VCCV	Silábica
42	MA	OD	ES	VC	Alfabética
43	MA	CP	ES	CC	Silábica (como abrev.)
44	MA	UCP	ES	VCC	Alfabética
45	MA	DBO5	ES	CCV	Alfabética
46	MA	NEI	ES	CVV	Silábica
47	MA	ACP	ES	VCC	Alfabética
48	MA	EDTA	EN	VCCV	Silábica
49	MA	PVC	EN	CCC	Alfabética
50	MA	DPD	EN	CCC	Alfabética
51	MA	PNPP	EN	CCCC	Alfabética
52	MA	HCFC	EN	CCCC	Alfabética
53	MA	CPOM	EN	CCVC	Silábica
54	MA	PCB	EN	CCC	Alfabética
55	MA	ADN	ES	VCC	Alfabética
56	MA	TLm	EN	CCC	Alfabética
57	MA	FCSE	ES	CCCV	Alfabética
58	MA	MEDOC	ES	CVCVC	Silábica
59	MA	MEDOR	ES	CVCVC	Silábica
60	MA	ICONA	ES	VCVCV	Silábica
61	MA	MARPOL	EN	CVCCVC	Silábica
62	MA	MINER	ES	CVCVC	Silábica
63	MA	AEDENAT	ES	VVCVCVC	Silábica
64	MA	PNUMA	ES	CCVCV	Silábica
65	MA	NASA	EN	CVCV	Silábica
66	MA	TRAGSA	ES	CCVCCV	Silábica
67	MA	FAPAS	ES	CVCVC	Silábica

Dado que el análisis fonético es el que menos pistas aporta a la hora de establecer patrones y criterios para la detección semiautomática de siglas, hemos limitado esta parte del análisis a dos aspectos. Por un lado, a la observación del tipo de pronunciación que prevalece en función del número de caracteres de cada sigla en inglés y español y, por otro lado, a la observación de estos en los ámbitos de GH y MA.

Los datos de la tabla anterior permiten inferir que, en inglés, la mayoría de las siglas deletreadas son de tres caracteres mientras que la mayoría de las siglas que se pronuncian silábicamente son de cuatro caracteres. Algo similar sucede en español, en donde, la mayoría de las siglas que se deletrean son de tres caracteres mientras que las que se pronuncian silábicamente son mayormente las siglas de cuatro y cinco caracteres.

En el caso de las siglas de dos caracteres presentes en la tabla 15, aunque bien podrían deletrearse, su pronunciación tiende a ser como la de una abreviatura; es decir, se tiende a pronunciar su forma desarrollada. Este es el caso de CP (Código penal), CE (Comisión Europea) y UE (Unión Europea).

Los gráficos 21, 22, 23 y 24 muestran, por una parte, los valores para cada tipo de pronunciación en inglés y español, y por otra parte, los gráficos 25, 26 y 27 muestran estos mismos valores con relación a GH y MA en conjunto y por separado.

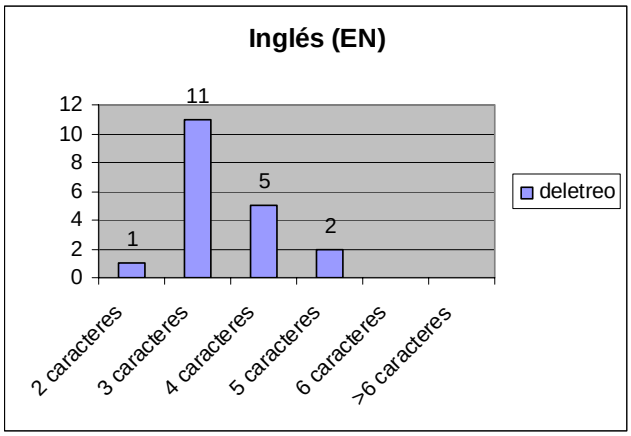


Gráfico 21. Número de siglas con pronunciación alfabética (deletreo) en inglés

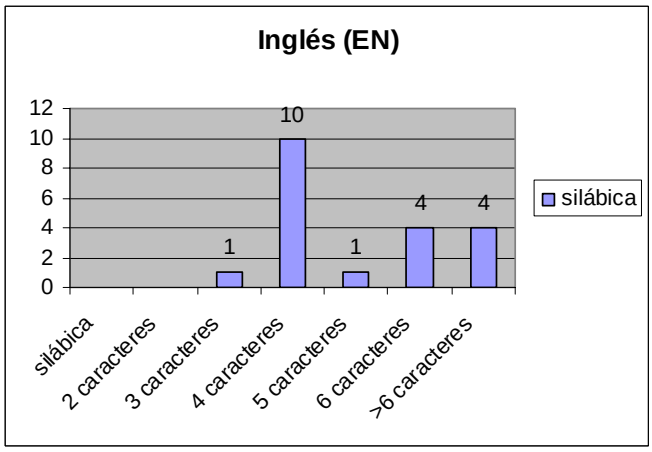


Gráfico 22. Número de siglas con pronunciación silábica en inglés

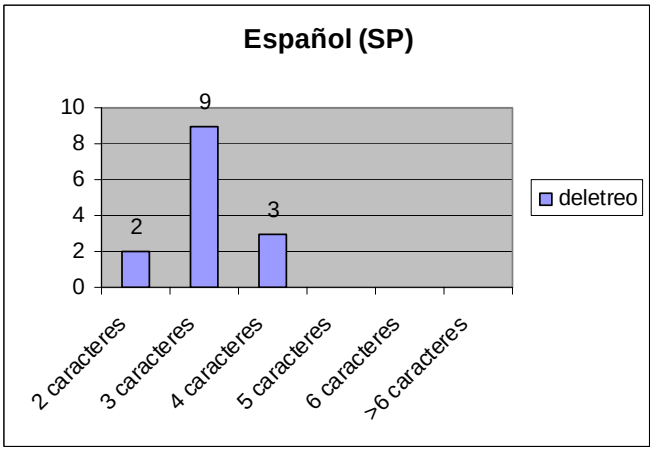


Gráfico 23. Número de siglas con pronunciación alfabética (deletreo) en español

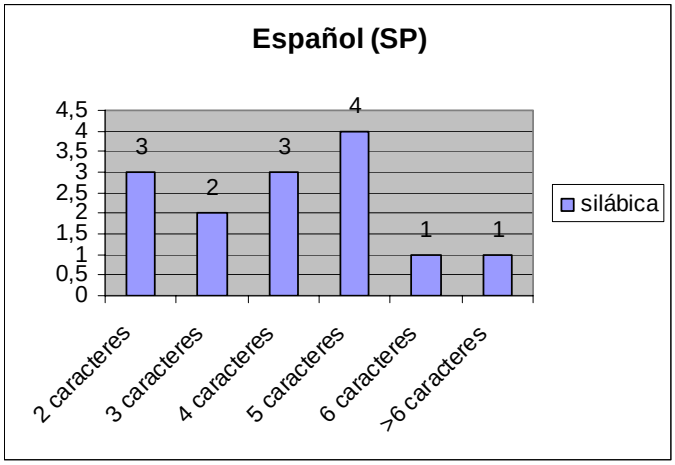


Gráfico 24. Número de siglas con pronunciación silábica en español

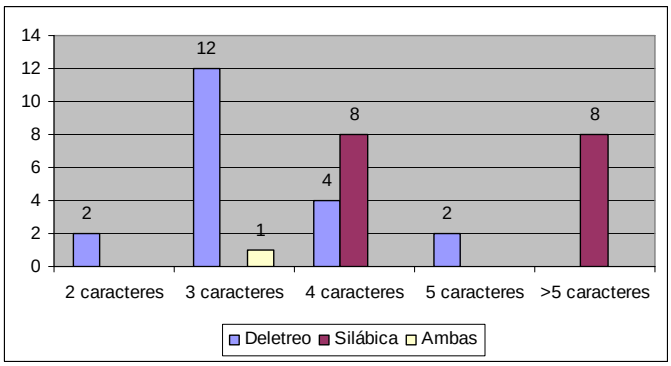


Gráfico 25. Número de siglas con pronunciación alfabética, silábica o ambas en el ámbito de GH

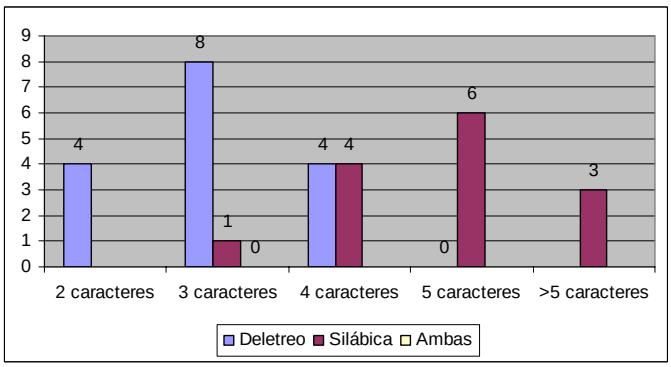


Gráfico 26. Número de siglas con pronunciación alfabética, silábica o ambas en el ámbito de MA

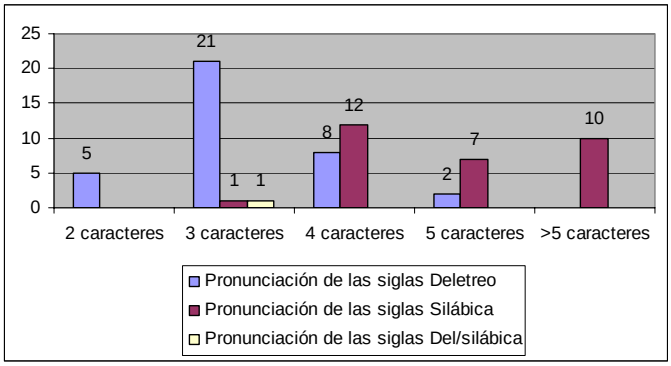


Gráfico 27. Valores totales para la muestra de siglas tomada

2. Morfología

2.1 Determinación de la categoría gramatical predominante en las siglas

La determinación de la categoría gramatical predominante en las siglas ha sido una cuestión ampliamente tratada en la literatura existente sobre el tema. Entre otros, destacamos autores como Rodríguez, 1981, 1993; Martínez de Sousa, 1984; Nakos, 1990; Ambadiang, 1999; Mestres & Guillén, 2001; Cardero, 2002; Fijo, 2003, y Rull, 2005.

Rodríguez (1981: 15) sostiene que "la referencia a sectores científicos, partidos políticos y organizaciones de diversas clases, determina el carácter sustantivo de éstas. A su vez, en virtud de la naturaleza predominantemente político-social de las denominaciones que representan, la mayoría de las siglas constituyen nombres propios con cierto carácter abstracto. En menor número se siglan designaciones de nombres muy concretos referidos a personas (PNN, ATS) u objetos (TV, OVNI)". Coinciden con esta posición Rull (2005:5) y Martínez de Sousa (1984: 39). Así mismo, Rodríguez (1981: 10) afirma que, al desglosar los elementos constitutivos de

la sigla se obtienen las iniciales de dos tipos de palabras ("lexemas"). Por un lado, se obtiene el lexema nuclear, denominado "nombre principal" y que funciona como elemento "determinado o base" de la sigla; y por otro lado, se obtienen los lexemas predicativos, con una función "determinante". La unión de los elementos determinados y determinantes constituye la representación sintagmática subyacente de la sigla, o sea, su forma desarrollada; por ejemplo:

PGH *(Proyecto Genoma humano)*
Elemento base o determinado: *Proyecto*
Elementos determinantes: *Genoma humano*

Nakos (1990: 407) en su estudio sobre las siglas y los nombres propios en francés afirma que existen tres relaciones posibles entre el nombre propio y la sigla, a saber:

1) La sigla formada originalmente por nombres comunes se convierte en nombre propio (el caso de las denominaciones de organismos oficiales)

   En esta primera categoría se encuentran sobre todo las siglas que designan organismos y sociedades de todo tipo.

   Las siglas que designan organismos nacionales no suelen tener una sigla correspondiente en otras lenguas; por ejemplo, AFNOR (*Association Française de Normalisation*). Adicionalmente, las siglas de organismos internacionales, a veces presentan diferencias gráficas según la lengua utilizada; por ejemplo, CEE (*Communauté économique européenne*) se traduce por EEC (*European Economic Community*). Así mismo, algunas siglas de organismos internacionales tienen sus formas desarrolladas traducidas; *e.g.*: UNICEF, cuya denominación oficial en inglés es "*United Nations International Children's Emergency Fund*" y en francés "*Fonds International de Sécours à l'enfance*".

En el caso de las sociedades, los imperativos económicos son los que priman. En efecto, una sociedad se identifica en el mundo de una sola manera, *e.g.*: IBM, BBVA, HSBC, etc.

2) El nombre propio se encuentra en la formación de siglas que son nombres propios (el caso de denominaciones de patronímicos o topónimos)

Esta categoría incluye las siglas que sirven para designar nombres de personajes célebres o nombres de lugares. En francés e inglés, por ejemplo, se han creado siglas de este tipo tales como: BB (*Brigitte Bardot*) y JFK (*John Fitzgerald Kennedy*).

En cuanto a la aplicación de siglas a topónimos, Nakos recoge ejemplos como UK (*United Kingdom*), GB (*Great Britain*) o USA (*United States of America*).

Tanto en el primer como en el segundo caso recién expuestos, ciertas siglas, resultantes de la asociación de nombres comunes o propios, pueden dar lugar a derivados nominales o adjetivales gracias a la lexicalización. Por este método se han obtenido en francés palabras como "*cégétiste*" de CGT (*Confédération Générale du Travail*) y "*onusien*" de ONU (*Organisation des Nations Unies*).

3) El nombre propio sirve para la formación de siglas que son nombres comunes (el caso del vocabulario técnico y científico)

Si se compara el conjunto de siglas relativas al campo de la técnica, sólo un número muy restringido de siglas técnicas utiliza el nombre propio.

La primera categoría mencionada (denominaciones de organismos oficiales) se emplea en este caso para calificar ciertos términos del vocabulario técnico; por ejemplo, "*ASTM supercharge ratings*", donde ASTM significa *American*

Standards for Testing and Materials. En este caso, la denominación oficial siglada se convierte a la vez en una palabra que sirve para calificar un sustantivo, cuya unión forma un sintagma que tiene un sentido específico.

En el campo científico los sintagmas largos y complejos favorecen la formación de siglas. De esta forma, en campos como la medicina, a mayor uso del nombre, mayor posibilidad de que se forme una sigla; por ejemplo, el *bacilo de Koch* (KB). Además, los científicos privilegian el empleo de una sigla única para designar la misma noción; por ejemplo, BCG conserva la misma grafía en inglés y francés.

En el caso de los patronímicos, se pueden dar los siguientes casos:

a) La sigla resulta de la reducción de un grupo de palabras formado por un núcleo de sintagma y topónimos o patronímicos en posición adjetiva (el caso de los epónimos).

Siglas como VWF (*Von Willebrand Factor*) ejemplifican casos de patronímicos en función adjetiva, en tanto que siglas como VEE (*Venezuela Equine Encephalitis*) ilustran casos de toponimia en función adjetiva.

b) En el interior de un sintagma el núcleo se escribe completo mientras que los topónimos o patronímicos en posición adjetiva se siglan.

Los siguientes ejemplos ilustran este caso: fiebre Q (donde Q designa Queensland, Australia) y banda H (donde H significa *Hensen*).

c) En el interior de un sintagma, el núcleo se sigla mientras que los topónimos o patronímicos en posición adjetiva se escriben completamente.

Dentro de este tipo se encuentran siglas como "*TAP-Schiff*", que significa "Reacción al tetracetato de plomo Schiff", y que da lugar a sintagmas como "material de TAP-Schiff positivo" o "reacción TAP-Schiff positiva".

d) La sigla resulta de la unión de un patronímico abreviado con un símbolo (el caso de las unidades de medida).

Algunos símbolos de uso internacional, como las unidades de medida, se crean a partir de la unión de patronímicos como *Hertz* (*Hz*), *Jule* (*J*) o *Watt* (*W*) con compuestos (prefijos) como *mega*, *deci*, *mili* o *kilo*. Algunos ejemplos son: *KPa* (*KiloPascal*), *MHz* (*Megahertz*), *Kw* (*Kilowatt*), etc.

De acuerdo con Nakos "existe un gran número de combinaciones posibles de siglas y nombres comunes (Fiebre Q) o de siglas y nombres propios (TAP-Schiff), o incluso los dos al mismo tiempo (reacción TAP-Schiff positiva) donde la sigla no es el resultado de un sintagma sino el 'origen' de uno nuevo. En este sentido, la sigla se comporta exactamente como otro nombre común que se vuelve el núcleo de un sintagma fijo y monosémico".

En su estudio, Nakos concluye que la sigla tiene la particularidad de ser parte integrante del conjunto del léxico, del vocabulario científico-técnico y de las denominaciones de instituciones oficiales. Ella deriva primero de un sintagma nominal dando lugar a otra clase de posibilidades. Por un lado, puede comportarse como nombre y, por otro lado, como adjetivo (virus SLE). Además, puede dar origen a derivados nominales (*bécégiste*) o adjetivales (*étude onusienne*). Su grafía tiende unas veces al nombre propio y otras veces al nombre común, dependiendo de su utilización precisa dentro del discurso.

En otras lenguas como el inglés se ha detectado que las siglas pueden representar una categoría gramatical diferente a la nominal. Rodríguez (1981: 15) en su estudio sobre siglas inglesas ha documentado el comportamiento de éstas como categorías diferentes a los nombres. Sostiene que "el inglés ofrece una casuística más variada:

verbos (to T.O; to O.C), adverbios (p.m.; a.m.),[37,38] etc. El escaso número de siglas no nominales se ve relativamente incrementado por medio del cambio funcional o metábasis; es decir, siglas que por su base son sustantivas "cambian" de categoría adoptando funciones adjetivas, verbales, etc. Un cambio muy frecuente en las siglas es el de nombres a adjetivos en función atributiva (neutralidad TV, televisión USA, políticos UCD, etc.)". Este último caso se retomará en el apartado 3.3 sobre la "Combinatoria de las siglas".

2.1.1 Análisis del núcleo de la sigla

El tratamiento de la información para este análisis ha implicado dos pasos. En el primero, se ha marcado y extraído la palabra que funciona como núcleo (N) en cada binomio de forma desarrollada-sigla. En el segundo, se ha agregado a cada núcleo la información sobre categoría gramatical, aspectos flexivos y tipo de nombre, por una parte y, por otra parte, se ha indicado la letra (o carácter) de la sigla que abrevia al núcleo de la forma desarrollada.

Este análisis se formaliza de la siguiente manera:

FD	↔	Sigla
Núcleo de la FD [*cat. gram., género, número*][*clase de nombre*]	↔	Núcleo de la sigla

(1) Polymerase chain **reaction** ↔ PC**R**
reacción [*n., f., s.*] [*común*] ↔ R

(2) **Proyecto** Genoma Humano ↔ **P**GH
proyecto [*n., m., s.*] [*común*] ↔ P

37 *to turn over*; *to overcharge deliberately*
38 *post meridiem*; *ante meridiem*

(3) Restriction fragment length **polimorphisms** ↔ RFL**P**
polimorfismos [*n., m., pl.*] [*común*] ↔ P

(4) **Virus** de la inmunodeficiencia humana ↔ **V**IH
virus [*n., m., s.* y *pl.*] [*común*] ↔ V

(5) Human leukocyte **antigen** ↔ HL**A**
antígeno [*n., m., s.*] [*común*] ↔ A

(6) Major histocompatibility **complex** ↔ MH**C**
complejo [*n., m., s.*] [*común*] ↔ C

(7) Variable **number** of tandem repeats ↔ V**N**TR
número [*n., m., s.*] [*común*] ↔ N

(8) Embryonic **stem** ↔ E**S**
célula [*n., f., s.*] [*común*] ↔ S

(9) **Fibrosis** quística ↔ **F**Q
fibrosis [*n., f., s.*] [*común*] ↔ F

(10) Linear energy **transfer** ↔ LE**T**
transferencia [*n., f., s.*] [*común*] ↔ T

(11) **Ácido** desoxirribonucleico ↔ **A**DN
ácido [*n., m., s.*] [*común*] ↔ A

(12) Deoxyribonucleic **Acid** ↔ DN**A**
ácido [*n., m., s.*] [*común*] ↔ A

(13) **Ácido** ribonucleico ↔ **A**RN
ácido [*n., m., s.*] [*común*] ↔ A

(14) Ribonucleic **Acid** ↔ RN**A**
ácido [*n., m., s.*] [*común*] ↔ A

(15) **Ácido** ribonucleico mensajero ↔ **A**RNm
ácido [*n., m., s.*] [*común*] ↔ A

(16) **Aberraciones** cromosómicas ↔ **A**Cs
aberraciones [*n., f., pl.*] [*común*] ↔ A

(17) Breast cancer susceptibility **gene** 1 ↔ BRCA1
gen [*n., m., s.*] [*común*] ↔ *****[39]

(18) Cystic fibrosis transmembrane conductance **regulator** ↔ CFT**R**
regulador [*n., m., s.*] [*común*] ↔ R

(19) **Adenosín(a)** trifosfato ↔ **A**TP
adenosín [*n., m., s.*] [*común*] ↔ A
adenosina [*n., f., s.*] [*común*] ↔ A

(20) Breast cancer susceptibility **gene** 2 ↔ BRCA2
gen [*n., m., s.*] [*común*] ↔ *****

(21) **Adenosín(a)** desaminasa ↔ **A**DA
adenosín [*n., m., s.*] [*común*] ↔ A
adenosina [*n., f., s.*] [*común*] ↔ A

(22) Long interspersed nuclear **elements** ↔ LIN**E**/LIN**Es**
elementos [*n., m., pl.*] [*común*] ↔ E/Es

(23) Regulated on activation, normal T **cells** expressed and secreted ↔ RAN**T**ES
células [*n., f., pl.*] [*común*] ↔ T

(24) Ethylenediamine Tetraacetic **Acid** ↔ EDT**A**
ácido [*n., m., s.*] [*común*] ↔ A

(25) Short interspersed **elements** ↔ SIN**E**/SIN**Es**
elementos [*n., m., pl.*] [*común*] ↔ E/Es

(26) Human Genome **Organization** ↔ HUG**O**
organización [*n., f., s.*] [*común*] ↔ O

(27) Yeast artificial **chromosomes** ↔ YA**Cs**

39 * el núcleo no está explícito en la sigla

cromosomas [*n., m., pl.*] [*común*] ↔ Cs

(28) Immunoreceptor tyrosine-based inhibition **motifs** ↔ ITI**M**
motivos [*n., m., pl.*] [*común*] ↔ M

(29) United Nations Educational, Scientific and Cultural **Organization** ↔ UNESC**O**
organización [*n., f., s.*] [*común*] ↔ O

(30) Janus **quinasa** 2 ↔ JA**K**2
quinasa [*n., f., s.*] [*común*] ↔ K

(31) Mega yeast artificial **chromosomes** ↔ MegaYA**Cs**
cromosomas [*n., m., pl.*] [*común*] ↔ Cs

(32) California **Institute** of Technology ↔ CalTech
instituto [*n., m., s.*] [*común*] ↔ *

(33) Pig gene mapping **project** ↔ PiGMap
proyecto [*n., m., s.*] [*común*] ↔ P

(34) Phragment Assembly **Program** ↔ PHRA**P**
programa [*n., m., s.*] [*común*] ↔ P

(35) GenBank Gene Products ↔ Genpept
banco [*n., m., s.*] [*común*] ↔ *****

(36) Public/Publisher MEDLINE ↔ PubMed
banco [*n., m., s.*] [*común*] ↔ *

(37) **Síndrome** de inmunodeficiencia adquirida ↔ **S**ida
síndrome [*n., m., s.*] [*común*] ↔ S

(38) **Comisión** europea ↔ **C**E
comisión [*n., f., s.*] [*común*] ↔ C

(39) **Demanda** bioquímica de oxígeno ↔ **D**BO
demanda [*n., f., s.*] [*común*] ↔ D

(40) **Unión** Europea ↔ **U**E
unión [*n., f., s.*] [*común*] ↔ U

(41) **Organización** de Cooperación y Desarrollo Económicos ↔ **O**CDE
organización [*n., f., s.*] [*común*] ↔ O

(42) **Oxígeno** disuelto ↔ **O**D
oxígeno [*n., m., s.*] [*común*] ↔ O

(43) **Código** penal ↔ **C**P
código [*n., m., s.*] [*común*] ↔ C

(44) **Unidad** de carbón piedra ↔ **U**CP
unidad [*n., f., s.*] [*común*] ↔ U

(45) **Demanda** bioquímica de oxígeno a los 5 días ↔ **D**BO5
demanda [*n., f., s.*] [*común*] ↔ D

(46) Nuevos **estados** independientes ↔ **N**EI
Estados [*n., m., pl.*] [*común*] ↔ N

(47) **Análisis** en componentes principales ↔ **A**CP
análisis [*n., m., s.* y *pl.*] [*común*] ↔ A

(48) Ethylenediamine Tetraacetic **Acid** ↔ EDT**A**
ácido [*n., m., s.* y *pl.*] [*común*] ↔ A

(49) Polyvinyl **Chloride** ↔ PV**C**
cloruro [*n., m., s.*] [*común*] ↔ C

(50) **Dietil**-p-fenilendiamina ↔ DP**D**
Dietil [*n., f., s.*] [*común*] ↔ D

(51) Para-nitrophenyl**phosphate** ↔ PNP**P**
fosfato [*n., m., s.*] [*común*] ↔ P

(52) Hydrochlorofluoro**carbon** ↔ HCF**C**
carbono [*n., m., s.*] [*común*] ↔ C

(53) **Materia** orgánica particulada gruesa ↔ CPO**M**
materia [*n., f., s.*] [*común*] ↔ M

(54) Polychlorinated **Biphenyl** ↔ PC**B**
bifenilo [*n., m., s.*] [*común*] ↔ B

(55) **Ácido** desoxirribonucleico ↔ **A**DN
ácido [*n., m., s.*] [*común*] ↔ A

(56) Threshold **limit** ↔ T**Lm**
límite [*n., m., s.*] [*común*] ↔ Lm

(57) **Ley** de fuerzas y cuerpos de seguridad ↔ FCSE
ley [*n., f., s.*] [*común*] ↔ *

(58) **Mediterráneo** occidental ↔ **MED**OC
Mediterráneo [*n., m., s.*] [*propio*] ↔ MED

(59) **Mediterráneo** oriental ↔ **MED**OR
Mediterráneo [*n., m., s.*] [*propio*] ↔ MED

(60) **Instituto** para la Conservación de la Naturaleza ↔ **I**CONA
instituto [*n., m., s.*] [*común*] ↔ I

(61) Marine pollution **convention** ↔ MARPOL
convención [*n., f., s.*] [*común*] ↔ *

(62) **Ministerio** de Industria y Energía ↔ **M**INER
ministerio [*n., m., s.*] [*común*] ↔ M

(63) **Asociación** Ecologista de Defensa de la Naturaleza ↔ **A**EDENAT
asociación [*n., f., s.*] [*común*] ↔ A

(64) **Programa** de Naciones Unidas para el Medio Ambiente ↔ **P**NUMA
programa [*n., m., s.*] [*común*] ↔ P

(65) National Aeronautics and Space **Administration** ↔ NAS**A**

administración [*n., f., s.*] [*común*] ↔ A

(66) **Transformación** agraria s.a. ↔ **TR**AGSA
transformación [*n., f., s.*] [*común*] ↔ TR

(67) **Fondo** para la Protección de los Animales Salvajes ↔ **F**APAS
fondo [*n., m., s.*] [*común*] ↔ F

Con base en los datos anteriores, se ha observado que la totalidad de las siglas tiene como núcleo a un nombre, hecho que nos ha permitido corroborar el carácter eminentemente nominal de estas unidades. Resumimos la afirmación anterior mediante el siguiente esquema:

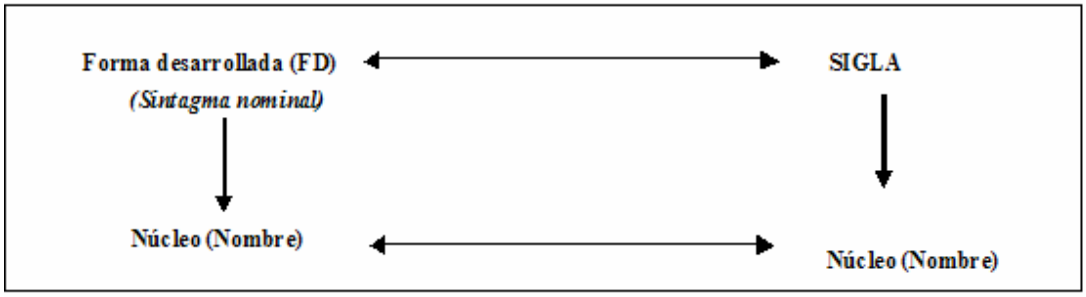


Fig. 4. Correspondencia de núcleo en forma desarrollada y sigla

La sintagmación y la siglación van de la mano, siendo las siglas el resultado lógico de la primera. A este respecto Koucurek (1982: 121) sostiene que *"[…] l'utilité des sigles et des acronymes est due au fait qu'ils combinent dans la communication spécialisée, la maniabilité syntagmatique d'un seul mot avec le caractère sémantique descriptif du syntagme sous-jacent étendu"*.

Del análisis de los núcleos de las siglas se desprende que 65 núcleos de forma desarrollada corresponden a nombres comunes (NC) en tanto que 2 corresponden a nombres propios (NP), tal y como puede apreciarse en la tabla siguiente.

Tabla 16. Listado de nombres que desempeñan la función de núcleo en el sintagma

Tipo de nombre	Núcleo del sintagma nominal	Frecuencia de aparición del nombre (núcleo de la sigla)
NP	Mediterráneo	2
NC	Reacción	1
	Proyecto	2
	Polimorfismo	1
	Virus	1
	Antígeno	1
	Complejo	1
	Variación	1
	Célula	2
	Fibrosis	1
	Transferencia	1
	Ácido	8
	Aberración	1
	Gen	2
	Regulador	1
	Adenosina	2
	Elemento	2
	Organización	3
	Cromosoma	2
	Motivo	1
	Quinasa	1
	Instituto	2
	Banco	2
	Programa	2
	Síndrome	1
	Comisión	1
	Demanda	2
	Oxígeno	1
	Código	1
	Unidad	1
	Estado	1
	Análisis	1
	Cloruro	1
	p-fenilendiamina	1
	Fosfato	1
	Carbono	1
	Materia	1
	Bifenilo	1
	Límite	1
	Ley	1
	Convención	1
	Ministerio	1
	Asociación	1
	Agencia	1
	Fondo	1
	Transformación	1
	Unión	1

Ahora bien, para la detección de los casos de siglas que actúan como nombres propios, hemos empleado la tipología de Nakos (1990), mencionada al comienzo de este apartado. Los resultados son los siguientes:

Clase de nombre	Siglas
1. La sigla formada originalmente por nombres comunes se convierte en nombre propio (el caso de las denominaciones de organismos oficiales).	1. PGH (Proyecto Genoma Humano 2. HUGO (Human Genome Organization) 3. UNESCO (Organización de las Naciones Unidas para la Educación, la Cultura y la Ciencia) 4. CalTech (California Institute of Technology) 5. PiGMap (Pig Gene Mapping Project) 6. PHRAP (Phragment Assembly Program) 7. Genpept 8. PubMed 9. CE (Comisión Europea) 10. UE (Unión Europea) 11. OCDE (Organización de Cooperación y Desarrollo Económicos) 12. ICONA (Instituto para la Conservación de la Naturaleza) 13. MARPOL (Convention for the prevention of pollution from ships) 14. MINER (Ministerio de Industria y Energía) 15. AEDENAT (Asociación Ecologista de Defensa de la Naturaleza) 16. PNUMA (Programa de las Naciones Unidas para el Medio Ambiente) 17. NASA (National Aeronautics and Space Administration) 18. TRAGSA (Transformación Agraria S.A.) 19. FAPAS (Fondo para la Protección de los Animales Salvajes)
2. El nombre propio se encuentra en la formación de siglas que son nombres propios (el caso de denominaciones de patronímicos o topónimos).	1. MEDOR (Mediterráneo oriental) 2. MEDOC (Mediterráneo occidental)
3. El nombre propio sirve para la formación de siglas que son nombres comunes (el caso del vocabulario técnico y científico).	---
3.1 La sigla resulta de la reducción de un grupo de palabras formado por un núcleo de sintagma y topónimos o patronímicos en posición adjetiva (el caso de los epónimos).	---
3.2 En el interior de un sintagma el núcleo se escribe completo y topónimos o patronímicos en posición adjetiva se siglan.	---
3.3. Al interior de un grupo de palabras, el núcleo del sintagma está siglado y el topónimo o patronímico en posición adjetiva está escrito completamente.	---

3.4 La sigla-símbolo resulta de la reducción de un patronímico (el caso de las unidades de medida).	---
4. Nombres comunes	1. PCR (Polymerase Chain Reaction) 2. RFLP (Restriction Fragment Length) 3. VIH (Virus de la inmunodeficiencia humana) 4. HLA (Human Leukocyte Antigen) 5. MHC (Major Histocompatibility Complex) 6. VNTR (Variable Number of Tandem Repeats) 7. ES (Embryonic Stem) 8. FQ (Fibrosis quística) 9. LET (Linear Energy Transfer) 10. ADN (Ácido desoxiribonucleico) 11. DNA (Deoxyribonucleic Acid) 12. ARN (Ácido ribonucleico) 13. RNA (Ribonucleic Acid) 14. ARNm (ARN mensajero) 15. ACs (Aberraciones cromosómicas) 16. BRCA1 (Breast Cancer Susceptibility Gene 1) 17. CFTR (Cystic Fibrosys Transmembrane Conductance Regulator) 18. ATP (Adenosine Triphosphate) 19. BRCA2 (Breast Cancer Susceptibility Gene 2) 20. ADA (Adenosin desaminase) 21. LINE (Long Interspersed Nuclear Elements) 22. RANTES (Regulated on Activation, Normal T cells Expressed and Secreted) 23. EDTA (Ácido etilendiaminotetracético) 24. SINE (Small Interspersed Repetitive Elements) 25. YACS (Yeast Artificial Chromosomes) 26. ITIM (Immunoreceptor Tyrosine-based Inhibition Motifs) 27. JAK2 (Janus kinase 2) 28. MegaYACs (Mega Yeast Artificial Chromosomes) 29. SIDA (Síndrome de inmunodeficiencia adquirida) 30. DBO (Demanda bioquímica de oxígeno) 31. OD (Oxígeno disuelto) 32. CP (Código penal) 33. UCP (Unidad de carbón piedra) 34. DBO5 (Demanda bioquímica de oxígeno a los 5 días) 35. NEI (Nuevos Estados independientes) 36. ACP (Análisis en Componentes Principales) 37. EDTA (Ethylenediamine Tetraacetic Acid) 38. PVC (Polyvinyl Chloride) 39. DPD (Diethyl-p-phenylene diamine) 40. PNPP (Para-nitrophenylphosphate) 41. HCFC (Hydrochlorofluorocarbon) 42. CPOM (Coarse and fine Particulate Organic Matter) 43. PCB (Polychlorinated Biphenyls) 44. ADN (Ácido desoxirribonucleico) 45. TLm (Threshold limit) 46. FCSE (Ley de fuerzas y cuerpos de seguridad)

Tabla 17. Función de las siglas como nombres propios o comunes

En lo que respecta al primer tipo de relación establecida en la tipología de Nakos (el caso de las denominaciones de organismos oficiales), se han encontrado 19 casos en los que se evidencia la transmutación de nombre común a nombre propio, a saber:

(1) Proyecto [n. común] → PGH [n. propio]
(2) Organización [n. común] → HUGO [n. propio]
(3) Organización [n. común] → UNESCO [n. propio]

(4) Instituto [n. común] → CalTech [n. propio]
(5) Proyecto [n. común] → PiGMap [n. propio]
(6) Programa [n. común] → PHRAP [n. propio]
(7) Banco [n. común] → Genpept [n. propio]
(8) Banco [n. común] → PubMed [n. propio]
(9) Comisión [n. común] → CE [n. propio]
(10) Unión [n. común] → UE [n. propio]
(11) Organización [n. común] → OCDE [n. propio]
(12) Instituto [n. común] → ICONA [n. propio]
(13) Convención [n. común] → MARPOL [n. propio]
(14) Ministerio [n. común] → MINER [n. propio]
(15) Asociación [n. común] → AEDENAT [n. propio]
(16) Programa [n. común] → PNUMA [n. propio]
(17) Administración [n. común] → NASA [n. propio]
(18) Transformación [n. común] → TRAGSA [n. propio]
(19) Fondo [n. común] → FAPAS [n. propio]

Con respecto al segundo tipo, es decir, a las siglas que originariamente han sido formadas por nombres propios y que se han convertido en nombres propios, se han registrado dos casos, a saber:

(1) Mediterráneo oriental → MEDOR
(2) Mediterráneo occidental → MEDOC

En lo concerniente a las siglas pertenecientes al tercer tipo, es decir, los nombres propios que sirven para formar siglas que son nombres comunes, no hemos encontrado ningún caso.

El conjunto de siglas resultante, es decir 46, pertenece a la categoría de nombres comunes (que hemos añadido en la tabla anterior bajo el numeral 4). Los casos detectados son:

(1) reacción [n. común] → PCR [n. común]
(2) polimorfismo [n. común] → RFLP [n. común]
(3) virus [n. común] → VIH [n. común]
(4) antígeno [n. común] → HLA [n. común]
(5) complejo [n. común] → MHC [n. común]
(6) número [n. común] → VNTR [n. común]
(7) célula [n. común] → ES [n. común]
(8) fibrosis [n. común] → FQ [n. común]

(9) transferencia [n. común] → LET [n. común]
(10) ácido [n. común] → ADN [n. común]
(11) ácido [n. común] → DNA [n. común]
(12) ácido [n. común] → ARN [n. común]
(13) ácido [n. común] → RNA[n. común]
(14) ácido [n. común] → ARNm [n. común]
(15) aberraciones [n. común] → ACs [n. común]
(16) gen [n. común] → BRCA1 [n. común]
(17) regulador [n. común] → CFTR [n. común]
(18) adenosín [n. común] → ATP [n. común]
(19) gen [n. común] → BRCA2 [n. común]
(20) adenosín [n. común] → ADA [n. común]
(21) elementos [n. común] → LINE [n. común]
(22) células [n. común] → RANTES [n. común]
(23) ácido [n. común] → EDTA [n. común]
(24) elementos [n. común] → SINE [n. común]
(25) cromosomas [n. común] → YACS [n. común]
(26) motivos [n. común] → ITIM [n. común]
(27) quinasa [n. común] → JAK2 [n. común]
(28) cromosomas [n. común] → MegaYACs [n. común]
(29) síndrome [n. común] → SIDA [n. común]
(30) demanda [n. común] → DBO [n. común]
(31) oxígeno [n. común] → OD [n. común]
(32) código [n. común] → CP [n. común]
(33) unidad [n. común] → UCP [n. común]
(34) demanda [n. común] → DBO5 [n. común]
(35) estados [n. común] → NEI [n. común]
(36) análisis [n. común] → ACP [n. común]
(37) ácido [n. común] → EDTA [n. común]
(38) cloruro [n. común] → PVC [n. común]
(39) dietil [n. común] → DPD [n. común]
(40) fosfato [n. común] → PNPP [n. común]
(41) carbono [n. común] → HCFC [n. común]
(42) materia [n. común] → CPOM [n. común]
(43) bifenilo [n. común] → PCB [n. común]
(44) ácido [n. común] → ADN [n. común]
(45) límite [n. común] → TLm [n. común]
(46) ley [n. común] → FCSE [n. común]

Los datos anteriores reflejan que dentro de la muestra de 67 siglas de GH y MA, que hemos seleccionado para el análisis lingüístico, predominan los nombres comunes. En efecto, se observa que 46 siglas corresponden a esta clase de nombres en tanto que 21 corresponden a nombres propios. Estos datos están en consonancia con lo dicho por Nakos (1990: 410), quien sostiene que sólo un número muy restringido de siglas científico-técnicas utiliza el nombre propio. Este hecho llama la atención, pues según otras investigaciones enfocadas hacia el estudio de siglas en el discurso general, la tendencia es a la inversa; es decir, que en el discurso general predominan los nombres propios (*cf*. Rodríguez, 1993: 13; Rull, 2005: 5; y Martínez de Sousa, 1984: 39).

Los gráficos 28, 29 y 30 muestran la proporción que ocupan los nombres propios (NP) y comunes (NC) en cada ámbito y en conjunto.

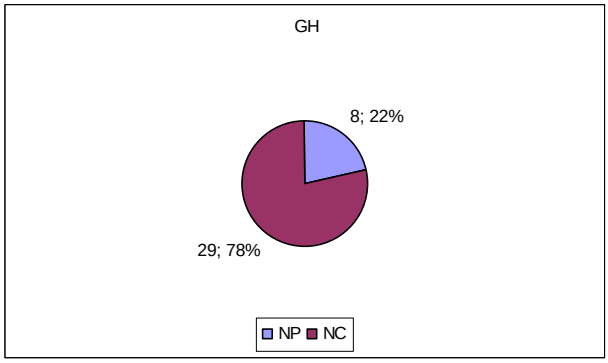


Gráfico 28. Siglas funcionando como NP y NC en las siglas del ámbito de GH

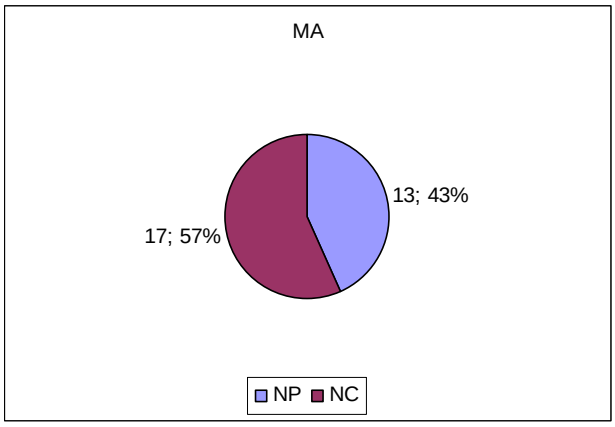


Gráfico 29. Siglas funcionando como NP y NC en las siglas del ámbito de MA

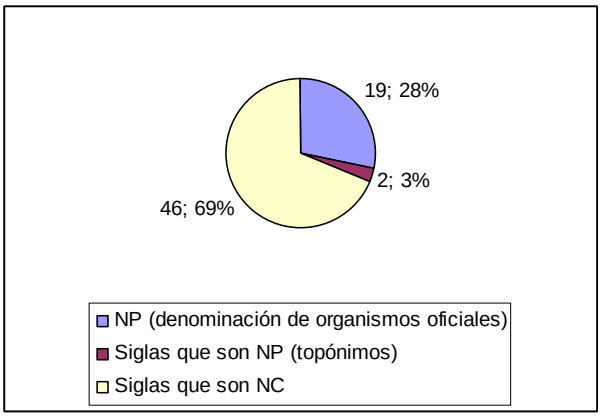


Gráfico 30. Siglas funcionando como NP y NC en toda la muestra

Dado que las siglas que se han escogido para este análisis constituyen sólo una pequeña muestra de nuestro corpus, convendría realizar otros estudios con miras a verificar estos hechos en corpus más extensos.

El carácter nominal de las siglas permite que sean modificadas por medio de adjetivos, tal y como se puede apreciar en los siguientes casos detectados en el corpus.

	GH		MA	
	Sigla	**Sigla+Adj.**	**Sigla**	**Sigla+Adj.**
1	PCR	PCR arbitraria PCR asimétrica PCR cuantitativa PCR estándar PCR inversa PCR multiplex PCR negativa PCR positiva	CE	---
2	PGH	---	DBO	---
3	RFLP	---	UE	---
4	VIH	---	OCDE	---
5	HLA	---	OD	OD consumido OD inicial
6	MHC	MHC humano	CP	---
7	VNTR	VNTR humanos	UCP	---
8	ES	ES humanas ES pluripotentes	DBO5	---
9	FQ	---	NEI	---
10	LET	---	ACP	---
11	ADN	ADN autosómico ADN bacteriano ADN bicatenario ADN celular ADN circular ADN circundante ADN clonado ADN complementario ADN cromosómico ADN extraño ADN genómico ADN híbrido ADN humano ADN libre ADN microsatélite ADN minisatélite ADN nuclear ADN paterno ADN recombinante ADN repetitivo ADN ribosómico ADN satélite ADN superenrollado ADN viral	EDTA	---
12	DNA	DNA anónimo DNA bacteriano DNA basura DNA circular DNA clonado DNA codificante DNA cromosómico DNA dañado DNA donante	PVC	---

		DNA egoísta DNA espaciador DNA eucariótico DNA exógeno DNA extraño DNA foráneo DNA genómico DNA heterodúplex DNA humano DNA minisatélite DNA mitocondrial DNA mutado DNA nuclear DNA operador DNA original DNA parásito DNA plasmídico DNA polimórfico DNA proviral DNA purificado DNA quimérico DNA repetitivo DNA retrotranscrito DNA satélite DNA sintético DNA subgenómico DNA superenrollado DNA telomérico DNA trasnformante DNA unicatenario DNArecombinante		
13	ARN	ARN bacteriano ARN bicatenario ARN celular ARN codificado ARN diana ARN genómico ARN intermediario ARN maduro ARN mensajero ARN nuclear ARN residual ARN ribosómico ARN sintetizado ARN transcrito ARN unicatenario ARN vírico	DPD	---
14	RNA	RNA celular RNA diana RNA funcional RNA informativo RNA inmaduro RNA maduro RNA mensajero	PNPP	---

		RNA nuclear RNA proviral RNA ribosómico RNA transferente RNA transferido RNA vírico		
15	ARNm	ARNm anómalo ARNm bacteriano ARNm común ARNm delecionado ARNm eucariótico ARNm híbrido ARNm humano ARNm isomorfo ARNm maduro ARNm monocatenario ARNm monocistrónico ARNm mutuo ARNm nuclear ARNm policistrónico ARNm quimérico	HCFC	---
16	ACs	---	CPOM	---
17	BRCA1	---	PCB	---
18	CFTR	---	ADN	---
19	ATP	---	TLm	---
20	BRCA2	---	FCSE	---
21	ADA	---	MEDOC	---
22	LINE/LINEs	---	MEDOR	---
23	RANTES	---	ICONA	---
24	EDTA	---	MARPOL	---
25	SINE/SINEs	---	MINER	---
26	HUGO	---	AEDENAT	---
27	YACS	---	PNUMA	---
28	ITIM	---	NASA	---
29	UNESCO	---	TRAGSA	---
30	JAK2	---	FAPAS	---
31	MegaYACs	---		
32	CalTech	---		
33	PiGMap	---		
34	PHRAP	---		
35	Genpept	---		
36	PubMed	---		
37	Sida	---		

Tabla 18. Siglas modificadas por adjetivos

Del análisis de la tabla anterior se deduce que los adjetivos más comunes que modifican a las siglas de nuestro estudio son: humano (6 ocurrencias), nuclear (5 ocurrencias), bacteriano (4 ocurrencias); genómico, celular, maduro y ribosómico (3 ocurrencias), y circular, clonado, codificado, cromosómico, diana, eucariótico,

extraño, híbrido, mensajero, minisatélite, proviral, quimérico, recombinante, repetitivo, superenrollado y unicatenario (2 ocurrencias). El contraste de los dos ámbitos de especialidad muestra que el fenómeno es mucho más recurrente en GH, donde 9 siglas aparecen modificadas por adjetivos, contra un solo caso que se registra en MA.

2.2 Aspectos flexivos de las siglas

Al asumir que toda sigla proviene de un sintagma nominal, cuyo núcleo es un nombre, asumimos también que las siglas tienen género y número.

2.2.1 El género en las siglas

El género, sirve para tres propósitos: 1) actualizar determinado morfema lexemático como nombre sustantivo o adjetivo; 2) marcar la concordancia (junto con el número y el artículo), y 3) aportar información sobre el sexo y otros aspectos de la realidad que representa el lexema. El principal elemento diferenciador del género es el artículo. (Cascón, 2004: 219).

Todas las siglas que comprenden el presente estudio han reflejado el género que portan sus respectivos núcleos en la forma desarrollada. Un par de ejemplos son PGH y MINER:

(1) Proyecto Genoma Humano ↔ PGH

N: proyecto [*n., m., s.*]

<m00190>: para pasar lo .</s> <s>El ## PGH ## hará posible el diagnóstico prenatal
<m00190>: original sobre financiamiento de l ## PGH ## .</s> <s>En el original ,
<m00190>: por el desierto de l ## PGH ## hasta la tierra prometida de
<m00225>: <s>Collins dirige el Proyecto del Genoma Humano (## PGH ##), que hasta ahora ha
<m00225>: todos nuestros genes .</s> <s>El ## PGH ## es un consorcio público integrado
<m00225>: <s>El método utilizado por el ## PGH ## se ha descrito como concienzudo
<m00225>: que los equipos de l ## PGH ## y de Celera tienen sobre
<m00225>: calificado muchas veces a l ## PGH ## y a Celera como competidores
<m00225>: incuestionable .</s> <s>Dado que el ## PGH ## es un proyecto público ,
<m00225>: l genoma humano .</s> <s>El ## PGH ## anunció a principios de mayo

(2) Ministerio de Industria y Energía ↔ MINER

N: Ministerio [*n., m., s.*]

<a00022>: <s>El Ministerio de Industria y Energía (## MINER ##), por citar un caso
<a00022>: <s>En el organigrama de l ## MINER ## se incluyen también algunos de
<a00022>: siempre su dependencia de l ## MINER ## .</s> <s>Con ellos sí hay
<a00022>: verdadero bosque .</s> <s>Desde el ## MINER ## tampoco se dieron demasiadas facilidades.
<a00022>: l mundo nuclear .</s> <s>El ## MINER ## es también , a través de l
<a00022>: entre el MOPTMA y el ## MINER ## , a l no coincidir
<a00022>: también se gestiona desde el ## MINER ## .</s> <s>Un buen porcentaje de

En la literatura sobre el tema se afirma con frecuencia que el género de las siglas en español está determinado por el género del nombre principal de la forma desarrollada, casi siempre representado por la letra inicial (*cf.* Rodríguez, 1981: 91; 1983: 277; 1984: 310; 1993: 14; Martínez de Sousa, 1984: 37; Mestres & Guillén, 2001: 29; Desrosiers, 2005: 18). Esta misma regla se observa en otras lenguas romances como el italiano o el francés. Para Rodríguez esta solución se debe a la estructura misma de las lenguas, según la cual el sustantivo precede en general al adjetivo y al complemento. Calvet (1980: 97) coincide con la afirmación anterior y señala que las estructuras más comunes en las formas desarrolladas de las siglas en francés son tres, a saber:

1) Nombre + adjetivo + complemento de nombre, *e.g.*: CGT;
2) Nombre + adjetivo + adjetivo, *e.g.*: CEE;

3) Nombre + complemento de nombre + adjetivo, *e.g.*: ONU, URSS. En este caso el adjetivo califica en este caso el complemento del nombre y no el nombre inicial.

Aunque como se ha señalado el nombre inicial o núcleo da el género al conjunto y, por consiguiente, es lógico que dé el género a la sigla; *e.g.*: *el* PGH (*el Proyecto* Genoma humano), Rodríguez afirma que las asociaciones semánticas en la mente del hablante son el principal aspecto que determina el género en las siglas. Estas asociaciones explicarían algunas fluctuaciones que a veces se detectan en el empleo del artículo; *e.g.*: *la* DINA (Dirección de Inteligencia Nacional), forma más frecuente, pero ocasionalmente *el* DINA. En el caso de *el* DINA se supone que lo que está implícito es un concepto como "organismo", "cuerpo represivo", etc. todos ellos de género masculino (Rodríguez, 1984: 310). En nuestro estudio no encontramos ninguna sigla que refleje este fenómeno.

De acuerdo con Fernández (1999: 84), en el caso de las siglas que hacen referencia a nombres propios, tales como nombres de instituciones o de cualquier producto de la actividad humana, es el nombre apelativo que especifican el que decide comúnmente el género; *e.g.*: *el* MEDOC (Mediterráneo), *el* ICONA (Instituto), *la* UE (Unión), etc.

Cuando se trata de siglas extranjeras con un orden sintáctico diferente, en inglés por ejemplo, lo que generalmente se señala como factor determinante del género es el nombre principal equivalente en español; *e.g.*: *la* CIA (*Central Intelligence Agency*), donde la A representa (la) *Agencia* (Rodríguez, 1983: 277; 1984: 310). En nuestro estudio se corroboran estos hechos en los casos que se citan a continuación:

La PCR [del inglés *Polymerase chain reaction*] = *La reacción* en cadena de la polimerasa; donde la R representa (la) *reacción*.

```
<m00740>:       cadena de la polimerasa ( ## PCR ## ) es un método que
<m00740>:       medicina , concretamente , la ## PCR ## ha tenido un mayor impacto
<m00740>:       de biología molecular , la ## PCR ## es una técnica rutinaria de
<m00740>:       deseado .</s> <s>El método de ## PCR ## fue desarrollado a mediados de
<m00740>:       simple como molde .</s> <s>La ## PCR ## utiliza dos fragmentos cortos de
<m00740>:       de reacción y comenzar la ## PCR ## lo más rápidamente posible para
<m00740>:       <s>En un principio , la ## PCR ## se realizaba manualmente , y
<m00740>:       las primeras « máquinas de ## PCR ## ».</s> <s>Los primeros termocicladores eran
```

El MHC [del inglés *Major histocompatibility complex*] = *El complejo* mayor de histocompatibilidad; donde la C representa (el) *complejo*.

```
<m00835>:       DE HISTOCOMPATIBILIDAD</head> <s>El Complejo Principal de Histocompatibilidad ( ## MHC ## , Major
<m00835>:       .</s> <s>Los genes de l ## MHC ## se encuentran situados en el
<m00835>:       el cromosoma 6 .</s> <s>El ## MHC ## humano se extiende a lo largo de unas
<m00835>:       ).</s> <s>Los genes de l ## MHC ## de clase I en humanos
<m00835>:       las cadenas xxx de l ## MHC ## de clase I .</s> <s>Las
<m00835>:       .</s> <s>Las moléculas de l ## MHC ## de clase I son glucoproteínas
<m00835>:       l complejo génico de l ## MHC ## .</s> <head>Estructura de las moléculas
<m00835>:       de las moléculas de l ## MHC ## de clase I</head> <s>N y
```

El PVC [del inglés *Polyvinyl Chloride*] = *El cloruro* de polivinilo; donde la C representa (el) *cloruro.*

```
<a00011>:       de atención .</s> <s>en cuanto a l ## PVC ## , las recomendaciones son moderadas
<a00011>:       <s>En cuanto a l uso de l ## PVC ## , las recomendaciones de l
<a00011>:       proceso oxiran .</s> <s>En cuanto a l ## PVC ## , la VCI discrepa de
<a00011>:       asociadas de sustituir a l ## PVC ## por otros materiales poliméricos .</s>
<a00011>:       plástico tales como en el ## PVC ## , el óxido de propileno
<a00022>:       , incineración , papeleras , ## PVC ## ( 72 ); Pesca (
<a00179>:       datos científicos relativos a l ## PVC ## y , si fuera necesario
<a00179>:       que la eliminación de l ## PVC ## mediante la incineración ( tanto
```

2.2.2 El número en las siglas

El número sirve para: 1) actualizar un determinado morfema lexemático como nombre sustantivo o adjetivo; 2) marcar la concordancia (junto con el género y el artículo), y 3) aportar, en la mayor parte de casos, una información de aumento sobre el contenido del lexema; es decir que se da la indicación de que hay más de un objeto. Este morfema afecta a todos los componentes del sintagma nominal (sustantivos, adjetivos, pronombres y determinantes) y al verbo (Cascón, 2004: 219).

Las siglas en español normalmente no tienen plural. Martínez de Sousa (1984: 34) afirma que "el morfema de plural en español, *-s* o *–es*, no tiene aplicación a las siglas en nuestro idioma, en algunos casos porque el propio enunciado no lo admite; por ejemplo, no se podría decir los COIS porque solo hay un Comité Olímpico Internacional; pero, por otro lado, porque, con esa grafía, la S podría interpretarse como parte de la sigla (como, por ejemplo, CIA y CIAS, dos siglas distintas). Para resolver este problema, los anglosajones han recurrido a soluciones que en español resultan peregrinas, como COEs, COE's, COE-s, etc. En consecuencia, el número de las siglas, si no queda explícito en el contexto, se indica mediante el artículo: el, la, singular; los, las, plural: los CON (Comités Olímpicos Nacionales)".

No obstante, en español se registran formaciones de plural en siglas mediante el uso de los morfemas *-ϕ* y *–(e)s* (*cf.* Rodríguez, 1981, 1983, 1993; Martínez de Sousa, 1984; Casado Velarde, 1999; Salvanyà, 2005; Ortega, 2005). En este sentido, Rodríguez ofrece una amplia lista de ejemplos. Para el caso del morfema *-ϕ* muestra ejemplos como *los LP y los ovni* mientras que para el morfema *–(e)s* presenta *los LPs, los ovnis; las Otanes*. Además, llama la atención sobre la existencia de otros mecanismos para la expresión del plural en la lengua escrita tales como la reduplicación (*e.g.*: PP.CC.) y del fenómeno de la "vacilación tipográfica" que

produce el morfema –*s*, por ejemplo: para LP se ha documentado: *los LP, los LPs, los LP's, los LPS* y *los elepés*.[40]

No obstante, Rodríguez (1983: 139) destaca igualmente que el caso más frecuente de plural es el de siglas cuya base es singular en su origen: *las* UVI, *los* LP, etc. Por los ejemplos aquí citados, se comprende que el morfema ϕ es un rasgo característico de la sigla. Todos estos plurales no flexionados comportan una discordancia formal de número. Para este autor, el morfema ϕ parece ser la forma más natural en las lenguas románicas como el francés y el español, al menos en las primeras etapas de la lexicalización de las siglas" (Rodríguez, 1981: 16).

Cascón (2004: 230) afirma que las siglas no suelen tener plural porque designan organismos, instituciones, etc., que normalmente son entes únicos. Sin embargo, hay casos en los que sí se da esa posibilidad, porque el nombre ya está en plural. Por tanto, éste puede originarse de varias maneras:

1) Por repetición de las iniciales: FF AA (Fuerzas Armadas); CC OO (Comisiones Obreras);

2) Por pluralización mediante el determinante: los MIR (médicos internos residentes); los PAU (programas de actuación urbanística).

40 En medios de comunicación como la radio se encuentran políticas como la aplicada por Ràdio Flaixbac y COMRàdio: *"...Trobem que és necessari fer la distinció, en primer lloc per desambiguar certs casos: si diem que en una festa 'hi haurà DJ, gogós i molt bon ambient', sembla que diguem que hi haurà un sol DJ, i en realitat n'hi haurà uns quants; per tant, parlem dels 'DJs'. I en segon lloc, per fer la llengua més espontània i fluïda: dir que 's'ha detingut una banda que venien CD pirates' o que 'hi ha hagut una trobada d'ONG' sona postís i forçat, i preferim parlar de 'CDs pirates' i de 'trobades d'ONGs'. I com que ho diem, també hem optat per escriure-ho (a les webs, per exemple): afegim una essa minúscula a la sigla: els CDs, els DJs (dit dijeis, a l'anglesa), els DVDs, etc."*. (Salvanyà, 2005).
"A COMRàdio apostem tranquil·lament pels plurals 'cedés', 'devedés', i si fóssim un mitjà escrit també ho escriuríem. Hi ha un cas anàleg, que és prou conegut i normatiu, 'elapé –és', que no estranya ni sorprèn ningú. Suposo que és una qüestió d'agosarament: qui primer s'atreveixi a escriure-ho s'endurà el peix al cove. En el cas de les ONG, el fet que siguem un mitjà orals ens dóna avantatge, ja que podem fer el plural col·loquial 'oenagés' sense haver-nos de plantejar si posem ONG, ONGs, ONG's, o oenagés. Personalment crec que la sigla està lexicalitzada". (Ortega, 2005).

> Esta forma alterna, en casos de siglas muy difundidas por el uso, con la adición de *–s*:
>
> Las AMPA/las AMPAs (asociaciones de madres y padres de alumnos); los GEO/los GEOs (miembros del Grupo Especial de Operaciones); las OPA/las OPAs (ofertas públicas de adquisición), etc.

Cuando se trata de siglas lexicalizadas, que se escriben ya normalmente con minúsculas por haberse convertido en sustantivos, forman el plural en *–s*: los ovnis, los elepés. En siglas que carecen de apoyo vocálico, o que no lo han desarrollado, se encuentran pluralizaciones con el alomorfo *–s* en minúscula, precedido de consonante y, en ocasiones, separado por apóstrofo del cuerpo de la sigla: ONG's, LP's, LPs, etc. (Casado Velarde, 1999: 5.083).

En definitiva, toda sigla tiene un género: el del sustantivo núcleo del sintagma que constituye su base. Así, ONU, cuya primera letra es la *o* de *organización*, es femenino y singular (la ONU). La especificación del número se basa fundamentalmente en la estructura de la base de la forma desarrollada, pero también en la significación general de la sigla (Rodríguez, 1981: 15).

En lo que respecta a nuestro estudio, hemos detectado que 12 de las 67 siglas bajo análisis presentan la marca de plural. Los casos encontrados son los siguientes:

(1) RFLP/RFLPs (Restriction fragment length polymorphism-*s*)

<m00294>: , que se identificaban como ## RFLPs ## (polimorfismos basados en la
<m00294>: las sondas usadas detectan frecuentemente ## RFLPs ## de más de 5Kb .</s>
<m00313>: Maine), incluyen datos sobre ## RFLPs ## , asignación a cromosomas ,
<m00576>: su aplicación en diagnóstico : ## RFLPs ## y SSCPs</item> <item>Otras técnicas de
<m00653>: <head>Fragmentos de restricción polimórfica (## RFLPs ##)</head> <s>Para realizar un primer

(2) VNTR/VNTRs (Variable number of tandem repeat-*s*)

```
<m00873>:       capaces de reconocer simultáneamente varias ## VNTRs ## presentes en el genoma .</s>
<m00873>:       caracteriza a los minisatélites o ## VNTRs ## .</s> <s>El entrecruzamiento desigual y
<m00873>:       de las secuencias minisatélite o ## VNTRs ## en eucariotas implicaría también la
<m00873>:       de una proteína con secuencias ## VNTRs ## fueron Kinzler y Vegelstein en
<m00873>:       revisando las secuencias de varias ## VNTRs ## encontraron que muchas de ellas
<m00873>:       en 12 de las 21 ## VNTRs ## examinadas incluyendo algunas de ellas
<m00873>:       evidencias de nuevos alelos de ## VNTRs ## surgen por mecanismos que no
```

(3) LET/LETs (Linear energy transfer-s)

```
<m00590>:       <s>Este tipo de radiaciones tienen ## LETs ## de 3-3,5 KeV / xxx
<m00590>:       de ciertos radioisótopos , tienen ## LETs ## comprendidas entre 10 a cientos de
<m00590>:       .</s> <s>Las radiaciones de alta ## LETs ## depositan gran cantidad de energía
```

(4) ADN/ADNs (Ácido-*s* desoxirribonucleicos)

```
<m00190>:       genética , así como los ## ADNs ## de partida , existían de
<m00598>:       genes estén mapeados y sus ## ADNs ## secuenciados , dispondremos de una
<m00784>:       genes estén mapeados y sus ## ADNs ## secuenciados , dispondremos de una
<m00900>:       competitivo humano , a los ## ADNs ## sometidos a restricción con Hinf I
```

(5) DNA/DNAs (Deoxyribonucleic acid-*s*)

```
<m00702>:       de 512 personas ( 65 ## DNAs ## de paternidades y en 485
<m00702>:       de paternidades y en 485 ## DNAs ## de donantes de sangre )
<m00702>:       en 487 personas ( 65 ## DNAs ## de paternidades y en 422
<m00702>:       de paternidades y en 422 ## DNAs ## de donantes de sangre )
<m00740>:       a cuantificar .</s> <s>Los dos ## DNAs ## compiten por los « primers
```

(6) ARN/ARNs (Ácido-*s* ribonucleicos)

<m00190>: " biblioteca " de tales ## ARNs ## .</s> <s>El 22 de mayo de 1961 , Mathei
<m00190>: obtenidos a partir de los ## ARNs ## mensajeros expresados en distintos tejidos
<m00190>: ribosomas), 40 genes para ## ARNs ## nucleares pequeños (implicados en
<m00190>:) y 275 genes para ## ARNs ## de transferencia (los implicados
<m00190>: de la concentración de estos ## ARNs ## en la célula .</s> <head>Herramientas
<m00784>: otros muchos genes que transcriben ## ARNs ## no codificantes (ncARNs)
<m00784>: de los ribosomas ;</s> <s>pequeños ## ARNs ## nucleolares requeridos en el procesamiento

(7) RNA/RNAs (Ribonucleic acid-s)

<m00728>: los RNA mensajeros son necesarios ## RNAs ## de transferencia , ribosomas ,
<m00873>: de genes que codifican pequeños ## RNAs ## , como los tRNAs y
<m00873>: por reversotranscripción a partir de ## RNAs ## poliadenilados y posterior inserción en
<m00873>: de un gen , por consiguiente ## RNAs ## con homologáis con ambas cadenas
<m00873>: los genes que codifican con ## RNAs ## ribosómicos xxx , xxx y

(8) LINE/LINEs (Long interspersed element-s)

<m00368>: los elementos dispersos largos o ## LINES ## (de l inglés «
<m00590>: subtipos histona xxx</cell> <cell>Ricos en ## LINEs ## (Long Intermediate Repetitive
<m00590>: , xxx) y las ## LINEs ## (large interspersed repetitive elements , xxx)
<m00638>: los más abundantes , los ## LINEs ## , SINEs y LTRs ,
<m00638>: los retrovirus endógenos , los ## LINEs ## y los SINEs carecen de
<m00638>: no vira les .</s> <s>Los ## LINEs ## son elementos repetidos autónomos que
<m00638>: .</s> <s>Los mRNAs de los ## LINEs ## poseen dos marcos de traducción
<m00638>: , la mayoría de los ## LINEs ## son elementos truncados no funcionales

(9) SINE/SINEs (Short interspersed nuclear element-s)

<m00590>: (Long Intermediate Repetitive DNA Sequences)</cell> <cell>Ricos en ## SINEs ## (Short
<m00590>: repetitivas , tales como las ## SINEs ## (small interspersed repetitive elements , xxx)
<m00638>: abundantes , los LINEs , ## SINEs ## y LTRs , sean los
<m00638>: , los LINEs y los ## SINEs ## carecen de LTRs , de
<m00638>: su propagación genómica .</s> <s>Los ## SINEs ## son retrotransposones que carecen de
<m00638>: una variante muy exitosa de ## SINEs ## exclusiva de la especie humana

(10) YAC/YACs (Yeast artificial chromosome-*s*)

<m00190>: para las levaduras se denominan ## YACs ## , que es la abreviatura
<m00307>: (yeast artificial chromosomes: YACs)</s></item> <item>Banco de ## YACs ## : cada pocillo
<m00313>: que derivan de levaduras (## YACs ##).</s> <s>Las bacterias son los
<m00313>: levadura como cromosomas artificiales (## YACs ##).</s> <s>Hasta que se desarrollaron
<m00313>: <s>Hasta que se desarrollaron los ## YACs ## , los vectores de clonación
<m00313>: de 20-40 Kb .</s> <s>Los ## YACs ## han permitido disminuir drásticamente el
<m00313>: hay que ordenar .</s> <s>Muchos ## YACs ## pueden cubrir genes humanos completos
<m00313>: fragmentos más grandes clonados en ## YACs ## , hay que subclonar esos

(11) ITIM/ITIMs (Immunoreceptor tyrosine-based inhibitory motif-*s*)

<m00906>: en humanos .</s> <s>NKG2AB presentan ## ITIMS ## en la región citoplasmática mientras
<m00906>: de la tirosina en los ## ITIMs ## , seguido de l reclutamiento

(12) PCB/PCBs (Polychlorinated Biphenyl-*s*)

<m00573>: <head>MUTACIONES EN EL GEN K-RAS Y CONCENTRACIONES SÉRICAS DE DDT , DDE , ## PCBS ## Y OTROS COMPUES
<m00573>: <s>Se detectaron diez congéneres de bifenilos policlorados (## PCBs ##).</s>

La revisión de los 12 casos de pluralización detectados en nuestra muestra indica que 10 siglas corresponden a casos de siglas en inglés. Los dos casos restantes (ADN y ARN) corresponden a siglas formadas en español. Todos los casos de pluralización registrados presentan el carácter “s” como marca de plural. No se han detectado casos de formación de plural por repetición de las iniciales.

2.3 Las siglas como base de una nueva unidad léxica: sufijación y prefijación

2.3.1 Sufijación

Como se ha visto en los apartados anteriores, la naturaleza nominal de la sigla hace que también pueda ser sometida a procesos de derivación y flexión. Según Casado Velarde (1999: 5.083) los derivados de carácter nominal son justamente los que más abundan en las siglas. Para él "la forma sufijal más común es *–ista*: acenepista ("miembro de la ACNP"), aprista (APRA), apriísta (PRI), cederista (CDR), etc.". Menos frecuentes son las formaciones con el sufijo *–ismo*: cenetismo (CNT), priísmo (PRI), otanismo (OTAN); *-ción*: ucedificación (UCD), otanización (OTAN), psuquisación (PSUC), ucedización (UCD), etc.

Basándonos en la tabla de sufijación nominal y adjetival de la Gramática descriptiva de la lengua española, hemos buscado casos de sufijación en nuestra muestra de siglas. Como resultado sólo se ha hallado un caso de sigla (lexicalizada) con presencia de sufijación. Se trata de "sida", la cual va unida al sufijo de carácter adjetival *–oso*, dando origen a "sidoso", tal y como se puede constatar en los siguientes contextos extraídos del corpus.[41]

```
<m00300>:       de varones homosexuales pertenecían a ## sidosos ## .</s> <s>La inclusión de unos
<m00300>:       <s>La inclusión de unos cuantos ## sidosos ## heterosexuales no basta para comprobar
<m00300>:       que experimentan casi todos los ## sidosos ## en los momentos finales de
<m00680>:       las células coadyuvantes de los ## sidosos ## pueden acarrear en sus membranas
<m00680>:       de virus .</s> <s>En los ## sidosos ## , las células inflamatorias de
<m00689>:       pacientes con cáncer avanzado , ## sidosos ## y diabéticos sin tratar .</s>
<m00690>:       detienen la pandemia ;</s> <s>los ## sidosos ## podrían seguir diseminando la infección
<m00697>:       producidas por el citomegalovirus en ## sidosos ## .</s> <s>Y hay en vías
```

41 Sigla lexicalizada

2.3.2 Prefijación

Además del fenómeno de la sufijación, las siglas también pueden presentar prefijaciones que dan origen a compuestos, *e.g.*: anti-CEE, anti-OTAN, antigrapos, superetas, ex-grapo, etc. (*cf.* Casado Velarde, 1999: 5.083). Se han analizado los prefijos a partir de la tabla existente en la Gramática descriptiva de la lengua española.[42,43] En nuestro caso, se han detectado 5 siglas prefijadas, a saber:

(1) Mega-yeast artificial chromosomes

[prefijo [*mega*] + SN [YAC]] → MegaYACs

```
<m00190>:       mayores , los modernos " ## MegaYACs ## " pueden albergar fragmentos d
<m00190>:       <s>Basta con 3000 de estos ## MegaYACs ## para completar una genoteca
```

(2) Retro-Polymerase chain reaction

[prefijo [*retro*] + SN [PCR]] → retro-PCR

```
<m00581>:       moléculas de ARN previo paso a ADNc ( RT-PCR o ## retro-PCR ## ); la PCR inversa ; 
```

(3) Pre-Messenger Ribonucleic Acid

[prefijo [*pre*] + [SN] mRNA]] → pre-mRNA

```
<m00368>:       <s>El resultado final de este procesamiento de l ## pre-mRNA ## es el mRNA .</s>
<m00368>:       <s>Los RNA nucleares pequeños ( snRNA ) están implicados en el procesamiento de los ## pre-mRNA ##
```

(4) Pre-ARN mensajero

[prefijo [*pre*] + [SN] ARNm]] → pre-ARNm

42 Véase lista de sufijos y prefijos analizados en anexo 1.
43 Véase Lacuesta, R., Bustos, E. (1999: 4.505-4.594).

<m00730>: precursores de ARN conocidas como ## Pre-ARNm ## .</s> <s>Estos precursores son procesados

(5) Antivirus de la inmunodeficiencia humana

[prefijo [*anti*] + [SN] VIH]] → anti-VIH

<m00218>: redujo la producción de anticuerpos ## anti-VIH ## en los ratones e intensificó

Se observa que, dentro de las 67 siglas analizadas, sólo 1 (1,5%) refleja el fenómeno de la sufijación, mientras que 5 siglas (7,5%) muestran casos de prefijación. Tanto el caso de sufijación como los de prefijación se han encontrado en el corpus de GH. Esto lleva a pensar que las siglas de GH y MA aún están en un estadio muy temprano de lexicalización. El gráfico 31 muestra la proporción de estos fenómenos en las siglas analizadas.

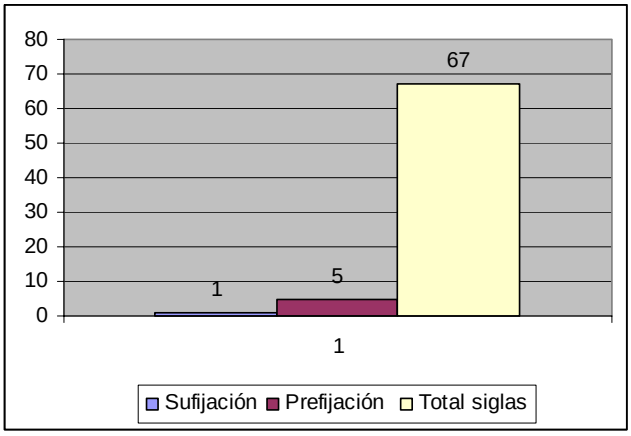


Gráfico 31. Casos de prefijación y sufijación registrados en la muestra de siglas analizada

3. Sintaxis

3.1 Relación entre los elementos de la forma desarrollada y los elementos de la sigla

Partimos de la premisa de que una sigla es el acortamiento de un conjunto de piezas léxicas más extenso. A este respecto, nos hemos preguntado si cada carácter inicial de los componentes de una forma desarrollada tiene su inicial correspondiente (letra, sílaba o una combinación de ambas) en su sigla. El análisis de la muestra de siglas ha permitido obtener una respuesta en tal sentido. Se ha encontrado que no hay uniformidad en la correspondencia entre los caracteres iniciales de la forma desarrollada y los caracteres que forman la sigla sino que, por el contrario, existen varios grados de correspondencia, que hemos denominado total, parcial y nula.

La "correspondencia total" se da cuando la letra o carácter inicial de cada uno de los elementos de la forma desarrollada está presente en la sigla. A esta categoría pertenecen las siguientes siglas:

(1) **F**ibrosis **q**uística	**FQ**
(2) **P**royecto **G**enoma **H**umano	**PGH**
(3) **N**uevos **e**stados **i**ndependientes	**NEI**
(4) **C**ódigo **p**enal	**CP**
(5) **O**xígeno **d**isuelto	**OD**
(6) **U**nión **E**uropea	**UE**
(7) **C**omisión **E**uropea	**CE**

Dentro de este grupo incluimos un subgrupo que hemos denominado "correspondencia total e irregular". A esta clase pertenecen aquellas siglas que reflejan más de una letra o carácter inicial de cada uno de los elementos de la forma desarrollada. Dentro de esta categoría se incluyen los siguientes casos:

(1) **Med**iterráneo **occ**idental **MEDOC**
(2) **Med**iterráneo **or**iental **MEDOR**
(3) **ARN m**ensajero **ARNm**
(4) **Tr**ansformación **ag**raria **s.a.** **TRAGSA**

La "correspondencia parcial" se presenta cuando falta una letra inicial o sílaba de alguno de los elementos de la forma desarrollada. En este tipo de correspondencia también suelen elidirse elementos de la forma desarrollada tales como preposiciones y conjunciones. En concreto, la correspondencia parcial se presenta bajo dos tipos, a saber:

1) Correspondencia parcial de tipo A: se da bien cuando en la sigla falta una letra inicial o sílaba de alguno de los elementos de la forma desarrollada (*e.g.*: **D**emanda **b**ioquímica de **o**xígeno a los **5** días → **DBO5**) , o bien cuando se eliden elementos de la forma desarrollada como preposiciones, conjunciones y artículos (*e.g.*: **P**rograma ~~de~~ ~~las~~ **N**aciones **U**nidas ~~para~~ ~~el~~ **M**edio **A**mbiente → **PNUMA**);
2) Correspondencia parcial de tipo B: se da cuando la sigla incluye tanto caracteres iniciales como internos de los elementos de la forma desarrollada (*e.g.*: **Á**cido **D**esoxirribo**n**ucleico → **ADN**).[44]

Se han hallado los siguientes casos de correspondencia parcial de tipo A:

(1) **D**emanda **b**ioquímica ~~de~~ **o**xígeno → **DBO**
(Correspondencia de iniciales, elisión de preposición)
(2) **O**rganización ~~de~~ **C**ooperación ~~y~~ **D**esarrollo **E**conómicos → **OCDE**
(Correspondencia de iniciales, elisión de preposición y conjunción)
(3) **U**nidad ~~de~~ **c**arbón **p**iedra → **UCP**
(Correspondencia de iniciales, elisión de preposición)
(4) **A**nálisis ~~en~~ **c**omponentes **p**rincipales → **ACP**
(Correspondencia de iniciales, elisión de preposición)

44 Para aquellas siglas provenientes del inglés se ha realizado un análisis doble; *i.e.*, la correspondencia de sus elementos con la forma desarrollada en inglés y español. Sin embargo, a la hora de cuantificar los datos sólo hemos tenido en cuenta los datos de correspondencia FD-sigla en español. Esto por dos razones: a) por ser esta la lengua en que se basa nuestro corpus, y b) para buscar pistas que nos ayuden a encontrar patrones de detección de siglas.

(5) **P**rograma ~~de~~ **N**aciones **U**nidas ~~para~~ ~~el~~ **M**edio **A**mbiente → **PNUMA**

(Correspondencia de iniciales, elisión de preposiciones y artículo)

(6) **I**nstituto ~~para~~ ~~la~~ **Co**nservación ~~de~~ ~~la~~ **Na**turaleza → **ICONA**

(Correspondencia con una inicial y dos sílabas, elisión de preposiciones y artículos)

(7) Fuerzas ~~y~~ **C**uerpos ~~de~~ **S**eguridad ~~del~~ **E**stado (Ley de) → **FCSE**

(Correspondencia de iniciales, elisión de conjunción y preposición + artículo (del))

(8) **J**anus quinasa **2** → **JAK2**

(Correspondencia parcial de iniciales. En inglés hay correspondencia de iniciales: **J**anus **A**ctivating **K**inase **2**)

(9) **D**emanda **b**ioquímica ~~de~~ **o**xígeno ~~a~~ ~~los~~ **5** días → **DBO5**

(Correspondencia parcial de iniciales, elisión de preposición y artículo)

(10) **F**ondo **A**sturiano ~~para~~ ~~la~~ **P**rotección ~~de~~ ~~las~~ ~~Especies~~ **S**alvajes → **FAPAS**

(Correspondencia parcial de iniciales, elisión de preposiciones y artículos)

(11) **V**ariación ~~en~~ ~~el~~ **n**úmero ~~de~~ repeticiones ~~en~~ tándem → **VNTR**

(Hay correspondencia parcial de iniciales, además se eliden preposiciones y artículos. En inglés hay correspondencia parcial de iniciales: **V**ariable **n**umber ~~of~~ **t**andem **r**epeats)

(12) **V**irus ~~de~~ ~~la~~ **i**nmunodeficiencia **h**umana → **VIH**

(Correspondencia parcial de iniciales, se elide la preposición y el artículo)

(13) **A**denosina **t**rifosfato → **ATP**

(Correspondencia parcial de iniciales. En inglés hay correspondencia total: **A**denosine **T**ri-**P**hosphate)

Respecto de la correspondencia parcial de tipo B, se han hallado los siguientes casos:

(1) **A**berraciones **c**romosómica**s** → **ACs**

(Correspondencia de iniciales más la inclusión de un carácter no inicial como lo es la marca de plural "s")

(2) **A**denosina-**d**es**a**minasa → **ADA**

(Correspondencia de inciales e inclusión del carácter interno "a")

(3) **Á**cido **d**esoxirribo**n**ucleico → **ADN** (en corpus de GH y MA)*****

(Correspondencia de inciales e inclusión del carácter interno "n")

(4) **Á**cido **r**ibo**n**ucleico → **ARN**

(Correspondencia de inciales e inclusión del carácter interno "n")

(5) **S**índrome ~~de~~ **i**nmuno**d**eficiencia **a**dquirida → **Sida**[45] (Correspondencia parcial de iniciales, elisión de preposición)

45 Sigla lexicalizada

(6) **M**inisterio ~~de~~ **I**ndustria ~~y~~ E**ner**gía → **MINER** (Correspondencia parcial de las iniciales, elisión de preposición y conjunción)

(7) **Mega**cromosomas artificiales de levadura → **MegaYACs** (Correspondencia parcial: coincide el prefijo. En inglés hay correspondencia total y presenta marca de plural: **Mega y**east **a**rtificial **c**hromosome**s**)

(8) **A**sociación **e**cologista ~~de~~ **de**fensa ~~de~~ ~~la~~ **nat**uraleza → **AEDENAT** (Correspondencia parcial de iniciales y sílabas, elisión de preposición y artículo)

(9) **D**ietil-p-fenilen**d**iamina → **DPD** (Correspondencia parcial de iniciales en español e inglés: **d**iethyl-p-**p**henylene **d**iamine).

La "correspondencia nula" indica que ninguna de las letras o sílabas iniciales de la forma desarrollada está alineada con los constituyentes de la sigla. Se ha detectado que, normalmente, este fenómeno se produce porque en los textos se traduce la forma desarrollada pero no la sigla. Los casos que pertenecen a esta categoría son:

(1) Elementos dispersos largos → **LINE**
(No hay correspondencia ni de iniciales ni de sílabas. En inglés hay correspondencia de iniciales: **L**ong **i**nterspersed **n**uclear **e**lements)

(2) Motivos de inhibición en inmunorreceptores basados en tirosina → **ITIM**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de iniciales: **I**mmunoreceptor **t**yrosine-based **i**nhibition **m**otifs)

(3) Public/Publisher MEDLINE → **PubMed**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia de sílabas)

(4) GenBank Gene Products → **Genpept**
(No hay correspondencia de iniciales ni de sílabas). En inglés hay correspondencia de la primera sílaba)

(5) Instituto de tecnología de California → **CalTech**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de sílabas: **Cal**ifornia ~~Institute~~ ~~of~~ **Tech**nology)

(6) Convención sobre la polución marina → **MARPOL**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de sílabas: **Mar**ine **Pol**lution ~~convention~~)

(7) Agencia espacial estadounidense → **NASA**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de iniciales: **N**ational **A**eronautics ~~and~~ **S**pace **A**dministration)

(8) Organización de las Naciones Unidas para la Educación, la Cultura y la Ciencia → **UNESCO** (No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia de iniciales y elisión de conjunción: **U**nited **N**ations **E**ducational, **S**cientific ~~and~~ **C**ultural **O**rganization)

(9) Organización Genoma Humano → **HUGO**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial en iniciales y en la primera sílaba de la primera palabra: **Hu**man **G**enome **O**rganization)

(10) Phragment Assembly Program → **PHRAP**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de iniciales: **Ph**ragment **A**ssembly **P**rogram)

(11) Pig Gene Mapping Project → **PiGMap**
(No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial: **Pi**g **g**ene **ma**pping **p**roject)

(12) Secuencias repetitivas cortas y dispersas → **SINE**
(No hay correspondencia de iniciales ni de sílabas. En inglés sí hay correspondencia de iniciales: (**S**hort **i**nterspersed **n**uclear **e**lements)

(13) Polimorfismo de longitud de los fragmentos por restricción → **RFLP**
(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales: **R**estriction **f**ragment **l**ength **p**olimorphisms)

(14) Ácido etilendiaminotetracético → **EDTA**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular: **E**thylene**d**iamine **T**etraacetic **A**cid)

(15) Célula T normal regulada tras la activación → **RANTES**
(No hay correspondencia de iniciales. En inglés la correspondencia es parcial, además se eliden preposiciones y conjunciones: **R**egulated ~~on~~ **a**ctivation, **n**ormal **T** ~~cells~~ **e**xpressed ~~and~~ **s**ecreted)

(16) Materia orgánica gruesa → **CPOM**
(No hay correspondencia de iniciales. En inglés la correspondencia es parcial: **C**oarse ~~and~~ ~~fine~~ **p**articulate **o**rganic **m**atter)

(17) Clorofluorocarbono hidrogenado → **HCFC**
(Correspondencia parcial de iniciales. En inglés hay correspondencia total e irregular aunque todas las iniciales están condensadas en una sola palabra, lo cual le da la forma de una abreviatura: **H**ydro**c**hloro**f**luoro**c**arbon)

(18) Umbral de concentración admisible → **TLm**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **T**hreshold **lim**it)

(19) p-nitrofenil fosfato → **PNPP**
(Correspondencia parcial de iniciales. En inglés hay correspondencia total e irregular: **p**-**n**itro**p**henyl **p**hosphate)

(20) Complejo mayor de histocompatibilidad → **MHC**
(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales: (**M**ajor **h**istocompatibility **c**omplex)

(21) Reacción en cadena de la polimerasa → **PCR**
(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales: **P**olymerase **c**hain **r**eaction)

(22) Células madre embrionarias → **ES**
(No hay correspondencia de iniciales. En inglés sí hay correspondencia de iniciales: **E**mbryonic **s**tem)

(23) Transferencia lineal de energía → **LET**

(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales: **L**inear **e**nergy **t**ransfer)

(24) Complejo antigénico leucocitario humano → **HLA**
(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales: **H**uman **l**eukocyte **a**ntigen)

(25) Cromosomas artificiales de levadura → **YACS**
(No hay correspondencia de iniciales. En inglés hay correspondencia de iniciales, además lleva la marca de plural: **Y**east **a**rtificial **c**hromosome**s**)

(26) Regulador de transmembrana de la fibrosis quística → **CFTR**
(No hay correspondencia de iniciales. En inglés hay correspondencia parcial: **C**ystic **f**ibrosis **t**ransmembrane ~~conductance~~ **r**egulator)

(27) Primer gen de susceptibilidad al cáncer de mama → **BRCA1**
(No hay correspondencia de iniciales. En inglés hay correspondencia parcial: **Br**east **ca**ncer ~~susceptibility~~ ~~gene~~ **1**)

(28) Cloruro de polivinilo → **PVC**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **P**oly**v**inyl **C**hloride)

(29) Bifenilo policlorado → **PCB**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **P**oly**c**hlorinated **B**iphenyls)

(30) Ácido etilendiamino tetraacético → **EDTA**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **E**thylene**d**iamine **T**etraacetic **A**cid)

(31) Ácido desoxirribonucleico → **DNA**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **D**eoxyribo**n**ucleic **A**cid)

(32) Ácido ribonucleico → **RNA**
(No hay correspondencia de iniciales. En inglés hay correspondencia total e irregular de iniciales: **R**ibo**n**ucleic **A**cid)

(33) Segundo gen de susceptibilidad al cáncer de mama → **BRCA2**
(No hay correspondencia de iniciales. En inglés hay correspondencia parcial: **Br**east **ca**ncer ~~susceptibility~~ ~~gene~~ **2**).

Del análisis anterior se desprende que 7 siglas se corresponden totalmente con los caracteres iniciales de su forma desarrollada, 4 se corresponden totalmente pero con alguna irregularidad en el orden; 23 se corresponden parcialmente y 33 presentan una correspondencia nula entre sus constituyentes y los caracteres iniciales de la forma desarrollada. En el gráfico 32 se aprecian los porcentajes de cada tipo de correspondencia.

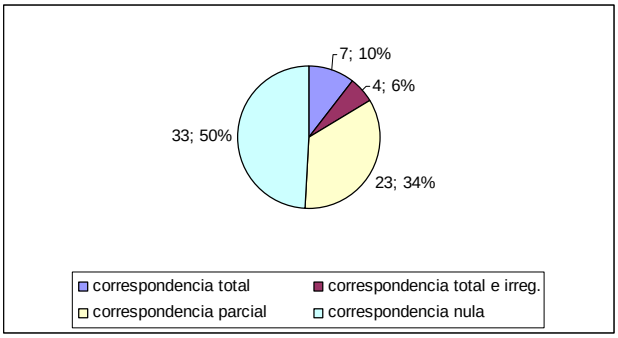


Gráfico 32. Valores para cada tipo de correspondencia dentro de la muestra

Como se ha expuesto anteriormente, los casos en los que no existe correspondencia total entre la forma desarrollada y la sigla obedecen a fenómenos como la elisión de palabras gramaticales (artículos, preposiciones y conjunciones); o al préstamo de la sigla; es decir, su importación desde una lengua extranjera como el inglés y su conservación y uso en español. Esta clasificación de los grados de correspondencia permite ver reflejada la tipología de siglas. Así, las siglas propias se ven reflejadas en la correspondencia total, mientras que las siglas mixtas se ven reflejadas en las correspondencias total e irregular, parcial y nula.

Coincidimos con Casado Velarde en que las palabras gramaticales no suelen ser un elemento relevante en la formación de las siglas. En efecto, este autor sostiene que éstas "no suelen trascender a la sigla, a no ser que se justifique su presencia por motivos de pronunciabilidad o de búsqueda de homonimia: ACUDE (Asociación de Consumidores y Usuarios de España), COPEL (Coordinadora de presos en lucha), PYME (Pequeña y Mediana Empresa), etc. En ocasiones, sin embargo, las preposiciones o conjunciones trascienden a la sigla, pero con letras minúsculas: PSdG, CiU (sigla esta formada en catalán)" Casado Velarde (1999: 5.081).

Finalmente, consideramos que el hecho de que no exista uniformidad en la correspondencia de las iniciales de la forma desarrollada con la sigla tiene implicaciones importantes a la hora de establecer los patrones para la detección de siglas en corpus. Así, por ejemplo, el grado de correspondencia total es el principio bajo el cual opera el sistema de Taghva & Gilbreth. Su sistema, denominado

Acronym Finder Program, se basa en el algoritmo *Longest common subsequence* (LCS), el cual busca justamente la alineación de cada carácter de la sigla con el primer carácter de cada palabra de la forma desarrollada. Ahora bien, cuando no se da una correspondencia total (del tipo **A**sociación **e**cologista de **de**fensa de la **nat**uraleza (AEDENAT)), el sistema es incapaz de reconocer como válido un candidato a sigla. Quizás la inclusión de otros tipos de correspondencia permitirían robustecer un sistema como este. De esta manera, por ejemplo, la correspondencia parcial permitiría dar como válido el candidato a sigla AEDENAT, lo cual aumentaría el rendimiento del sistema desde el punto de vista de la precisión y la exhaustividad. Este tema se trata con mayor detalle en el capítulo 8.

3.2 Combinatoria de las siglas

Las siglas pueden combinarse con otras categorías gramaticales. Para Rodríguez (1987: 143) "el escaso número de siglas no nominales se ve relativamente incrementado por medio del 'cambio funcional'; es decir, siglas que por su base son sustantivas 'cambian' de categoría adoptando funciones adjetivas, verbales, etc. Un cambio muy frecuente en las siglas es el de nombre a adjetivo en función atributiva; *e.g.*: nombres UCD, conductores EMT, etc. Construcciones similares se repiten en otras lenguas*:* fr. *émmissions tv*, it. *televisiones USA*".

Un ejemplo de este fenómeno en siglas en textos técnicos puede apreciarse en el siguiente fragmento analizado por Cardero (2002):

"**RS232** es un proceso de utilería que lee telemetría almacenada trama por trama y la manda a un **puerto RS232** en un formato que puede ser leído por una unidad seleccionable de despliegue".

Aquí la sigla RS232 cumple dos funciones diferentes. Por un lado, funciona como nombre (núcleo del sujeto) que, por contigüidad espacial incluye en su significado a

proceso. Por otro lado, RS232 también funciona como adjetivo. Esta autora coincide con lo dicho antes por Rodríguez al afirmar que "en relación con el comportamiento sintáctico de estas formas, éstas funcionan indistintamente como sustantivos o adjetivos. Pueden aparecer las siglas como modificadores directos del núcleo sustantivo como en: enlace RF. A modo de modificadores del núcleo o de los complementos del nombre: frecuencia de rango por T&C. Y como modificadores directos del núcleo sustantivo en cualquiera de los complementos del nombre: proceso de utilería QLOAD" (Cardero, 2002: 6). A pesar de esto, Cardero establece la categorización de acuerdo al tipo de objetos que designan las siglas. Así, por ejemplo, "las que designan asociaciones profesionales funcionan siempre como sustantivos. Las que designan marcas de equipo sufren variación de categoría gramatical como sustantivos o adjetivos; ejemplo: la BL, o una cámara BL. De la misma manera el significado se transfiere por continuidad sintagmática. Las mismas variaciones de cambio de categoría de significado sufren las que designan clasificación de materiales. Una característica de estas iniciales es que intervienen los números como especificadores de los significados. Las que designan procesos o características de los materiales presentan esquemas muy estables tanto en relación con el significado como con relación a la categoría gramatical".

De las 67 siglas analizadas se han registrado 22 en función adjetiva (19 en GH y 3 en MA). En orden decreciente las palabras que combinan con mayor frecuencia con las siglas son: secuencia (7 ocurrencias); gen (5 ocurrencias); locus/loci (6 ocurrencias); virus y polimorfismo (4 ocurrencias); marcador y genoma (3 ocurrencias); y célula, complejo, familia, región, sistema y tecnología (2 ocurrencias). En síntesis, la muestra de siglas analizada refleja que las siglas con categoría nominal tienen las mismas propiedades sintácticas que los nombres. La tabla 19, que aparece a continuación, detalla cada caso.

	GH		**MA**	
	Sigla	**N+sigla**	**Sigla**	**N+sigla**
1	PCR	Estudios PCR Tecnología PCR	CE	---
2	PGH	---	DBO	Prueba DBO Valor DBO

3	RFLP	Marcador RFLP Patrones RFLP	UE	---
4	VIH	Anticuerpos VIH Estirpe VIH Secuencia VIH Virus VIH	OCDE	---
5	HLA	Alelo HLA Antígenos HLA Complejo HLA Donante HLA Embriones HLA Examen HLA Genes HLA Haplotipos HLA Loci HLA Locus HLA Marcadores HLA Molécula HLA Región HLA Sistema HLA Tipaje HLA	OD	---
6	MHC	Complejo MHC Componente MHC Genes MHC Locus MHC Moléculas MHC Región MHC	CP	---
7	VNTR	Loci VNTR Marcadores VNTR Polimorfismos VNTR Secuencias VNTR Sistemas VNTR Sonda VNTR	UCP	---
8	ES	Célula ES	DBO5	---
9	FQ	Célula FQ Cromosomas FQ Fenotipo FQ Gen FQ Individuos FQ Locus FQ Mutación FQ	NEI	---
10	LET	---	ACP	---
11	ADN	Ciudad ADN Copia ADN Genoma ADN Polimorfismo ADN Secuencia ADN Transposones ADN Vector ADN	EDTA	Complejo EDTA
12	DNA	Fracción DNA Polimorfismo DNA Tecnología DNA Virus DNA	PVC	---

13	ARN	Enzima ARN Genoma ARN Híbrido ARN Molde ARN Secuencias ARN Virus ARN	DPD	Análisis DPD Método DPD
14	RNA	Genoma RNA Secuencia RNA Virus RNA	PNPP	---
15	ARNm	---	HCFC	---
16	ACs	---	CPOM	---
17	BRCA1	Gen BCRA1 Proteínas BCRA1	PCB	---
18	CFTR	Estructura CFTR Gen CFTR Locus CFTR Proteína CFTR	ADN	---
19	ATP	Enzima ATP Nucleótido ATP	TLm	---
20	BRCA2	---	FCSE	---
21	ADA	Actividad ADA Déficit ADA Paciente ADA Terapia ADA	MEDOC	---
22	LINE/LINEs	Familia LINE Secuencia LINE	MEDOR	---
23	RANTES	---	ICONA	---
24	EDTA	---	MARPOL	---
25	SINE/SINEs	Familia LINE Secuencias SINE Tipo LINE	MINER	---
26	HUGO	---	AEDENAT	---
27	YACS	---	PNUMA	---
28	ITIM	Motivos ITIM	NASA	---
29	UNESCO	---	TRAGSA	---
30	JAK2	---	FAPAS	---
31	MegaYACs	---		
32	CalTech	---		
33	PiGMap	---		
34	PHRAP	---		
35	Genpept	---		
36	PubMed	---		
37	Sida	---		

Tabla 19. Siglas en función adjetiva

3.3 Siglas como sujeto u objeto de verbo

El análisis de los contextos donde aparecen las siglas analizadas evidencia, como se acaba de ver con los adjetivos, que éstas pueden desempeñar las mismas funciones sintácticas de los nombres.

En lo que se refiere a la relación entre siglas y verbos, hemos encontrado 25 siglas en función de sujeto de verbo en GH y 15 en MA, las cuales aparecen en la siguiente tabla.

GH	Sigla sujeto de verbo	Frecuencia
1	PCR	
	consistir	1
	poder	1
	tener	1
	permitir	3
	haber	4
	ser	5
2	PGH	
	contribuir	1
	incluir	1
	iniciar	1
	anunciar	1
	poder	1
	requerir	1
	salir	1
	constituir	2
	tener	2
	haber	5
	ser	8
3	RFLP	
	constituir	1
	haber	1
	poder	1
	proporcionar	1
	ser	3
4	VIH	
	entrar	1
	necesitar	1
	pertenecer	1
	seguir	1
	sintetizar	1
	utilizar	1
	Poder	2
	ser	2

5	HLA	
	contener	1
6	VNTR	
	ser	2
7	ES	
	diferir	1
	tener	1
8	LET	
	definir	1
	ser	1
9	ADN	
	Constar	1
	constituir	1
	deber	1
	dominar	1
	entrar	1
	envolver	1
	extraer	1
	haber	1
	llevar	1
	ofrecer	1
	provenir	1
	recobrar	1
	regular	1
	resultar	1
	sufrir	1
	tener	1
	estar	2
	poder	2
	contener	3
	formar	3
	ser	10
10	DNA	
	aportar	1
	codificar	1
	contener	1
	deber	1
	existir	1
	fabricar	1
	formar	1
	ofrecer	1
	poseer	1
	soler	1
	poder	2
	ser	2
	tener	2
	estar	4
11	ARN	
	deber	1
	estar	1
	inducir	1
	realizar	1

	tener	1
	ser	5
12	RNA	
	actuar	1
	codificar	1
	constar	1
	deber	1
	diferir	1
	estar	1
	formar	1
	penetrar	1
	poder	1
	portar	1
	radicar	1
	ser	1
	tener	1
	poder	2
13	ARNm	
	tener	2
	ser	3
	Acs	
	afectar	1
	ser	1
	poder	3
14	BRCA1	
	contener	1
	CFTR	
	formar	1
	poder	1
	ser	2
15	ATP	
	ser	1
	poder	1
16	BRCA2	
	conferir	1
	contener	1
	ser	1
	estar	2
	tener	3
17	ADA	
	catalizar	1
	extender 1	1
18	LINE	
	ser	2
19	RANTES	
	inducir	1
	ser	1
20	EDTA	
	actuar	1
	ser	1
21	SINE	
	carecer	1

	ser	1
22	HUGO	
	ser	1
23	YACs	
	haber	1
	poder	1
	presentar	1
24	UNESCO	
	publicar	1
25	SIDA	
	haber	1
	ser	1
MA	**Sigla sujeto de verbo**	**Frecuencia**
1	CE	
	haber	2
	tender	1
2	UE	
	Adoptar	1
	regular	1
3	NEI	
	afrontar	1
4	EDTA	
	haber	1
	poder	1
	ser	1
5	PVC	
	ser	1
6	DPD	
	producir	1
	ser	1
7	CPOM	
	presentar	1
8	ADN	
	contener	1
	estar	1
9	MEDOC	
	constituir	1
	ser	1
	tener	1
10	MEDOR	
	tener	1
11	ICONA	
	contar	1
	haber	1
	invertir	1
	presentar	1
12	MINER	
	ser	1
13	AEDENAT	
	controlar	1
14	PNUMA	
	considerar	1

	Formar	1
	hacer	1
	poder	1
	haber	2
	deber	3
15	NASA	
	contratar	1

Tabla 20. Casos de siglas en función sujeto de verbo

Lorente (2007: 373) estudia los verbos desde la óptica del discurso especializado y establece una clasificación de los mismos, que retomamos aquí con el fin de identificar la clase de verbos más frecuente con la que ocurren las siglas en función de sujeto. Esta autora clasifica los verbos en cuatro categorías, a saber:

1) Verbos cuasi-términos (*i.e.*, verbos que son unidades de conocimiento especializado);
2) Verbos fraseológicos (*i.e.*, verbos que forman parte de unidades de conocimiento especializado sintagmáticas fijadas o colocaciones);
3) Verbos de relación lógica (*i.e.*, verbos que forman parte de unidades de conocimiento especializado de carácter oracional), y
4) Verbos performativos del discurso (*i.e.*, verbos que no forman parte de unidades de conocimiento especializado).

Las siglas antes mencionadas aparecen como sujetos de diferentes tipos de verbos. A continuación citamos los más frecuentes:

1) Cuasi-términos (catalizar, contratar e invertir);
2) Fraseológicos (actuar, carecer, constituir, controlar, extender, formar, producir y regular);
3) De relación lógica (contener, diferir, estar, haber, inducir, poder, presentar, ser, tender y tener), y

4) Performativos del discurso (adoptar, afrontar, contar, deber, definir, permitir y publicar).[46]

Destaca el hecho de que la mayoría de las siglas aparecen como sujetos del verbo de relación lógica "ser"; concretamente, en 18 siglas de GH y 5 de MA.[47]

Los casos contenidos en la tabla siguiente confirman igualmente que las siglas ocurren en función objeto; por ejemplo:

1) Verbos fraseológicos (agregar, construir, contener, desarrollar, emplear, fabricar, formar, introducir, producir, secuenciar, sintetizar, tener y transcribir);
2) Verbos de relación lógica (generar, inducir y ser) y,
3) Verbos performativos del discurso (denominar, detectar, encontrar, expresar, realizar, seguir y utilizar).

Las siglas analizadas indican que mayormente son objeto de los verbos "contener" y "denominar", los cuales aparecen en 4 y 3 siglas, respectivamente. Así mismo, se ha encontrado que, dentro de la muestra de siglas analizada no hay ningún verbo cuasi-término.

El listado de siglas en posición objeto es el siguiente:

GH	Siglas objeto de verbo [V+sigla]	Frecuencia
1	PCR	
	realizar	1
	utilizar	1
2	RFLP	
	denominar	1

46 Lorente (2007) ofrece una explicación detallada de esta clasificación.
47 Las ocurrencias de siglas como sujetos y objetos de verbo se recogen en el anexo 2.

	emplear	1
	encontrar	1
	utilizar	1
3	VNTR	
	amplificar	1
	denominar	2
	detectar	11
4	ADN	
	absorber	1
	aislar	1
	analizar	1
	centrifugar	1
	combinar	1
	comparar	1
	compartir	1
	comprobar	1
	cortar	1
	denominar	1
	empaquetar	1
	emplear	1
	exponer	1
	haber	1
	hidrolizar	1
	intercambiar	1
	marcar	1
	mezclar	1
	multiplicar	1
	obtener	1
	polimerizar	1
	poseer	1
	recibir	1
	recuperar	1
	sintetizar	1
	usar	1
	añadir	2
	clonar	2
	degradar	2
	inyectar	2
	introducir	3
	contener	5
	ser	8
5	DNA	
	centrifugar	1
	encontrar	1
	identificar	1
	incorporar	1
	inyectar	1
	mantener	1
	producir	1
	requerir	1
	sintetizar	1
	utilizar	1

	añadir	2
	cortar	2
	denominar	2
	digerir	2
	haber	2
	obtener	2
	tener	2
	construir	3
	contener	3
	emplear	3
	unir	3
	usar	3
	secuenciar	5
6	ARN	
	analizar	1
	codificar	1
	copiar	1
	degradar	1
	generar	1
	multiplicar	1
	preferir	1
	proporcionar	1
	transcribir	1
	usar	1
	contener	2
	producir	2
	ser	2
	fabricar	4
	sintetizar	7
7	RNA	
	albergar	1
	formar	1
	identificar	1
	obtener	1
	presentar	1
	sintetizar	1
	utilizar	2
	contener	3
	transcribir	3
	producir	4
8	ARNm	
	buscar	1
	detectar	1
	fabricar	1
	leer	1
	utilizar	1
	tener	2
9	ACs	
	demostrar	1
	formar	1
	generar	2
	inducir	7

10	CFTR	
	denominar	2
	expresar	2
11	ATP	
	consumir	1
	convertir	1
	fabricar	1
	generar	1
	hidrolizar	1
	ligar	1
	regenerar	1
	requerir	1
	sintetizar	1
	formar	2
	producir	2
	generar	5
12	EDTA	
	contener	1
	agregar	2
13	YACS	
	denominar	1
14	SIDA	
	desarrollar	1
	tener	1
MA	**Siglas objeto de verbo**	**Frecuencia**
1	DBO	
	seguir	1
2	NASA	
	ser	1

Tabla 21. Casos de siglas en función objeto de verbo

4. Semántica

La gran mayoría de los ámbitos de especialidad no tienen estandarizados sus procedimientos de denominación, ni siquiera en el caso de conceptos fundamentales (Nenadić, 2002). Tradicionalmente, la teoría clásica de la terminología ha abogado por la monoreferencialidad, esto es, la correspondencia unívoca entre un término y un concepto. Sin embargo, en la práctica, el discurso refleja problemas como la polisemia, la sinonimia y la homonimia. En concreto, las siglas suelen reflejar casos de sinonimia y homonimia como puede apreciarse a continuación.

4.1 Sinonimia

Según Rodríguez (1980: 334) "en la relación de sinonimia entre siglas, a la inversa que en la homonimia, dos lexemas siglares coinciden por alusión a una misma denominación (ONU/UNO; USA/EE.UU.; NATO/OTAN), si bien en alguna ocasión el sistema subyacente a la siglación puede ser distinto (MCE 'Mercado Común Europeo'/ CEE 'Comunidad Económica Europea'). En cuanto a los primeros, el tipo más importante, la sinonimia se basa en el doble sistema empleado en la elección de la morfología de la sigla, bien por préstamo (ONU, USA, NATO), bien por traducción al español siguiendo un orden sintáctico autóctono cual es la característica de la secuencia progresiva del compuesto (ONU, EE.UU., OTAN). En la lengua escrita el caso más frecuente es la alternancia de las siglas USA y EE.UU.".

En la muestra de siglas escogida se han detectado los siguientes casos de sinonimia:

GH	Sigla prestada	Sigla traducida
(1)	PCR (Polymerase chain reaction)	RCP (Reacción en cadena de la polimerasa)
(2)	RFLP (Restriction fragment length polimorphisms)	FRLP (Fragmento de restricción de longitud polimórfica)
(3)	HIV (Human inmunodeficiency virus)	VIH (Virus de la inmunodeficiencia humana)
(4)	MHC (Major histocompatibility complex)	CMH (Complejo mayor de histocompatibilidad)
(5)	EM (Embryonic stem)	ME (Células madre embrionarias)
(6)	DNA (Deoxyribonucleic acid)	ADN (Ácido desoxirribonucleico)
(7)	RNA (Ribonucleic acid)	ARN (Ácido ribonucleico)
(8)	mRNA (Messenger ribonucleic acid)	ARNm (ARN mensajero). Adicionalmente, se han registrado otras dos formas como son: RNAm y mARN
(9)	YACs (Yeast artificial chromosomes)	CAL (Cromosomas artificiales de levadura)
MA	**Sigla prestada**	**Sigla traducida**
(1)	DNA (Deoxyribonucleic Acid)	ADN (Ácido desoxirribonucleico)
(2)	RNA (Ribonucleic acid)	ARN (Ácido ribonucleico)
(3)	US/USA (United States)/(United States of America)	EEUU/EUA (Estados Unidos)/(Estados Unidos de América)
(4)	WHO (World Health Organization)	OMS (Organización Mundial de la Salud)
(5)	OPEC (Organization of the Petroleum Exporting Countries)	OPEP (Organización de Países Exportadores de Petróleo)

Tabla 22. Casos de sinonimia en la muestra analizada

Todos los casos de sinonimia anteriormente citados corresponden a siglas prestadas y traducidas del inglés al español. Adicionalmente, se han encontrado un par de casos más que corresponden a sinonimia dentro de la misma lengua, producida por cierta variación en alguno de los caracteres que componen la sigla como, por ejemplo, la alternancia de uso entre números arábigos y romanos o la alternancia de caracteres en mayúscula y minúscula. Los casos detectados son:

(1) BCRA1 / BRCA1 / BRCA I (Breast Cancer Gene 1)
(2) YACs / YACS (Yeast artificial chromosomes)

Por último, se ha observado que todos los casos de sinonimia hallados (16 en total) representan el 24% de la muestra.

4.2 Homonimia

De acuerdo con Rodríguez (1981: 327) "la sigla recién creada puede así mismo formar homónimo con otras siglas de idéntica contextura… en español peninsular MIR designa al colectivo de 'Médicos internos y residentes' mientras que el MIR de Venezuela y Chile es el 'Movimiento de la Izquierda Revolucionario' y el MIR de Argentina es el 'Movimiento de Intransigencia Radical'".

La homonimia genera casos de ambigüedad cuando no aparece la forma desarrollada de las siglas dentro del texto. De ahí que, como podrá apreciarse en el capítulo siguiente, algunos sistemas de detección y extracción de siglas hayan implementado módulos de desambiguación de siglas. Se entiende por desambiguación de siglas al mecanismo de selección de la forma desarrollada apropiada para una ocurrencia específica de una sigla en un contexto dado. Por ejemplo, si se pretende recuperar documentos relacionados con APC, con el sentido de "*Antigen presenting cell*", no deberían recuperarse aquellos documentos que contengan la sigla APC con un significado diferente como "*Adenomatous polyposis coli*". El problema de la

ambigüedad que puede llegar a producir la homonimia es tal que, según estudios llevados a cabo por Liu *et al.* (2001, 2002), el 81% de las siglas encontradas en los *abstracts* de Medline publicados en 2001 eran ambiguas y tenían 16 sentidos en promedio.

Entre las siglas de la muestra seleccionada sólo se han encontrado homónimos para el ámbito de MA. Debido a que este es un factor muy importante en lo que se refiere a la detección y desambiguación de siglas, se ha decidido ampliar la búsqueda a todo el corpus, con el fin de observar qué tan recurrente es el fenómeno de la homonimia en nuestro caso. Los resultados son los siguientes:

	GH	**Forma desarrollada**
(1)	ACS	American Cancer Society || American Chemical Society
(2)	APC/ApC ó APC	Antigen presenting cell || Adenomatous polyposis coli
(3)	CDC	Centers for Disease Control and Prevention || Cell division cycle
(4)	DM	Depresión mayor || Distrofia miotónica || Diabetes mellitus
(5)	ER	Endoplasmic reticulum || Estrogen receptor
(6)	ME	Células madre embrionarias || Membrana externa
(7)	NSF	Natural Science Foundation || N-ethylmaleimide-sensitive factor
(8)	OR	Odds ratio || Oak ridge
(9)	PCC	Premature chromosome condensation || Propionil-CoA carboxile
(10)	RE	Retículo endoplasmático || Receptor de estrógenos
(11)	RF	Replicative form || Recombinant frequency
(12)	RM	Retraso mental || Razón metabólica || Resonancia magnética
(13)	TC	Tomografía computarizada || Toxina colérica
	MA	**Forma desarrollada**
(1)	CE	Comisión Europea || Constitución española
(2)	CP	Capa profunda || Código penal
(3)	CRA	Capacidad de retención de agua || Coeficiente de radiación

Tabla 23. Casos de homonimia en la muestra analizada

5. Conclusiones

El análisis lingüístico de las siglas sirve por un lado para la descripción de las unidades, y por otro, para aportar refinamiento a los sistemas de identificación de siglas. En lo concerniente a la descripción se puede concluir que:

1) La mayoría de las siglas deletreadas son de tres caracteres mientras que la mayoría de las siglas que se pronuncian silábicamente son de cuatro caracteres o más.

2) La totalidad de las siglas tiene como núcleo a un nombre, hecho que corrobora la naturaleza nominal de estas unidades.

3) La muestra de siglas analizada indica que la mayoría de los nombres que conforman el núcleo de las siglas son comunes. Este hecho coincide con lo afirmado por autores como Nakos (1990: 410), quien sostiene que sólo un número muy restringido de siglas científico-técnicas utiliza el nombre propio.

4) Al asumir que toda sigla proviene de un sintagma nominal, cuyo núcleo es un nombre, asumimos también que las siglas tienen género y número.

5) El carácter nominal de la sigla hace que también pueda ser sometida a procesos de derivación y flexión, aunque con más limitaciones. La utilización de estos mecanismos en una sigla es un indicador del grado de lexicalización alcanzado por ella. Dentro de la muestra analizada, los valores de sufijación y prefijación son bajos 1,5% y 7,5%, respectivamente.

6) Los casos en los que no existe correspondencia total entre la forma desarrollada y la sigla se deben a fenómenos como la elisión de palabras gramaticales y al préstamo de la sigla, normalmente del inglés.

7) El hecho de que no siempre exista una correspondencia total entre las iniciales de la forma desarrollada y la sigla tiene implicaciones importantes a la hora de establecer los patrones para la detección de siglas en corpus.

8) Las siglas son de categoría nominal y tienen las mismas propiedades sintácticas que los nombres. Por tanto, pueden combinarse con otras categorías gramaticales como son los adjetivos y ser sujetos u objetos de verbo.

9) El discurso presenta fenómenos como la sinonimia y la homonimia, los cuales también se ven reflejados en las siglas. Normalmente, la sinonimia se da por el préstamo o traducción de siglas, en su mayoría del inglés.

10) El 81% de las siglas encontradas en los *abstracts* de Medline publicados en 2001 eran homónimas (16 sentidos o formas desarrolladas en promedio). La homonimia produce ambigüedad cuando no aparece la forma desarrollada de las siglas dentro del texto. De ahí que algunos sistemas de detección y extracción de siglas hayan implementado módulos de desambiguación de siglas.

En lo que respecta al refinamiento a los sistemas de identificación de siglas, consideramos que los aspectos que se pueden tener en cuenta son:

1) Desde la morfología
    a) Categoría gramatical (predominancia total de N);
    b) N (sigla) en función de Adj (en vista de la alta frecuencia de siglas en posición adjetiva cabe la posibilidad de analizar la coocurrencia de N+ADJ con un mínimo de dos letras mayúsculas iniciales como posible pista para la detección de un candidato a sigla);
    c) Número (para formar el plural de la sigla puede añadirse una “s”, caso recurrente en siglas creadas en inglés como *retinoid x receptors* (RXRs); o duplicarse los caracteres de la sigla, caso propio de lenguas como el español; *e.g.*: EEUU.).

2) Desde la sintaxis

   a) Para buscar las formas desarrolladas de las siglas;
   b) Para buscar la coincidencia sigla-forma desarrollada;
   c) Para siglas en inglés la estrategia es buscar en diccionarios en inglés puesto que no habrá coincidencia con la forma desarrollada. En español se espera que la coincidencia sigla-forma desarrollada sea total o parcial (cuando se eliden palabras gramaticales).

3) Desde la semántica

   Se presenta la necesidad de crear métodos que permitan la desambiguación de las siglas cuando éstas son homónimas y aparecen sin sus respectivas formas desarrolladas dentro de los textos. Hasta el momento, el método predominante ha sido la desambiguación manual; es decir, la llevada a cabo por humanos y que consiste en la lectura de los contextos de aparición de las siglas para establecer su verdadero significado. Algunos sistemas de detección de siglas incorporan diccionarios de siglas como estrategia para la desambiguación.

Capítulo 8

Capítulo 8

Sistemas de detección y extracción semiautomática de siglas: estado de la cuestión

Introducción

Hoy en día dominios como la informática, las telecomunicaciones, la biología molecular y la genética presentan una rápida evolución, la cual se refleja en la generación de conceptos y denominaciones nuevos.

En los textos especializados muchas denominaciones suelen acortarse mediante procesos como la siglación para ajustarse a criterios estilísticos o editoriales. Normalmente, una sigla va acompañada de su forma desarrollada (FD) la primera vez que aparece dentro de un documento. A partir de allí, suele aparecer sola; lo cual puede dificultar la comprensión a aquellos lectores que no son expertos en el tema.

Existen muchos diccionarios de siglas de carácter especializado, tanto en formato papel como electrónico. Los diccionarios en papel no pueden actualizarse automáticamente, por lo que, inevitablemente, llegan desactualizados al público. Autores como Gehénot (1990: 105) han investigado sobre la producción de este tipo de recursos a lo largo de la historia. En su estudio destaca obras como el *Tractatus de*

Siglis Veterum (1703),[48] *Abréviations de sociétés, conventionnelles et usuelles* (1926),[49] o el *Dictionnaire d'abréviations françaises et étrangères, techniques et usuelles, anciennes et nouvelles* (1951).[50]

Como se indicaba antes, el auge de la ciencia y la tecnología ha sido un factor clave para que el número de recursos sobre siglas continúe en aumento. En las últimas décadas han aparecido obras como: *Dictionnaire international d'abréviations scientifiques et techniques* (1978), *Diccionario internacional de siglas y acrónimos* (1984), *Dictionnaire des abréviations et acronymes scientifiques, techniques, médicaux, économiques et juridiques* (1992), *Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols* (1997), y *Acronyms, Initialisms and Abbreviations Dictionary*, 32ª edición (2003).

Los diccionarios electrónicos de siglas, también denominados diccionarios *on line*, surgieron a finales de los años ochenta. Funcionan como bases de datos consultables, cuyo objetivo es responder a la necesidad de conocer las formas desarrolladas de la gran cantidad de siglas que se producen en todos los campos de conocimiento. Su ventaja respecto de los diccionarios en papel radica en que permiten almacenar y actualizar rápidamente la información.

Existen diversos diccionarios de abreviaciones (principalmente siglas) disponibles en Internet entre los que sobresalen *Acronym Server* (1988), *Acronym Finder* (1996), *Wiley InterScience* (1999), *Abbreviations.com* (2001) y *Acronyma* (2004). Generalmente, esta clase de diccionarios funciona mediante una interfaz de consulta donde se pueden buscar dos tipos de información: una forma desarrollada o una palabra específica dentro de todas las formas desarrolladas existentes en el diccionario.

48 Esta obra escrita en latín, comprende 314 páginas y 49 capítulos que estudian en detalle el uso de las siglas en un tema particular: derecho, medicina, aritmética, gramática, música, numismática, etc.
49 Abreviaturas marítimas, bursátiles, comerciales en francés, inglés, alemán y español etc.
50 Este diccionario reúne 8.000 abreviaturas de artes, automoción, aviación, banca, cartografía, química, comercio, derecho, electricidad, finanzas, impuestos, industria, jurisprudencia, marina, matemáticas, mecánica, medicina, etc.

A pesar de lo anterior, este tipo de diccionarios tampoco es ajeno a la desactualización, inevitable a causa de la gran cantidad de siglas que se crean a diario.[51] La puesta al día de recursos como *Acronym Finder*, *Abbreviations.com* o *Acronyma* depende del envío de nuevas siglas por parte de los usuarios, las cuales se someten primero a un proceso de edición, que puede tardar varias semanas.[52]

Este capítulo se divide en tres apartados mayores, a saber: 1) sistemas de detección y extracción de siglas; 2) sistemas de desambiguación, y 3) criterios para el diseño de un modelo de detector de siglas en español.

1. Sistemas de detección y extracción de pares de sigla-forma desarrollada

Desde finales de la década de los noventa han surgido diversos sistemas para el tratamiento automático de siglas. La creación de estos sistemas ha llevado a los investigadores a buscar paralelamente soluciones para dos tipos de problemas: la detección y la desambiguación.

Se denomina detección y extracción al proceso mediante el cual las siglas se recopilan, manual o automáticamente, a partir de corpus textuales.[53] Se denomina desambiguación al mecanismo mediante el cual se selecciona la forma desarrollada apropiada de una sigla en un contexto dado.

Durante la última década se ha dado un desarrollo vertiginoso en lo que se refiere a los sistemas de detección y extracción de siglas. Este periodo se ha destacado por dos

51 En 2001, un estudio de Pustejovsky *et al*. mostró que cerca de 12.000 siglas se creaban mensualmente en los *abstracts* de *Medline*.
52 A 19 de octubre de 2006, *Abbreviations.com* tenía cerca de 100.000 siglas por editar e incorporar definitivamente.
53 A este proceso también se le conoce como identificación, reconocimiento o adquisición.

hechos: en primer lugar, el fomento de la investigación en este campo por parte de la biomedicina y la informática. Y, en segundo lugar, el predominio del inglés como lengua objeto de estas investigaciones.

Los sistemas de detección y extracción de siglas actuales emplean métodos basados en patrones, estadística, aprendizaje máquina, o en una combinación de éstos (híbridos).

En este apartado se describen estos métodos y los principales sistemas que los emplean. Dentro de cada sistema se analizan los patrones y la técnica de detección, así como la evaluación del rendimiento, medida en términos de precisión y exhaustividad.

La precisión consiste en medir el porcentaje de siglas correctas con respecto al número total de siglas extraídas por el sistema, mientras que la exhaustividad mide el porcentaje de siglas correctas identificadas por el sistema con respecto al número total de siglas existente en el corpus.

1.1 Métodos basados en patrones

Los métodos más tradicionales para hallar pares de sigla-forma desarrollada se basan en la coincidencia de patrones. Estos métodos difieren unos de otros en cuanto al tipo de información que codifican en sus reglas, las cuales son cruciales en el rendimiento del sistema.

Entre los sistemas que emplean métodos basados en patrones se destacan: *Acronym Finder Program*, *Three Letter Acronym*, *Acrophile*, *Acromed*, *A simple algorithm* y *Sistema para la variación terminológica*.

1.1.1 Acronym Finder Program (AFP)

Acronym Finder Program (Taghva & Gilbreth, 1999) es una herramienta que usa el algoritmo *Longest Common Subsequence* (LCS) para hallar todas las alineaciones posibles entre un candidato a sigla y su forma desarrollada. AFP se evaluó en un corpus de documentos oficiales sobre el Proyecto de disposición de basuras de *Yucca Mountain*.

El trabajo de Taghva & Gilbreth es de gran importancia por cuanto es el pionero en el estudio de los sistemas de detección y extracción de siglas; de ahí que sea referencia frecuente en todas las publicaciones del área.

a. Patrones

1) Todas las palabras mayúsculas, desde tres hasta diez caracteres, son aceptadas como candidatos a sigla;
2) Cada carácter de la sigla debe coincidir con el primer carácter de cada palabra de la forma desarrollada.

b. Técnica de extracción de siglas

El proceso de extracción de siglas de AFP consta de cuatro fases: inicialización, filtro, *parser* y aplicación del algoritmo.

1) Inicialización

   La entrada de datos para el algoritmo consta de los siguientes componentes:

   a) Palabras vacías o *stopwords*. Consiste en una lista de las palabras que generalmente se omiten en la formación de siglas; es decir, artículos, preposiciones y conjunciones.

b) Palabras rechazadas. Consiste en una lista opcional de palabras frecuentes en los documentos, y que se sabe que no son siglas; por ejemplo: *TABLE*, *FIGURE*, números romanos, etc.).
c) Base de datos de siglas. Esta información puede ser usada para ignorar la rutina de búsqueda del programa o para repetir el proceso cuando la búsqueda no arroje resultados. Se trata de un mecanismo opcional.
d) Corpus. Consiste en el texto o conjunto de textos a rastrear.

2) Filtro de datos

La entrada de datos se procesa para descartar líneas de texto en mayúsculas como pueden ser los títulos. Cuando el programa identifica un candidato a sigla consulta la lista de palabras rechazadas. Si el candidato no aparece en dicha lista, entonces el proceso continúa con la búsqueda de su forma desarrollada en el texto que lo rodea. Para ello, el programa crea una ventana de texto formada por dos subventanas llamadas ventana anterior y ventana posterior. La longitud de cada ventana (medida en palabras) se establece al multiplicar por dos el número de caracteres de la sigla.

3) Parser

El sistema prioriza diferentes tipos de palabras para que el algoritmo encuentre un número razonable de formas desarrolladas, así:

a) Palabras vacías. No pueden eliminarse del proceso de búsqueda de las formas desarrolladas. Si el algoritmo ignora por completo este tipo de palabras, muchas siglas pueden pasarse por alto. Aunque el sistema da prioridad a los elementos que no son palabras vacías, Taghva & Gilbreth reconocen que hay casos en los que estas deben tenerse en cuenta; *e.g.*: *Department* ***of*** *Energy* (D**O**E). Pero, por el contrario, hay ocasiones en que las palabras vacías deben ignorarse; *e.g.*: *Office* ***of*** *Nuclear Waste Isolation* (OWNI).

b) Palabras separadas por guiones. Las formas desarrolladas contienen a menudo palabras separadas por guiones. En este sentido, puede darse uno de los siguientes casos:

- Que la primera palabra separada por guión pertenezca a la forma desarrollada; *e.g.*: ***X**-ray photoelectron spectroscopy* (**X**PS);
- Que todas las partes de la unidad separada por guión pertenezcan a la forma desarrollada; *e.g.*: ***n**on-**h**igh-**l**evel solid waste* (**NHL**SW).

c) Siglas. Dentro de los textos, las siglas pueden estar cerca unas de otras como cuando las siglas incluyen otras siglas en sus formas desarrolladas; *e.g.*: ***A**RINC **C**ommunications **a**nd **R**eporting **S**ystem* (ACARS).

d) Palabras que no pertenecen a ninguno de los tipos anteriores. Esta clase de unidades constituye la mayor parte de las palabras de las formas desarrolladas y no necesita de un tratamiento especial durante el proceso.

Cuando se aplica el *parser* a una subventana, se generan dos patrones de símbolos. El primer patrón se denomina "líder" (o carácter inicial de cada palabra) y el "tipo" (o clase de palabra presente en la subventana). Los tipos de palabras se representan así:

s = *stopword*
H = parte inicial de una palabra separada por guión
h = partes contiguas a la palabra separada por guión
a = sigla
w = palabra normal

Por ejemplo, dado el texto:

[…] spent fuel and recycling the recovered uranium and plutonium results in the generation of transuranic (TRU) non-high-level solid waste (NHLSW). Volumes and characteristics of these wastes, and methods for [...]

La ventana anterior para la sigla NHLSW es:

*[**r**esults **i**n **t**he **g**eneration **o**f **t**ransuranic (**T**RU) **n**on-**h**igh-**l**evel **s**olid **w**aste]*

Los patrones líder y tipo son:

[r	i	t	g	o	t	t	n	h	l	s	w]
líderes											
[w	s	s	w	s	w	a	H	h	h	w	w]
tipo											

4) Aplicación del algoritmo

Para encontrar un candidato a forma desarrollada, el algoritmo identifica una subsecuencia común de letras de la sigla y del patrón líder. Una subsecuencia es justo una secuencia dada con algunos elementos removidos. Para las secuencias X y Y, decimos que una secuencia Z es una subsecuencia común de X y Y si es una subsecuencia tanto de X como de Y; *e.g.*:

Si *X*= *acbceac* y *Y*= *cebaca*, entonces *cba* es una subsecuencia común de *X* y *Y* de longitud 3.

Obsérvese que *ceac* y *cbca* también son subsecuencias comunes de *X* y *Y* (longitud 4). Obsérvese además que no hay subsecuencias comunes mayores a longitud 4; es decir, *ceac* es una subsecuencia común de máxima longitud. La subsecuencia común más larga (LCS) de cualquiera de las dos cadenas *X* y

Y es una subsecuencia común con la mayor longitud entre todas las subsecuencias comunes.

c. Evaluación

Para evaluar AFP se tomó un corpus de 17 documentos. El sistema identificó correctamente 398 siglas, que en términos de rendimiento implica 98% de precisión y 86% de exhaustividad.

A partir de los resultados anteriores, los autores excluyeron las siglas menores o iguales a dos caracteres. De esta manera, lograron aumentar la exhaustividad hasta 93%, mientras que la precisión se mantuvo en el mismo nivel, es decir, en 98%.

AFP sólo considera como candidatos a sigla las cadenas de tres o más caracteres en mayúscula, dejando por fuera un gran número de candidatos que pueden estar formados por sólo dos caracteres. Adicionalmente, si no se da una correspondencia exacta entre los caracteres de la sigla y las iniciales de cada una de las palabras de la forma desarrollada; *e.g.*: *Teledyne Wachang Albany* (TWCA), el sistema será incapaz de reconocerlo como un candidato válido.

1.1.2 Three Letter Acronym (TLA)

TLA es un sistema creado por Stuart Yeates (1999) para la detección y extracción de pares de sigla-forma desarrollada a partir de documentos de la Biblioteca digital de Nueva Zelanda.

a. Patrones

1) Las siglas son más cortas que sus formas desarrolladas;
2) Las siglas contienen las iniciales de la mayoría de las palabras de sus formas desarrolladas;

3) Las siglas se forman con letras mayúsculas;
4) Las siglas más cortas tienden a tener palabras más largas en sus formas desarrolladas;
5) Las siglas más largas tienden a tener más palabras vacías.

b. Técnica de extracción de siglas

Inicialmente, un analizador léxico toma una secuencia de texto sin procesar, de la que selecciona los candidatos a sigla y sus formas desarrolladas. Luego, estos candidatos se pasan por un revisor heurístico, el cual aplica un número de reglas para descartar las concordancias falsas provenientes del analizador léxico. Posteriormente, en la fase de refinamiento, se eliminan los duplicados de las siglas resultantes. El siguiente gráfico muestra la estructura general del sistema.

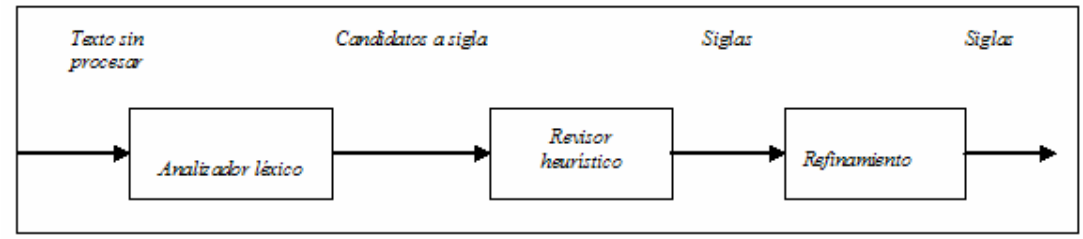


Gráfico 33. Estructura general del extractor de siglas (Yeates, 1999)

El analizador léxico ejecuta dos funciones. De un lado, remueve todos los caracteres que no son alfabéticos y divide el texto en trozos *"chunks"* basándose en la ocurrencia de los caracteres de la coma (,) y el punto (.), los cuales indican el final de un *chunk* y el comienzo de otro; por ejemplo:

el texto: **Ab cde (fgh ijk) lmn o p. Qrs**

se divide en: **Ab cde | fgh ijk | lmn o p | Qrs**

Posteriormente, en cada *chunk* se tiene en cuenta cada palabra para determinar si es un candidato a sigla. Se compara con los *chunks* anterior y posterior para buscar una forma desarrollada concordante. De esta manera se generan los siguientes pares:

Ab fgh ijk

cde fgh ijk

fgh Ab cde

ijk Ab cde

fgh lmn o p

ijk lmn o p

... ...

Si se encuentra un candidato a forma desarrollada, el par sigla-forma desarrollada se convierte en un candidato y se pasa por el revisor heurístico.

El analizador léxico usa un algoritmo al momento de buscar las formas desarrolladas, como puede apreciarse en este gráfico:

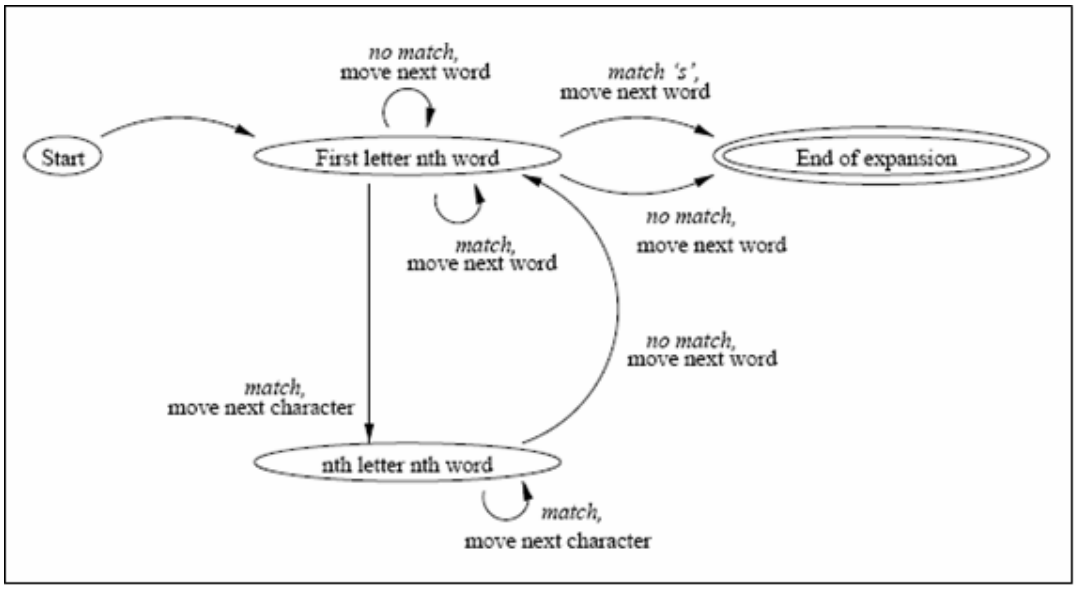


Gráfico 34. Algoritmo de concordancia de siglas (Yeates, 1999)

c. Evaluación

El sistema TLA no logra reconocer palabras dentro de la forma desarrollada que contienen más de una letra mayúscula como DBMS (*DataBase Management System*).

El corpus de evaluación de TLA constaba de 10 informes técnicos de informática. El rendimiento alcanzado fue de 68% de precisión y 91% de exhaustividad.

1.1.3 Acrophile

Larkey *et al.* (2000) desarrollaron *Acrophile*, un sistema que contiene una colección de siglas y sus formas desarrolladas, recogidas de gran cantidad de páginas web mediante un proceso heurístico de extracción. Para llevar a cabo este proyecto se evaluaron y compararon cuatro algoritmos.

Acrophile permite a los usuarios hacer dos tipos de consultas. Por un lado, se puede buscar las formas desarrolladas correspondientes a una sigla como IRS, y, por otro lado, se puede buscar la sigla correspondiente a una forma desarrollada como *Internal Revenue*. El sistema produce listas de siglas y formas desarrolladas clasificadas por puntaje de calidad.

Como prestación adicional, este sistema ofrece una casilla para introducir direcciones de páginas web, las cuales son rastreadas con el fin de encontrar pares de candidatos sigla-forma desarrollada.

a. Patrones

Los patrones para la identificación de siglas se detallan en el subapartado “Detección de siglas” y en la tabla 24 “Características de las siglas y sus formas desarrolladas”.

b. Técnica de extracción de siglas

Como se dijo anteriormente, los algoritmos de *Acrophile* utilizan el analizador léxico *flex* y el *parser yacc* para procesar textos y extraer sus siglas. Las formas desarrolladas se detectan en el texto por medio de una combinación de reglas contextuales y canónicas, las cuales coinciden con los patrones con que se expresan comúnmente las formas desarrolladas en el inglés estándar escrito.

Larkey *et al.* implementaron y probaron cuatro algoritmos de extracción diferentes. Todas las versiones trabajan bajo el principio general de que una secuencia de caracteres es una sigla si cumple con ciertos patrones, confirmándola como sigla cuando se encuentra cerca una forma desarrollada coincidente.

Después de realizada la extracción, se lleva a cabo un proceso de normalización para todos los algoritmos: dos siglas se consideran equivalentes si sólo difieren en el uso de mayúsculas o en la presencia o ausencia de puntos, guiones o espacios. Dos formas desarrolladas de una sigla se consideran equivalentes si sólo difieren en el uso de mayúsculas.

Los cuatro algoritmos, denominados contextual, canónico, canónico-contextual y simple difieren en cuanto a:

1) Los patrones que toman para indicar los candidatos a sigla;
2) Los tipos de forma desarrollada que pueden encontrarse, y
3) Los patrones de texto que indican un candidato par sigla-forma desarrollada.

Los algoritmos contextual, canónico y canónico-contextual están relacionados y surgen de la modificación de un algoritmo contextual previo. El algoritmo simple se desarrolló de forma independiente para probar un enfoque más limitado y lograr mayor precisión en los candidatos a sigla. Cada algoritmo se probó y perfeccionó

mediante la aplicación de una prueba piloto que consistió en procesar un corpus de 12.380 artículos del *Wall Street Journal*.

El algoritmo simple es el más riguroso; sólo busca aquellos pares de sigla-forma desarrollada que se ajusten a un pequeño grupo de formas canónicas tales como: "forma desarrollada (SIGLA)" o "SIGLA or forma desarrollada".

El algoritmo contextual, mucho menos estricto, busca una forma desarrollada en el contexto de la sigla sin requerir ningún patrón canónico (o, paréntesis, comas, etc.) que indique su relación.

Los algoritmos canónico-contextual y canónico se encuentran a medio camino entre los dos anteriores. En la tabla 24 pueden observarse las características de los cuatro algoritmos así como las características de las siglas y las formas desarrolladas.

1) Detección de siglas

   Para la identificación de los candidatos a sigla, los algoritmos buscan los patrones indicados en la primera fila de la tabla 1. Esta fila presenta una notación de expresión seudo-regular en la que:

   a) + indica una o más ocurrencias de un símbolo
   b) Indica 0 ó más ocurrencias
   c) Los *superscripts* numerados indican un número específico o rango de ocurrencias
   d) U significa una letra mayúscula
   e) L significa una letra minúscula
   f) D significa un dígito
   g) S significa un final opcional con *s* o *'s*
   h) {sep} es un punto o punto seguido por un espacio, y
   i) {dig} es un número entre 1 y 9, opcionalmente seguido de un guión. Los términos en corchetes son alternativos.

Seguidamente se indican los patrones que reconoce cada uno de los algoritmos:

- Algoritmo contextual. Acepta siglas que tienen:

  a) Mayúsculas sostenidas como USA
  b) Puntos como U.S.A.
  c) Una secuencia de letras minúsculas, cuyos caracteres pueden aparecer al final del patrón siguiendo al menos tres caracteres en mayúscula, como COGSNet, o internamente siguiendo al menos dos caracteres en mayúscula, como AchemS
  d) Un patrón de mayúsculas también puede tener cualquier número de dígitos en cualquier posición.

- Algoritmos canónico-contextual y canónico. Aceptan una amplia gama de patrones de siglas. Tienen menos restricciones respecto de las secuencias de letras minúsculas para permitir patrones como DoD**.** Permiten barras y guiones en las siglas para obtener patrones como AFL-CIO y 3-D. No tienen en cuenta las siglas que terminan en letras minúsculas excepto la *s*, y sólo detectan siglas con un dígito.

- Algoritmo simple. Usa un enfoque minimalista; excluye siglas con dígitos, puntos y espacios. Busca unidades que comiencen por una letra mayúscula, seguida de cero a ocho letras mayúsculas minúsculas, barras, guiones y que finalicen con una letra mayúscula.

2) Detección de formas desarrolladas

Los elementos que tiene en cuenta cada algoritmo para la detección de formas desarrolladas son:

- Algoritmo contextual. Encuentra las formas desarrolladas buscando coincidencias desde el último carácter de una sigla hacia atrás. Siempre guarda la

ventana con las últimas 20 palabras que preceden la sigla, de manera que, cuando identifica un candidato a sigla, trata de encontrar la forma desarrollada dentro de esta. En caso de no encontrarla, continuará la búsqueda en el texto que hay después de la sigla. No requiere de formas canónicas, por tanto, trata con éxito textos como: ...“*is three dimensional. In 3D images...*”

Las reglas de formas desarrolladas remiten a una lista de 35 de palabras vacías como *and*, *for*, *of* y *the*, las cuales suelen omitirse en la formación de las siglas; *e.g.*: CIIR (*Center* ***for*** *Intelligent Information Retrieval*). El algoritmo trata de hallar una secuencia de palabras tal que parte de los primeros cuatro caracteres de cada palabra, que no sean una palabra vacía, concuerden con los caracteres de la sigla; *e.g.*: *Bureau of Personnel* (BUPERS). Además:

a) Un carácter inicial de una palabra vacía puede coincidir con un carácter interno de una sigla; *e.g.*: *Department of Defense* (DOD).
b) Una palabra vacía puede omitirse; *e.g.*: *Research Experience for Undergraduates* (REU).
c) El primer, cuarto, quinto o sexto carácter de las palabras del candidato a forma desarrollada puede concordar con los caracteres de la sigla; *e.g.*: *PostScript* (PS).

Las siglas que poseen dígitos reciben un tratamiento especial. El algoritmo intenta reemplazar el dígito y el carácter anterior o posterior con *n* repeticiones del carácter; *e.g.*: MMM por 3M. Si no puede encontrar una forma desarrollada para esta sigla transformada, entonces trata de encontrar la concordancia entre el dígito y el número escrito en letras, por ejemplo: *three dimensional* por 3D. Los puntos en las siglas se ignoran cuando se buscan las formas desarrolladas.

Uno de los principales problemas del algoritmo contextual es su tendencia a tratar de hacer coincidir más de un carácter inicial de las palabras de la forma desarrollada. Esto lleva al algoritmo a expandir NIST como *National Institute of*

Standards, tomando la *t* de *Standards*, en lugar de ***N**ational **I**nstitute of **S**tandards and **T**echnology*. Otro problema, en particular con las siglas de dos letras, es su tendencia a hallar secuencias de palabras en minúscula con una concordancia falsa para la sigla; *e.g.*: *story from* para SF.

- Algoritmo canónico-contextual. Es una modificación del algoritmo contextual para tratar los dos problemas antes mencionados. En primer lugar, incluye reglas canónicas para restringir la aceptación de palabras minúsculas en la forma desarrollada. Solamente se permite una forma desarrollada en minúsculas si un par sigla-forma desarrollada cumple con una de las formas que se describen en la tabla 24. Una forma desarrollada hallada por medio de reglas contextuales debe estar en mayúscula a excepción de las *stopwords*. En segundo lugar, el algoritmo trata de buscar, con criterio conservador, las concordancias entre múltiples caracteres en una forma desarrollada, solucionando de esta forma el problema ilustrado anteriormente con NIST. Además, los guiones y las barras se tienen en cuenta dentro de las siglas, pero se pasan por alto al expandirlas. Si una forma desarrollada está separada por guiones; *e.g.*: *Real-Time*, que hace parte de la sigla CRICCS (*Center for **Real-Time** and Intelligent Complex Computing Systems*), el algoritmo puede tratar *Real Time* como dos palabras o como una sola palabra, sin necesidad de que exista una *T* en la sigla.

- Algoritmo canónico. Es un derivado del algoritmo canónico-contextual del cual sólo toma los pares de sigla-forma desarrollada que se encuentren en la forma canónica.

- Algoritmo simple. Busca suprimir gran parte de la complejidad del algoritmo contextual y sus derivados. Al igual que el algoritmo canónico, el algoritmo simple requiere que la sigla se encuentre en ciertos contextos, aunque acepta menos patrones canónicos para los pares de sigla-forma desarrollada y menos patrones de sigla. El algoritmo busca las formas presentadas en la forma desarrollada canónica de la tabla 24, en el orden en que se listan.

Al momento de revisar la validez de un candidato a forma desarrollada, el algoritmo tiene varios esquemas de concordancia de sigla-forma desarrollada. Cada uno de los cuales revisa repetitivamente las formas desarrolladas más cortas primero. Los esquemas de concordancia se llevan a cabo de la siguiente manera:

a) Mayúscula estricta. Cada letra en la sigla debe estar representada, en orden, por una letra mayúscula en la forma desarrollada. Esta debe comenzar con la primera letra de la sigla.
b) Minúscula estricta. Cada letra en la sigla debe estar representada, en orden, por la primera letra de una palabra en la forma desarrollada. Esta debe comenzar con la primera letra de la sigla y no debe contener letras mayúsculas.
c) Mayúscula flexible. La primera palabra debe comenzar con la primera letra de la sigla y la última palabra debe comenzar con una letra de la sigla. Este esquema es sumamente flexible, pudiendo llevar a formas desarrolladas donde algunas letras de la sigla no coincidan en absoluto.

A continuación se presenta la tabla 24, que presenta las características de las siglas y sus formas desarrolladas según el tipo de algoritmo (Larkey *et al.*, 2000).[54]

54 El *superescript* + indica una o más ocurrencias de un símbolo; * indica 0 ó más ocurrencias; los *superescripts* numerados indican un número específico o rango de ocurrencias; U significa una letra mayúscula; L significa una letra minúscula; D significa un dígito; S significa un final opcional s o 's;{sep} es un punto o punto seguido por un espacio, y {dig} es un número entre 1 y 9, opcionalmente seguido de un guión. Los términos en corchetes son alternativos.

SIGLAS	**Algoritmos**		
	Contextual	***Canónico-contextual / Canónico***	***Simple***
Patrones para las siglas	(U {sep})$^{+}$ *e.g.*: U.S.A U^{+} *e.g.*: USA D*U[DU]* *e.g.*: 3D,62A2A UUU^{+}L^{+} *e.g.*: JARtool UU^{+}L^{+}U^{+} *e.g.*: AChemS	(U {sep})$^{2-9}$ S *e.g.*: U.S.A, U.S.A. 's U$^{2-9}$S *e.g.*: USA, USA's U*{dig}U^{+} *e.g.*: 3D, 3-D, I3R U^{+}L^{+}U^{+} *e.g.*: DoD U^{+}[/-]U^{+} *e.g.*: AFL-CIO	U[UL/-]$^{0-8}$U *e.g.*: USA, DoD, AFL-CIO
Mayúsculas *vs* minúsculas	Los dos primeros caracteres deben ser U, luego cualquier número de L en alguna parte, pero adyacente	L interna, o *s* final o *'s* DOD, DoD, DOD's	Debe comenzar y finalizar con U Puede tener L en otro sitio DOD, DoD
Dígitos	Cualquier número de dígitos en cualquier lugar	Sólo un dígito, en cualquier posición que no sea la final; *e.g.*: 3M, 2ATAF	Ninguno
Espacios y puntos	Después de letras mayúsculas	"." o ". + espacio" debe estar después de cada caracter; *e.g.*: N.A.S.A, N. A. S. A.	Ninguno
Barra o guión (/ ó -)	Ninguno. Tratados como espacio en la *tokenización*	Una barra o guión interior; *e.g.*: OB/GYN, CD-ROM	Cualquier número de barras o guiones en el interior; *e.g.*: OB/GYN, CD-ROM
Longitud máxima	No está explícita	9 caracteres alfanuméricos, más cualquier puntuación incluida o *s* final	10 caracteres incluyendo cualquier puntuación
FORMAS DESARROLLADAS			
Palabras vacías o *Stopwords* (*and*, *for*, *of*, *the*)	Lista fija de 35 *stopwords*	Lista fija de 40 *stopwords*	Ninguna
Palabras omitidas	Sólo *stopwords*	*Stopwords* o palabras que se encuentran después de guiones	Sólo las primeras y últimas palabras tienen que coincidir con caracteres en la sigla
Caracteres de *stopwords*	Como máximo uno, únicamente caracteres internos en la sigla		No aplicable
Prefijos	Si, asume que cualquier inicial hasta la quinta posición puede ser un prefijo		No aplicable
Caracteres procedentes de palabras que no son *stopwords*	Hasta 4 caracteres. Algoritmo *"greedy"*, prefiere tomar más	Hasta 4 caracteres. Algoritmo "conservador", prefiere tomar menos	Prefiere hasta 1 caracter. Puede tomar más si la palabra comienza por mayúscula
Forma desarrollada canónica	No aplicable	(se buscan en desorden) SIGLA (Forma desarrollada), Forma desarrollada (SIGLA) (Forma desarrollada) SIGLA, (SIGLA) Forma desarrollada SIGLA or Forma desarrollada, Forma desarrollada or SIGLA, SIGLA stands for Forma desarrollada SIGLA {is} an acronym for Forma desarrollada known as the SIGLA Forma desarrollada "SIGLA", "SIGLA" Forma desarrollada	(se buscan en orden) Forma desarrollada (SIGLA) Forma desarrollada or SIGLA Forma desarrollada, or SIGLA Forma desarrollada, SIGLA SIGLA (Forma desarrollada) SIGLA, Forma desarrollada
Uso de mayúsculas o minúsculas	La forma desarrollada puede aparecer en su totalidad en letras minúsculas	Canónico: todas pueden ser minúsculas Contextual: sólo las *stopwords* pueden ser minúsculas, el resto deben ser mayúsculas	Se permiten las minúsculas, pero con reglas más estrictas que las de las mayúsculas; la letra inicial de cada palabra de la forma desarrollada debe coincidir con la letra que ocupa idéntica posición en la sigla.
Números	En letras o en dígitos		Sin números

Características de las siglas y sus FD según el tipo de algoritmo (Larkey *et al.*, 2000)

c. Evaluación

Para la evaluación de los algoritmos, Larkey *et al.* tomaron un corpus de 936.550 páginas web de instituciones militares y gubernamentales de los Estados Unidos, el cual procesaron para incluir en la base de datos *Acrophile*. De este conjunto se escogieron al azar 170 páginas para buscar los pares de siglas-formas desarrolladas. Como resultado se obtuvieron 353 pares, de los cuales 10 tenían símbolos como & y /. Ninguna de las siglas presentaba números o guiones. Las variaciones en las formas desarrolladas consideradas como correctas fueron la omisión o adición de la 's' y las diferencias en la puntuación.

Los siguientes son los valores de precisión y exhaustividad para los 4 algoritmos en las 353 siglas del test (328 de las cuales tienen una longitud igual o mayor a 3 caracteres).

Algoritmo	**Todas las siglas**		**Siglas de longitud > 2 caracteres**	
	Precisión	Exhaustividad	Precisión	Exhaustividad
Contextual	0.89	0.61	0.96	0.60
Canónico-contextual	**0.87**	**0.84**	0.92	0.84
Canónico	0.96	0.57	0.99	0.59
Simple	0.94	0.56	0.99	0.57

Tabla 25. Precisión y exhaustividad en el corpus de evaluación (Larkey *et al.*, 2000)

Los cuatro algoritmos fallaron en la detección de 16 casos debido a que la forma desarrollada estaba a una distancia superior a 20 palabras; es decir, demasiado lejos de su sigla correspondiente. Sin embargo, los autores dejan claro que no estaba dentro de sus expectativas que alguno de sus algoritmos lo consiguiera.

El algoritmo de mejor rendimiento es el canónico-contextual. No obstante, para los autores, sus resultados no son comparables con la precisión y la exhaustividad alcanzadas por los sistemas de Taghva & Gilbreth y de Yeates, puesto que éstos emplean corpus y criterios de exactitud diferentes.

1.1.4 Acromed

Putstejovsky *et al*. (2001) desarrollaron un sistema llamado *Acromed*, una de las herramientas diseñadas para procesar y extraer información de los *abstracts* de la base de datos Medline. Estos autores afirman que este sistema se diferencia de los sistemas de extracción de siglas preexistentes (Taghva & Gilbreth, 1999; Yeates, 1999 y Larkey, 2000) en que el algoritmo de reconocimiento de siglas incluye un análisis sintáctico superficial o *shallow parsing* de los textos.

a. Patrones

Los pares de candidatos deben coincidir con el patrón "FD (sigla)". El carácter inicial de la primera palabra de la forma desarrollada debe coincidir con el carácter inicial de la sigla.

b. Técnica de extracción de siglas

La estrategia de extracción de siglas de *Acromed* tiene dos vías. La primera considera el problema del reconocimiento de pares de forma desarrollada-sigla como el problema de encontrar dos cadenas de caracteres en un texto que coincidan con ciertas expresiones regulares, a lo que se denomina "algoritmo de expresión regular".

Los autores han desarrollado un patrón muy restringido para el par de forma desarrollada-sigla:

$$\#\ String_i(String_j)$$

Donde
significa el límite de una oración

$String_i$ representa la forma desarrollada

($String_j$) representa la sigla

Debido a que el patrón usado era demasiado limitado, Pustejovsky *et al.*, decidieron incluir la posibilidad de buscar las formas desarrolladas en la ventana o contexto derecho además del izquierdo, como medida para mejorar la exhaustividad.

La segunda vía consiste en el refinamiento del paso anterior. Aunque el problema básico es el mismo (es decir, dos cadenas de caracteres se comparan para decidir si una es la forma desarrollada de la otra), la extensión y límites del contexto donde se busca la forma desarrollada son totalmente diferentes.

Acromed usa dos mecanismos para la extracción de información como son el preproceso de textos y la anotación sintáctica. Posteriormente, busca una sigla objetivo en un contexto tal y como se ilustra a continuación:

EXP_i, EXP_j, T_ACRONYM, EXP_k, EXP_m, donde las expresiones son cadenas de caracteres etiquetadas y T_ACRONYM es otra expresión (generalmente una cadena etiquetada) como en:

1) [[‘the’, ‘DT’], [‘performance’, ‘NN’], [‘of’, ‘IN’], [‘an’, ‘DT’], [‘automatic’, ‘JJ’], [‘speech’, ‘NN’], [‘recognition’, ‘NN’], [‘NX’],
2) [‘(‘,’(‘],
3) [[‘ASR’, ‘NN’], ‘NNX’]
4) [‘)’, ‘)’]

Con el ejemplo anterior Pustejovsky *et al*. muestran 4 expresiones que son la entrada para el detector de pares de forma desarrollada-sigla. Bajo este modelo, las cadenas de caracteres: “*The performance of an automatic speech recognition*, *ASR*” serán usadas como entrada para la expresión regular para el reconocimiento del par forma desarrollada-sigla.

Según Pustejovsky, este diseño permite restringir considerablemente el contexto de búsqueda de la forma desarrollada. En un algoritmo que considere únicamente las cadenas de caracteres y su contexto, se debe establecer una ventana o límite arbitrario. Con el análisis sintáctico superficial, el límite se establece naturalmente gracias a las propiedades del lenguaje. Con esta estrategia, se especifica que la forma desarrollada es un sintagma nominal que está cerca del candidato a sigla (o sigla "objetivo"). Dentro del contexto de un candidato a sigla se pueden establecer restricciones como signos de puntuación y coordinación de sintagmas nominales.

Estos autores utilizan un autómata de estado finito que usa las expresiones, verificando sus tipos; es decir, sintagma nominal, sintagma verbal o signo de puntuación. Si se encuentra un candidato par forma desarrollada-sigla, entonces las cadenas de caracteres correspondientes a ambas expresiones se suministran a la cadena de búsqueda de siglas (la estrategia previa), la cual decidirá si una subcadena de la forma desarrollada concuerda con la sigla. Si es así, se considerará como una identificación positiva y se almacenará en la BD de siglas.

Los primeros experimentos con el mecanismo de restricciones sintácticas usaron sólo el siguiente patrón:

$$\text{T\_LF\_Noun Phrase}_1\ (\text{T\_A\_Noun Phrase}_2)$$

donde T_LF significa objetivo donde hallar la forma desarrollada y T_A significa objetivo o candidato a sigla.[55]

Los experimentos posteriores usaron restricciones sintácticas modificadas y adicionaron los siguientes patrones:

55 LF (*long form*) es equivalente a FD (forma desarrollada de la sigla). T_A equivale a *"target acronym"* (sigla objetivo o candidato a sigla).

1) T_A_Noun Phrase$_1$ (T_LF_ Phrase$_2$)
2) T_A_Noun Phrase$_1$, T_LF_Noun Phrase$_2$
3) (T_A_Noun Phrase$_1$) T_LF_Noun Phrase$_2$

c. Evaluación

Para la fase de evaluación del sistema se empleó un corpus de entrenamiento y otro de evaluación.[56] El primero contenía 86 *abstracts* de la base de datos de *Medline* y 155 pares de sigla-forma desarrollada, mientras que el segundo contenía 100 *abstracts* y 173 pares de sigla-forma desarrollada.

En el corpus de entrenamiento, *Acromed* recuperó 123 pares de forma desarrollada-sigla, de los cuales 106 eran correctos. Los resultados de precisión y exhaustividad son similares a los obtenidos en otras investigaciones sobre el tema.

En el corpus de evaluación, *Acromed* empleó los algoritmos de expresión regular y restricción sintáctica. Con el primer algoritmo recuperó 117 pares de sigla-forma desarrollada, de los cuales 106 eran correctos. Con el segundo algoritmo, recuperó 105 pares, de los que 104 eran correctos.

Algunos errores en la detección de pares de forma desarrollada-sigla se asocian al problema de la delimitación de la ventana para la búsqueda de la forma desarrollada; *e.g.*: *p16= products*, extraído de: "*which encodes two gene products (p16(INK4a) and p19(ARF))*".

Los resultados obtenidos por *Acromed* con respecto a la precisión y la exhaustividad fueron los siguientes:

56 Estos autores toman como punto de referencia para la evaluación de su sistema el *Gold Standard*, formado por 149 pares de sigla-forma desarrollada.

	Corpus de entrenamiento		Corpus de evaluación	
	Precisión	Exhaustividad	Precisión	Exhaustividad
Expresión regular	88,1%	73,2%	90%	63%
Restricciones sintácticas	97,2%	72,5%	99%	61,9%
Restricciones sintácticas modificadas	94,6%	82,5%	**98,3%**	**72%**

Tabla 26. Precisión y exhaustividad de Acromed (Pustejovsky *et al.*, 2001)

1.1.5 Sistema para gestión de variación terminológica

Nenadić *et al.* (2002) consideran las siglas como un fenómeno de variación terminológica muy común. Las siglas pueden ser variantes terminológicas de tipo léxico-semántico, puesto que se usan como sinónimos de sus formas desarrolladas correspondientes; o pueden ser variantes terminológicas de tipo pragmático, ya que facilitan la lectura de los textos científicos.

Estos autores consideran que expertos como los biólogos moleculares crean frecuentemente siglas específicas que usan localmente (dentro de un artículo científico) o dentro de todo su campo de especialidad.

Las siglas, al igual que los términos, adolecen de los siguientes problemas:

- Variación. Un mismo término puede tener varias siglas; *e.g.*: NF kappa, NF kB;
- Ambigüedad. Una misma sigla puede referirse a conceptos diferentes; *e.g.*: GR puede referirse tanto a *glucocorticoid receptor* como a *glutathione reductase*.

Nenadić *et al.* sólo tratan el problema de variación de las siglas. Consideran que el problema de la ambigüedad puede resolverse simplemente con el uso de la última forma desarrollada introducida en el texto, en caso de que haya una. Si no hay una forma desarrollada introducida, entonces deben usarse los métodos generales para la desambiguación de términos.

a. Técnica de extracción de siglas

Para localizar los candidatos a forma desarrollada, el sistema de Nenadić *et al.* se basa en los patrones sintácticos que se usan principalmente en los artículos científicos. Una vez se detecta que una secuencia de palabras coincide con un patrón *x*, se recupera y se analiza morfológicamente con el propósito de descubrir la relación entre la sigla y su forma desarrollada.

El método de adquisición de siglas consta de tres pasos, a saber:

1) Recuperación de las formas desarrolladas

   En este paso se explora el texto para la búsqueda de candidatos a forma desarrollada. Varios patrones de forma desarrollada se han identificado manualmente para describir varios contextos de introducción de una sigla, a saber:

   a) Forma desarrollada a la izquierda; *e.g.*: *9-cis retinoic acid* (9cRA);
   b) Forma desarrollada a la derecha; *e.g.*: MIBP (*Myc-intron-binding peptide*).

Para estos autores más del 90% de las ocurrencias corresponden al patrón "forma desarrollada a la izquierda"; suelen introducirse mediante el uso de paréntesis; *e.g.*: *tumor necrosis factor alpha* (TNF-alpha) y, raras veces, mediante el uso de un formato similar a la aposición; *e.g.*: ...*enzyme-linked immunosorbent assay*, ELISA, ...

2) Concordancia entre siglas y forma desarrollada

   En este paso se aplica un conjunto de patrones de formación de siglas para lograr la concordancia entre un candidato a forma desarrollada y su sigla. En general, las siglas se forman mediante la selección de caracteres iniciales de las palabras de la forma desarrollada. Sin embargo, se ha observado que en el campo de la biología

molecular las iniciales de las formas de combinación también se usan con el mismo propósito. Las formas de combinación son afijos específicos (principalmente prefijos e infijos como: *acetyl*, *trans*, *di* e *hydro*), que se usan con regularidad en los patrones de formación de términos; *e.g.*: ***c****hloramphenicol* ***a****cetyl****t****ransferase* (*CAT*). En el momento de buscar la concordancia entre la forma desarrollada y la sigla se usa un diccionario de formas de combinación del ámbito de la biología molecular.

El método básico de concordancia entre siglas y sus formas desarrolladas mejora si se tienen en cuenta los siguientes fenómenos relacionados con las formas desarrolladas:

a) Inserción. Una palabra está presente en la forma desarrollada, pero no ha sido usada en la formación de la sigla; *e.g.*: ***t****hyroid* ***hormone r****eceptor* (TR);
b) Omisión. Una palabra falta en la forma desarrollada, aunque se usa al momento de formar la sigla; *e.g.*: *[****human****]* ***e****strogen* ***r****eceptor* (hER);
c) Sigla en plural. Se establece un plural para una sigla; *e.g.*: ***r****etinoid* ***x*** ***r****eceptor****s*** (RXRs);
d) Sigla recursiva. La forma desarrollada de una sigla contiene a su vez otra sigla o abreviatura; *e.g.*: ***CREB-b****inding* ***p****rotein* (CCBP);
e) Siglas coordinadas. Las siglas se definen dentro de una estructura coordinada; *e.g.*: ***e****strogen* (ER) *and* ***p****rogesterone* (PR) ***r****eceptors*;
f) Sigla parcial. Una sigla contiene una parte de su forma desarrollada, normalmente palabras griegas o latinas; *e.g.*: ***r****etinoid* ***x*** ***r****eceptor* ***alpha*** (RXR alpha);
g) Variación estructural. Se define una sigla y posteriormente se realiza una transformación morfológica/estructural en su forma desarrollada; *e.g.*: ***d****ay of* ***h****atching* (HD);
h) Siglas fórmula. Una sigla contiene una parte de una fórmula química; *e.g.*: ***1****alpha****,25-dihydroxy****vitamin* ***D3*** [1,25 (OH) 2D3].

Los fenómenos anteriormente listados se consideran cuando el método básico de concordancia (es decir, la concordancia de caracteres de la sigla con los constituyentes de un candidato a forma desarrollada) no arroja un resultado positivo.

Finalmente, el paso anterior produce una lista de siglas que concuerdan con sus forma desarrollada.

3) Agrupamiento de siglas

   Por último, en el tercer paso, los autores tratan de establecer las clases de variantes de una sigla. En primer lugar, tanto las siglas como sus formas desarrolladas se normalizan con respecto a sus rasgos ortográficos, morfológicos, sintácticos y léxico-semánticos. En particular, las siglas en plural como NRs (*nuclear receptors*) se hacen concordar con la correspondiente sigla en singular NR (*nuclear receptor*). Todas las siglas que comparten una forma desarrollada normalizada conforman un clúster o agrupación de siglas.

b. Evaluación

La evaluación se llevó a cabo a partir de dos corpus creados a partir de la BD *Medline*, conformados por 2.008 y 6.323 *abstracts*, respectivamente.

Para la evaluación sobre la adquisición de siglas se tomó una muestra aleatoria de 50 *abstracts* del primer corpus. La siguiente tabla muestra algunos ejemplos de siglas reconocidas automáticamente.

Sigla(s)	FD
RAR alpha	Retinoic acid receptor alpha
RAR-alpha	
RARA	
RARa	
RARs	Retinoic acid receptors
RAR	Retinoic acid receptor
RT-PCR	Reverse transcription PCR

TR	Thyroid hormone receptor
TRs	Thyroid hormone receptors Thyroid receptor
9-c-RA	9-cis-retinoic acid
9cRA	9-cis retinoic acid
ES	Ewing sarcoma Ewing's sarcoma Ewings sarcoma

Tabla 27. Ejemplos de siglas reconocidas por el sistema de Nenadić *et al.* (2002)

La precisión de este método es muy alta, ubicándose en un rango entre 94% y 99%, dependiendo del tamaño del corpus. Aunque la exhaustividad del 73% no es un resultado despreciable, Nenadić sostiene que podría mejorarse, dado que se han identificado patrones adicionales durante la fase de evaluación manual.

En la siguiente tabla se presentan los resultados de la evaluación en los diferentes corpus.

Corpus / **Siglas**	**2.008 *abstracts***	**6.323 *abstracts***	**50 *abstracts***
Número de siglas diferentes reconocidas	1.015	2.343	66
Número de siglas reconocidas correctamente	992	2.314	62
Número de siglas introducidas	-	-	85
Precisión	97,73%	98,76%	93,94%
Exhaustividad	-	-	72,94%

Tabla 28. Evaluación en los diferentes corpus con el sistema de Nenadić et al.

1.1.6 A simple algorithm

Schwartz & Hearst (2003) implementaron un algoritmo simple para extraer pares de siglas-forma desarrolladas presentes en textos biomédicos.

El sistema ejecuta dos tareas: la primera consiste en la extracción de los pares de candidatos sigla-forma desarrollada, mientras que la segunda consiste en la identificación de la forma desarrollada correcta a partir de los candidatos presentes en el contexto de la sigla.

a. Patrones

Los patrones para seleccionar un candidato a sigla son:

1) 2 a 10 caracteres
2) Máximo 2 palabras
3) Mínimo una letra
4) Primer carácter alfanumérico.

Los patrones para seleccionar un candidato a forma desarrollada son:

1) Una forma desarrollada debe aparecer inmediatamente antes o después de su sigla correspondiente, es decir, en la misma oración y no debe tener más de (|A|+5), (|A|*2) palabras. Donde |A| es el número de caracteres de la sigla.

   El método de selección de los candidatos a sigla, al igual que en muchos de los métodos existentes, se determina por su adyacencia a un paréntesis; es decir:

   a) FD (sigla)
   b) Sigla (FD)

Según estos autores, en la práctica la mayoría de los pares de candidatos se ajustan al patrón FD (sigla), de ahí que sea el patrón que emplean en su estudio. Además, subrayan que los candidatos a forma desarrollada contiguos a la sigla son los únicos que se tienen en cuenta.

b. Técnica de extracción de siglas

Cuando se detecta un candidato a sigla, el algoritmo busca su forma desarrollada en el contexto que se encuentra a derecha e izquierda. El algoritmo trata de encontrar la concordancia entre cada carácter de la sigla y de la forma desarrollada moviéndose a la izquierda, comenzando desde el final de ambas cadenas de caracteres. El algoritmo acierta si el primer carácter de la sigla coincide con el primer carácter de cada palabra de la forma desarrollada.

c. Evaluación

El algoritmo considera dos tipos de patrones, a saber: "FD (sigla)" y "sigla (FD)". La forma desarrollada debe estar contigua a la sigla. El algoritmo se evaluó en dos corpus diferentes. Por un lado, se probó con una versión corregida del "*Gold Standard*" de Pustejovsky *et al.*, que contiene 168 pares de sigla-forma desarrollada. En este caso el sistema identificó 143 pares, de los cuales 137 eran correctos. Los 31 pares restantes no fueron identificados, pues no presentaban una coincidencia exacta entre sus caracteres; *e.g.*: CNS1 (*cyclophilin seven suppressor*); ATN (*anterior thalamus*). Este resultado se traduce en 96% de precisión y 82% de exhaustividad.[57]

Por otro lado, el algoritmo se evaluó en un corpus de 1.000 *abstracts* extraídos de *Medline*, los cuales contenían 954 pares sigla-forma desarrollada. El rendimiento fue de 95% de precisión y 82% de exhaustividad.

A modo de síntesis, se presenta la siguiente tabla comparativa del rendimiento de los sistemas de detección de siglas basados en patrones.

57 Schwartz & Hearst manifiestan que los resultados de su algoritmo son bastante parecidos a los alcanzados por otros sistemas más complejos como los de Pustejovsky (72% de exhaustividad y 98% de precisión) o Chang (83% de exhaustividad y 80% de precisión).

Sistema	Autor	Año	Corpus	Ámbito	Precisión	Exhaustiv.
AFP	Taghva & Gilbreth	1999	17 documentos técnicos (463 siglas-FD)	Medio ambiente	98%	93%
TLA	Yeates	1999	10 documentos técnicos	Informática	68%	91%
Acrophile	Larkey *et al.*	2000	170 páginas web (353 siglas-FD)	Militar-gubernamental	87%	84%
Acromed	Pustejovsky *et al.*	2001	100 *abstracts* de la BD *Medline* (173 siglas-FD)	Biomedicina	98%	72%
Sistema de gestión de variación terminológica	Nenadić *et al.*	2002	50 *abstracts* de *Medline* (85 siglas-FD)	Biomedicina	94%	73%
A Simple Algorithm	Schwartz & Hearst	2003	100 *abstracts* de *Medline* (168 siglas-FD)	Biomedicina	96%	82%
			1.000 *abstracts Medline* (954 siglas-FD)		95%	82%

Tabla 29. Rendimiento de los sistemas de detección de siglas basados en patrones

Aunque los resultados logrados por el sistema de Taghva & Gilbreth son superiores, la mayoría de los autores coinciden en afirmar que dichos resultados no son comparables, dado que su algoritmo no tuvo en cuenta las siglas de dos caracteres y se evaluó en un corpus muy pequeño, de tan sólo de 17 textos. Puede decirse entonces que los dos métodos con mejor rendimiento dentro del grupo de sistemas basados en patrones son los de Schwartz & Hearst y Larkey *et al.*, respectivamente.

1.2. Métodos basados en estadística y aprendizaje máquina

Las investigaciones más recientes sobre técnicas de extracción de siglas apuntan al desarrollo de métodos estadísticos y de aprendizaje máquina.

Los métodos estadísticos se basan en la frecuencia de las siglas en un corpus. Dentro de esta clase destacan los trabajos de Chang *et al.* (2002) y Adar (2004).

Los algoritmos de aprendizaje usan ejemplos, atributos y valores para reconocer y clasificar siglas. Tienen la capacidad de mejorar con la experiencia (entrenamiento).

Entre los autores que han desarrollado sistemas basados en este método destacan Young (2004), Zahariev (2004), Dannélls (2005) y Nadeau & Turney (2005).

Además de los dos métodos anteriores, existen los métodos híbridos, es decir, aquellos que combinan la estadística con el aprendizaje máquina. Dentro de estos métodos destaca el trabajo de Park & Byrd (2001).

1.2.1 Métodos basados en técnicas estadísticas

1.2.1.1 Diccionario de abreviaciones en línea

Chang *et al*. (2002) desarrollaron un método de detección de siglas basado en un algoritmo de regresión logística, que usa un conjunto de rasgos para describir los diferentes patrones presentes en las siglas.

El conjunto de rasgos empleados por Chang *et al*. para la descripción de las siglas es el siguiente:

Rasgo	Descripción
Minúscula *vs.* mayúscula	% de letras en la sigla en minúscula
Comienzo de palabra	% de letras alineadas al comienzo de una palabra
Final de palabra	% de letras alineadas al final de una palabra
Límite de sílaba	% de letras ubicadas en un límite de sílaba
Después de letra alineada	% de letras alineadas inmediatamente después de otra letra
Letras alineadas	% de letras que están alineadas en la sigla
Palabras omitidas	Número de palabras en la FD no alineadas con la sigla
Letras alineadas por palabra	Número promedio de letras alineadas por palabra
CONSTANTE	Normalización constante por algoritmo de regresión logística

Tabla 30. Vector de rasgos de una sigla (Chang *et al.*, 2002)

a. Heurística

1) El algoritmo sólo considera los candidatos que aparezcan entre paréntesis de acuerdo con el patrón “FD (sigla)”;

2) Dentro del paréntesis se recuperan las palabras que se encuentren antes de una coma o un punto y coma;
3) Una sigla debe contener una letra como mínimo;
4) El contexto o ventana de búsqueda de la forma desarrollada es de 2*|A| palabras.

b. Técnica de extracción de siglas

Para Chang *et al.* el proceso de detección de las siglas consta de cuatro pasos, a saber:

1) Búsqueda de candidatos a sigla. Se emplea el patrón "FD (sigla)". Dentro de los paréntesis sólo se recuperan las cadenas de caracteres que están antes de una coma o un punto y coma. Se rechazan los candidatos a sigla que no posean ninguna letra. Para cada candidato se salvan las palabras que se encuentran antes del paréntesis (prefijo), de manera que se pueda buscar en ellas la forma desarrollada de la sigla.

2) Alineación de los candidatos a sigla con el texto que los precede (ventana). Las letras del candidato a sigla se alinean con las del prefijo o texto que lo precede. Este paso es equivalente al procedimiento que ejecuta el algoritmo *Longest common subsequence* (LCS), empleado en investigaciones como la de Taghva & Gilbreth.

3) Conversión de las alineaciones en un vector de rasgos. Se calculan los vectores de rasgos que describen cuantitativamente al candidato par sigla-forma desarrollada. Para llevar a cabo esta tarea, Chang *et al.* han usado las 9 características que más información aportan sobre el candidato. Estas características se establecieron a partir de las observaciones realizadas sobre un corpus de *abstracts* de *Medline*.

4) Puntuación de las alineaciones mediante un algoritmo de aprendizaje máquina. Se utiliza la puntuación de las alineaciones mediante un algoritmo de aprendizaje máquina. Para el entrenamiento de este algoritmo se empleó un

corpus de 1.000 candidatos a sigla seleccionados aleatoriamente de *abstracts* de *Medline*. A partir de estos *abstracts* se identificaron 93 siglas y se anotó manualmente la alineación entre las siglas y sus prefijos. Luego se generaron todas las alineaciones posibles en el corpus de los 1.000 candidatos, lo que llevó a la creación del corpus de experimentación. Los tipos de alineaciones que se encontraron fueron:

a) Alineación de siglas incorrectas;
b) Alineación correcta de siglas correctas;
c) Alineación incorrecta de siglas correctas.

Todas estas alineaciones se convirtieron en vectores de rasgos que sirvieron para entrenar el clasificador de regresión logística. Como resultado se obtuvo una lista de candidatos a sigla junto con sus formas desarrolladas y puntajes. El sistema considera correcto un par sigla-forma desarrollada cuando coincide exactamente con el *Gold Standard* de *Medstract*. Adicionalmente, el sistema sólo toma el puntaje más alto para cada sigla.

Chang *et al.* almacenaron en una BD relacional aquellas siglas con un puntaje mayor o igual a 0,001. La BD se encuentra en un servidor web, permite búsquedas por sigla o por palabra además de buscar siglas en textos directamente suministrados por el usuario.[58]

[58] *cf*. http://abbreviation.stanford.edu/

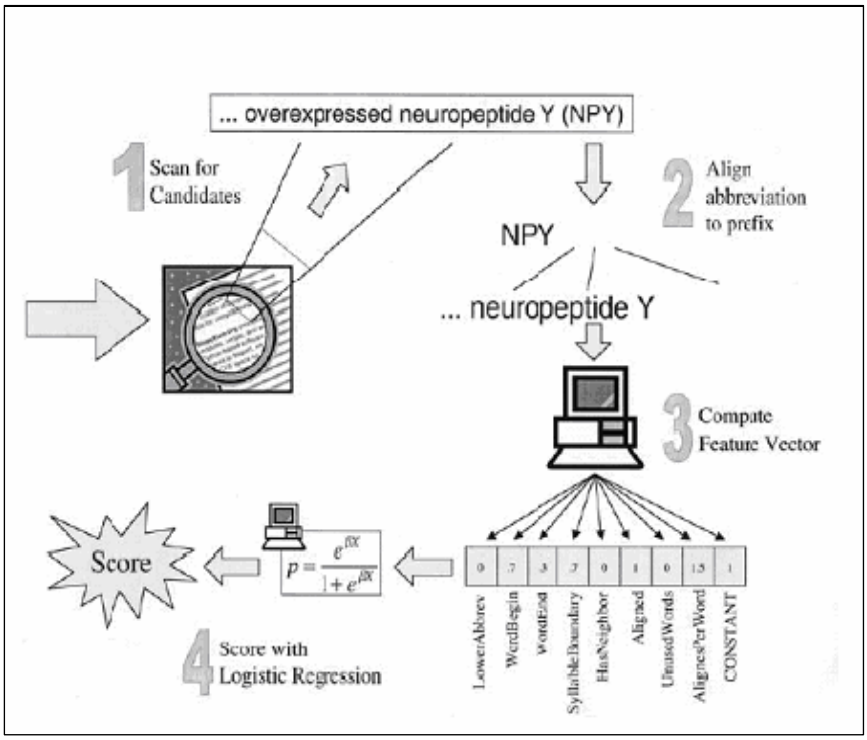


Gráfico 35. Arquitectura del sistema (Chang *et al.*, 2002)

c. Evaluación

Chang *et al*. evaluaron su algoritmo contra el *Gold Standard* de *Medstract*, el cual contiene 168 siglas extraídas de *Medline* y anotadas por expertos. De éstas el sistema identificó correctamente 140, alcanzando 80% de precisión y 83% de exhaustividad.

Los aportes de Chang *et al.* a la investigación en este campo son:

1) El desarrollo de un nuevo algoritmo para la identificación de siglas;
2) La elaboración de un conjunto de rasgos descriptivos de las siglas, y
3) La creación de un nuevo diccionario de abreviaciones en línea a partir de los *abstracts* de la BD *Medline*.

1.2.1.2 A Simple and Robust Abbreviation Dictionary (SaRAD)

SaRAD es el sistema desarrollado por Adar (2002; 2004) orientado a la creación de entradas agrupadas y a la generación de reglas de clasificación para la desambiguación de forma desarrollada. El sistema emplea una técnica modificable que limita el espacio de búsqueda y optimiza el proceso de creación del diccionario.

a. Heurística

1) El patrón más común es "FD (sigla)";[59]
2) Una sigla es una palabra o un conjunto de palabras separadas por guiones donde hay al menos una letra mayúscula;
3) Para determinar la ventana de la forma desarrollada se seleccionan como máximo un número de palabras igual a *n* + palabras *buffer* antes del paréntesis; donde *n* es el número de caracteres de la sigla y el *buffer* es cuatro.[60]

b. Técnica de extracción de siglas

En general, puede decirse que la creación del diccionario SaRAD implica el uso de varios módulos. El primero de ellos se encarga de procesar el corpus para la extracción de las siglas y la generación de los candidatos a forma desarrollada. El segundo toma todas las forma desarrollada y las agrupa por sigla. Los resultados de este paso se usan para hacer remisiones entre siglas y para agrupar las formas desarrolladas.

Posteriormente, se efectúa un segundo agrupamiento de los documentos de *Medline* basado en el análisis de los *Medical Subject Headings* (MeSH) para desambiguar las siglas y completar el diccionario.

59 Según el autor esta conclusión se desprende de un análisis de 5.000 documentos de la base de datos *Medline*.
60 El *buffer* se usa para dar cuenta de las palabras que a veces no se incluyen dentro de la sigla. Esto es importante en casos como el de la sigla AABB donde la FD (*American Association of Blood Banks*) incluye la preposición "*of*".

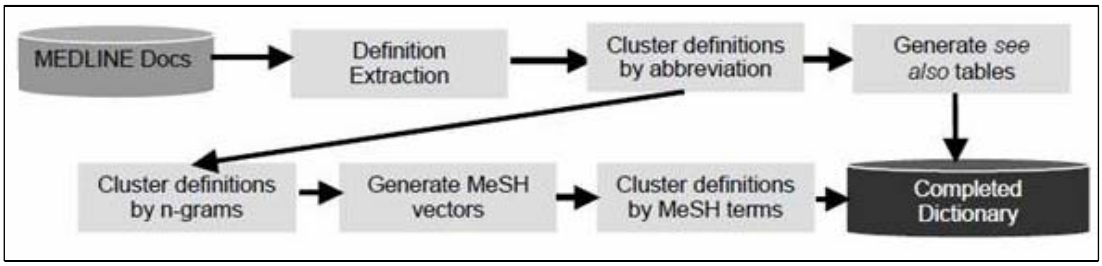


Gráfico 36. Arquitectura general del sistema (Adar, 2002)

En particular, el proceso de extracción de siglas empleado en la construcción de SaRAD consta de los siguientes pasos:

1) Búsqueda de las siglas en el texto fuente;
2) Búsqueda de las formas desarrolladas en la ventana de texto que rodea a la sigla;
3) Generación de las rutas (*paths*) a través del texto que pueden definir la sigla;
4) Puntuación de las rutas para determinar cuál es el mejor candidato a forma desarrollada.

En cuanto al establecimiento de las ventanas para la búsqueda de las forma desarrollada, existen muchas formas posibles de expandir una sigla en un texto, siendo la más común la que corresponde al siguiente patrón:

<texto> FD (sigla) <texto>

De acuerdo con las observaciones hechas en el corpus de *Medline*, se desarrolló el módulo de extensión para localizar una sigla dentro de un paréntesis y una ventana de texto antes de este, tal y como se representa en el siguiente ejemplo:

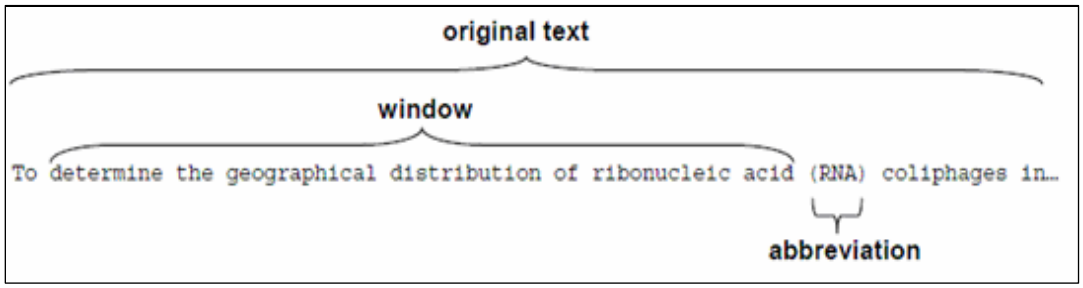


Gráfico 37. Muestra de una ventana de texto para la búsqueda de la forma desarrollada de RNA (Adar, 2002)

Después de establecer la ventana de texto, se buscan los candidatos a forma desarrollada, tarea que se ejecuta en la fase denominada "generación de las rutas o *paths*". Esto se logra mediante la búsqueda hacia adelante de los caracteres que coinciden "en orden" con los de la sigla. Este proceso permite crear la ruta que conduce a la ubicación de los caracteres de la sigla dentro del texto. La siguiente figura muestra tres rutas que corresponden a igual número de candidatos a forma desarrollada de la sigla RNA dentro de la ventana "*determine the geographical distribution of ribonucleic acid*".

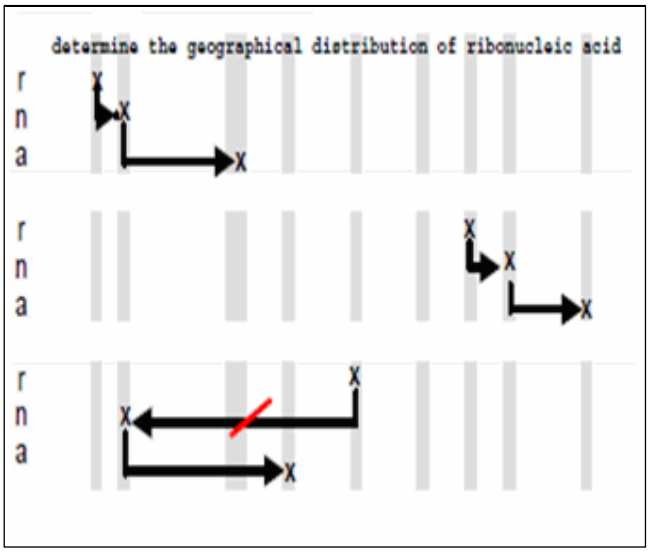


Gráfico 38. Procedimiento del algoritmo para hallar la forma desarrollada en la ventana de texto adyacente (Adar, 2002)

Posteriormente, a las rutas encontradas se les asigna un puntaje para determinar si existe o no una forma desarrollada adecuada. Dado que la aplicación de esta regla no es suficiente por sí sola para la identificación de las siglas, se requiere la aplicación de otras reglas como:

1) Por cada carácter de la sigla que se encuentre al comienzo de una palabra de la forma desarrollada se agrega 1 punto;
2) Por cada palabra extra entre la forma desarrollada y el paréntesis donde se encontró la sigla se resta 1 punto;
3) A las formas desarrolladas que se encuentren justo al lado del paréntesis se agrega 1 punto extra;
4) El número de palabras de la forma desarrollada debería ser menor o igual que el número de caracteres de la sigla. Por cada palabra extra se resta 1 punto.

El punto de equilibrio para las formas desarrolladas correctas es cero. Si el puntaje de un candidato a forma desarrollada es mayor que cero, se considera que éste tiene una alta probabilidad de ser la forma desarrollada correcta. El gráfico 39 muestra un caso de puntuación para los candidatos a forma desarrollada de la sigla RNA.

	Score
determine the geographical distribution of ribonucleic acid	
deteRmiNe the geogrAphical distribution of ribonucleic acid	-4
determine the geographical distribution of ribonucleic acid	-4
determine the geographical distribution of ribonucleic acid	-2
deteRmine the geographical distributioN of ribonucleic Acid	-2
determine the geographical distribution of riboNucleic acid	-2
determine the geographical distribution of ribonucleic acid	0
determine the geogRaphical distribution of riboNucleic Acid	0
determine the geographical distribution of ribonucleic acid	1
determine the geographical distribution of ribonucleic acid	1
determine the geographical distribution of riboNucleic Acid	3*

Gráfico 39. Ejemplo de puntuación para la selección de la forma desarrollada (Adar, 2002)

Una vez realizados los pasos que corresponden al primer módulo, cabe agrupar las formas desarrolladas relacionadas, que normalmente son las que están en plural; ***i.e.***, *Estrogen Receptor* y *Estrogen Receptors*.

El sistema de Adar emplea dos técnicas de *clustering* diferentes. La primera, basada en n-gramas, se encarga de buscar definiciones con raíces similares. La segunda utiliza la lista de descriptores *Medical Subject Headings* (MeSH) para la creación del *cluster* de formas desarrolladas.

La técnica específica de n-grama que se utiliza es la de trigramas, la cual segmenta cada forma desarrollada en grupos de tres letras que se comparan unas con otras, por ejemplo, la forma desarrollada "ABCDE" contiene el grupo de trigramas (ABC, BCD, CDE).

La segunda técnica consiste en tomar cada *cluster*, encontrar los documentos iniciales de los que se extrajeron las definiciones y generar un vector que representa los términos de MeSH.

La parte final del análisis para crear el diccionario SaRAD consiste en crear las remisiones o referencias cruzadas entre las formas desarrolladas relacionadas. Esto es particularmente útil para unidades con variantes tipográficas como el uso de mayúsculas, por ejemplo: ACH, AcH, ACh, Ach, abreviaturas de *acetylcholine*.

Para efectuar las remisiones de las formas desarrolladas y generar la lista de "véase también" en el diccionario, se buscan todas las formas desarrolladas equivalentes en un rastreo a través de todo el diccionario.

Finalmente, para la desambiguación de las siglas se reutilizan los vectores de *MeSH* generados anteriormente para el agrupamiento de las formas desarrolladas.

c. Evaluación

El sistema es bastante limitado puesto que sólo tiene en cuenta los candidatos que se ajustan al patrón "FD (sigla)". Al aplicarse el algoritmo al *Gold Standard* de *Acromed*, el sistema halló 144 siglas. Teniendo en cuenta que el punto de equilibrio para la selección de las formas desarrolladas correctas es cero, el sistema alcanzó una precisión de 86% y una exhaustividad de 88%.

1.2.2 Métodos basados en algoritmos de aprendizaje máquina

1.2.2.1 Teoría universal de la formación de siglas

En la "Teoría universal de la formación de siglas", Zahariev (2004), las siglas se consideran un fenómeno universal cuya formación se rige por preferencias lingüísticas basadas en reglas a nivel de caracteres, fonemas, palabras y frases.

La teoría universal de la formación de siglas se desarrolla a partir de los ejemplos tomados de 15 lenguas con sistemas de escritura diferentes como son: inglés, español, francés, alemán, finlandés, italiano, húngaro, rumano, ruso, búlgaro, hebreo, árabe, farsi, chino y japonés.

El trabajo de Zahariev apunta a la solución de la adquisición y la desambiguación, los dos principales problemas en el tratamiento automático de las siglas. Para la solución de cada problema, se emplea un algoritmo de aprendizaje máquina.

En general, se propone un enfoque modular para el tratamiento de las siglas; para ello ejecutan las siguientes tareas:

1) Identificación de siglas
2) Identificación de formas desarrolladas
3) Concordancia de siglas-formas desarrolladas

4) Desambiguación de siglas.

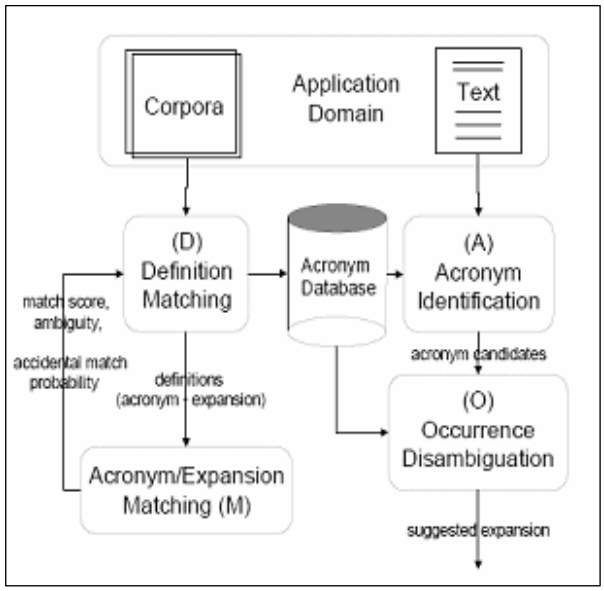


Gráfico 40. Enfoque modular para la adquisición automática de siglas (Zahariev, 2004)

a. Patrones

El sistema de Zahariev reconoce candidatos a sigla que cumplen, en general, con algunas de las siguientes características:

1) Letras mayúsculas;
2) Puntuación interior; *e.g.*: U.S.A., TCP/IP, S-MIME, MIB-s, etc.

Sin embargo, el sistema descarta aquellas siglas que contienen dígitos y caracteres diferentes a la barra y el guión; *e.g.*: 3M, MP+, etc.

A pesar de que el conjunto de reglas de formación de siglas se considera universal, para cada lengua el orden de importancia de las reglas varía. Además, en ciertos casos puede haber reglas inaplicables, por ejemplo, la "concordancia de sílabas" no es aplicable en lenguas como el chino.

Zahariev establece el siguiente conjunto de reglas:

1) Concordancia inicial. Concordancia de un carácter inicial de una palabra de la forma desarrollada con un carácter equivalente en la sigla.
2) Concordancia de morfemas. Concordancia de un carácter inicial de un morfema al interior de una palabra de la forma desarrollada con un carácter equivalente en la sigla.
3) Concordancia de sílabas. Concordancia de un carácter inicial de una sílaba al interior de una palabra de la forma desarrollada con un carácter equivalente en la sigla.
4) Concordancia de grupos de caracteres. Concordancia de un grupo de caracteres consecutivos en una palabra de la forma desarrollada con un grupo equivalente de caracteres consecutivos en la sigla.
5) Concordancia de caracteres internos. Concordancia de un carácter interno en una palabra de la forma desarrollada con un carácter equivalente en la sigla.
6) Omisión de palabras gramaticales o *stopwords*. Omisión en la sigla de una palabra gramatical presente en la forma desarrollada.
7) Omisión de una palabra precedida por signos de puntuación. Omisión en la sigla de una palabra precedida por ciertos signos de puntuación (guión y barra) presentes en la forma desarrollada.
8) Omisión de una palabra. Omisión en la sigla de un carácter que representa una de las palabras de la forma desarrollada.
9) Formación del plural mediante duplicación. Duplicación en la sigla de los caracteres equivalentes de una forma desarrollada pluralizada.
10) Concordancia simbólica. Concordancia de un símbolo, carácter, morfema, grupo de caracteres, palabra o expresión en la forma desarrollada con un carácter o grupo de caracteres en la sigla, siguiendo reglas *ad-hoc*, las cuales son reconocibles generalmente por determinados grupos sociales (por ejemplo: CU por "*see you*"*;* XMAS por "*Christmas*").
11) Migración. En las lenguas donde algunos acentos u otros signos pueden acompañar a los caracteres del alfabeto, los caracteres de la forma desarrollada pueden "migrar" a la sigla como elementos no acentuados. Por ejemplo, en rumano "Î, Ş, Ţ" migran a "I, S, T", respectivamente. De igual modo, en francés

la "É" migra a "E" como se comprueba en el siguiente caso: "*Électricité de France* (EDF)".

12) Flexión. En lenguas con morfología aglutinante, las concordancias de los grupos que representan morfemas enteros pueden flexionarse en la sigla.
13) Concordancia consecutiva. Los caracteres de una sigla pueden concordar consecutivamente, en la misma dirección de los símbolos, caracteres o palabras pertenecientes a la forma desarrollada. Este es el principio que siguen algoritmos como *Longest common subsequence* (LCS), empleado por Taghva & Gilbreth.
14) Inversión. Hay situaciones en que los caracteres constitutivos de una sigla concuerdan con caracteres de la forma desarrollada pero en orden inverso.
15) Préstamo. Las siglas pueden prestarse directamente de otras lenguas en lugar de crearse a partir de la traducción de sus forma desarrollada. Este caso es muy común en las siglas de áreas científico-técnicas. Algunos ejemplos son: HTTP, URL, DNA, etc.

En lo que respecta a la lengua española, Zahariev sostiene que las reglas predominantes en la formación de siglas son:

1) Concordancia inicial; *e.g.*: TLCAN (**T**ratado de **L**ibre **C**omercio de **A**mérica del **N**orte);
2) Omisión de palabras gramaticales; *e.g.*: SICAV (**S**ociedad de **i**nversión de **ca**pital **v**ariable);
3) Concordancia de morfemas; *e.g.*: SIDA (**s**índrome de **i**nmuno**d**eficiencia **a**dquirida);
4) Formación del plural mediante duplicación; *e.g.*: EEUU (Estados Unidos)
5) Concordancia de grupos de caracteres*;* según el autor, esta regla se usa también para combinar nombres propios *e.g.*: Marisa (**Mar**ia **Isa**bel).

b. Técnica de extracción de siglas

Para la detección de siglas Zahariev propone un algoritmo que opera en dos fases sucesivas, a saber: detección de pares de sigla-forma desarrollada y detección de formas desarrolladas. Los resultados de ambas fases se listan.

En la fase de detección de los pares de sigla-forma desarrollada, el proceso comienza con cada ocurrencia de una sigla. A partir de aquí el algoritmo busca el candidato a forma desarrollada en el contexto que la rodea.

El método para hallar los candidatos a forma desarrollada es similar al empleado por el algoritmo canónico-contextual de Larkey.

La búsqueda de la concordancia de letras entre la sigla y la forma desarrollada se realiza mediante un conjunto de reglas flexibles que se aplican sucesivamente en diferente orden en el contexto que rodea al candidato a sigla. Las reglas son:

1) Concordancia inicial. Cuando el carácter inicial de la palabra concuerda con el carácter equivalente en la sigla.
2) Omisión de palabras gramaticales. Se aplica una lista de exclusión de palabras gramaticales (artículos, preposiciones y conjunciones).
3) Concordancia de subsiglas. Cuando una sigla entera está incluida dentro de otra sigla; *e.g.*: VLAN por "*virtual LAN*".
4) Concordancia de prefijo morfológico. Se aplica una lista de prefijos del inglés para buscar concordancias con el prefijo de determinada palabra de la forma desarrollada. Cuando hay coincidencia, tanto la inicial del prefijo como el resto de la palabra, se consideran parte de la sigla; *e.g.*: "*hypertext*" concuerda con "HT", el comienzo de "HTML", la sigla correspondiente a "*Hypertext Markup Language*".
5) Concordancia de prefijo. Cuando las letras del comienzo de una palabra en la forma desarrollada son idénticas al grupo de letras correspondiente en la sigla;

e.g.: las primeras cuatro letras de la palabra "***boot****strap*" concuerdan, al comienzo de la sigla "BOOTP", sigla correspondiente a "*Bootstrap Protocol*".

6) Concordancia orientada a la sílaba. Cuando la o las primeras letras o consonantes en cada sílaba concuerdan con las letras correspondientes en la sigla; *e.g.*: la palabra "*connectionless*" se descompone en las sílabas *con-nec-ti-on-less*, y las iniciales de las sílabas "*con*" y "*less*" concuerdan con "CL" del comienzo de CLNP, una sigla para "*Connectionless Network Protocol*".
7) Omisión de palabras. Cuando las palabras introducidas por signos de puntuación como la barra o el guión se omiten en la concordancia de letras de la sigla; *e.g.*: la palabra "*Level*" se omite en la forma desarrollada "*High-Level Data Link Control*", correspondiente a la sigla HDLC.
8) X. Cuando las concordancias incluyen la letra X dentro de una sigla, se considera que pueden concordar con la palabra "ex" al comienzo de palabras en la forma desarrollada, como es el caso de "XML", la sigla correspondiente a "*Extensible Markup Language*".

c. Evaluación

Zahariev evalúa el rendimiento de su algoritmo por medio de un corpus creado con los *abstracts* de la *Internet Engineering Task Force* (IETF) *Request for Comments* (RFC). El corpus de evaluación, conformado por 681 pares de candidatos sigla-forma desarrollada, arrojó 98,58% de precisión y 93,19% de exhaustividad.

1.2.2.2 Automatic Acronym Identification and Creation of an Acronym Database

Young (2004) desarrolló una técnica de identificación de siglas y una base de datos para su almacenamiento. La investigación se centra en dos dominios: general (noticias) y especializado (biomedicina). En el primer caso emplea el sitio de la BBC y el corpus de la *Agencia Reuters* mientras que en el segundo utiliza la base de datos de *Medline*.

Para la constitución del corpus de evaluación del sistema, Young creó los siguientes módulos:

1) Unidad de compilación de páginas web (*Harvest unit*). Busca y reúne para su procesamiento todos los documentos útiles provenientes de las páginas web;
2) Unidad de filtro para remoción de las etiquetas HTML. Elimina todas las etiquetas HTML de los documentos, dejándolos en texto plano.

a. Patrones

Young basa su trabajo en el vector de rasgos sugerido por Chang *et al*. (2002), el cual se describe a continuación:

1) Minúscula *vs*. mayúscula. Gran parte de las de siglas contienen más letras mayúsculas que minúsculas.
2) Comienzo de palabra. Los listados de siglas existentes muestran que las siglas se crean, generalmente, a partir de las primeras letras de las palabras de la forma desarrollada.
3) Final de palabra. Existen otros sitios en la forma desarrollada de los cuales se pueden tomar letras como son los finales de las palabras.
4) Límite de sílaba. Es un sitio lógico para escoger letras que formen una sigla, en especial cuando el límite de la sílaba es el límite entre dos palabras.
5) Después de letra alineada. Las letras que siguen a la letra previa escogida también pueden hacer parte de la sigla.
6) Letras alineadas. Si hay un alto número de letras no alineadas es probable que el candidato a forma desarrollada no sea el correcto.
7) Palabras omitidas. Es el número de palabras de la forma desarrollada que no concuerdan con ninguna letra de la sigla. El vector de rasgos utilizado descartará cualquier palabra vacía (o *stopword*) en el cálculo de este rasgo.
8) Letras alineadas por palabra. Las listas de siglas muestran que generalmente se alinea una letra de la sigla por una palabra de la forma desarrollada. Un gran número de letras alineadas por palabra muestra que se toman demasiadas letras de palabras individuales, lo cual indica una concordancia adicional.

Aparte de los rasgos establecidos por Chang, Young sugiere la inclusión de un rasgo adicional: la clasificación (*ranking*) por medio de Internet. Para calcular dicha clasificación se emplea un buscador. El cálculo se basa en el porcentaje de páginas halladas para todos los candidatos sigla-forma desarrollada. Esta clasificación es útil puesto que usa Internet como un sistema de votación; cada página que presenta el par sigla-forma desarrollada cuenta como un voto para la elección de la forma desarrollada correcta.

b. Técnica de extracción de siglas

El extractor identifica los candidatos a sigla mediante el uso de la siguiente expresión regular:

[a-zA-Z0-9] * [([A-Z] {2}) ([A-Z] [0-9]) ([0-9] [A-Z])] [a-zA-Z0-9]*

Esta se aplica a una lista de palabras *tokenizadas* por espacios en blanco. La parte central de la expresión; *i.e.*, [([A-Z] {2}) ([A-Z] [0-9]) ([0-9] [A-Z])] indica que una porción del texto debe tener bien dos letras mayúsculas [([A-Z] {2}), una letra mayúscula y un número ([A-Z] [0-9]), o bien un número y una letra mayúscula ([0-9] [A-Z]).

El sistema analiza la ventana de texto que hay a la derecha e izquierda del candidato a sigla, sin importar si hay o no mayúsculas. Cuando se encuentra una sigla se registra su posición y se extrae junto con su ventana de texto para pasarla luego por el módulo de deducción de la forma desarrollada.

La extensión de la ventana de texto varía de acuerdo con la longitud de la sigla. Si la sigla contiene menos de 7 caracteres, la extensión equivaldrá al número de letras de la sigla multiplicado por 2. Si la sigla contiene más de 7 caracteres, se multiplicará por 1,2 con el fin de limitar el número de posibles candidatos a forma desarrollada. En

particular, las siglas extensas producen ventanas de texto extensas lo que significa que podrían inferirse más pares de candidatos sigla-forma desarrollada.

Para la extracción de las formas desarrolladas el sistema sigue los siguientes pasos:

1) Identificación de la forma desarrollada. El módulo de extracción de las formas desarrolladas infiere los candidatos a partir del contexto donde aparece la sigla. Este módulo emplea el algoritmo *Longest common subsequence* (LCS), por su facilidad para generar todos los candidatos sigla-forma desarrollada sin necesidad de usar heurística *ad-hoc*. Los pares de sigla-forma desarrollada se envían posteriormente al módulo de clasificación (*ranking*).
2) Clasificación (*ranking*). El módulo de clasificación asigna un puntaje a los pares sigla-forma desarrollada. De esta manera, el par que tenga el mayor puntaje se convertirá en el correcto. Este módulo se basa en un algoritmo de aprendizaje máquina supervisado. El factor clave en el éxito de este tipo de algoritmo radica en la selección de un vector de rasgos que describe la estructura de la sigla y su contexto.

Para el almacenamiento de los pares de sigla-forma desarrollada, Young ha creado una interfaz de Internet compuesta por la BD de siglas y el sitio web.[61]

La BD almacena todos los pares de sigla-forma desarrollada correctos junto con el porcentaje de probabilidad de exactitud.

La interfaz permite tanto la consulta de la BD de siglas como la sugerencia de nuevas siglas por parte de los usuarios.

En síntesis, Young presenta un diseño de sistema modular, con el que logra que cada sección pueda ejecutarse independientemente. La arquitectura general del sistema es la siguiente:

61 Actualmente no se encuentra disponible para su consulta

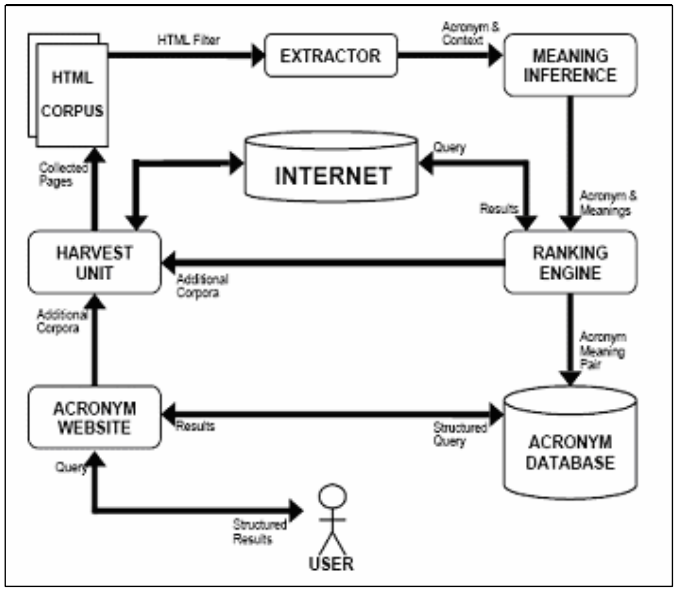


Gráfico 41. Aspecto general del sistema de extracción y generación de la base de datos de siglas (Young, 2004)

c. Evaluación

Para la evaluación del sistema, Young tuvo en cuenta dos casos. En el primero, el sistema procesó artículos del corpus de *Reuters* seleccionados aleatoriamente. En el segundo, el sistema procesó el *Gold Standard* de *Medstract* con el fin de realizar comparaciones directas con los resultados de otros trabajos.

1) Evaluación del corpus de *Reuters*

   El sistema procesó 100 artículos procedentes del corpus de *Reuters* dando como resultado 94% de precisión y 77% de exhaustividad. No obstante, esta autora advierte que, de momento, estos resultados no son comparables ya que no se cuenta con otros trabajos que hayan empleado corpus de noticias.

2) Evaluación contra el *Gold Standard* de *Medstract*

> Young evaluó su sistema con el "*Gold Standard*" de *Medstract* para poder establecer comparaciones con los resultados de otros trabajos.[62] Sin embargo, por limitaciones de tiempo, no se pudo entrenar el sistema que ejecuta el *Gold Standard* de *Medstract* con textos biomédicos sino con artículos noticiosos.

Teniendo en cuenta esta limitación, el sistema de Young logró identificar correctamente 47 pares y marcó otros 5 pares como correctos cuando en realidad se trataba de falsos positivos. El sistema no detectó 38 pares y el resto los descartó bien porque se trataba de siglas duplicadas o porque no se ajustaban a la definición de sigla establecida para este trabajo.

El sistema de Young obtuvo 92,2% de precisión y 55,3% de exhaustividad; porcentajes que, según la autora, son favorables si se considera que su sistema no pudo entrenarse con literatura biomédica.

1.2.2.3 A supervised learning approach

Nadeau & Turney (2005) proponen un sistema de detección de siglas basado en aprendizaje supervisado.

a. Heurística

Un candidato a sigla es un *token* conformado por *1-n* caracteres alfanuméricos, los cuales pueden incluir puntos. La primera letra de la forma desarrollada debe coincidir con la primera letra de la sigla. La forma desarrollada no debe contener signos de puntuación tales como: [] , ; : ¿? ¡!

El sistema emplea 17 rasgos identificables en un candidato a sigla-forma desarrollada, como son:

62 El *Gold Standard* de *Medstract* contiene 168 pares de siglas-forma desarrollada.

1) Número de letras que coinciden con la primera letra de una palabra de la forma desarrollada;
2) La heurística anterior se basa en la longitud de la sigla;
3) Número de letras de la forma desarrollada que están en mayúscula;
4) La heurística anterior se basa en la longitud de la sigla;
5) Longitud (en palabras) de la forma desarrollada;
6) Distancia (en palabras) entre la sigla y la forma desarrollada;
7) Número de palabras de la forma desarrollada que no participan;
8) La heurística anterior se basa en la longitud de la forma desarrollada;
9) Tamaño medio de las palabras en la forma desarrollada que no participan;
10) Categoría gramatical de la primera palabra de la forma desarrollada (es decir, si es preposición, conjunción o determinante);
11) Categoría gramatical de la última palabra de la forma desarrollada (es decir, si es una preposición, conjunción o determinante);
12) Número de preposiciones, conjunciones o determinantes en la forma desarrollada;
13) Número máximo de letras que participan en una sola palabra de la forma desarrollada;
14) Número de letras de la sigla que no participan;
15) Número de dígitos de la sigla y puntos que no participan;
16) Presencia de paréntesis en la sigla o en la forma desarrollada;
17) Número de verbos en la forma desarrollada.

b. Técnica de extracción

El sistema emplea un conjunto de heurísticas tanto para identificar los candidatos a forma desarrollada como para limitar el contexto (ventana) de búsqueda. Luego, los candidatos a sigla se convierten en vectores. Cada vector consiste en 17 características que describen cada miembro de los pares de siglas.

Para determinar si un par sigla-forma desarrollada es correcto, el algoritmo lo confronta con un corpus anotado. Un par sigla-forma desarrollada se etiqueta como válido si hay una coincidencia exacta entre la sigla y la forma desarrollada en el corpus.

c. Evaluación

Una heurística flexible permite la identificación de un gran número de candidatos sigla-forma desarrollada lo cual se traduce en un alto grado de exhaustividad. Estos autores encontraron que los rasgos más productivos en estricto orden son: 1) distancia entre la forma desarrollada y la sigla; 2) número de letras de la sigla que concuerdan con las primeras letras de las palabras de la forma desarrollada y 3) uso de paréntesis.

El algoritmo empleado es el SVM (*Support Vector Machine*) SMO de WEKA, el cual se probó contra el *Gold Standard* de *Medstract*, usado como corpus de evaluación. Con esta prueba se detectaron correctamente 126 pares de sigla-forma desarrollada de un total de 168, lo que implica en términos de rendimiento 92,5% de precisión y 88,4% de exhaustividad.

1.2.2.4 Recognizing acronyms in Swedish texts

Dannélls (2006) implementó un sistema para el reconocimiento de siglas en textos de biomedicina escritos en sueco. El sistema usa una heurística general para identificar y extraer los candidatos sigla-forma desarrollada. Posteriormente, un algoritmo de aprendizaje máquina se encarga de clasificar estos candidatos. Y, por último, los pares clasificados correctamente se almacenan en una base de datos.

a. Heurística

Un candidato a sigla es una cadena de caracteres alfabéticos, numéricos y especiales (guiones o barras).

Un candidato se considera válido cuando cumple con las condiciones 1 y 2 y 3 ó 4, que se enumeran a continuación:

1) Contener al menos dos caracteres;
2) No pertenecer a la lista de palabras rechazadas (*stopwords*);
3) Contener al menos una letra mayúscula;
4) Poseer como carácter final una minúscula o un número.

b. Técnica de extracción de siglas

El método empleado por Dannélls es similar al algoritmo utilizado por Schwartz & Hearst, pero con la ventaja de que puede reconocer pares de sigla-forma desarrollada que no están marcados por paréntesis.

Cuando se encuentra una sigla, el algoritmo busca su forma desarrollada correspondiente en las palabras aledañas. El candidato a forma desarrollada debe cumplir con todas y cada una de las siguientes condiciones:

1) Que al menos una letra de las palabras concuerde con una letra en la sigla;
2) Que la cadena de caracteres de la sigla no contenga signos como: “;”, “:”, “?”, “!”;
3) Que la longitud máxima de la cadena de caracteres sea min(|A|+5, |A|*2), donde |A| es la longitud de la sigla;
4) Que la cadena de caracteres no contenga sólo letras mayúsculas.

De acuerdo con estas condiciones, el proceso de extracción de pares sigla-forma desarrollada consta de dos fases, a saber:

1) Concordancia de paréntesis. En la práctica la mayoría de los pares sigla-forma desarrollada se ajustan a alguno de estos patrones: “FD (sigla)” o “sigla (FD)”. Por lo tanto, el algoritmo extrae los candidatos sigla-forma desarrollada que cumplen con esta condición.

2) Ausencia de concordancia de paréntesis. El algoritmo busca candidatos a sigla que cumplan las cuatro condiciones antes mencionadas y que no se encuentren entre paréntesis. Una vez se detecta un candidato a sigla, el algoritmo rastrea el contexto (ventana) anterior y posterior en busca de un candidato a forma desarrollada. El tamaño de la ventana corresponde al resultado de multiplicar cuatro palabras por el número de letras del candidato a sigla.

La selección del candidato sigla-forma desarrollada correcto se hace mediante la reducción de la ventana donde aparece el candidato a forma desarrollada, así:

1) El algoritmo busca caracteres idénticos entre el candidato a sigla y el candidato a forma desarrollada comenzando desde el final de ambas cadenas de caracteres. El candidato par sigla-forma desarrollada es correcto si satisface las siguientes condiciones:

    a) Que al menos un carácter de la sigla concuerde con un carácter de la forma desarrollada;
    b) Que el primer carácter en la sigla concuerde con el primer carácter de la primera palabra de la forma desarrollada, independientemente de que esté en mayúscula o minúscula.

Dannélls emplea un algoritmo de aprendizaje máquina, el cual requiere que los candidatos a sigla-forma desarrollada se representen como vectores de rasgos. La selección de estos rasgos es importante tanto para el proceso de aprendizaje como para la selección del algoritmo y método de entrenamiento del clasificador.

El cálculo de los vectores de rasgos para describir los pares de sigla-forma desarrollada se basa en los diez rasgos siguientes:

1) La sigla o la forma desarrollada están entre paréntesis (0 falso, 1 verdadero);
2) La forma desarrollada aparece antes de la sigla (0 falso, 1 verdadero);
3) La distancia en palabras (*offset*) entre la sigla y la forma desarrollada;

4) El número de caracteres de la sigla;
5) El número de caracteres de la forma desarrollada;
6) El número de minúsculas en la sigla;
7) El número de minúsculas en la forma desarrollada;
8) El número de mayúsculas en la sigla;
9) El número de mayúsculas en la forma desarrollada;
10) El número de palabras en la forma desarrollada.

Además, Dannélls menciona un rasgo adicional, relacionado con el tipo de predicción; *i.e.*, candidato verdadero (+), candidato falso (-).

La siguiente es una representación de un par sigla-forma desarrollada mediante un vector de rasgos:

Sigla-FD	**Vector de rasgos**										
	Rasgo1	Rasgo2	Rasgo3	Rasgo4	Rasgo5	Rasgo6	Rasgo7	Rasgo8	Rasgo9	Rasgo10	Rasgo11
VCJD-variant CJD	0	0	1	4	11	1	7	3	3	2	+

Tabla 31. representación de un par sigla-forma desarrollada mediante un vector de rasgos

c. Evaluación

El corpus para evaluar el sistema consta de 861 pares de sigla-formas desarrollada, extraídos del corpus MEDLEX (textos de medicina en sueco). Dicho corpus se anotó manualmente con etiquetas XML.

El algoritmo detectó 671 pares, de los cuales 47 eran incorrectos, lo que supone 93% de precisión y 72,5% de exhaustividad. El sistema detectó erróneamente 47 porque:

1) Las palabras que aparecen en la forma desarrollada no tienen una letra correspondiente en la sigla;
2) Las letras en la sigla no tienen una palabra correspondiente en la forma desarrollada; *e.g.*: "*PGA, glycol alginate lösning*";

3) Los caracteres en la forma desarrollada no concuerdan con los caracteres de la sigla.

El análisis de errores mostró las causas por las que el sistema no identificó los 190 pares de sigla-forma desarrollada restantes. Estas son:

1) Las letras de la forma desarrollada no aparecen en la sigla (eso se debe básicamente a que la forma desarrollada aparece traducida al sueco mientras que la sigla se mantiene en inglés);
2) La mezcla de números arábigos con romanos; *e.g.*: "*USH3, Usher type III*";
3) La posición de números/letras;
4) Las siglas de tres caracteres que aparecen en minúsculas.

El algoritmo de aprendizaje máquina de mejor resultado es el IB1, pues clasificó correctamente el 98,8% de los pares sigla-forma desarrollada.

A modo de síntesis, se presenta la siguiente tabla comparativa del rendimiento de los sistemas de detección de siglas basados en métodos estadísticos y de aprendizaje máquina.

Sistema	Método	Autor	Año	Corpus	Ámbito	Precisión	Exhaustiv.
Dicc. en línea de abreviaciones	Estadístico	Chang *et. al*	2002	*Gold Standard* de *Medstract* (168 siglas)	Biomedicina	80%	83%
SaRAD	Estadístico	Adar	2002	*Gold Standard* de *Medstract* (168 siglas)	Biomedicina	86%	88%
s.n.	Aprendizaje máquina	Zahariev	2004	*Abstracts* de *Internet Engineering Task Force-Request for comments* (681 siglas)	Internet-Informática	99%	93%
s.n.	Aprendizaje máquina	Young	2004	100 artículos corpus Agencia *Reuters*	General	94%	77%
				Gold Standard de *Medstract* (168 siglas)	Biomedicina	92%	55%
s.n.	Aprendizaje máquina	Nadeau & Turney	2005	*Gold Standard* de *Medstract* (168 siglas)	Biomedicina	92%	88%
s.n.	Aprendizaje máquina	Dannélls	2006	Corpus de textos en sueco *"MEDLEX"* (861 siglas)	Medicina	93%	72%

Tabla 32. Rendimiento de los sistemas de detección de siglas basados en métodos estadísticos y de aprendizaje máquina

De los datos recogidos en la tabla anterior se desprende que el sistema basado en estadística con mejor desempeño es el de Adar. Así mismo, los sistemas basados en aprendizaje máquina con mejor rendimiento son los de Zahariev y Nadeau & Turney.

1.3 Métodos híbridos

Park & Byrd (2001) emplean un método basado en tres tipos de conocimiento: reglas de formación de siglas, marcadores textuales y palabras clasificadoras.

Las reglas de formación de siglas describen cómo se forma una sigla a partir de su forma desarrollada. Los marcadores textuales son símbolos especiales que se usan para indicar la relación de siglas y forma desarrollada en los textos; *e.g.*: "()", "[]" o "=". Y las palabras clasificadoras indican una fuerte relación entre la sigla y su forma desarrollada; *e.g.*: "*or*", "*short*", "*acronym*", "*stand*", etc.

El sistema realiza cinco procesos, a saber:

1) Detección de siglas
2) Búsqueda de forma desarrollada
3) Aplicación de reglas
4) Concordancia de siglas
5) Selección del mejor candidato a sigla-forma desarrollada.

a. Patrones

Park & Byrd consideran como candidato a sigla aquella cadena de caracteres que cumple con las siguientes condiciones:

1) Primer carácter alfanumérico
2) Longitud entre 2 y 10 caracteres
3) Mínimo una letra mayúscula.

Y se ajusta a las siguientes restricciones:

1) No es una palabra recogida en un diccionario ni aparece como la primera palabra de una oración;
2) No es nombre de persona o lugar;
3) No es una *stopword*.

b. Técnica de extracción de siglas

Cuando se encuentra un candidato a sigla, el algoritmo determina el contexto de búsqueda de la forma desarrollada, el cual tiene una longitud máxima de +10 palabras a la derecha e izquierda del candidato a sigla.

Cuando se encuentra un par sigla-forma desarrollada, se generan los patrones que describen la sigla y la forma desarrollada, así:

1) Patrones para la sigla

   a) Los caracteres alfabéticos se reemplazan con una “c”;
   b) Los caracteres numéricos se reemplazan con una “n”.

Por ejemplo:

Sigla	Patrones
2MASS	ncccc
NEXT	cccc
R&D	cc
SN1987A	ccnc

2) Patrones para la forma desarrollada

   a) *Word* (w)
   b) *Stopword* (s)
   c) *Prefix* (p)
   d) *Headword* (h)
   e) *Number* (n).

Por ejemplo:

FD	**patrones**
Supernova 1987A	phnw
Two-Micron All Sky Survey	wwwww
U.S. Department of Agriculture	wwsw

Teniendo en cuenta lo anterior, los patrones para el par sigla-forma desarrollada “*X2B* (*Hexadecimal to Binary*)” son: (“cnc”, “phsw”).

Posteriormente, se generan las reglas de siglas, las cuales se usan para describir cómo se forma una sigla a partir de su forma desarrollada. Una regla de formación consiste en: un patrón de sigla, un patrón de forma desarrollada y una regla de formación.

Una regla de formación define cómo se forma cada carácter de una sigla a partir de su forma desarrollada. Un elemento en una regla de formación tiene dos valores; *i.e.*, un número de palabra (ubicación de la palabra dentro de la forma desarrollada) y un método de formación (existen 5 métodos de formación).

Los métodos de formación son: "F" (primer carácter), "I" (carácter interior), "L" (último carácter), "E" (carácter exacto, para caracteres numéricos) y "R" (reemplazo de concordancia).

El sistema cuenta inicialmente con una base de 45 reglas de formación de siglas, extraídas a partir del análisis de un corpus de 4.500 siglas del ámbito de la informática.

A continuación se muestran dos ejemplos de reglas de formación de siglas.

	Par sigla-FD
Regla de formación	2-MASS Two-micron All Sky Survey <ncccc, wwwww, (1,R) (2,F) (3,F) (4,F) (5,F)>
Regla de formación	CONTOUR Comet Nuclear Tour <ccccccc, www, (1,F) (1,I) (2,F) (3,F), (3,I) (3,I) (3,L)>

c. Evaluación

El sistema se probó en tres corpus diferentes, ingeniería automotriz, farmacéutica y boletines de prensa de la NASA. El sistema no detectó algunas siglas, básicamente porque:

1) Las formas desarrolladas se encontraban fuera del contexto de búsqueda establecido;
2) El etiquetador de POS hizo una mala interpretación.

La siguiente tabla indica el rendimiento de este sistema en cada uno de los corpus.

Sistema	Método	Autor	Año	Corpus	Ámbito	Precisión	Exhaustiv.
s.n.	Híbrido	Park & Byrd	2001	20.379 palabras (33 siglas)	Ing. automotriz	96,9%	93,9%
				97.000 palabras (63 siglas)	Farmacéutica	100%	95,2%
				83.539 palabras (81 siglas)	Boletines prensa de la NASA	97,4%	93,8%

Tabla 33. Rendimiento del sistema (Park & Byrd, 2001)

El análisis de los datos de esta tabla muestra que los resultados del sistema son bastante parecidos a los obtenidos por los métodos de aprendizaje máquina de Zahariev, mencionados en el apartado anterior.

2. Sistemas de desambiguación de siglas

Se entiende por desambiguación de siglas al mecanismo de selección de la forma desarrollada apropiada para una ocurrencia específica de una sigla en un contexto dado. Por ejemplo, si se pretende recuperar documentos relacionados con SRF, con el sentido de "*Serum Response Factor*", no deberían recuperarse aquellos documentos que contengan la cadena SRF con un significado diferente como *"Spatial Receptive Field"*. A menudo, no es posible desambiguar el sentido de la sigla por medio de expresiones buleanas porque simplemente no existe la forma desarrollada en el documento.

Diferentes estudios llevados a cabo por Liu *et al.* (2001, 2002) muestran que el 33% de las siglas listadas en el *Unified Medical Language System* (UMLS) en 2001 son ambiguas. En un estudio posterior, estos autores demostraron que el 81% de las siglas encontradas en los *abstracts* de Medline eran ambiguas y tenían 16 sentidos en promedio.

Entre los autores que han desarrollado sistemas de desambiguación se encuentran: Pustejovsky (2001), Pakhomov (2002), Yu (2003), Adar (2004), Zahariev (2004), Bracewell *et al.* (2005), Gaudan *et al.* (2005) y Joshi *et al.* (2006).

En general, los métodos de desambiguación de siglas realizan el siguiente proceso:

1) Uso de un lexicón para la compilación de las siglas y sus sentidos (o formas desarrolladas);
2) Cómputo del contexto de uso para cada sentido;
3) Entrenamiento de un algoritmo de aprendizaje máquina con el contexto de cada sentido.

Estos pasos pueden observarse en el siguiente gráfico de Gaudan *et al.* (2005).

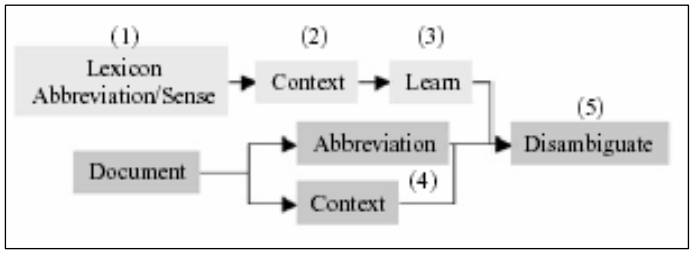


Gráfico 42. Proceso de desambiguación de siglas descrito por Gaudan *et al.* (2005)

A continuación se describen dos de los sistemas de desambiguación con los que se cuenta en la actualidad.

2.1 Polyfind

Pustejovsky *et al.* (2001) desarrollaron el algoritmo para la desambiguación de siglas *Polyfind*. Para la evaluación de este se escogió la sigla SRF con diez formas desarrolladas diferentes y se recogieron todos los *abstracts* en Medline que contuvieran estos pares de sigla-forma desarrollada. Los *abstracts* se agruparon de acuerdo con la

forma desarrollada que contenían, resultando diez grupos, que servirían como plantillas de documento contra las cuales se evaluarían las ocurrencias ambiguas de SRF. Posteriormente, se recogieron 42 *abstracts* en los cuales esta sigla aparecía sin la forma desarrollada. Estas ocurrencias de *SRF* se desambiguaron manualmente y se dividieron en cuatro grupos, de acuerdo con la forma desarrollada a la que correspondían: *Serum Response Factor*, *Subretinal Fluid*, *Surfactin* y *C Elegans surface antigen gene mutations.*

Los *abstracts* de cada uno de estos cuatro grupos se usaron más adelante como consultas, y se evaluaron contra los diez conjuntos de formas desarrolladas. Tanto la búsqueda como los vectores de la plantilla del documento incluían los valores de los *tokens* del título, los nombres de los autores, el título del *journal* y el cuerpo del *abstract*.

Pustejovsky *et al.* emplearon como estándar el método de puntuación "atc". Posteriormente, se computó la medida de similitud entre la búsqueda y cada uno de los conjuntos de formas desarrolladas. La consulta se consideraba desambiguada correctamente si el conjunto con la forma desarrollada correcta de SRF obtenía el mayor puntaje de similitud.

En definitiva, los resultados preliminares de *Polyfind* demuestran que la aplicación de un modelo de vector espacial para la desambiguación de los sentidos de siglas polisémicas es prometedor, *Polyfind* alcanzó 97,2% de exactitud en la desambiguación.

2.2 Automatic resolution of ambiguous abbreviations in Biomedical texts

Yu *et al*. (2003) han implementado un algoritmo para desambiguar las siglas en los *abstracts* de la base de datos Medline. Este sistema se basa en el uso de máquinas de soporte vectorial (SVM) y en la hipótesis de que "todas las ocurrencias de una sigla

dentro de un *abstract* tienen la misma forma desarrollada". En la evaluación, este sistema alcanzó una exactitud de 87%.

Las SVM emplean como corpus de entrenamiento un corpus etiquetado. Para las tareas de desambiguación, el corpus contiene vectores, donde cada vector es una descripción de la ocurrencia de una abreviación. Dicho vector tiene la forma de rasgos (*feature*$_1$, *feature*$_2$....*feature*$_n$, *label)*, donde *label* representa qué forma desarrollada está siendo usada en una ocurrencia de una sigla y *feature*$_1$, *feature*$_2$....*feature*$_n$ describen el contexto donde aparece la sigla; es decir, las dos palabras a la derecha e izquierda de cada ocurrencia de la sigla. Yu *et al.* lo ilustran así:

"OBJECTIVES: The aim of the present study was to assess the **contribution of** angiotensin-converting enzyme (ACE) **inhibitor therapy** to bradykinin-induced tissue type plasminogen activator (t-PA) release in patients with heart failure (HF) secondary to ischemic heart disease. BACKGROUND: Bradykinin is a potent endothelial cell stimulant that causes vasodilatation and t-PA release. In large-scale clinical **trials, ACE inhibitor therapy** prevents ischemic events....".

Los vectores extraídos de este fragmento de *abstract* para la FD "*angiotensin-converting enzyme*" de la sigla "ACE" son los siguientes:

L2= contribution, L1=of, R1=inhibitor, R2=therapy

L2=trials, L1= ",", R1=inhibitor, R2=therapy

En síntesis, El método de Yu *et al.* consiste en:

1) Emplear el algoritmo SVM, de clasificación supervisada, para predecir la forma desarrollada probable de una sigla;
2) Extraer mediante SVM los datos usando las formas desarrolladas de las siglas desambiguadas;
3) Utilizar la hipótesis " un sentido por documento".

El algoritmo propuesto es el siguiente:

Input:

(1) A set of Medline abstracts (SMA)

(2) Abbreviation dictionary containing the abbreviations and their long forms disambiguated (AD)

Output:

Set of vectors in form (Feature$_1$, Feature$_2$, …, Feature$_n$, A, LF), where A is an abbreviation, and LF is one of the long forms of the abbreviation A (every vector represents an occurrence of the abbreviation A in context (Feature$_1$, Feature$_2$, …, Feature$_n$,) and the abbreviation A in the context has the sense (long form) LF)

Algorithm:

```
FOR (X ∈ SMA) DO  //X is an abstract in the set SMA
{  FOR (A ∈ AD) DO
   //A is an abbreviation in the dictionary AD
   { IF ((a long forms of abbreviation A is found in X)
       AND (the found long form is LF)
       AND (no other long form of A is also found in X))
     {  FOR (each of the occurrences of A in X) DO
       {  Generate vector (Feature1, Feature2, …, Featuren,
                           A, LF),
          where Feature1, Feature2, …, Featuren, is the
          context of the occurrence;
       }
     }
}
```

Fig. 5. Algoritmo empleado (Yu *et al.*, 2003)

Hasta aquí se han tratado dos de los tres apartados principales de que consta este capítulo; *ie.*, los sistemas de reconocimiento de siglas y los sistemas de desambiguación más relevantes. El tercer y último apartado se deriva de la información recopilada principalmente del estudio de los sistemas de reconocimiento de siglas, pero, más importante aún de la observación minuciosa de las concordancias (o ventanas de texto) donde aparecía cada una de la siglas de nuestro corpus. Esta observación ha permitido, por un lado, conocer en profundidad cuáles elementos se suelen tener en cuenta a la hora de formar una sigla y, por otro lado, conocer de qué manera suelen aparecer las siglas y sus formas desarrolladas dentro de los textos.

El tercer apartado, que se presenta a continuación, se ha dedicado a los elementos que debe incorporar un detector de siglas en español. En primer lugar, un detector de siglas debe conocer previamente las reglas de formación de siglas para identificar los candidatos a sigla. En segundo lugar, debe conocer las reglas de concordancia de caracteres para determinar si se trata de un candidato a par sigla-forma desarrollada. En tercer lugar, debe conocer los patrones de reconocimiento tanto de siglas como de pares de sigla-forma desarrollada dentro de las ventanas de texto. Todos estos elementos se detallan a continuación.

3. Criterios para el diseño de un modelo de detector de siglas para el español

El primer paso que se debe cumplir durante el diseño de un sistema de reconocimiento de siglas consiste en determinar las reglas que rigen estas unidades; es decir, la "heurística". A partir del análisis de los trabajos citados anteriormente, se pueden deducir las reglas de formación y de concordancia, así como los patrones de identificación de las siglas.

3.1 Reglas de formación de siglas

Partiendo de la concepción de Park & Byrd (2001), consideramos que una regla de formación es la manera como se forma cada carácter de una sigla a partir de su forma desarrollada. En este sentido, puede decirse que existen 2 tipos de reglas de formación de siglas: básicas y complementarias.

3.1.1 Reglas básicas de formación de siglas

1) Mínimo 2 caracteres y máximo 10
2) Máximo 2 palabras

3) Mínimo una letra mayúscula
4) Primer carácter alfanumérico
5) Exclusión de los signos ; : ? !

3.1.2 Reglas complementarias de formación de siglas

1) Inclusión. Una sigla puede incluir signos como barras, puntos o guiones;
2) Plural. Para formar el plural de la sigla puede añadirse una "s", caso recurrente en siglas creadas en inglés como *retinoid x receptors* (RXRs); o duplicarse los caracteres de la sigla, caso propio de lenguas como el español; ***e.g.***: EEUU;
3) Caracteres alfanuméricos. El primer o último carácter de la sigla puede ser una letra o un número.

3.2 Reglas de concordancia de pares sigla-forma desarrollada

3.2.1 Concordancia de caracteres

1) Concordancia inicial. Concordancia de un carácter inicial de una palabra de la forma desarrollada con un carácter equivalente en la sigla.
2) Concordancia de morfemas. Concordancia de un carácter inicial de un morfema al interior de una palabra de la forma desarrollada con un carácter equivalente en la sigla.
3) Concordancia de sílabas. Concordancia de un carácter inicial de una sílaba al interior de una palabra de la forma desarrollada con un carácter equivalente en la sigla. Es decir, la primera letra(s) o sílaba(s) en cada sílaba concuerda(n) con las letras equivalentes de la sigla; ***e.g.***: la palabra "*connectionless*" se descompone en las sílabas *con-nec-ti-on-less*, y las iniciales de las sílabas "*con*" y "*less*" concuerdan con "CL" del comienzo de CLNP, la sigla de "*Connectionless Network Protocol*".

4) Concordancia de grupos de caracteres. Concordancia de un grupo de caracteres consecutivos en una palabra de la forma desarrollada con un grupo equivalente de caracteres consecutivos en la sigla.
5) Concordancia de caracteres internos. Concordancia de un carácter interno en una palabra de la forma desarrollada con un carácter equivalente en la sigla.
6) Concordancia simbólica. Concordancia de un símbolo, carácter, morfema, grupo de caracteres, palabra o expresión en la forma desarrollada con un carácter o grupo de caracteres en la sigla, siguiendo reglas *ad-hoc*, las cuales son generalmente reconocibles por determinados grupos sociales (por ejemplo, CU por "*see you*"*;* XMAS por "*Christmas*").
7) Concordancia consecutiva. Concordancia de caracteres de la sigla en la misma dirección de los símbolos, caracteres o palabras pertenecientes a la forma desarrollada. Este es el principio que siguen algoritmos como *Longest common subsequence* (LCS), empleado por Taghva & Gilbreth.
8) Concordancia de subsiglas. Una sigla entera puede estar incluida dentro de otra sigla; *e.g.*: VLAN por "*virtual LAN*".
9) Concordancia de prefijo morfológico. Se aplica una lista de prefijos del inglés para buscar concordancias con el prefijo de determinada palabra en la forma desarrollada. Cuando hay coincidencia, tanto la inicial del prefijo como el resto de la palabra se consideran parte de la sigla; *e.g.*: "*hypertext*" concuerda con "HT", el comienzo de "HTML", la sigla correspondiente a "*Hypertext Markup Language*".

3.2.2 Inversión

Hay situaciones en que los caracteres constitutivos de una sigla concuerdan con caracteres de la forma desarrollada pero en orden inverso.

3.2.3 Inserción

Una palabra está presente en la forma desarrollada de una sigla, pero no ha sido usada en la formación de la sigla; *e.g.*: ***t****hyroid* ***hormone r****eceptor* (TR).

3.2.4 Omisión

Existen tres casos típicos de omisión, a saber:

1) Omisión de palabra. Una palabra inexistente en la forma desarrollada de una sigla se usa al momento de formar la sigla; *e.g.*: *[**human**] estrogen receptor* (hER).
2) Omisión de palabra gramatical. Un artículo, preposición o conjunción presente en la forma desarrollada puede desaparecer al formar la sigla; *e.g.*: VIH (**V**irus de la **i**nmunodeficiencia **h**umana).
3) Omisión de palabras separadas por signos de puntuación. Palabras introducidas por signos de puntuación como la barra o el guión pueden omitirse; *e.g.*: la palabra "*Pacific*" presente en "*Asia-Pacific Association for Machine Translation*" se omite en su sigla AAMT.

3.2.5 Sigla recursiva

La forma desarrollada de una sigla contiene a su vez otra sigla o abreviatura; *e.g.*: ***CREB-b****inding* ***p****rotein* (CBP).

3.3 Patrones para la identificación de pares sigla-forma desarrollada

Un patrón es aquello que se toma como punto de referencia para valorar otras cosas de la misma especie. En este sentido, el presente trabajo considera dos tipos de patrones, a saber: patrones para identificación de candidatos a sigla y patrones para identificación de candidatos pares sigla-forma desarrollada.

3.3.1 Patrones para la identificación de candidatos a sigla

Con base en el análisis de los trabajos reseñados aquí, en especial en Larkey *et al.* (2000), y en el análisis de nuestro corpus, se establecen los siguientes patrones de identificación de siglas.

1) Patrón: (U {sep})2-9S

   U= *Uppercase* o mayúscula

   {sep}= punto o punto seguido por un espacio

   2-9= rango de caracteres

   S= marca de plural (No todas las siglas la contienen. Es muy frecuente en las siglas creadas en inglés).

   Una sigla puede presentar entre 2 y 9 caracteres en mayúscula, puede contener puntos y puede emplear la marca de plural; *e.g.*: U.S.A, U.S.A.'s.

2) Patrón: U2-9S

   U= mayúscula

   2-9= rango de caracteres

   S= marca de plural (No todas las siglas la contienen. Es muy frecuente en las siglas creadas en inglés).

   Una sigla puede presentar entre 2 y 9 caracteres e ir acompañada de la marca de plural; *e.g.*: USA, USA's.

3) Patrón: U*{dig}U+

   U= mayúscula

   *= 0 o más ocurrencias

   {dig}= número entre 1 y 9, opcionalmente seguido de un guión

   + = 1 o más ocurrencias de un carácter.

Una sigla puede presentar cero o más ocurrencias de caracteres en mayúscula seguidos de un número entre 1 y 9 y una o más ocurrencias de caracteres en mayúscula; *e.g.*: 3D, 3-D, I3R.

4) Patrón: U+L+U+

   U= mayúscula

   + = 1 o más ocurrencias de un carácter

   L= *Lowercase* o minúscula.

   Una sigla puede contener uno o más caracteres en mayúscula seguidos de uno o más caracteres en minúscula seguidos de uno o más caracteres en mayúscula; *e.g.*: DoD.

5) Patrón: U+[/-]U+

   U= mayúscula

   + = 1 o más ocurrencias de un carácter

   [/-]= carácter separador barra o guión.

   Una sigla puede estar formada por uno o más caracteres en mayúscula separados por un guión o barra seguido de uno o más caracteres en mayúscula; *e.g.*: AFL-CIO.

3.3.2 Patrones para la identificación de pares sigla-forma desarrollada hallados en los trabajos analizados

Para establecer los patrones más frecuentes para la identificación de pares de candidatos sigla-forma desarrollada, partimos del análisis de los trabajos reseñados. La mayoría de los autores, concretamente, Larkey *et al.* (2000), Pustejovsky *et al.* (2001), Schwartz & Hearst (2003), Nenadić *et al.* (2002), Adar (2004), Nadeau & Turney (2005) y Dannélls

(2006), coincide en que el patrón más frecuente es **FD (SIGLA)**, hecho que igualmente hemos constatado en nuestro corpus. Los patrones encontrados son los siguientes.[63]

1) FD (SIGLA)
2) FD, SIGLA,
3) SIGLA (FD)
4) FD, SIGLA.
5) SIGLA, FD,
6) "FD" (SIGLA)
7) SIGLA = FD
8) SIGLA – FD
9) (FD) SIGLA
10) (SIGLA) FD
11) SIGLA or FD
12) FD or SIGLA
13) SIGLA stands for FD
14) SIGLA is an acronym for
15) FD known as the SIGLA
16) FD "SIGLA"
17) "SIGLA" FD

3.3.3 Patrones para la identificación de pares sigla-forma desarrollada hallados en el corpus de este estudio

El análisis de nuestro corpus ha permitido identificar los siguientes patrones:

	Patrón	GH	MA	Presente en trabajos de otros autores
		Frecuencia	Frecuencia	
1	FD (SIGLA)	295	192	Sí
2	SIGLA (FD)	102	52	Sí
3	SIGLA ("FD")	13	1	No
4	SIGLA, FD	12	6	No
5	"FD" (SIGLA)	9	8	Sí
6	(SIGLA, FD)	9	2	No
7	FD o SIGLA	8	-	Sí
8	FD (SIGLA,)	6	-	No
9	(FD, SIGLA)	5	1	No

63 Es de notar que estos autores han hallado estos patrones en corpus de textos en inglés.

10	FD, SIGLA	4	4	No
11	(SIGLA, del inglés FD)	3	-	No
12	(SIGLA; FD)	3	-	No
13	FD, o SIGLA,	3	-	No
14	SIGLA, o FD	3	-	No
15	(“FD” o SIGLA)	2	-	No
16	(SIGLA o FD)	2	-	No
17	FD (abreviado, SIGLA)	2	-	No
18	La abreviatura para FD es SIGLA	2	-	No
19	SIGLA (del inglés “FD”)	2	-	No
20	SIGLA (del inglés FD)	2	-	No
21	SIGLA (siglas en inglés de FD)	2	-	No
22	SIGLA o FD	2	1	Si
23	SIGLA, que es la abreviatura de “FD”	2	-	Si
24	“FD”, SIGLA	1	-	No
25	(FD, o SIGLA)	1	-	No
26	(FD/SIGLA)	1	-	No
27	(FD: SIGLA)	1	1	No
28	(SIGLA del inglés FD)	1	-	No
29	(SIGLA, “FD”)	1	-	No
30	(SIGLA, de “FD”)	1	-	No
31	(SIGLA, por “FD”)	1	-	No
32	(SIGLA: FD)	1	1	No
33	FD (“SIGLA”)	1	-	No
34	FD (abreviado como SIGLA)	1	-	No
35	FD (abreviado SIGLA)	1	-	No
36	FD, comúnmente conocido por SIGLA	1	-	No
37	FD, o de forma abreviada, SIGLA	1	-	No
38	La abreviatura SIGLA significa FD	1	-	No
39	SIGLA “FD”	1	-	No
40	SIGLA (abreviatura de FD)	1	-	No
41	SIGLA (acrónimo del inglés FD)	1	-	No
42	SIGLA (del inglés, “FD”)	1	-	No
43	SIGLA (que abrevia FD)	1	-	No
44	SIGLA significa FD	1	-	Sí
45	SIGLA, abreviatura para “FD”	1	-	No
46	SIGLA, acrónimo de FD	1	-	No
47	SIGLA, de las iniciales inglesas de FD	1	-	No

48	SIGLA, FD.	1	-	Sí
49	SIGLA; FD	1	-	No

Tabla 34. Patrones de identificación de pares de sigla-forma desarrollada para el español

Algunos de estos patrones coinciden con los de los autores analizados más arriba; otros sólo se han hallado en nuestro corpus, lo cual puede deberse a las características propias del discurso de GH y MA en lengua española. Se destaca igualmente la amplia variedad de patrones empleados en los textos sobre genoma humano frente a los empleados en los textos de medio ambiente.

Tanto las reglas de formación como los patrones para la identificación de siglas son elementos esenciales a la hora de diseñar un extractor de siglas. Además, cabe tomar decisiones sobre el método que se empleará para la adquisición. Como se ha mencionado antes, un sistema de detección-extracción puede basarse en: patrones, estadística, aprendizaje máquina o en una combinación de estos.

Independientemente del método que se emplee, el sistema deberá establecer la longitud de la ventana (contexto) a la derecha e izquierda del candidato a sigla, de modo que pueda identificarse el candidato a par sigla-forma desarrollada. En la mayoría de los estudios se ha adoptado como longitud de ventana la siguiente: (|A|+5) o (|A|*2) palabras; aunque otros como Larkey *et al.* han establecido como criterio las 20 palabras anteriores y posteriores a la sigla.

4. Conclusiones

A lo largo de este capítulo se ha mostrado el estado de la cuestión en lo referente a los sistemas de detección y extracción de siglas. Del análisis realizado se derivan las siguientes conclusiones:

1) Las siglas son un fenómeno presente en todas las lenguas escritas, de ahí su carácter universal.
2) La motivación para investigar sobre técnicas y métodos de detección y extracción de este tipo de unidades proviene de las necesidades de ámbitos como: inteligencia artificial (IA), minería de datos (DM), recuperación de información (RI) y procesamiento del lenguaje natural (PLN).
3) El campo de la biomedicina es el que más esfuerzos ha dedicado a la investigación para el desarrollo de sistemas de extracción de siglas. Autores como Chang, Nadeau, Schwartz, Adar, Dannélls y Zahariev, han trabajado en esta área. Otros ámbitos en los que se ha experimentado con extracción de siglas son: automoción, farmacéutica (Park & Byrd), biología molecular (Nenadić), prensa (Young) y medio ambiente (Taghva & Gilbreth).
4) Casi todos los sistemas de extracción de siglas estudiados se han creado para aplicarse a la lengua inglesa. A excepción del estudio de Dannélls (2005; 2006), orientado a las siglas de textos médicos en sueco, y del estudio de Zahariev (2004), la revisión de la bibliografía no ha arrojado luz sobre la existencia de sistemas similares para otras lenguas.
5) En la literatura revisada sobresalen dos motivaciones principales a la hora de crear un extractor de siglas: a) para alimentar automáticamente BD de siglas y, de esta forma, mantenerlas actualizadas; y b) para facilitar la tarea de extracción o recuperación de información en un campo de conocimiento dado.
6) El proceso de extracción de siglas consta de tres fases principales: a) identificación de las siglas; b) identificación de los pares sigla-forma desarrollada, y c) desambiguación.
7) Los sistemas de detección y extracción de siglas actuales se basan en tres métodos diferentes, a saber: a) patrones; b) estadística, y c) aprendizaje máquina. También puede darse el caso de que se empleen combinaciones de los sistemas anteriores, es decir, sistemas híbridos.
8) Los sistemas basados en aprendizaje máquina junto con los híbridos se perfilan como los de mejor rendimiento. Sin embargo, no se debe pasar por alto que todos estos sistemas han sido pensados para analizar textos en lengua inglesa, por lo que se desconoce la eficacia de su aplicación en lenguas como el español.

9) A raíz de la circunstancia antes mencionada, se debe tener en cuenta que, por ejemplo, para el desarrollo de los patrones de detección de siglas en español, es necesario considerar un conjunto de patrones mixto, porque se pueden dar los siguientes casos:

    a) La sigla y la forma desarrollada aparecen en español
    b) La sigla aparece en lengua extranjera y la forma desarrollada en español
    c) La sigla aparece en español y la forma desarrollada en lengua extranjera

10) Los patrones más productivos para la identificación de pares sigla-forma desarrollada son:

    a) FD (SIGLA)
    b) SIGLA (FD)

11) Por último, un sistema para el español debe incorporar un número mayor de patrones de identificación de pares de sigla-forma desarrollada, tal y como se evidencia en la tabla 32.

Capítulo 9

Capítulo 9

Conclusiones generales y posibles líneas de trabajo futuro

1. Conclusiones

Las siglas son un mecanismo de reducción léxica creado en la antigüedad, popularizado por los romanos y potenciado por la ciencia y la técnica a partir de las revoluciones industrial, tecnológica y mediática acaecidas durante el siglo XX. Su uso responde a cuestiones de economía lingüística, mnemotecnia, estilística o editoriales tanto en el discurso general como en el especializado. De ahí que sean objeto de interés de diversos campos de conocimiento como la traducción, la lexicología, la terminología, la redacción técnica, los LSP o la lingüística computacional.

Como se ha expuesto al comienzo de esta tesis, nuestro objetivo general ha sido doble. Por un lado, se ha pretendido analizar los aspectos teórico-descriptivos del fenómeno de la siglación; y por otro, proponer los principios básicos para el diseño de una aplicación que permita el reconocimiento de siglas en corpus en español.

En el plano teórico hemos encontrado que, aunque el fenómeno de la siglación está presente prácticamente en todas las lenguas, en español el tema ha sido tratado de manera tangencial, en especial si se mira desde la óptica del discurso especializado. La mayoría de los estudios se han llevado a cabo desde la lexicología, donde destaca el trabajo de Rodríguez (1981). En él se coteja el comportamiento lingüístico de las siglas con el de los vocablos normales de la lengua. Para ello se basa en textos escritos

tomados de diarios, revistas y glosarios pertenecientes al discurso general. Y desde la terminología, donde sobresale el trabajo de Fijo (2003). En él se intenta de describir los aspectos relacionados con la creación, uso y traducción de las siglas como términos pertenecientes al ámbito de la enfermería. Se trata de un estudio contrastivo inglés-español cuya metodología consiste en la recopilación de un corpus paralelo de 50 textos para el análisis, aplicación del método terminográfico.

A pesar de los trabajos realizados hasta el momento, no se cuenta con estudios contrastivos sobre siglas en el discurso especializado en español y menos aún con criterios para sistemas de detección y extracción a partir de textos en esta lengua. Nuestra investigación constituye pues un intento por iniciar el estudio de este campo en español a la vez que pretende despertar el interés para que se desarrollen trabajos similares, en especial en la parte aplicada.

Aparte de la escasez de investigaciones teóricas y aplicadas sobre el tema de la siglación en español, cabe destacar el problema de la falta de consenso en la delimitación de los conceptos de sigla y demás unidades de reducción léxica. Este hecho ha supuesto una dificultad metodológica en este trabajo a la hora de seleccionar las unidades para el corpus. De ahí que nuestra investigación partiera del estado de la cuestión para determinar cuál era la definición y tipología de siglas más integradora a seguir.

En el plano descriptivo, nos hemos propuesto observar las siglas y sus características lingüísticas y estadísticas. En cuanto a los rasgos lingüísticos, observamos nuestro objeto de estudio desde una perspectiva terminológica. Para ello, nos apoyamos en la Teoría Comunicativa de la Terminología (TCT), la cual tipifica las siglas que corresponden a la reducción de un término de origen complejo (*i.e.*, ADN = ácido desoxirribonucleico) como unidades terminológicas, que son a su vez unidades de conocimiento especializado (UCE).

El análisis lingüístico de las siglas ha servido tanto para la descripción de las unidades como para el aporte de pistas para el refinamiento de los sistemas de identificación de siglas.

En lo concerniente a la descripción se puede concluir que la totalidad de las siglas tiene como núcleo a un nombre, hecho que corrobora el carácter nominal de estas unidades. Los datos han mostrado que la mayoría de las siglas tienen como núcleo a un nombre común. Por tanto, se constata que las siglas, al igual que los nombres, tienen género y número, pueden combinarse con otras categorías gramaticales como los adjetivos, pueden ser sujetos u objetos de verbo y pueden ser sometidas a procesos de derivación y flexión. Si bien es cierto que la presencia de estos últimos es muy baja en las siglas, su utilización constituye un indicio de que la sigla ha iniciado el proceso de lexicalización. En síntesis, las siglas son elementos léxicos funcionales producto de la creatividad léxica. Son propias tanto del discurso general como del especializado. Su flexibilidad radica en que pueden funcionar como nombres, algunas veces como adjetivos y, por tanto, dar origen a compuestos y derivados.

A partir del análisis lingüístico se han detectado algunos problemas que deben ser tenidos en cuenta durante la fase de diseño de un detector de siglas. En primer lugar, aunque la forma canónica de creación de las siglas establece que éstas se deben crear con los caracteres iniciales de cada uno de los elementos de la forma desarrollada, este no siempre es el caso. De hecho, un amplio número de siglas se crea elidiendo palabras gramaticales y/o agregando más de un carácter inicial. Adicionalmente, en algunos casos, las siglas que han sido prestadas de otra lengua aparecen en los textos con su forma desarrollada traducida o viceversa. Todo esto lleva a la falta de correspondencia entre los caracteres que forman el par sigla-forma desarrollada, y que suponen una dificultad para los sistemas de detección de siglas, en especial para los basados en algoritmos como el *Longest common subsequence*. Por consiguiente, la inclusión de los diferentes tipos de correspondencia sigla-forma desarrollada puede ayudar a un determinado sistema de extracción a mejorar sus medidas de rendimiento, es decir, su precisión y exhaustividad.

Las siglas tampoco son ajenas a fenómenos semánticos como la sinonimia y la homonimia. Los casos de sinonimia suelen darse a causa del préstamo o traducción de siglas, generalmente del inglés. Los casos de homonimia son más problemáticos por cuanto producen ambigüedad e impiden determinar cuál es el verdadero significado de una sigla cuando ésta no va acompañada de su forma desarrollada. La magnitud del problema ha quedado al descubierto en corpus como el de *abstracts* de Medline donde, sólo en el año 2001, se encontró que cada sigla podía corresponder en promedio a 16 formas desarrolladas. Hasta el momento, el método predominante ha sido la desambiguación manual, es decir, la llevada a cabo por humanos y que consiste en la lectura de los contextos de aparición de las siglas para establecer su verdadero significado. No obstante, empiezan a aparecer sistemas de detección de siglas que incorporan módulos de desambiguación, basados en su mayoría en diccionarios de siglas previamente confeccionados.

Las siglas de cada ámbito de conocimiento resultan de difícil comprensión para el lector, en especial para el lego. Generalmente, esta limitación se da por dos motivos: 1) porque la sigla encapsula un sintagma pleno (forma desarrollada), hecho que genera opacidad cuando se desconoce la relación de equivalencia entre dicho sintagma y la sigla, y 2) porque la sigla puede generar ambigüedad cuando se desconoce el verdadero significado dentro del contexto en el que se encuentra. A pesar de que este problema se puede evitar con la inclusión de la forma desarrollada, ésta puede producir otro tipo de fenómenos como la variación denominativa y la redundancia dentro del texto.

De lo expuesto anteriormente, se infiere que las siglas inciden en el discurso especializado porque introducen variación denominativa, reflejada tanto por la expresión de sus formas desarrolladas (variantes formales) como por sus variantes por traducción. Este hecho, confirma uno de los principios de la Teoría Comunicativa de la Terminología (TCT) que establece que: "Todo proceso de comunicación comporta inherentemente variación, explicitada en formas alternativas de denominación del mismo concepto (sinonimia) o en apertura significativa de una misma forma (polisemia). Este principio es universal para las unidades terminológicas, si bien admite

diferentes grados según las condiciones de cada tipo de situación comunicativa...” (Cabré, 1999: 85).

Los aspectos lingüísticos que pueden ayudar en el refinamiento de los sistemas de detección de siglas son de diversa índole. En primer lugar, la morfología proporciona elementos para la identificación de las unidades como son la categoría gramatical nombre (N) y el número plural, puesto que muchos candidatos a sigla usan el sistema anglosajón de añadir una “s” para la formación del plural en lugar de la duplicación de caracteres como se ha documentado en el sistema español. En segundo lugar, la sintaxis puede ayudar a crear un tipo de “sintaxis especial” que se ocupe de las relaciones entre un candidato a par sigla-forma desarrollada. Un intento de desarrollar una técnica de este tipo ha quedado plasmado en los grados de correspondencia total, parcial y nula encontrados entre los elementos de la forma desarrollada y los elementos que constituyen la sigla. Adicionalmente, la sintaxis permite conocer las tendencias combinatorias de las siglas, en especial la sigla en posición adjetiva. En este sentido, podría estudiarse la coocurrencia N+ADJ con un mínimo de dos letras mayúsculas iniciales como posible pista para la detección de un candidato a sigla. Por último, la sintaxis también ayuda a observar el comportamiento de las siglas dentro de las oraciones, encontrándose que pueden funcionar tanto como sujeto como objeto de verbo.

El análisis estadístico de las siglas o siglometría ha procurado determinar la incidencia de las siglas en el discurso de genoma humano y medio ambiente. A partir de aquí, se ha confirmado la hipótesis de que la aparición de las siglas varía de un ámbito de especialidad a otro. Además, se ha intentado explicar las consecuencias que este factor tiene en el discurso.

Los análisis estadísticos descriptivos han revelado que el peso de las siglas en un corpus es relativamente bajo. En ninguno de los casos estudiados ha superado el 1,1% del total de palabras del corpus. No obstante, su creación y uso van en constante aumento tal y como puede comprobarse en las estadísticas de incorporación de nuevas siglas en diccionarios en línea como *Acronym Finder*, el cual reporta la inclusión de cerca de

5.000 unidades mensuales entre abreviaturas y siglas. Otra forma de constatar el aumento progresivo del empleo de las siglas radica en los estudios diacrónicos. Bloom (2004), por ejemplo, ha comprobado la manera como ha aumentado el número de siglas y su frecuencia de uso en el campo de la urología a lo largo del siglo XX.

El análisis estadístico inferencial, realizado mediante la técnica del ANOVA, ha demostrado que las variables o rasgos de las siglas creadas en los ámbitos de GH y MA tienen en su mayoría valores opuestos. GH supera en proporción a MA en cantidad de siglas mixtas, híbridas, alfanuméricas, pluralizadas, especializadas y creadas en inglés. Y, a su vez, MA supera en proporción a GH en cantidad de siglas propias, generales, creadas en español, de 2 caracteres y de más de seis caracteres. Se evidencian, por tanto, grandes diferencias entre ambos campos de especialidad. La más importante, desde el punto de vista del análisis del discurso, tiene que ver justamente con la tendencia de cada ámbito; mientras GH se inclina más por la siglación de términos, MA tiende a siglar más palabras del discurso general o nombres de organismos e instituciones. Como consecuencia de lo anterior, puede pensarse que los textos de GH tienen una concentración mayor de terminología que los de MA. Esta particularidad puede explicarse, entre otras razones, porque GH es un área de especialidad en evolución constante, más dinámica que otras desde el punto de vista de la generación de conocimiento, y por consiguiente, más generadora de denominaciones nuevas, muchas de las cuales terminan siendo abreviadas.

Un análisis estadístico descriptivo complementario, aplicado a otros dos ámbitos de especialidad como son informática y economía, ha permitido conocer que algunos de los rasgos son comunes a los de GH y MA. De esta forma, se ha encontrado que los cuatro ámbitos se pueden agrupar en dos según su grado de afinidad. GH tiene mayor afinidad con INF mientras que MA se asemeja más a ECON.

Los análisis estadísticos refrendan la hipótesis formulada por Baudet (2001: 34), que expone que la frecuencia de las siglas varía significativamente de un ámbito a otro e incluso de una lengua a otra.

En síntesis, puede decirse que el análisis descriptivo revela que las siglas son:

1) Un mecanismo de reducción léxica cada vez más frecuente;
2) Un fenómeno de frecuencia variable según los ámbitos de especialidad y las lenguas;
3) Unidades representativas de términos y de palabras, por tanto, están presentes tanto en el discurso especializado como en el general;
4) Unidades que corresponden a nombres comunes y, a veces, a nombres propios;
5) Unidades que tienden a aparecer sin su forma desarrollada cuando están fijadas en el discurso (el caso de ADN, ARN, ONU, etc.);
6) Un fenómeno universal; es decir, que están presentes en todas las lenguas.

En el plano aplicado hemos pretendido establecer los criterios básicos para el diseño de un sistema de detección de siglas en español. Para ello se han llevado a cabo dos tareas. Por una parte, se han identificado las reglas de formación de siglas y, por otra parte, se han establecido los patrones de detección.

Por su repercusión en la lengua, muchas áreas de conocimiento se han interesado por el estudio de las siglas; entre las que se cuentan inteligencia artificial, minería de datos, recuperación de información y procesamiento del lenguaje natural (PLN). Así mismo, el campo de la biomedicina es el que más esfuerzos ha dedicado a la investigación para el desarrollo de sistemas de extracción de siglas. En la actualidad no se cuenta con sistemas de este tipo para el español. Los existentes han sido creados todos para analizar textos en lengua inglesa, a excepción del propuesto por Dannélls (2005; 2006), orientado al reconocimiento de siglas en textos médicos en sueco.

Los sistemas de detección y extracción de siglas cumplen fundamentalmente dos tareas. De un lado, sirven para alimentar automáticamente las bases de datos de siglas y mantenerlas de esta forma actualizadas y, de otro lado, sirven para facilitar las tareas de extracción o recuperación de información en un campo de conocimiento dado.

Los sistemas basados en aprendizaje máquina junto con los híbridos se perfilan como los de mejor rendimiento. Sin embargo, no se debe pasar por alto que todos estos sistemas han sido pensados para analizar textos en lengua inglesa, por lo que se desconoce la eficacia de su aplicación en lenguas como el español. Se debe tener en cuenta que, por ejemplo, para el desarrollo de los patrones de detección de siglas en español, es necesario considerar un conjunto de patrones mixto, porque se pueden dar los siguientes casos: 1) la sigla y la forma desarrollada aparecen en español; 2) la sigla aparece en lengua extranjera y la forma desarrollada en español, y 3) la sigla aparece en español y la forma desarrollada en lengua extranjera.

Se han encontrado 49 patrones diferentes para el reconocimiento de pares de sigla-forma desarrollada en español, los cuales deberían incorporarse a un sistema diseñado para el español. Sin embargo, los patrones más productivos han resultado ser dos, los cuales indicamos a continuación: 1) la forma desarrollada seguida de la sigla entre paréntesis; es decir, "FD (SIGLA)", y 2) la sigla seguida de la forma desarrollada entre paréntesis; es decir, "SIGLA (FD)".

Algunos de estos patrones coinciden con los establecidos en la literatura sobre el tema; sin embargo, otros sólo se han hallado en nuestro corpus, lo cual puede deberse a las características propias del discurso de GH y MA en lengua española. En el ámbito de GH la variedad de patrones de identificación de pares de sigla-forma desarrollada es mayor que la de MA.

En síntesis, desde el punto de vista aplicado, un sistema de detección de siglas debe tener en cuenta que:

1) Las siglas son unidades compuestas en su mayoría por 3 ó 4 caracteres alfanuméricos;
2) Las siglas pueden ser préstamos, generalmente procedente del inglés y, por consiguiente, unidades con una estructura sintáctica diferente. De ahí que no se correspondan exactamente con las iniciales de su forma desarrollada en español;

3) El paréntesis es la forma más común de aparición de las formas desarrolladas de las siglas. Sin embargo, en español se han documentado hasta 49 formas diferentes de introducir pares de sigla-forma desarrollada en los textos. Además, existen otras estrategias de introducción bien de la sigla o bien de la forma desarrollada como son palabras clasificadoras del tipo "*abreviatura*".

Finalmente, para el diseño de un sistema de identificación de siglas se deben tener en cuenta tres aspectos básicos: 1) las reglas de formación y los patrones para la identificación de pares de sigla-forma desarrollada; 2) la longitud de la ventana (contexto) a la derecha e izquierda del candidato a sigla para la identificación de los candidatos a par sigla-forma desarrollada, y 3) el método para el reconocimiento de los candidatos (*i.e.*, patrones, estadística, aprendizaje máquina o una combinación de estos).

2. Posibles líneas de trabajo futuro

A partir de los resultados de esta investigación se perfilan varias líneas de trabajo futuro como son:

1) Continuación del análisis y descripción de las siglas en: a) otros ámbitos de especialidad como medicina, telecomunicaciones, derecho, biomedicina, etc., y b) en distintos géneros de texto (historias médicas, manuales de usuario, etc.).

2) Continuación del estudio sobre la calidad de los bancos de datos de siglas. En la primera exploración llevada a cabo en esta tesis se analizaron los diccionarios en línea de siglas de corte más general (*Acronym Finder*, *Acronyma* y *Abbreviations.com*). Dicho análisis podría ampliarse a otros diccionarios del mismo tipo como *Acronym server* y *All-Acronyms.com*.

Así mismo, en una segunda fase, podrían analizarse las ventajas y desventajas de los diccionarios en línea de siglas especializadas como son: *AcroMed*, *ARGH: Biomedical Acronym Database*, *SaRAD* (*A Simple and Robust Dictionary of Biomedical Abbreviations*), *Stanford Biomedical Abbreviation Server*, *Wiley InterScience Acronym Finder*, etc.

3) Creación de un observatorio de siglas y un BD de siglas generales y especializadas para el español. Esta posible vía de trabajo tendría una finalidad doble. Por una parte, hacer un seguimiento constante de las siglas mediante la observación de su evolución y fijación en el discurso. Y, por otra parte, crear una BD de siglas generales y especializadas en inglés, español y catalán aplicando el modelo propio de ficha para el vaciado de estas unidades. Un proyecto de estas características se justifica porque:

   a) Los fenómenos de abreviación y especialmente la siglación interesan a diversas áreas: lingüística, neología, traducción, lexicología, terminología, redacción técnica y enseñanza de lenguas con fines específicos.

   b) No existe un recurso de tales características en español y, al parecer, en ninguna de las lenguas romances.

   Se trata de un proyecto de investigación interdisciplinario que vincularía como mínimo a expertos de las áreas de lingüística, informática y bibliotecología.

   El proyecto constaría de las siguientes fases principales:

   a) Conformación del corpus. Los textos para la detección de los candidatos a pares sigla-forma desarrollada se tomarían de la web.
   b) Diseño e implementación de un sistema de detección de siglas para el español. Para esta fase se partirá de los criterios establecidos anteriormente en el capítulo 8.

c) Diseño e implementación de una ficha para el vaciado de las siglas que recoja el mayor tipo de información útil al mayor número de usuarios posible. Esto se debe a que los diccionarios de abreviaciones actuales se preocupan por recoger la mayor cantidad de unidades, pero suelen dejar de lado mucha información relacionada con las siglas, la cual podría ser de gran utilidad para un perfil específico de usuarios que consultan estos recursos como es el colectivo de traductores, intérpretes, terminólogos, profesores de LSP y redactores técnicos.

d) Creación de una interfaz de consulta. Finalmente, cualquier usuario podría interrogar el BD de siglas mediante una interfaz de consulta, que podría incluir dos modos, a saber: consulta de la forma desarrollada o consulta de la sigla.

Bibliografía

Bibliografía

Abreu, J. M. (1997). «Las siglas y los acrónimos en el lenguaje técnico». En *Actas III Simposio Iberoamericano de Terminología*. San Millán de la Cogolla, La Rioja. 19-27.

Adar, E. (2002). «S-RAD: A Simple and Robust Abbreviation Dictionary». [En línea]. *http://cond.org/s-rad-090502.pdf* [Consulta: octubre 30 de 2007].

Adar, E. (2004). «SaRAD: A simple and Robust Abbreviation Dictionary». *Bioinformatics* 20 (4). 527-533. [En línea]. *http://www.hpl.hp.com/research/idl/papers/srad/s-rad-090502.pdf* [Consulta: 4 de septiembre de 2005].

Akira, T.; Tokunaga, T. (2001). «Automatic disabbreviation by using context information». [En línea]. *http://www.afnlp.org/nlprs2001/WS-Paraphrase/pdf/03-terada.pdf* [Consulta: 16 de marzo de 2006].

Alcaraz, E.; Martínez, Mª A. (1997). *Diccionario de Lingüística moderna.* Barcelona: Editorial Ariel.

Alcaraz, Mª A. (2003). «Las siglas del discurso biomédico escrito en inglés: análisis y aplicaciones didácticas». *The ESP* 23 (1). 37-51. [En línea]. *http://lael.pucsp.br/especialist/23_1_2002/AlcarazAriza.pdf* [Consulta: 5 de octubre de 2004].

Alcina, J., Blecua, J. M. (1991). *Gramática española*. Barcelona: Instrumenta. 513-547.

Algeo, J. (1978). «The Taxonomy of Word Making». *Word: Journal of the Linguistic Association* 29 (2). 122-131.

Algeo, J. (1991). *Fifty years among the new words.* Cambridge: Cambridge University Press. 8-12.

Algeo, J. (2003). «Abbreviation». En Frawley, W. (ed.). *International Encyclopedia of Linguistics.* Vol. 1. New York: Oxford University Press.

Alvar, M.; Miró, A. (1983). *Diccionario de siglas y abreviaturas.* Madrid: Alhambra. 1-25.

Álvarez de Miranda, P. (2007). *Acrónimos, acronimia: revisión de un concepto*. [En línea].

http://www.cervantesvirtual.com/servlet/SirveObras/12937622007074869643624/p0000001.htm#I_0_ [Consulta: 6 de junio de 2007].

Ambadiang, T. (1999). «La flexión nominal. Género y número». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Vol. 3. Madrid: Espasa. 4.876-4.896.

Ananiadou, S. *et al.* (2002). *Term-based Literature Mining from Biomedical Texts.* [En línea]. *http://www.pdg.cnb.uam.es/BioLink/Ananiadou.doc* [Consulta 4 de septiembre de 2004].

Ao, H.; Takagi, T. (2003). «An Algorithm to Identify Abbreviations from MEDLINE». *Genome Informatics* 14. 697-698.

Ao, H. (2005). «ALICE: An Algorithm to Extract Abbreviations from MEDLINE». *Journal of the American Medical Informatics Association* 12 (5). 576-586. [En línea]. *http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1205607&blobtype=pdf* [Consulta: 27 de octubre de 2006].

Arntz, R.; Picht, H. (1995). *Introducción a la terminología.* Madrid: Fundación Germán Sánchez Ruipérez.

Bach, C. *et al.* (1997). *El Corpus de l'IULA: descripció.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Papers de l'IULA, Informes, 17).

Bastons, C.; Font, J. V. (2001). *Guía práctica de la lengua castellana.* Barcelona: Promociones y Publicaciones Universitarias.

Baudet, J. C. (2001). «La siglométrie: outil de linguistique comparée». *La Banque des mots* 62. 34-36.

Baudet, J. C. (2002). «Les sigles et la science en français». *La Banque des mots* 64. 93-96.

Bauer, R. (1990). «Parlons a bit du bit: Les acronymes dans le français de l'informatique». En *Terminologie et Traduction*. Nº 2. Luxembourg: Office des publications officielles des Communautés européennes. 171-193.

Baum, S. V. (1955). «From "Awol" to "Veep": The Growth and Specialization of the Acronym». *American Speech* 30 (2). 103-110.

Belda, J. R. (2003). *El lenguaje de la informática e Internet y su traducción.* Alicante: Publicaciones Universidad de Alicante.

Bezos, J. (2007). «Reflexiones abreviadas». *Donde dice…* 6. 10-13.

Bloom, D. A. (2000). «Acronyms, abbreviations and initialisms». *BJU International* 86. 1-6.

Bombardieri, S. *et al.* (1998). «A unifed list of acronyms for the rheumatology literature». *Arthritis Rheum* 41. 1.901-1.905.

Bosque, I.; Demonte, V. (eds.) (1999). *Gramática descriptiva de la lengua española*. Madrid: Espasa.

Bosque, I. (1999). «El nombre común». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Vol. 1. Madrid: Espasa. 5-75.

Bracewell, D. *et al*. (2005). «Identification, Expansion, and Disambiguation of Acronyms in Biomedical Texts». En Chen, G., *et al*. (eds.). *ISPA Workshops 2005.* Berlin/Heidelberg: Springer-Verlag. 186-195. [En línea]. *https://troia.upf.edu/http/www.springerlink.com/content/p34858l370315765/fulltext.pdf* [Consulta: 2 de febrero de 2006].

Bracho, Ll. (2004). «La siglació terminològica en la traducció mediambiental en català, anglés i espanyol». En *Actas del IX Simposio Iberoamericano de Teminología: 'Contribución a la cultura de la paz, la diversidad y la sostenibilidad'*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Bradley, J. (2004). «The Acronym addiction». *Texas Heart Institute Journal* 31 (1). 108-109.

Brusaw, Ch. *et al.* (1987). *Handbook of Technical Writing.* 3rd edition. New York: St. Martin's Press.

Burnett, R. (1994). *Technical Communication.* Belmont: Wadsworth.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics.* London: Routledge.

Cabré, Mª T. (1993). *La terminología: teoría, metodología, aplicaciones.* Barcelona: Antártida/Empúries.

Cabré, Mª T. (1999). *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cabré, Mª T. *et al.* (2000). «La néologie ibérique». En Chevalier J. C.; Delport M. F. (eds.). *La Fabrique des mots.* Paris: Presses de l'Université de Paris-Sorbonne. 110-111.

Cabré, Mª T. *et al.* (2000). «Nombre propio y formación de palabras». En Wotjak, G. (ed.). *En torno al sustantivo y adjetivo en el español actual.* Frankfurt: Vervuert Verlag. 191-206.

Cabré, Mª T. (2002). «Análisis textual y terminología, factores de activación de la competencia cognitiva en traducción». En Alcina, A.; Gamero, S. (eds.). *La traducción científico-técnica y la terminología en la sociedad de la información*. Castellón: Publicacions de la Universitat Jaume I. 87-105.

Cabré, Mª T. (2002). «El conocimiento especializado y sus unidades de representación: diversidad cognitiva». *Sendebar* 13. 141-153.

Cabré, Mª T. (2002). «Terminologie et linguistique: la théorie des portes». *Terminologies nouvelles* 21. [En línea]. http://elies.rediris.es/elies16/Cabre.html [Consulta: 5 de octubre de 2004].

Cabré, Mª T. (2002). «Textos especializados y unidades de conocimiento: metodología y tipologización». En García Palacios, J., Fuentes Morán, Mª T. (eds.) *Texto, terminología y traducción.* Salamanca: Almar. 15-36.

Cabré, Mª T. (2003). «Teorías de la terminología: de la prescripción a la descripción». En Olschki, L. S. (ed.). *Lessico Intellettuale Europeo*. 169-188.

Cabré, Mª T. (2003). «Theories of terminology: their description, prescription and explanation». *Terminology* 9 (2). 163-200.

Cabré, Mª T. (dir.) (2004). *Curso on line de terminología: Terminología, ingeniería lingüística y lingüística computacional*. [cd-rom]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Calvet, J. L. (1980). *Les Sigles.* Paris: Presses Universitaires de France.

Cannon, G. (1989). «Abbreviations and Acronyms in English Word-Formation». *American Speech* 64 (2). 99-127.

Capó, J.; Veiga, M. (1997). *Abreviacions.* Barcelona: Direcció General de Política Lingüística, Generalitat de Catalunya.

Cardero, A. Mª (2002). «Las terminologías y los procesos de acortamiento: abreviaturas, acrónimos, iniciales y siglas. Algunas puntualizaciones». En *Actas del VIII Simposio Iberoamericano de Terminología* [cd-rom]. Cartagena de Indias.

Cardero, A. Mª (2003). *Terminología y procesamiento.* México: Universidad Nacional Autónoma de México, Campus Acatlán.

Cardona, G. (1991). *Diccionario de Lingüística.* Barcelona: Ariel.

Carton, J. (1987). *Dictionnaire de sigles nationaux et internationaux.* Paris: La maison du dictionnaire.

Casado Velarde, M. (1985). *Tendencias en el léxico español actual.* Madrid: Coloquio.

Casado Velarde, M. (1999). «Otros procesos morfológicos: Acortamientos, formación de siglas y acrónimos». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Madrid: Espasa. 5.075-5.096.

Cascón, E. (2004). *Manual del buen uso del español*. Madrid: Castalia.

Castro, E.; Ruiz, C. M. (2000). *Índice de acrónimos y siglas comunes en bioquímica y biología molecular*. [En línea]. *http://www.personales.ulpgc.es/ecastro.dbbf/Descargas/index00.pdf* [Consulta: 4 de agosto de 2004].

Cerdà, R. (1986). *Diccionario de lingüística.* Madrid: Anaya.

Chang, J. *et al.* (2002). «Creating an Online Dictionary of Abbreviations from MEDLINE». *Journal of the American Medical Informatics Association* 9 (6). 612-620. [En línea]. *http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12386112* [Consulta: 22 de julio de 2004].

Cheng, T. (1997). «Non-English acronyms must be explained in their native languages». *International Journal of Cardiology* 61. 199.

Cheng, T. (2002). «Acronyms must be defined». *Atherosclerosis* 165. 383.

Cheng, T. (2002). «Every acronym should be defined when it first appears in a publication». *Circulation* 106. 134.

Cheng, T. (2005). «Celestial acronyms». *International Journal of Cardiology* 101. 307-308.

Ciapuscio, G.; Kugel, I. (2002). «Hacia una tipología del discurso especializado». En García Palacios, J., Fuentes Morán, Mª T. (eds.). *Texto, terminología y traducción*. Salamanca: Almar. 37-73.

Ciapuscio, G. (2003). *Textos especializados y terminología.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Colás, J. (1994). *Diccionario de lengua y lingüística Gran Vox.* Barcelona.

Cruz Piñol, M. (1999). *El léxico del español de la Internet. Las siglas.* [En línea]. http://elies.rediris.es/elies1/62.htm [Consulta: 22 de diciembre de 2003].

Crystal, D. (2003). *A Dictionary of Linguistics & Phonetics.* 5th edition. Oxford: Blackwell.

Dannélls, D. (2005). *Classifying Swedish Acronyms with MBT*. [En línea]. http://www.cling.gu.se/~cl2ddoyt/pub/mbl_project.pdf [Consulta: 4 de julio de 2006].

Dannélls, D. (2005). *Recognizing Swedish acronyms and their definitions in biomedical literature*. [En línea]. http://www.cling.gu.se/~cl2ddoyt/acronyms/report.pdf [Consulta: 4 de julio de 2006].

Dannélls, D. (2006). *Automatic Acronym Recognition*. [En línea]. http://www.cling.gu.se/~cl2ddoyt/pub/automatic.pdf [Consulta: 16 de marzo de 2006].

Dannélls, D. (2006). Acronym Recognition: Recognizing acronyms in Swedish texts. Göteborg, Department of Linguistics, Göteborg University. [Tesis de máster dirigida por: Lars Borin, Dimitrios Kokkinakis]. [En línea]. http://www.cling.gu.se/~cl2ddoyt/pub/masterThes.pdf [Consulta: 22 de octubre de 2006].

De Granda, J. I. (2003). «Las siglas: ¿debemos aceptarlas?». *Arch Bronconeumol* 39 (6). 286.

Desrosiers, J. (2005). «Le genre des sigles». *L'actualité langagière* 2 (4). 18-20.

Díaz, M. T. (1998). *La categoría lingüística sustantivo*. Cádiz: Universidad de Cádiz.

Domènech, M. (1998). Unitats de coneixement i textos especialitzats: primera proposta d'anàlisi. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Tesina dirigida por: Mª Teresa Cabré].

Dominich, S. *et al*. (2003). «A Study of the Usefulness of Institutions' Acronyms as Web Queries». En Sebastiani. F. (ed.). *ECIR 2003*. Berlin/Heidelberg: Springer-Verlag. 580-587. [En línea]. https://troia.upf.edu/http/www.springerlink.com/content/09dhtwfntgh5lq9k/fulltext.pdf [Consulta: 19 de diciembre de 2005].

Dubois, J. *et al.* (1994). *Dictionnaire de linguistique et des sciences du langage.* Paris: Larousse.

Estopà, R. (2000). Extracción de terminología: elementos para la construcción de SEACUSE (Sistema de Extracción Automática de Candidatos a Unidades de Significacón Especializada). Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pomepu Fabra. [Tesis doctoral dirigida por: Mª Teresa Cabré].

Fallowfield, L. (2002). «Acronymic trials: the good, the bad, and the coercive». *The Lancet* 360. 1.622.

Farber, H. (2002). «On the abuse of acronyms». *American Journal of Respiratory and Critical Care Medicine* 166. 1.607-1.608.

Fernández, G. (2002). *La invasión de las siglas y acrónimos en las publicaciones científicas: el fenómeno lexicogenésico.* [En línea]. http://www.oftalmo.com/seo/2002/07jul02/01.htm [Consulta: 11 de febrero de 2004].

Fernández, M. J. (1999). «El nombre propio». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Vol. 1. Madrid: Espasa. 79-128.

Fernández-Pampillón, A.; Matesanz, M. (2006). «Los diccionarios electrónicos: hacia un nuevo concepto de diccionario». *Estudios de Lingüística del Español (ELiEs)* 24. [En línea]. http://elies.rediris.es/elies24/pampillon.htm [Consulta: 29 de agosto de 2006].

Fijo, Mª I. (2003). Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas. Sevilla, Departamento de humanidades, Universidad Pablo de Olavide. [Tesis dirigida por: Antonio Garnica].

Fontanillo, E. (1986). *Diccionario de Lingüística.* Madrid: Anaya.

Fred, H. (2003). «Acronymesis. The Exploding Misuse of Acronyms». *Texas Heart Institute Journal* 30 (4). 255-257.

Fuentes, X.; Castiñeiras, Mª J. *et al.* (2003). *Diccionario inglés-español de ciencias de laboratorio clínico.* [En línea]. http://www.leeds.ac.uk/ifcc/PD/dict/spandict.html [Consulta: 27 de febrero de 2004].

Garner. B. A. (2000). *The Oxford Dictionary of American Usage and Style*. Oxford: Oxford University Press.

Gaudan, S. *et al*. (2005). «Resolving abbreviations to their senses in Medline». *Bioinformatics* 21 (18). 3.658-3.664.

Gehénot, D. (1990). «Siglomanía: una aproximación al problema». En *Terminologie et Traduction*. Nº 2. Luxembourg: Office des publications officielles des Communautés européennes. 103-135.

Gelpí, C. (2000). *La lexicografía*. Barcelona: Santillana.

Germain, C. (1988). «Le sigle. Définition, caractéristiques et emploi». *Cahiers de Lexicologie* 53 (2). 55-74.

Gilbert, L. (1975). *La créativité lexicale.* Paris: Librairie Larousse.

Giraldo, J. (2004). Siglas y variación vertical en textos sobre genoma humano y medio ambiente. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Trabajo de línea de investigación en terminología dirigido por: M. Teresa Cabré].

Giraldo, J.; Cabré, M. T. (2004). «Las siglas en la producción de textos especializados: hacia una propuesta de recuperación asistida mediante BwanaNet». En *Actas del GLAT: La producció de textos especialitzats: estructura i ensenyament.* Barcelona: GLAT (Groupe de Linguistique appliquée des télécommunications). 305-315.

Giraldo, J. (2005). Analisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente. Barcelona, Institut Universitari de Lingüística Aplicada. [Proyecto de tesis dirigido por: M. Teresa Cabré].

Giraldo, J.; Cabré, Mª T. (2006). «Importancia de las siglas en dos ámbitos temáticos: genoma humano y medio ambiente». En *Actas del IX Simposio Iberoamericano de Teminología: 'Contribución a la cultura de la paz, la diversidad y la*

sostenibilidad'. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 817-826.

Giraldo, J. (2007). «Tratamiento de las siglas en los principales diccionarios de abreviaciones en Internet». En Lorente, M.; Estopà, R.; Freixa, J.; Martí, J.; Tebé, C. (eds.). *Estudis de lingüística i de lingüística aplicada*. Barcelona: Institut Universitari de Lingüística Aplicada. 323-336.

Gómez de Enterría, J. (1992). «Las siglas en el lenguaje de la economía». *Revista de filología románica* 9. 267-274.

Gómez, L. (2004). *Nuevo manual de Español correcto*. Madrid: Arco Libros.

González, V. (2002). *Expresiones regulares. Curso básico de UNIX*. [En línea]. http://iie.fing.edu.uy/~vagonbar/unixbas/index.htm [Consulta: 19 de febrero de 2004].

Green, W. (1990). «Abbs. in Js». *Canadian Medical Association Journal* 142 (4). 287.

Grudin, J.; Barnard, P. (1985). «When does an abbreviation become a word? And related questions». En *CHI'85 Proceedings*. 121-125.

Guardiola, E.; Baños, J. E. (2003). «Sobre la correcta utilización de las siglas: reflexiones a propósito de AINE e IECA». *MEDIFAM* 13 (4). 325-328.

Gutiérrez, B. M. (1998). *La ciencia empieza en la palabra: análisis e historia del lenguaje científico.* Barcelona: Península.

Hacohen-Kerner, Y. *et al*. (2004). «Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents». En Vicedo, J. L., *et al*. (eds.). *EsTAL 2004*. Berlin/Heidelberg: Springer-Verlag. 58-69. [En línea]. https://troia.upf.edu/http/www.springerlink.com/content/y9ev3m8crhd7q12d/fulltext.pdf [Consulta: 2 de febrero de 2006].

Hahn, U. *et al*. (2005). «Cross-Language Mining for Acronyms and their Completions from the Web». En Hoffmann, A., *et al*. (eds.). *DS 2005*. Berlin/Heidelberg: Springer-Verlag. 113-123. [En línea]. http://www.coling.uni-freiburg.de/~marko/publications/hahn-ds2005.pdf [Consulta: 16 de marzo de 2006].

Hartnack, V. (2004). «Short forms, long search: Trying to make sense of abbreviations». *Confluências* 1. [En línea]. http://www.confluencias.net/n1/hartnack.html [Consulta: octubre 31 de 2007]

Hatch, E.; Lazaraton, A. (1991). *The Research Manual. Design and Statistics for Applied Linguistics.* Los Angeles: Newbury House.

Heller, L. G.; Macris, J. (1968). «A Typology of Shortening Devices». *American Speech* 43 (3). 201-208.

Hoffmann, L. (1998). *Llenguatges d'especialitat.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Huddleston, R. *et al.* (2002). *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Isaacs, D. (2007). «Acronymophilia: an update». *ADC* 83. 517-518.

Jack, D. (2003). «The cardiology SCANDAL». *The Lancet* 361. 538.

Joshi, M. *et al*. (2006). *A Comparative Study of Supervised Learning as Applied to Acronym Expansion in Clinical Reports*. [En línea]. *http://wsdgate.sourceforge.net/pubs/AMIA06JoshiM.pdf* [Consulta: diciembre 5 de 2006].

Jurafsky, D. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech.* Upper Saddle River: Prentice Hall.

Kiss, T.; Strunk, J. (2002). *Scaled log likelihood ratios for the detection of abbreviations in text corpora*. [En línea]. *http://www.linguistics.rub.de/~kiss/publications/abbrev.pdf* [Consulta: 16 de marzo de 2006].

Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante.* Wiesbaden: Oscar Brandstetter Verlag.

Lacuesta, R., Bustos, E. (1999). «La derivación nominal». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Vol. 3. Madrid: Espasa. 4.505-4.594.

Lader, E. (2002). «Acronym mania». *The Lancet* 160. 576.

Larkey, L. *et al.* (2000). *Acrophile: An Automated Acronym Extractor and Server.* [En línea]. *http://delivery.acm.org/10.1145/340000/336664/p205-larkey.pdf?key1=336664&key2=7455896901&coll=GUIDE&dl=GUIDE&CFID=28595879&CFTOKEN=50021223* [Consulta: 29 de marzo de 2003].

Lázaro Carreter, F. (1990). *Diccionario de términos filológicos.* Madrid: Gredos.

Liu, H. *et al*. (2001). *A Study of Abbreviations in the UMLS*. [En línea]. http://adams.mgh.harvard.edu/PDF_Repository/D010001239.pdf [Consulta: 29 de noviembre de 2006].

Liu, H. *et al*. (2002). *A Study of Abbreviations in MEDLINE Abstracts*. [En línea]. http://lhncbc.nlm.nih.gov/lhc/docs/published/2002/pub2002051.pdf [Consulta: 16 de marzo de 2006].

Loma-Osorio, M. (2004). Estructura y función del texto económico: fundamentos de una léxico-gramática del discurso económico en español y en inglés. Madrid, Departamento de filología inglesa I, Facultad de filología. Universidad Complutense de Madrid. [Tesis doctoral dirigida por: Ana Pinto y Valerio Báez].

López Rúa, P. (2000). English Acronyms and Alphabetisms. A prototype based approach with special reference to their method of formation, realization, and connections with other morphological devices. Santiago de Compostela, Facultade de filoloxía, Universidad de Santiago de Compostela. [Tesis doctoral].

López Rúa, P. (2004). «Acronyms & Co.: A typology of typologies». *Estudios Ingleses de la Universidad Complutense* 12. 109-129.

Lorente, M. (2004). «Construcciones verbales en el discurso de la genómica. Tipología verbal y discurso científico». *Studia romanica Posnaniensia* XXXXI. 353-359.

Lorente, M. (2007). «Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació». En Lorente, M.; Estopà, R.; Freixa, J.; Martí, J.; Tebé, C. (eds.). *Estudis de lingüística i de lingüística aplicada*. Barcelona: Institut Universitari de Lingüística Aplicada. 365-380.

Losson, G. (1990). «De l'emploi des formes abrégées».En *Terminologie et Traduction*. Nº 2. Luxembourg: Office des publications officielles des Communautés européennes. 7-33.

Maldonado, C. *et al.* (2002). *Diccionario Clave. Diccionario del uso del español actual.* Madrid: SM.

Márquez, M. (2004). El anglicismo terminológico integral en los textos especializados: pautas para su tratamiento automatizado. Barcelona, Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. [Tesis doctoral dirigida por: Mercè Lorente].

Martínez de Sousa, J. (1984). *Diccionario internacional de siglas y acrónimos.* Madrid: Pirámide.

Martínez de Sousa, J. (1993). *Diccionario de redacción y estilo*. Madrid: Pirámide.

Martínez de Sousa, J. (2002). *Neologismos en el Diccionario de la Academia.* [En línea]. *http://europa.eu.int/comm/translation/bulletins/puntoycoma/almagro/html/MARTINEZ-L* [Consulta: 19 de septiembre de 2004].

Matthews, P. H. (1997). *The Concise Oxford Dictionary of Linguistics.* Oxford: Oxford University Press.

Mayor Serrano, B. Mª (2003). «Tratamiento de las siglas en los textos de divulgación médica, inglés-español». *Panace@* IV (13-14). 261-265. [En línea]. *http://www.medtrad.org/pana.htm* [Consulta: enero 21 de 2004].

McArthur, T. (1998). *Concise Oxford Companion to the English Language.* [En línea]. *http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t29.000610* [Consulta: 17 de junio de 2003].

Mejía, J. (1980). «Abreviaturas y siglas. Una definición». *Yelmo* 44-45. 30-31.

Mestres, J. Mª (1985). «Abreviacions: un assaig de classificació tipològica». *Revista de Llengua i Dret* 3. 13-22.

Mestres, J. Mª *et al.* (1995). *Manual d'estil. La redacció i l'edició de textos.* Barcelona: Eumo.

Mestres, J. Mª (1996). «La problemàtica de les abreviacions i els diccionaris». *Revista de Llengua i Dret* 26. 9-28.

Mestres, J. Mª; Guillen, J. (2001). *Diccionari d'abreviacions. Abreviatures, sigles i símbols.* Barcelona: Enciclopèdia catalana.

Mima, H. *et al*. (2002). *A Methodology for Terminology-based Knowledge Acquisition and Integration*. [En línea]. *http://acl.ldc.upenn.edu/coling2002/proceedings/data/area-16/co-228.pdf* [Consulta: 5 de octubre de 2004].

Mitterand, H. (1986). *Les mots français.* Paris: Presses universitaires de France.

Morgan, P. (1985). «A quick look at medical abbreviations». *Canadian Medical Association Journal* 132. 897.

Mossman, J. (1992). *Acronyms, Intialisms and Abbreviations Dictionary.* Detroit/London: Gale.

Mounin, G. (1982). *Diccionario de Lingüística.* Barcelona: Labor.

Nadeau, D.; Turney, P. (2005). «A Supervised Learning Approach to Acronym Identification». En *The Eighteenth Canadian Conference on Artificial Intelligence (AI'2005)*. National Research Council Canada. 1-10. [En línea]. http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48121.pdf [Consulta: 1 de diciembre de 2006].

Nakos, D. (1990). «Sigles et noms propres». *Meta* 35 (2). 407-413. [En línea]. http://www.erudit.org/revue/meta/ [Consulta: 14 de julio de 2003].

Narayanaswamy, M *et al*. (2003). «A Biological Named Entity Recognizer». En *Pacific Symposium on Biocomputing* 8. 427-438 [En línea]. http://www.ccs.neu.edu/home/futrelle/bionlp/psb2003/narayanaswamy.pdf [Consulta: 3 de marzo de 2004].

Nenadić, G. *et al.* (2002). «Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts». En *Actas del 3rd International Conference on Language, Resources and Evaluation, LREC-3*. Las Palmas. 2.155-2.162.

Nenadić, G. *et al.* (2002). *Automatic Discovery of Term Similarities Using Pattern Mining.* [En línea]. http://acl.ldc.upenn.edu/coling2002/workshops/data/w05/w05-08.pdf [Consulta: 12 de marzo de 2004].

Nenadić, G. *et al.* (2003). *Terminology-driven Mining of Biomedical Literature.* [En línea]. http://bioinformatics.oupjournals.org/cgi/reprint/19/8/938 [Consulta: 8 de mayo de 2004].

Nenadić, G. *et al.* (2006). «Towards a terminological resource for biomedical text mining». En *Actas del International Conference on Language, Resources and Evaluation, LREC.* Génova. 1.071-1.076.

Okazaki, N.; Ananiadou, S. (2006). «Building an Abbreviation Dictionary Using a Term Recognition Approach». *Bioinformatics*. 1-7. [En línea]. http://bioinformatics.oxfordjournals.org/cgi/reprint/btl534v1 [Consulta: 28 de noviembre de 2006].

Okazaki, N.; Ananiadou, S. (2006). «Clustering acronyms in biomedical text for disambiguation». [En línea]. http://hnk.ffzg.hr/bibl/lrec2006/pdf/351_pdf.pdf [Consulta: 30 de mayo de 2006].

Okazaki, N.; Ananiadou, S. (2006). «Term Recognition Approach to Acronym Recognition». En *Proceedings of the COLING/ACL 2006.* Sydney: Association for Computational Linguistics. 643-650. [En línea]. http://www.chokkan.org/publication/okazaki_COLACL2006.pdf [Consulta: 28 de noviembre de 2006].

Ortega, R. (2005). El plural de las siglas. COMRàdio. [correo electrónico].

Pakhomov, S. (2002). «Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts». En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* Philadelphia: ACL. 160-167. [En línea]. http://nlp.cs.nyu.edu/nycnlp/P02-1021.pdf [Consulta: 21 de enero de 2004].

Park, Y.; Byrd, R. (2001). *Hybrid Text Mining for finding Abbreviations and their Definitions.* En línea]. http://www.research.ibm.com/talent/documents/emnlp2001_48.pdf [Consulta: 27 de octubre de 2004].

Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins.

Pedersen, J. K. (2001). *El manual del editor de expresiones regulares.* [En línea]. http://docs.kde.org/es/HEAD/kdeutils/KRegExpEditor/index.html [Consulta 22 de marzo de 2004].

Percebois, J. (2001). «Fonctions et vie des sigles et acronymes en contextes de langues anglaise et française de spécialité». *Meta* 46 (4). 627-645. [En línea]. http://www.erudit.org/revue/meta/ [Consulta: 30 de abril de 2004].

Pérez Saldanya, M. *et al.* (1998). *Diccionari de Lingüística.* Barcelona: Colomar.

Pustejovsky, J. *et al.* (2001). *Linguistic Knowledge Extraction from Medline: Automatic Construction of an Acronym Database.* [En línea]. http://www.medstract.org/papers/bioinformatics.pdf [Consulta: 24 de mayo de 2004].

Quirk, R. *et al.* (1985). *A Comprehensive Grammar of the English Language.* London/New York: Longman.

Rodríguez, F. (1975). «El 'Clipping' en la lengua inglesa y española: sus accidentes gramaticales». *Publicaciones del Departamento de inglés de la Universidad de Valladolid* 5. 251-264.

Rodríguez, F. (1981). Análisis lingüístico de las siglas: especial referencia al español e inglés. Salamanca, Facultad de filología, Universidad de Salamanca. [Tesis doctoral dirigida por: Antonio Llorente].

Rodríguez, F. (1982). «Variaciones fonotácticas en siglas: condicionamientos lingüísticos y sociolingüísticos». *Revista española de lingüística*, año 12, fasc. 2. 357-374.

Rodríguez, F. (1983). «Morfología del número en las siglas». *Lingüística española actual.* 137-151.

Rodríguez, F. (1983). «On the coigning of Acronyms by Homonymy». *Anglo-American Studies* 3. 209-221.

Rodríguez, F. (1983). «Problemas planteados en la asignación del género de siglas extranjeras». En *Actas del Primer Congreso Nacional de Lingüística Aplicada.* Murcia. 277-286.

Rodríguez, F. (1984). «El género de las siglas». *Revista española de lingüística*, año 14, fasc. 2. 311-366.

Rodríguez, F. (1986). «Apuntes lexicográficos: Reflexiones a propósito de un diccionario general de siglas». *Revista española de lingüística aplicada*. 127-149.

Rodríguez, F. (1987). «Naturaleza sintáctica de las formas siglares. El cambio funcional». *Estudios de lingüística* 4. 139-148.

Rodríguez, F. (1989). «La derivación de las siglas». *Boletín de la Real Academia Española* tomo LXIX, cuaderno CCXLVII. 211-255.

Rodríguez, F. (1990). «La traducción de las siglas inglesas». En Rodríguez, F.(ed.) *Estudios de filología inglesa. Homenaje al Dr. Pedro Jesús Marcos Pérez*. Alicante: Departamento de filología inglesa, Universidad de Alicante. 169-181.

Rodríguez, F. (1990). «Valor metasémico de la sigla: la metáfora y otros cambios de sentido». *Ramanische Forschungen* 102 (4). 414-424.

Rodríguez, F. (1991). «Translation and Borrowing of Acronyms: Main Trends». *International Review of Applied Linguistics in Language Teaching (IRAL)* XXIX (2). 161-170.

Rodríguez, F. (1993). «Las siglas como procedimiento lexicogenésico». *Estudios de lingüística* 9. 9-24.

Rodríguez, F. (1993). «Morphovariation and synonymy of acronyms». *Meta* 38 (2). 275-292. [En línea]. http://www.erudit.org/revue/meta/1993/v38/n2/index.html [Consulta 11 de noviembre de 2004].

Rodríguez, F. (2002). *Variación tipográfica en el uso de las "abreviaturas dobles"*. [En línea]. http://www.ucm.es/info/especulo/cajetin/abreviat.html [Consulta: 12 de febrero de 2004].

Rowe, R. (2003). «Abbreviation Mania and Acronymical Madness». *DDT* 8 (16). 732-733.

Rull, X. (2005). La lexicalització de sigles: pautes i propostes. (s.l) (s.n).

Sager, J. C. (1990). *A Practical Course in Terminology Processing.* Amsterdam/Philadelphia: John Benjamins.

Salvanyà, J. (2005). Plural de sigles. Radio Fleixbac. [correo electrónico].

Sánchez Gijón, P. (2003). Els documents digitals especialitzats: utilització de la lingüística de corpus com a font de recursos per a la traducció especialitzada. Barcelona, Departament de Traducció i interpretació. Universitat Autònoma de Barcelona [Tesis doctoral].

Sánchez-Gijón, P. (2004). *L'ús de corpus en la traducció especialitzada*. Barcelona: Institut Universitari de Lingüística Aplicada.

Santoyo, J. C. (1980). «Análisis lingüístico de las siglas inglesas usadas en español». *Yelmo* 9 (1). 17-19.

Schwartz, A.; Hearst, M. (2003). «A Simple Algorithm for identifying Abbreviation Definitions in Biomedical Text». En *Pacific Symposium on Biocomputing.* 451-462([En línea]. http://biotext.berkeley.edu/papers/psb03.pdf [Consulta: 26 de febrero de 2004].

Suárez, Mª M. (2004). Análisis contrastivo de la variación denominativa en textos especializados: del texto original al texto meta. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Tesis doctoral dirigida por: Mª Teresa Cabré].

Taghva, K.; Gilbreth, J. (1999). «Recognizing acronyms and their definitions». *International Journal on Document Analysis and Recognition* 1. 191-198. [En línea].

https://troia.upf.edu/http/www.springerlink.com/content/u6c9ymd1v8jflerh/fulltext.pdf [Consulta: 30 de noviembre de 2006].

Torii, M. *et al*. (2006). A Comparison Study of Biomedical Short Form Definition Detection Algorithms. En *TMBIO'06*. Arlington: ACM. 52-59.

Torres, E.; Schulz, K. (2005). «Stable methods for recognizing acronym-expansion pairs: from rule sets to hidden Markov models». *International Journal of Document Analysis* 8 (1). 1-14.

Tsuruoka, Y. *et al*. (2005). «A Machine Learning Approach to Acronym Generation». En *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*. 25-31. [En línea]. http://acl.ldc.upenn.edu/W/W05/W05-1304.pdf [Consulta: 30 de noviembre de 2006].

Tsuruoka, Y.; Tsujii, J. (2003). *Probabilistic Term Variant Generator for Biomedical Terms.* [En línea]. http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/papers/sigir03.pdf [Consulta: 26 de febrero de 2004].

Valderrábanos, A. *et al.* (2002). *Textractor: a Multilingual Terminology Extraction Tool.* [En línea]. http://liquid.sema.es/document_pdf/textractor_a_multilingual_terminology_extraction_tool.pdf [Consulta: 26 de febrero de 2004].

Vandaele, S.; Pageau, M. (2006). «Dynamique discursive et traduction des signes abréviatifs en biomédecine». *Équivalences* 33 (1-2). 165-190.

Varela, S. (1999). «La prefijación». En Bosque, I.; Demonte, V. (eds.). *Gramática descriptiva de la lengua española*. Vol. 3. Madrid: Espasa. 4.993-5.040.

Vivaldi, J. (2003). *BwanaNet: Programa d'explotació del Corpus Tècnic de l'IULA.* [En línea]. http://www.iula.upf.edu/materials/031204iulaterm.pdf [Consulta: 2 de septiembre de 2004].

Vivaldi, J.; Bach, C. (2003). «Explotación de corpus para el trabajo terminológico». En *Memorias de la IV Escuela Internacional de Verano de Terminología*. [cd-rom]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

VV.AA. (1998). *Recomendaciones para el diseño y la configuración de bancos de datos terminológicos*. Medellín: Universidad de Antioquia.

VV.AA. (1999). *Norma ISO 12620. Computer applications in terminology-Data categories.* Geneva: ISO.

VV.AA. (2000). *Norma ISO 1087-1. Terminology work-Vocabulary- Part 1. Theory and application.* Geneva: ISO.

VV.AA. (2000). [En línea]. Grammaire reverso. *http:// grammaire.reverso.net/index_alpha/Fiches/fiche347.htm* [Consulta: 29 de mayo de 2003].

VV.AA. (2003). *Norma ISO 639-1/ISO 639-2 Codes for the representation of Names of Languages.* [En línea]. *http://www.loc.gov/standards/iso639-2/langcodes.html#top* [Consulta: 29 de julio de 2004].

VV.AA. (1999). «Ortografía de la lengua española». Madrid: Real Academia de la Lengua Española.

VV.AA. (s.d.). Human Genome Project [En línea]. *http://www.ornl.gov/hgmis/acronym.html* [Consulta: 29 de mayo de 2003].

Walling, H. (2001). «When will the MEK inherit the ERK? Acronym alphabet soup». *TRENDS in Pharmacological Sciences* 22 (1). 14.

Wren, J.; Garner, H. (2002). «Heuristics for Identification of Acronym-Definition Patterns Within Text: Towards an Automated Construction of Comprehensive Acronym-Definition Dictionaries». *Methods of Information in Medicine* 41 (5). 426-34. [En línea] *http://www.schattauer.de/index.php?id=739&no_cache=1&artikel=413* [Consulta: 12 de febrero de 2007].

Xu, J.; Huang, Y. (2005). «A Machine Learning Approach to Recognizing Acronyms and their Expansion». En *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*. [En línea]. *http://nkxujun.googlepages.com/AcronymExtraction-ICMLC2005.pdf* [Consulta: diciembre 1 de 2006].

Xu, J.; Huang, Y. (2006). «Using SVM to extract acronyms from text». *Soft Comput* 11. 369-373. [En línea] *http://www.springerlink.com/content/276028782h150080/fulltext.pdf* [Consulta: 12 de febrero de 2007].

Yeates, S. (1999). *Automatic Extraction of Acronym from Text.* [En línea]. *http://www.cs.waikato.ac.nz/~nzdl/publications/1999/yeates-Auto-Extract.pdf* [Consulta: 5 de julio de 2004].

Yeates, S. *et al.* (2000). *Using Compression to identify Acronyms in Text.* [En línea]. http://arxiv.org/PS_cache/cs/pdf/0007/0007003.pdf [Consulta: 20 de marzo de 2004].

Yi, J.; Sundaresan, N. (1999). *Mining the Web for Acronyms Using the Duality of Patterns and Relations*. [En línea]. https://troia.upf.edu/http/delivery.acm.org/10.1145/320000/319782/p48-yi.pdf?key1=319782&key2=1164994611&coll=portal&dl=ACM&CFID=7731795&CFTOKEN=76752898 [Consulta: 16 de febrero de 2004].

Yoshida, M. *et al*. (2000). «PNAD-CSS: A workbench for constructing a protein name abbreviations dictionary». *Bioinformatics* 16 (2). 169-175. [En línea]. http://bioinformatics.oxfordjournals.org/cgi/reprint/16/2/169 [Consulta: 27 de octubre de 2006].

Young, A. (2004). Automatic Acronym Identification and the Creation of an Acronym Database. Sheffield, Department of Computer Science The University of Sheffield [Tesis de maestría dirigida por: Mark Stevenson]. [En línea]. http://www.dcs.shef.ac.uk/intranet/teaching/projects/archive/ug2004/pdf/u1ay.pdf [Consulta: 31 de marzo de 2005].

Yu, H. *et al.* (2002). *Mapping Abbreviations to Full Forms in Biomedical Articles.* [En línea]. http://www1.cs.columbia.edu/~hongyu/paper/JAMIA_02_Mapping.pdf [Consulta: 2 de abril de 2004]

Yu, H.; Agichtein, E. (2003). «Extracting Synonimous Gene and Protein Terms from Biological Literature». *Bioinformatics* 1 (1). 1-10.

Yu, H. *et al.* (2006). «A Large Scale, Corpus-Based Approach for Automatically Disambiguating Biomedical Abbreviations». *Journal of the American Medical Informatics Association* 9 (3). 262-272. [En línea]. http://delivery.acm.org/10.1145/1170000/1165778/p380-yu.pdf?key1=1165778&key2=5451994611&coll=ACM&dl=ACM&CFID=15151515&CFTOKEN=6184618 [Consulta: 7 de noviembre de 2006].

Yu, Z. *et al.* (2003). *Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypotesis.* [En línea]. http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/papers/sigir03bio.pdf [Consulta: septiembre 18 de 2004].

Zahariev, M. (2003) *Efficient Acronym-Expansion Matching for Automatic Acronym Acquisition.* [En línea]. http://www.cs.sfu.ca/~manuelz/personal/p/da.pdf [Consulta: octubre 5 de 2004].

Zahariev, M. (2004). «A Linguistic Approach to Extracting Acronym Expansions from Text». *Knowledge and Information Systems* 6. 366-373. [En línea] http://www.springerlink.com/content/mt4q8d4rhk8f667r/fulltext.pdf [Consulta: 10 de febrero de 2006].

Zahariev, M. (2004). A(Acronyms). Burnaby, School of Computing Science, Simon Fraser University. [Tesis doctoral]. [En línea]. http://www.cs.sfu.ca/~manuelz/personal/p/f.pdf [Consulta: octubre 5 de 2004].

Zhou, W. *et al*. (2006). «ADAM: another database of abbreviations in Medline». *Bioinformatics* 22 (22). 2.813-2.818.

Zolondek, D. (1991). «La siglaison». *Terminogramme* 62. 1-5.

Otras obras citadas

Abbreviations.com [En línea]. http://www.abbreviations.com/about.asp [Consulta: 6 de junio de 2007].

Abreviaturas de genes [En línea]. http://www.personales.ulpgc.es/ecastro.dbbf/descargas/index00.htm [Consulta: 23 de mayo de 2006].

AcroMed [En línea]. Massachusetts: Brandeis University. http://medstract.med.tufts.edu/acro1.1/ [Consulta: 25 de mayo de 2007].

Acronym Finder [En línea]. http://www.acronymfinder.com/ [Consulta: 2 de junio de 2007].

Acronym Server [En línea]. http://silmaril.ie/cgi-bin/uncgi/acronyms [Consulta: 2 de junio de 2007].

Acronyma [En línea]. http://www.acronyma.com/ [Consulta: 7 de junio de 2007].

Acronyms, Initialisms and Abbreviations Dictionary. New York: Thomson-Gale, 2003.

All-acronyms.com [En línea]. http://www.all-acronyms.com [Consulta: 11 de junio de 2007].

Compendium of Environmental and professional Acronyms [En línea].

http://web.umr.edu/~aeg/arco/arco.html [Consulta: 23 de mayo de 2006].

Diccionari de Lingüística Termcat. Barcelona: Termcat, 1992

Diccionario de la lengua española. Real Academia de la lengua española, 22ª edición. Madrid: Espasa-Calpe, 2001.

Diccionario de siglas y abreviaturas. Madrid: Ed. Alhambra, 1983.

Diccionario de la contaminación [En línea]. *http://www.edu365.cat/aulanet/comsoc/Lab_tecno/vincles/Diccionario_conta.htm* [Consulta: 23 de mayo de 2006].

Diccionari Enciclopèdic de Medicina [En línea]. Barcelona: Enciclopèdia catalana. *http://www.grec.cat/home/cel/mdicc.htm* [Consulta: 3 de noviembre de 2007].

Diccionario internacional de siglas y acrónimos. Madrid: Ed. Pirámide, 1984.

Dictionnaire des abréviations et acronymes scientifiques, techniques, medicaux, économiques et juridiques. 2ª ed. Paris: Tec & Doc-Lavoisier, 1992.

Dictionnaire international d'abréviations scientifiques et techniques. Paris: La maison du dictionnaire, 1978.

Dictionnaire Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française. Paris: Dictionnaires Le Robert, 2001.

EEA Multilingual Environmental Glossary [En línea]. *http://glossary.eea.eu.int/EEAGlossary* [Consulta: 23 de mayo de 2006].

Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols. Amsterdam: Elsevier, 1997.

Genetics home reference [En línea]. *http://ghr.nlm.nih.gov/* [Consulta: 23 de mayo de 2006].

Glosarios de Biotecnología [En línea]. *http://www.monsanto.es/noticias-y-recursos/prensa/definiciones-de-t-rminos-t-cnicos/definiciones-de-t-rminos-t-cnicos#d* [Consulta: 23 de mayo de 2006].

Glosarios de Genética [En línea].
http://ejb.ucv.cl/gmunoz/genweb/genetica/frame/textos/anexos/glosario.html
[Consulta: 23 de mayo de 2006].

Gouvernement de Québec [En línea].
http://w3.oqlf.gouv.qc.ca/BDL/gabarit_bdl.asp?impr=1&id=1385&T2.x=&T3.x=&t1 [Consulta: 29 de mayo de 2003].

Gran Diccionari de la Llengua Catalana [En línea].
http://www.grec.net/home/cel/dicc.htm [Consulta: 22 de julio de 2003].

Human Genome Acronym List [En línea].
http://www.ornl.gov/sci/techresources/Human_Genome/acronym.shtml
[Consulta: 25 de mayo de 2007].

Le trésor de la langue française informatisé [En línea]. http://zeus.inalf.fr/tlf.htm
[Consulta: 16 de junio de 2004].

List of Acronyms used on the Literature on Genome Research [En línea].
http://www.dpw.wau.nl/pv/aflp/acronyms.html [Consulta: 23 de mayo de 2006].

Medical Dictionary on-line [En línea]. http://www.online-medical-dictionary.org/
[Consulta: 25 de mayo de 2007].

Merck Source [En línea].
http://www.mercksource.com/pp/us/cns/cns_hl_dorlands.jspzQzpgzEzzSzppdocszSzuszSzcommonzSzdorlandszSzdorlandzSzdmd_a-b_00zPzhtm [Consulta: 23 de mayo de 2006].

Merriam-Webster Dictionary Online [En línea]. http://www.m-w.com/home.htm
[Consulta: 22 de julio de 2003].

Nombres de proteínas y genes [En línea].
http://www.biology.duke.edu/class/bio118/handouts/nomenclature.htm [Consulta: 23 de mayo de 2006].

Pugh's dictionary of acronyms and abbreviations: abbreviations in management, technology and information science. Chicago: Library Association Publishing, 1987.

Siglas [En línea]. http://www.siglas.com.br/ [Consulta: 2 de octubre de 2006].

The Oxford Dictionary of Abbreviatons. Oxford: Oxford University Press, 1993.

U.S. Environmental Protection Agency [En línea]. *http://www.epa.gov/iris/acronyms.htm* [Consulta: 23 de mayo de 2006].

Wiley Interscience [En línea]. *http://www3.interscience.wiley.com/cgi bin/home?CRETRY=1&SRETRY=0* [Consulta: 6 de junio de 2007].

Anexos

Anexo 1

Sufijos y prefijos

a) Sufijos nominales

-a
-ada
-ado ~ -ato
-aje$_1$
-aje$_2$
-al ~ -ar
-azo
-ción ~ sión ~ ión ~ón
-dad ~ -idad ~ -edad ~ -tad
-dero ~ -dera ~ -deras
-do ~ -da
-dor ~ -sor ~ -tor ~ -o
-dura
-e
-ería
-erío
-ero ~ -a
-ez
-eza
-ía
-ido
-io
-ismo
-ista
-itud
-m(i)ento
-ncia ~ -nza
-o ~ -eo
-or
-ura

b) Sufijos adjetivales

-'neo/a
-(t)o/a
-(t)orio/a
-lento/a
-ndero/a
-ndino
-ndo/a
-no/a
-ntio/a
-o/a
-ófago/a
-ófilo/a
-ógeno/a
-oide
-ol
-olatra
-ómano
-ón/a
-orro/a
-oso/a
-ota
-ote
-ple
-plo/a
-tario/a
-'tico/a
-uco/a
-udo/a
-uence
-ueño/a
-üeño/a
-ujo/a
-'ulo
-uncho/a
-uno/a
-úpeto/a
-'uple
-'uplo
-urno/a
-usco/a
-abl*o
-áceo/a
-aco/a
-'aco/a
-(i)aco/a
-ado/a
-al
-án
-anco/a
-ano/a
-ante
-ar
-ardo/a
-ario/a
-arra
-asco/a
-ata
-átil
-ato/a
-avo/a
-az
-ble
-bundo/a
-cio
-cundo/a
-dio
-dero/a
-do
-eco/a
-ego/a
-ejo/a
-el
-enario/a
-enco/a
-endo/a
-engo/a
-eno/a
-ense
-eño/a
-eo/a
-'eo/a
-erno/a
-ero/a
-és/a
-esco/a
-ésimo/a
-este
-estre
-eta
-eyo/a
-i
-iano/a
-íbundo/a
-ica
-icida
-icio/a
-ico/a
-icola
-icundo
-ida
-ido/a
-iego/a
-ién
-iento/a
-ífero/a
-ífico/a
-iforme
-ífugo/a
-ígeno/a
-'igo/a
-ígrado/a
-ijo/a
-il
-ílocuo/a
-'ímo/a
-in/a
-indo/a
-ino/a
-io/a
-iondo/a
-ípeto/a
-isco/a
-ista
-ístico
-isto/a
-ita
-itimo
-ito/a
-(t)ivo/a
-ívoro/a
-izante
-izo/a

Prefijos en español

A-
Ab-
Ambi-
Anfi-
Ante-
Anti-
Apo-
Archi-
Auto-
Bi-
Bien-
Casi-
Centi-
Circun-
Cis-
Citra-
Con-
Contra-
Cuatri-
Deca-
Deci-
Des-
Dia-
Dodeca-
Ecto-
En-
Endeca-
Endo-
Enea-
Entre-
Epi-
Equi-
Ex-
Exo-
Extra-
Fuera-
Hecto-
Hemi-
Hepta-
Hetero-
Hexa-
Hiper-
Hipo-
Homo-
In-
Infra-
Inter-
Intra-
Intro-
Iso-
Macro-
Mal-
Maxi-
Medio-
Mega-
Meta-
Micro-
Mili-
Mini-
Mono-
Multi-
Neo-
No-
Octa-
Octo-
Paleo-
Para-
Penta-
Per-
Peri-
Pluri-
Poli-
Post-
Pre-
Pro-
Re-
Recién-
Retro-
Semi-
Seudo-
Sex-
Sin-
Sobre-
Sub-
Super-
Supra-
Tetra-
Todo-
Trans-
Tri-
Ultra-
Uni-
Vice-

Anexo 2

1. Las siglas como sujetos de verbo

En el campo de genoma humano se han encontrado las siguientes siglas funcionando como sujetos de verbo:

(1) PCR

La PCR es una técnica || La PCR puede amplificar || La PCR es especialmente valiosa ||La PCR ha tenido || La PCR es un proceso || La PCR ha simplificado || la PCR permite detectar || la PCR ha revolucionado || la PCR consiste en || La PCR ha sustituido || La PCR es un método || La PCR permite clonar DNA || La PCR permite la amplificación || La PCR tiene numerosas || la PCR es una herramienta ||

(2) PGH

El PGH tiene muchas connotaciones positivas || El PGH fue formulado por un administrador || El PGH había emprendido su marcha || el PGH sale bien parado del trance || el PGH tiene bastantes puntos de toque || el PGH es excesivamente costoso || el PGH es ciencia pequeña || El PGH constituye el primer programa || El PGH hará posible el diagnóstico prenatal || El PGH es un consorcio público || Dado que el PGH es un proyecto público || El PGH anunció a principios de mayo que || El PGH es un asalto frontal al ||El PGH ha sido posible gracias al || El PGH constituye la mayor aventura || El PGH ha sido diseñado como un programa || El PGH es un proyecto de investigación || El PGH requiere de la colaboración estrecha entre || El PGH inició con secuenciación de DNA que || el PGH ha invertido también en el desarrollo || el PGH podría resultar más peligroso que benéfico || El PGH es una realidad || El PGH incluye la elaboración de diferentes mapas || Finalmente, el PGH contribuirá al desarrollo de ||

(3) RFLP

Los RFLP proporcionan una provisión || Los RFLP están dispersos abundantemente || Los RFLP han sido una herramienta útil || Los RFLP son importantes por tres razones || los RFLP pueden utilizarse para medir la || Por ello, los RFLP son importantes en los estudios evolutivos || Los RFLP constituyen marcadores moleculares útiles para ||

(4) VIH

El VIH podría convertir se en vehículo eficaz || el VIH necesita un factor de transcripción || el VIH sería un virus antiquísimo || El VIH pertenece a la familia de || el VIH utiliza su capacidad de variar para || El VIH puede tardar ese intervalo en agotar || Pero cuando el VIH entra en acción || el VIH sintetiza grandes cantidades de virus || El VIH sigue en escena || El VIH es uno de estos virus ||

(5) HLA

El HLA contiene múltiples loci de predisposición ||

(6) VNTR

los VNTR son regiones de ADN repetitivo || Los VNTR son secuencias repetitivas ||

(7) ES

Las ES difieren de las células llamadas || Las ES tienen la hardwarea para diferenciarse ||

(8) LET

La LET define la energía media depositada por || La LET es un parámetro importante ||

(9) ADN

el ADN había sido aislado, || el ADN resultaba completamente funcional || el ARN y el ADN son macromoléculas importantes || El ADN está asociado con histonas || Tanto el ARN como el ADN contienen cuatro bases cíclicas diferentes || Los ARN codificados en el ADN son los ARN de transferencia || El ADN formado a partir del genoma || El ADN formado a partir del genoma || El ADN es el material genético || El ADN está en la localización correcta || el ADN contiene A, G, C || El ADN es algo más que una secuencia || el ADN recobra su estado de cadena doble || El ADN permanece en el núcleo || el ADN forma complejos con unas proteínas muy || El ADN constituye una forma magnífica de almacenar || El ADN era tratado primero con un enzima || El ADN dominaba las operaciones de la célula || El ADN contiene un plan codificado || El ADN envolvía entonces al tetrámero|| Si el ADN proviene de células tumorales, || El ADN entra en la célula bien a través de || Si el ADN sufre cualquier alteración,|| El ADN consta de cuatro bloques de construcción || El ADN es la mayor de las moléculas || el ADN puede crear una hélice || Ese ADN ofrece densidades variables || El ADN lleva la información genética || Sólo el ADN es capaz de transmitir la información || El ADN regula la síntesis de determinados productos || El ADN es la molécula responsable del || Estructuralmente, el ADN es una molécula de doble cadena || Por ello, el ADN debía estar formado por dos hebras || Aparentemente, el ADN tiene un protagonismo mínimo || El ADN extraído de las muestras se ha || El ADN es una molécula extraordinariamente simple || El ADN puede degradarse ||

(10) DNA

El DNA suele obtenerse de los linfocitos || El DNA está localizado predominantemente en el núcleo || El DNA es una molécula enorme || el DNA tiene un triple papel biológico ||En tercer lugar, el DNA ofrece la base molecular para la || En términos químicos, el DNA contiene información || Aunque el DNA está compactado, no sufre la || el DNA fabrica el RNA || La inmensa mayoría del DNA codifica información genética || Su DNA está asociado a varios tipos de || El DNA aporta el «molde» para || El DNA tiene una estructura en doble hélice || En el DNA existen dos bases || El DNA posee características especiales || El DNA es una Hélice doble. || El DNA está compuesto por dos cadenas || El DNA forma una molécula bicatenaria || El DNA puede separarse según el tamaño || El DNA debe encontrar se fijado en un || El DNA es una molécula relativamente vulnerable || El DNA puede actuar como cebador ||

(11) ARN

el ARN es un polímero || Generalmente, el ARN es de una sola hebra. || El ARN desempeña una importante función || Los ARN codificados en el ADN son los || Luego, el ARN debe traducirse || El ARN introducido en las células induce fácilmente || El ARN está enrollado helicoidalmente || El ADN y el ARN realizan más actividades químicas vitales || Este ARN será procesado y convertido en ARN || ADN y ARN son dialectos de un mismo lenguaje || El ARN tiene una sola cadena ||

(12) RNA

Como éste, el RNA consta de glúcidos || tanto el DNA como el RNA son ácidos nucleicos || el RNA difiere en varios aspectos || El RNA está constituido por una sola cadena || Como consecuencia, el RNA puede adoptar una gama mucho mayor || El RNA tiene un azúcar ribosa || Los RNA pueden agruparse en dos clases || Algunos RNA actúan de intermediarios || El RNA forma estructuras unicatenarias || El RNA puede ser estudiado también mediante FISH || el RNA debe extraerse del tejido ||

(13) ARNm

Posteriormente, el ARNm penetra en el citoplasma || La importancia del ARNm radica en || Luego, el ARNm porta esta información a los ribosomas || Por tanto, los ARNm tienen una serie de ribosomas || Cada ARNm es sintetizado por transcripción || El ARNm es degradado rápidamente y los ribosomas || Puesto que cada ARNm codifica una proteína || el ARNm es traducido inmediatamente || De este modo, el ARNm tiene que obtenerse mediante transcripción ||

(14) ACs

Asimismo, las ACs pueden afectar tanto a los autosomas || Las ACs pueden ser inducidas al exponer || Las ACs pueden inducirse por || Por lo tanto, las ACs pueden originarse por diferentes tipos || Las ACs son convenientemente clasificadas ||

(15) BRCA1

BRCA1 contiene un dominio ||

(16) CFTR

Existen pruebas de que el CFTR es el mismo canal || el CFTR podría servir directamente de canal || el CFTR formaba por sí mismo un canal || De este modo, CFTR sería un canal de cloro ||

(17) ATP

El ATP es el compuesto rico en energía || El ATP podría hacer falta directamente para fosforilar ||

(18) BRCA2

BRCA2 está además implicado en el cáncer || Al igual que BRCA1, BRCA2 tiene una estructura genética compleja. || en el caso anterior, BRCA2 tiene un dominio de dedos de || BRCA2 contiene 8 repeticiones BRC internas || BRCA1 y BRCA2 tienen patrones de expresión similares || Desde que BRCA1 y BRCA2 fueron aislados, se han descrito || || Por el contrario, BRCA2 está muy relacionado con la aparición || Las mutaciones en BRCA2 confieren un riesgo incrementado ||

(19) ADA

El ADA extiende una clara y abierta prohibición || ADA cataliza la deaminación de adenosina ||

(20) LINEs

Los LINEs son elementos repetidos autónomos || la mayoría de los LINEs son elementos truncados ||

(21) RANTES

RANTES es inducible por mitógenos o antígenos || El RANTES induce la migración transendotelial de ||

(22) EDTA

El EDTA actúa quelando cationes divalentes normalmente necesarios || la función principal del EDTA es inhibir las nucleasas endógenas presentes ||

(23) SINEs

los LINEs y los SINEs carecen de LTRs, || Los SINEs son retrotransposones ||

(24) HUGO

HUGO es la "Organización de las Naciones Unidas para ||

(25) YACs

Los YACs han permitido disminuir drásticamente el número || Muchos YACs pueden cubrir genes humanos completos. || Sin embargo, los YACs presentan un elevado porcentaje de quimerismo ||

(26) UNESCO

La UNESCO publicó en 1997 una "Declaración ||

(27) Sida

El virus del sida es uno de estos retrovirus || Sólo en Estados Unidos, el sida ha matado a más de 350.000 personas ||

En el campo de medio ambiente se han encontrado las siguientes siglas funcionando como sujetos de verbo:

(1) CE

La CE ha elaborado una lista de 23 || La CE tiende, sin embargo, a exportar || La CE ha estado manteniendo contactos ||

(2) UE

La UE regula la gestión y el tratamiento || La UE ha adoptado un amplio programa ||

(3) NEI

los NEI afrontan los problemas derivados de una mala gestión de los residuos ||

(4) EDTA

El EDTA es un ácido || el EDTA puede producir enlaces de || Después que el EDTA ha formado complejos con ||

(5) PVC

Considerando que el PVC es un material normalmente presente en los vehículos ||

(6) DPD

la DPD produce un color rojo instantáneo || La DPD es tóxica, evite su ingestión. ||

(7) CPOM

la CPOM presenta una tendencia ||

(8) ADN

El ADN contiene toda la información necesaria || En algunas células, el ADN está recubierto por una membrana ||

(9) MEDOC

El MEDOC constituye una de las mayores || se puede ver que el MEDOC tiene una circulación termohalina complicada en || La posición del MEDOC es más ambigua debido a que ||

(10) MEDOR

El MEDOC y el MEDOR tienen aguas profundas muy constantes ||

(11) ICONA

El ICONA presenta el libro del Oso Pardo || El ICONA invierte 8.352 millones en la campaña contra incendios || Dentro del MAPA, el ICONA ha publicado hasta hace poco || El ICONA cuenta además con otro proyecto denominado «pueblos abandonados», ||

(12) MINER

El MINER es también, a través del IDAE, responsable del desarrollo ||

(13) AEDENTAT

Aedenat controla perfectamente el sector energético ||

(14) PNUMA

el PNUMA considera prioritario el fortalecimiento || Con ese fin, el PNUMA ha incluido un componente de formación || En todos los casos el PNUMA hace hincapié en el fortalecimiento de las instituciones || Por último el PNUMA forma especialistas ambientales || La UNESCO y el PNUMA habrían de impulsar prioritariamente el estudio conjunto de || El PNUMA debería prestar su apoyo para facilitar || El PNUMA debería prestar su apoyo continuo || Con este mismo fin, la UNESCO y el PNUMA deberían buscar nuevas fuentes de financiamiento || La UNESCO y el PNUMA deben trabajar conjuntamente || La UNESCO y el PNUMA pueden coordinar los programas ||

(15) NASA

NASA contrató en 1977, con la compañía Westinghouse, la fabricación ||

2. Siglas como objetos de verbo

En el campo de genoma humano se han encontrado las siguientes siglas funcionando como objeto de verbo:

(1) PCR

se realizan PCR múltiples || los autores utilizan PCR ||

(2) RFLP

Este tipo de marcadores se denomina || este método es muy efectivo para encontrar RFLP. || se muestra un ejemplo de cartografía genética utilizando RFLP como marcadores. || argumentaban que era posible construir un mapa de ligamiento completo del genoma humano empleando RFLP. ||

(3) VNTR

por lo que también se denominan VNTR || Estas repeticiones en tándem se denominan VNTR (del inglés «variable number tándem repeats») || son suficientes para amplificar VNTR específicos mediante la reacción en cadena de la polimerasa. || Una de las primeras sondas que detectaron VNTR humanos se obtuvo a partir de ||

(4) ADN

primero, introdujeron ADN de la peligrosa bacteria Staphylococcus || utilizaba agujas de cristal muy pequeñas para inyectar ADN directamente en el núcleo || Una de cada tres o cinco células recibía ADN funcional || si bien todos los seres humanos comparten ADN, no todos compartirán sus beneficios || se consiguió in vitro añadiendo ADN extraído de la bacteria lisa || el material genético de un procariota es ADN || el material genético es ADN, || Cuando se añade ADN a poblaciones de células|| Pero desde entonces se ha introducido ADN en huevos de ratón por microinyección || de hecho, es ADN excepto en los virus de ARN. || El material genético del fago T2 es ADN. || después de absorber ADN sobre un filtro, || Las células sólo emplean ADN. || puede utilizar para aislar ADN unido a proteínas. || En cambio, cuando no hay ADN de cadena simple || el primer virión podría contener ADN || el nucleoide puede contener ADN de más de un genoma. || cuando se centrifugaba ADN en determinadas condiciones || Métodos avanzados para obtener ADN de huesos fósiles más viejos || El segundo grupo de científicos usó una nueva técnica para unir ADN de bacteriófagos || Usando ADN de un sapo, unieron un segmento de un gen || se obtuvo ADN a partir de los leucocitos y se trató con enzimas || || Los plásmidos que contienen ADN para ser clonado || sirve para polimerizar ADN a 72 grados C. || moléculas con capacidad para cortar ADN || tras exponer ADN desnudo || que contienen ADN adyacente al telómero || la mayoría de organismos unicelulares poseen ADN no codificador || inyectaron ADN que contenía un gen || lo que se denomina ADN satélite. || Las células transformadas contienen ADN vírico, || quienes se esfuerzan por recuperar ADN de ese resto. || Clonaron ADN a partir de pieles de quagga, || clonó ADN extraído de una momia egipcia || multiplicaban ADN procedente de un hueso humano || analizaron ADN procedente de mitocondrias, || Wilson y su equipo compararon ADN mitocondrial || esperan dirimir la cuestión analizando ADN que nos ha llegado en los huesos antiguos || comprobar directamente la teoría secuenciando ADN procedente de restos || Endonucleasa que hidroliza ADN de hebra simple y doble. || No degrada ADN de hebra sencilla || Degrada ADN de hebra sencilla y ARN || para sintetizar ADN in vitro o crear sondas específicas. || El fago lambda sólo empaqueta ADN de un cierto tamaño, || La microinyección nos permite introducir ADN de manera directa e individual, || podría extraer y marcar ADN de una muestra de tejido infectado || En ese momento, el tampón contiene ADN y todo un surtido de restos celulares || desarrollar técnicas que combinaran ADN con moléculas que neutralizaran || Como los genes son ADN, el genoma humano es en realidad una molécula de ADN || reinventar e intercambiar ADN entre especies, etc., || Para introducir ADN extraño en mamíferos hay que transferirlo || éstos se basan en mezclar ADN de un organismo no eucariótico ||

(5) DNA

en un trabajo sobre poblaciones de Asia Central usando DNA mitocondrial (mtDNA) || Por ello se ha llegado a denominar DNA parásito || fragmentos del tamaño de megabases al digerir DNA humano || Aunque el proceso genético de la recombinación produce DNA recombinante, || Si hay DNA insertado en el sitio de clonación múltiple, || obtenido tras cortar DNA genómico con la misma enzima || se corta DNA de distintas fuentes con EcoRI || la transcriptasa inversa permite sintetizar DNA tomando como molde RNA || A menudo las bacterias tienen DNA extracromosómico || huevos fecundados que tienen DNA foráneo || Cuando se inyecta DNA que tiene sólo el Sry de ratón || asociada sólo a cromosomas que contenían DNA recién sintetizado || Si no se añade DNA molde, || En eucariotas también se encuentra DNA superenrollado y topoisomerasas || Las bacterias sin tratar no incorporarán DNA de forma significativa, || usando DNA de cadena sencilla como molde || El procedimiento que se sigue para obtener DNA vector depende de su naturaleza || Incluso los extremos romos pueden utilizarse para construir DNA recombinante || Cuando se centrifuga DNA genómico en un gradiente de cloruro de cesio || Esta clase de DNA repetido se denomina DNA microsatélite || A pesar de ello, se ha inyectado DNA recombinante || Para ciertas aplicaciones en que se requiere DNA no fragmentado || permiten obtener DNA a partir de una gota de sangre || Cuando se digiere DNA genómico se generan miles de fragmentos || Se han empleado dos métodos diferentes para extraer DNA plasmídico || resulta óptimo para secuenciar DNA de doble cadena || Para secuenciar DNA de doble cadena procedente de una PCR, || Se empleó DNA cromosómico total para el análisis por RFLP. || Utilizando DNA cromosómico, || ya que si a la reacción se le añadía DNA del plásmido || La PCR permite clonar DNA en pocas horas, || PCR permite la amplificación de secuencias específicas de material que contiene DNA muy degradado || y para intentar identificar DNA de muestras || No todas las pruebas mutacionales emplean DNA || y para mantener DNA extracromosomal || la generación de tecnología automatizada para secuenciar DNA || y baratas de secuenciar DNA, || unificar los términos siendo DNA microarray || Teniendo en cuenta que otras que unen DNA de cadena sencilla, || se desconoce si estas proteínas unen DNA de cadena doble o sencilla. || a la hora de construir DNA || entre las proteínas con homeodominio y las proteínas procarióticas que unen DNA ||

(6) ARN

quien había descubierto un modo de fabricar ARN artificial || junto con 140 genes que codifican ARN ribosómico || además de fabricar ARN mensajero || ciertos agentes con capacidad para degradar ARN || que analiza ARN mensajeros, || no fabricaban ARN mensajero || Algunos virus usan ARN || se mueve a lo largo de él, transcribiendo ARN. || La transcripción no es la única forma de sintetizar ARN || enzimas que son capaces de sintetizar ARN a partir de un molde || de sintetizar ARN a partir de un molde que también es ARN. || de un molde que también es ARN || Dichas reacciones producen ARN mensajeros que rigen proteínas || y proporcionan ARN genómicos para perpetuar el ciclo infectivo || Sin embargo, la célula sólo sintetiza ARN por transcripción || Probablemente entre 2000 y 5000 enzimas están sintetizando ARN en un momento dado || tiene la capacidad de sintetizar ARN sobre un molde de ADN, || Sintetizan ARN muy rápidamente || El núcleo de la enzima sintetiza ARN || Cuando se está produciendo ARN, || Una interacción similar tiene lugar en las cápsidas esféricas que contienen ARN unicatenario || que contienen ARN en lugar de ADN || estos experimentos consistieron

en sintetizar ARN en el tubo de ensayo, || se pueden combinar las dos técnicas, la de multiplicar ARN y la de la reacción en || pero también pueden copiar ARN con menor eficiencia, || excepto por la transcriptasa inversa que prefiere ARN como molde || y a sintetizar ARN mensajero, || y generar ARN mensajeros completos.|| la polimerasa viaja hacia el final del gen, sintetizando ARN en el camino hasta ||

(7) RNA

la transcriptasa inversa permite sintetizar DNA || ya que éstos producen RNA de doble filamento || son capaces de utilizar la transcriptasa inversa para transcribir RNA en DNA || la propiedad singular de transcribir RNA en DNA, || la transcripción de la región adyacente que tiene capacidad para formar RNA || una cadena de DNA utilizando RNA como molde || Esta reacción utiliza RNA como sustrato || El armazón contiene RNA nuclear heterogéneo || en los que producen RNA no codificante || cuando se trata de obtener RNA || que son virus que presentan RNA como genoma || Los genes en el humano se dividen en los que producen RNA no codificante || Estos virus contienen RNA en sus viriones, || el uso de microarrays para identificar RNA mensajeros || Además, alberga RNA, relacionado con la síntesis de proteínas específicas. || y transcribir el DNA para producir RNA que resultará ||

(8) ARNm

las células para fabricar ARNm a partir del ADN || HGS y sus asociados han analizado ARNm de muchas muestras de tejidos || Los genes que rigen los receptores han sido identificados buscando ARNm cuyas secuencias || sus órdenes moleculares se leen ARNm mediante || No se detecta ARNm del gen, || se observó que algunos pacientes no tenían ARNm, || se caracterizaban por no tener ARNm detectable en el Northern blot. ||

(9) ACs

Karl Sax demostró que las radiaciones ionizantes inducen ACs. || Estos agentes inducen ACs a partir de la || cómo las radiaciones pueden inducir ACs, || entre la eficiencia en generar ACs y la capacidad de inducir DSBs|| más eficientes en inducir ACs que las radiaciones || para formar ACs, || tienen potencial para inducir ACs || la HRR conservativa puede inducir ACs || la reparación por SSA y NHEJ puede inducir ACs de tipo intercambio || Si la proporción de reparación de las lesiones que generan ACs es rápida ||

(10) CFTR

de conductancia de transmembrana de la fibrosis quística, se denomina CFTR, || que normalmente no expresan CFTR provoca la activación || que denominaron CFTR (Cystic Fibrosis Transmembrane conductance Regulator), || sin que se sepa aún exactamente cuáles expresan CFTR. ||

(11) ATP

se utiliza para generar ATP, || utilizan la energía de la luz para fabricar ATP y para || Se forman ATP y la coenzima reducida NADH. || se libera CO2 y se produce ATP y las coenzimas || Las reacciones a la luz también generan ATP || y dicha enzima requiere ATP para llevar a cabo la unión || Los mutantes incapaces de convertir ATP en AMPc no pueden || utilizan la energía de la luz para generar ATP || se utilizan para generar ATP a partir de compuestos || dominios enlazantes de nucleótido, captan e hidrolizan ATP || En el ratón el ejercicio que consume ATP, necesita ácidos grasos y glucosa para aporte energético || Además de producir ATP mediante la fosforilación oxidativa, || incapaz de sintetizar ATP || generando ATP para el aparato contráctil || La cabeza contiene una región capaz de ligar ATP e hidrolizar un enlace fosfato || los hidratos de carbono son utilizados por el organismo para formar ATP || Ambas reacciones regeneran ATP de forma muy rápida || Las fibras musculares durante el ejercicio isquémico no generan ATP a través de la vía ||

(12) EDTA

insertos en un medio que contiene EDTA, sarcosil o || A medida que se agrega EDTA en la solución || Al agregar EDTA al complejo ||

(13) YACS

Los vectores adecuados para las levaduras se denominan YACs, que es la abreviatura de ||

(14) Sida

a los que tienen sida y a los ancianos || muchos hemofílicos han desarrollado SIDA como consecuencia del ||

En el campo de medio ambiente se han encontrado las siguientes siglas funcionando como objeto de verbo:

(1) DBO

los constituyentes serían los que siguen DBO ||

(2) NASA

patrocinado por el Departamento de Energía, siendo NASA el organismo técnico que administra ||