Caracterización de las siglas especializadas: el caso de los ámbitos de genoma humano y medio ambiente[1]

John Jairo Giraldo
Universidad de Antioquia
(Medellín, Colombia)

El propósito de este artículo es avanzar en el estudio y la caracterización de las siglas en el discurso especializado en español. La metodología empleada en este trabajo parte de la constitución de un corpus de siglas de las áreas de genoma humano y medio ambiente que permita observar el fenómeno desde dos ópticas diferentes: la lingüística y la estadística. El análisis de los datos ha permitido observar que el patrón de comportamiento de las siglas desde la óptica lingüística es similar en ambas áreas; sin embargo, desde el punto de vista estadístico el fenómeno es completamente diferente.

Palabras clave: siglas, terminología, español, discurso especializado, lingüística, estadística, genoma humano, medio ambiente

Characterization of Specialized Abbreviations: The Case of Human Genome and Environment Areas

The aim of this study is to advance the analysis and characterization of abbreviations in specialized Spanish discourse. The methodology used in this work is based on the constitution of an abbreviation corpus from human genome and environment areas that allows the examination of the phenomenon from two different points of view: linguistics and statistics. The data analysis has allowed the observation that the behavior pattern of abbreviations from the linguistic point of view is similar in both areas; however, from the statistical point of view the phenomenon is completely different.

Key words: abbreviations, terminology, Spanish, specialized discourse, linguistics, statistics, human genome, environment

---

1 Este artículo se deriva de la investigación *Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente,* que se hizo de marzo de 2008 a julio de 2011, financiado por la Universitat Pompeu Fabra y la Universidad de Antioquia.

Caractérisation des abréviations spécialisées : le cas des domaines de Génome Humain et d'Environnement

Le but de cette étude est d'avancer dans l'analyse et la caractérisation des abréviations dans le discours spécialisé en espagnol. La méthodologie employée dans ce travail part de la constitution d'un corpus d'abréviations des domaines de génome humain et d'environnement qui permet d'observer le phénomène de deux points de vue différents: linguistique et statistique. L'analyse des données a permis d'observer que le patron de comportement des abréviations du point de vue linguistique est similaire dans les deux domaines; cependant, du point de vue statistique le phénomène est complètement différent.

Mots clés : abréviations, terminologie, espagnol, discours spécialisé, linguistique, statistique, génome humain, environnement

INTRODUCCIÓN

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), 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 esto, Baudet sostiene que la siglación amerita un estudio cuidadoso por parte de lexicógrafos y, sobre todo, de terminólogos.

Aunque un fenómeno de reducción léxica, como el 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 se cuenta con trabajos como los de Rodríguez (1981), Alvar y Miró (1983), Martínez de Sousa (1984) y Casado Velarde (1985). Y desde la terminología, donde destacan los trabajos de Cardero (2002) y Fijo (2003).

Sin embargo, a pesar de los trabajos realizados hasta el momento, no se cuenta con estudios contrastivos sobre la incidencia de las siglas en el discurso especializado en español. En este artículo nos proponemos pues caracterizar, cualitativa y cuantitativamente, las siglas especializadas a

partir del análisis de un par de ámbitos de especialidad como son genoma humano (GH) y medio ambiente (MA).

Marco conceptual

A continuación se presenta el grupo de conceptos principales que se abordarán a lo largo de este artículo tanto desde el punto de vista lingüístico (cualitativo) como cuantitativo (estadístico). Así pues, 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; ejemplo: PCR, TS, TEP, Grb2, etc. (Giraldo, 2010). Distinguimos dos tipos de siglas, a saber:

- Siglas propias. Unidades de reducción formadas exclusivamente a partir de las iniciales de unidades léxicas de estructura sintagmática; ejemplo: PCR (*Polymerase chain reaction*), DMD (*Distrofia muscular de Duchenne*), etc.
- 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 (o significado de la sigla). 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; ejemplo: Grb2 (*Growth factor receptor-bound protein 2*), SRY (*Sex determining region Y*), SEF (*Superficie eficaz*), 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; ejemplo: HUGO (*Human Genome Organization*), ICONA (*Instituto para la conservación de la naturaleza*), 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), los cruces pueden adoptar formas diferentes de acuerdo con los

segmentos que los integran. Por tanto, pueden combinar los segmentos iniciales del primer y segundo elemento del sintagma; ejemplo: GeneBio (*Geneva Bioinformatics*), PubMed (*Public access to MEDLINE*), Agrimed (*Agricultura mediterránea*), etc. O pueden combinar el segmento inicial de la primera unidad y el segmento final de la segunda; ejemplo: Informática (*información automática*). También 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); ejemplo: Tergal (*poliéster galo*).

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; ejemplo: *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.

Por último, el término "siglometría" se emplea para denominar el estudio de 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, ejemplo: PCR consta de tres caracteres (letras) o BRCA1, que 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 con fines de creación de diccionarios.

METODOLOGÍA

Como se ha mencionado anteriormente, el objetivo de este trabajo consistió en observar las siglas en textos de especialidad de genoma humano y medio ambiente y establecer sus características lingüísticas y estadísticas. Para ello se estableció la siguiente metodología:

Constitución del corpus

Para cada ámbito de especialidad, GH y MA, se constituyó un corpus textual y, a partir de cada uno de ellos, el corpus de siglas.

- Corpus textual. A partir del Corpus Técnico del Instituto Universitario de Lingüística Aplicada (CT-IULA) se recogieron 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.
- Corpus de siglas. A partir de los corpus textuales se conformó el corpus de siglas respectivo. Dicho corpus cuenta con 800 siglas en GH y 317 en MA. Como se explicará más adelante, se fusionaron ambos corpus para el análisis lingüístico mientras que para el análisis estadístico se trataron tanto juntos como por separado.

Recolección de los datos

Para la recolección de datos se establecieron los siguientes pasos:

- Depuración manual. Ambos corpus presentaron 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 mayúsculas. Se eliminaron manualmente del corpus todas aquellas unidades que no fueran estrictamente siglas. El criterio de eliminación ha sido la definición y la tipología de siglas indicadas anteriormente. Algunos ejemplos del ruido eliminado en el corpus corresponden a elementos como: apellidos (ejemplo: McLeod, DeLisi, McCArty, DiGeorge); fórmulas químicas (ejemplo: NaCl); secuencias de bases químicas (ejemplo: ATGC, AGC, GAT); palabras en mayúscula sostenida (ejemplo: CIENCIA, PROYECTO, GENOMA); otras unidades de reducción léxica, especialmente abreviaturas, (ejemplo: Kmh), que no constituyen ruido propiamente dicho, pero que no son el foco de nuestro estudio, etc.

Una vez depurado el corpus, surgió la dificultad de que la mayoría de las siglas carecía de su forma desarrollada o significado. 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 fue el siguiente:

En primer lugar, se procedió a buscar la sigla y su forma desarrollada en el resto del CT-IULA. En aquellos casos en los que

no se logró confirmar por este medio, se continuó la búsqueda en otras fuentes, así:

- Búsqueda en fuentes específicas, a saber: en todo el corpus técnico del IULA correspondiente a los campos de especialidad tratados (GH o MA)
- En bases de datos de siglas, diccionarios y glosarios electrónicos en Internet sobre genoma humano, a saber: AcroMed, Medical Dictionary on-line, Diccionari Enciclopèdic de Medicina, Human Genome Acronym List, Glosarios de Biotecnología, Glosarios de Genética, Genetics home reference y Merck Source.
- En bases de datos de siglas, diccionarios y glosarios electrónicos en Internet sobre medio ambiente, a saber: Compendium of Environmental and professional Acronyms, U.S. Environmental Protection Agency,U.S. Global Change Research Information Office, Diccionario de la contaminación, EEA Multilingual Environmental Glossary.
- En páginas web, a saber: List of Acronyms on the Literature on Genome Research, Abreviaturas de genes, Nombres de proteínas y genes e Índice de acrónimos y siglas comunes en Bioquímica y Biología molecular.
- Búsqueda en fuentes generales, a saber: buscadores de siglas (Acronym finder, Acrophile, Acronym server, Abbreviations. com y Acronym search.
- En buscadores, a saber: Google y Scirus.

- Confección del corpus definitivo. Después de la depuración manual y de completar las formas desarrolladas de las siglas, se procedió a la creación de una Base de Datos (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.

Análisis de los datos

En primera instancia se analizaron las siglas del corpus de GH. Posteriormente, los resultados derivados de dicho análisis se cotejaron 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. Para realizar la descripción lingüística de las

siglas en el corpus unificado de GH y MA, se realizó un análisis desde los siguientes puntos de vista: morfológico, sintáctico y semántico. Para realizar el análisis cuantitativo general de las siglas en el corpus de GH y de MA, se siguieron los siguientes pasos:

- Análisis estadístico descriptivo (porcentaje de siglas que presentan su forma desarrollada en el corpus de GH y MA; porcentaje de siglas que presentan variantes formales en el corpus de GH y MA; tipo de siglas más frecuente en el corpus de GH y MA; porcentaje de siglas que representan términos (siglas especializadas) en el corpus de GH y MA; porcentaje de siglas creadas en inglés en el corpus de GH y MA; y porcentaje de siglas creadas en español en el corpus de GH y MA).
- Análisis estadístico inferencial. Con este análisis se buscaba conocer cuáles son los rasgos que establecen las diferencias más significativas en los dos ámbitos de especialidad.

Se diseñó una base de datos para el almacenamiento de las unidades que conforman el corpus de siglas de GH y MA. A continuación se indica el modelo de ficha empleado para la recopilación, análisis y presentación de algunos de los datos obtenidos.

ANÁLISIS Y RESULTADOS

Análisis lingüístico

Para este análisis tomamos de los corpus de GH y MA una muestra de las siglas más frecuentes hasta un máximo de diez por cada tipo de sigla. Partiendo de esta precisión inicial, se ha procedido al estudio de los rasgos de las siglas, que por razones de espacio limitaremos aquí solo a los aspectos morfológico, sintáctico y semántico.

La muestra de 67 siglas para este análisis se ha obtenido de la siguiente manera:

- Unión de los corpus de GH y MA
- Extracción de las siglas por tipo (propias, mixtas)
- Selección de las siglas más frecuentes de cada tipo (hasta un máximo de 10).

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

Tabla 1. Muestra para el análisis

Ámbito	Tipo de sigla	Sigla	FD
GH	Propias	PCR	Polymerase Chain Reaction
		PGH	Proyecto Genoma humano
		RFLP	Restriction Fragment Length Polimorphisms
		VIH	Virus de la inmunodeficiencia humana
		HLA	Human Leukocyte Antigen
		MHC	Major Histocompatibility complex
		VNTR	Variable Number of Tandem Repeats
		ES	Embryonic Stem
		FQ	Fibrosis quística
		LET	Linear Energy Transfer
	Mixtas típicas	ADN	Ácido desoxirribonucleico
		DNA	Deoxyribonucleic Acid
		ARN	Ácido ribonucleico
		RNA	Ribonucleic Acid
		ARNm	ARN mensajero
		Acs	Aberraciones cromosómicas
		BRCA1	Breast Cancer Susceptibility Gene 1
		CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
		ATP	Adenosine Triphosphate
		BRCA2	Breast Cancer Susceptibility Gene 2
	Mixtas acrónimos	ADA	Adenosin Desaminase
		LINE/LINEs	Long Interspersed Nuclear Element
		RANTES	Regulated on Activation, normal T Cells expressed and secreted
		EDTA	Ácido etilendiaminotetracético
		SINE/SINEs	Small interspersed repetitive elements
		HUGO	Human Genome Organization
		YACS	Yeast artificial Chromosomes
		ITIM	Immunoreceptor Tyrosine-based Inhibition Motifs
		UNESCO	Organización de las Naciones Unidas para la Educación, la Cultura y la Ciencia
		JAK2	Janus Kinase 2
	Mixtas cruces	MegaYACs	Mega Yeast artificial Chromosomes
		CalTech	California Institute of Technology
		PiGMap	Pig Gene Mapping Project
		PHRAP	Phragment Assembly Program
		Genpept	GenBank Gene Products
		PubMed	Public MEDLINE
	Siglónimos	Sida	Síndrome de inmunodeficiencia adquirida

Ámbito	Tipo de sigla	Sigla	FD
MA		CE	Comisión Europea
		DBO	Demanda bioquímica de oxígeno
		UE	Unión Europea
		OCDE	Organización de Cooperación y Desarrollo Económicos
		OD	Oxígeno disuelto
		CP	Código penal
		UCP	Unidad de Carbón Piedra
		DBO5	Demanda bioquímica de oxígeno a los 5 días
		NEI	Nuevos Estados independientes
GH	*Propias*	PCR	Polymerase Chain Reaction
		PGH	Proyecto Genoma humano
		ACP	Análisis en Componentes Principales
		EDTA	Ethylenediamine Tetraacetic Acid
	Mixtas típicas	PVC	Polyvinyl Chloride
		DPD	Diethyl-p-phenylene diamine
		PNPP	Para-nitrophenylphosphate
		HCFC	Hydrochlorofluorocarbon
		CPOM	Coarse and fine particulate organic matter
		PCB	Polychlorinated Biphenyls
		ADN	Ácido desoxirribonucléico
		TLm	Threshold Limit
		FCSE	Ley de fuerzas y cuerpos de seguridad
		MEDOC	Mediterráneo occidental
		MEDOR	Mediterráneo oriental
	Mixtas acrónimos	ICONA	Instituto para la Conservación de la Naturaleza
		MARPOL	Convention for the prevention of pollution from ships
		MINER	Ministerio de Industria y Energía
		AEDENAT	Asociación Ecologista de Defensa de la Naturaleza
		PNUMA	Programa de Naciones Unidas para el Medio Ambiente
		NASA	National Aeronautics and Space Administration
		TRAGSA	Transformación agraria s.a.
		FAPAS	Fondo para la Protección de los Animales Salvajes

A partir del análisis lingüístico hecho a la muestra tomada, se obtuvieron los siguientes hallazgos:

- La totalidad de las siglas tiene como núcleo a un nombre, hecho que corrobora la naturaleza nominal de estas unidades. En efecto, se observa que 46 siglas corresponden a nombres comunes en tanto que 21 corresponden a nombres propios. Estos datos están

en consonancia con lo dicho por Nakos (1990), quien sostiene que sólo un número muy restringido de siglas científico-técnicas utiliza el nombre propio.

- Al asumir que toda sigla proviene de un sintagma nominal, cuyo núcleo es un nombre, asumimos también que las siglas tienen género y número. Así pues, toda sigla tiene un género: el del sustantivo núcleo del sintagma que constituye su base. Por tanto, PGH, cuya primera letra es la p, de proyecto, es masculino y singular (el PGH). 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). En lo que respecta a nuestro estudio, hemos detectado que 12 de las 67 siglas bajo análisis presentan la marca de plural. A continuación se presentan algunos casos: RFLP/RFLPs (*Restriction fragment length polymorphism-s*); VNTR/VNTRs (*Variable number of tandem repeat-s*); ADN/ADNs (*Ácido-s desoxirribonucleicos*), etc.
- 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. 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. Basándonos en la tabla de sufijación nominal y adjetival de la Gramática descriptiva de la lengua española, se buscaron 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". También se analizaron los prefijos a partir de la tabla existente en la Gramática descriptiva de la lengua española. En nuestro caso, se han detectado 5 siglas prefijadas, a saber: MegaYACs, retro-PCR, pre-mRNA, pre-ARNm, y anti-VIH.
- 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. Este fenómeno lleva a la detección de 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. Ejemplos de esta categoría son: *Fibrosis quística* (FQ), *Proyecto Genoma Humano* (PGH), *Nuevos estados independientes* (NEI). 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: *Mediterráneo occidental* (MEDOC), *Mediterráneo oriental* (MEDOR), *ARN mensajero* (ARNm).

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: a) 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 (ejemplo: *Demanda bioquímica de oxígeno a los 5 días* → DBO5), o bien cuando se eliden elementos de la forma desarrollada como preposiciones, conjunciones y artículos (ejemplo: *Programa de las Naciones Unidas para el Medio Ambiente* → PNUMA); y, b) Correspondencia parcial de tipo B: se da cuando la sigla incluye tanto caracteres iniciales como internos de los elementos de la forma desarrollada (ejemplo: *Ácido Desoxirribonucleico* → ADN).

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: *Elementos dispersos largos* → LINE (No hay correspondencia ni de iniciales ni de sílabas. En inglés hay correspondencia de iniciales: *Long interspersed nuclear elements*), *Motivos de inhibición en inmunorreceptores basados en tirosina* → ITIM (No hay correspondencia de iniciales ni de sílabas. En inglés hay correspondencia parcial de iniciales: *Immunoreceptor tyrosine-based inhibition motifs*).

Del análisis de la muestra tomada para este estudio 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. 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.

- 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. 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). Algunos ejemplos extraídos del corpus son: secuencia ADN, gen FQ, virus ARN, marcadores HLA, polimorfismo ADN, etc.
- 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. Algunos ejemplos de sinonimia de siglas por traducción son: MHC (*Major histocompatibility complex*) / CMH (*Complejo mayor de histocompatibilidad*); WHO (*World Health Organization*) / OMS (*Organización Mundial de la Salud*). Sin embargo, también se puede dar el fenómeno de sinonimia de siglas en el interior de una misma lengua, ejemplo: (BRCA1 / BRCA I (*Breast Cancer Gene 1*), y YACs / YACS (*Yeast artificial chromosomes*).

- De acuerdo con Rodríguez (1981) 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. El problema de ambigüedad que puede llegar a producir la homonimia es tal que, según estudios llevados a cabo por Liu, H.; Lussier, Y.; Friedman, C. (2001) y Liu, H.; Aronson, A.; Friedman, C. (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. A continuación se presentan algunos ejemplos: ACS (*American Cancer Society* o *American Chemical Society*); APC (*Antigen presenting cell* o *Adenomatous polyposis coli*); CDC (*Centers for Disease Control and Prevention* o *Cell division cycle*); DM (*Depresión mayor* o *Distrofia miotónica* o *Diabetes mellitus*); ER (*Endoplasmic reticulum* o *Estrogen receptor*), etc.

Análisis estadístico

En este apartado se presentan los resultados del análisis siglométrico del corpus. En primer lugar, se tomaron todos los individuos (documentos) que conforman las poblaciones (corpus) de GH y MA. Así, las poblaciones de GH y MA están formadas por 158 y 47 individuos, respectivamente. En segundo lugar, se aplicaron los dos tipos de análisis estadísticos mencionados, es decir, descriptivo e inferencial. A partir de los resultados obtenidos se presentan las conclusiones generales y específicas para cada ámbito de especialidad.

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 se seleccionó el conjunto de variables 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.

Tabla 2. Análisis estadístico descriptivo de las siglas en GH

	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%

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%). 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.

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, se tuvo en cuenta las mismas variables, que se recogen en la siguiente tabla.

Tabla 3. Análisis estadístico descriptivo de las siglas en MA

	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%

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%). Así mismo, se observa 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 de ellas han sido creadas en inglés, pero en MA sobresalen las siglas creadas en español.

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*[2]. Para la realización del ANOVA se tomaron 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 Statgraphics es un programa creado en 1980 por Neil Polhemus para gestionar y analizar valores estadísticos. Este programa se destaca por sus capacidades para la representación gráfica de todo tipo de estadísticas y el desarrollo de experimentos, previsiones y simulaciones en función del comportamiento de los valores. Véase http://www.statgraphics.com/

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 algunas de 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.

- Total de siglas. 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.[3] El ANOVA muestra que el número de siglas es superior en GH. El P-valor es 0.0322.

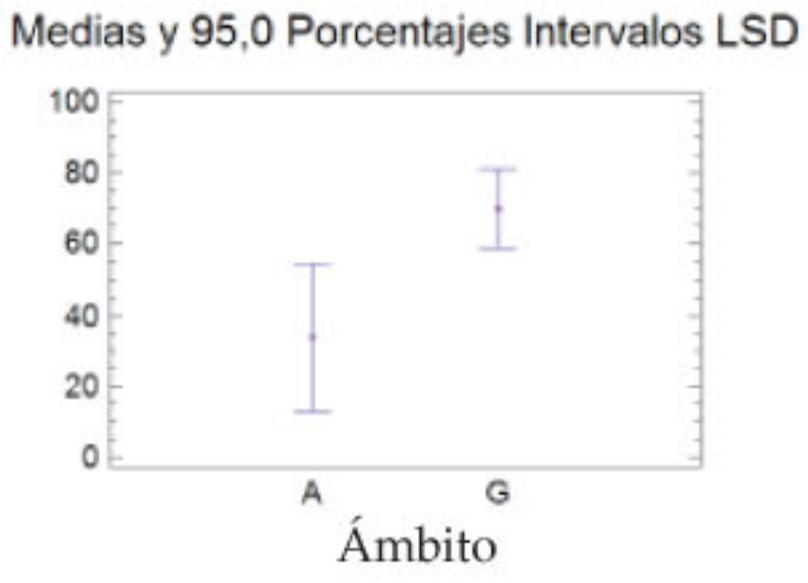


Gráfico 1 Total de siglas

- 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).

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

Por el contrario, en GH se emplean con mayor frecuencia las siglas mixtas (i.e., aquellas siglas que están formadas por caracteres internos de los componentes de la forma desarrollada, cifras, o bien omiten partes fundamentales de la forma desarrollada).[4] Los P-valores para las siglas propias y para las siglas mixtas son 0.0082 y 0.096, respectivamente.

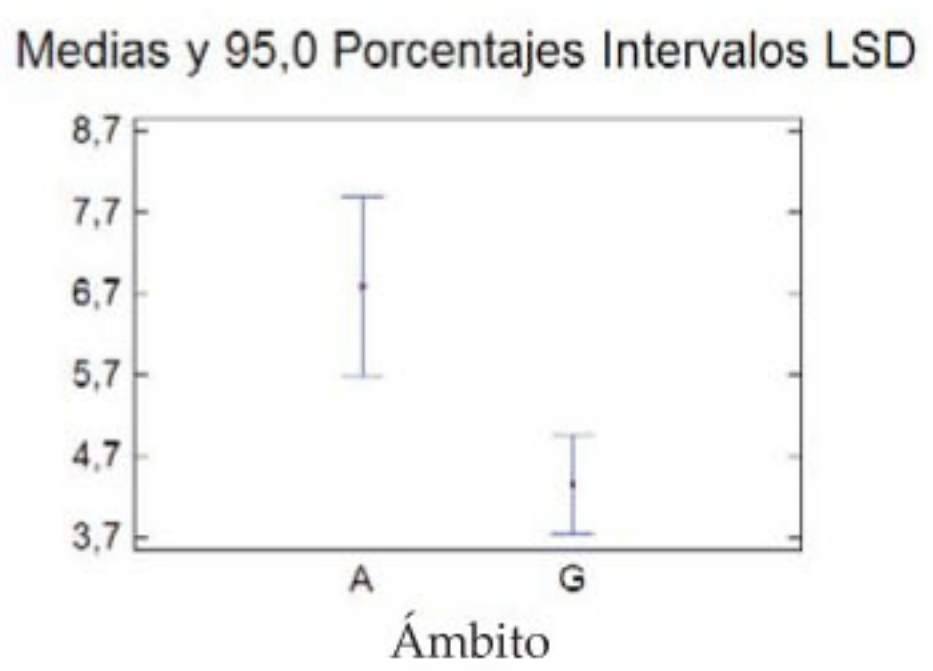


Gráfico 2 Número de siglas propias

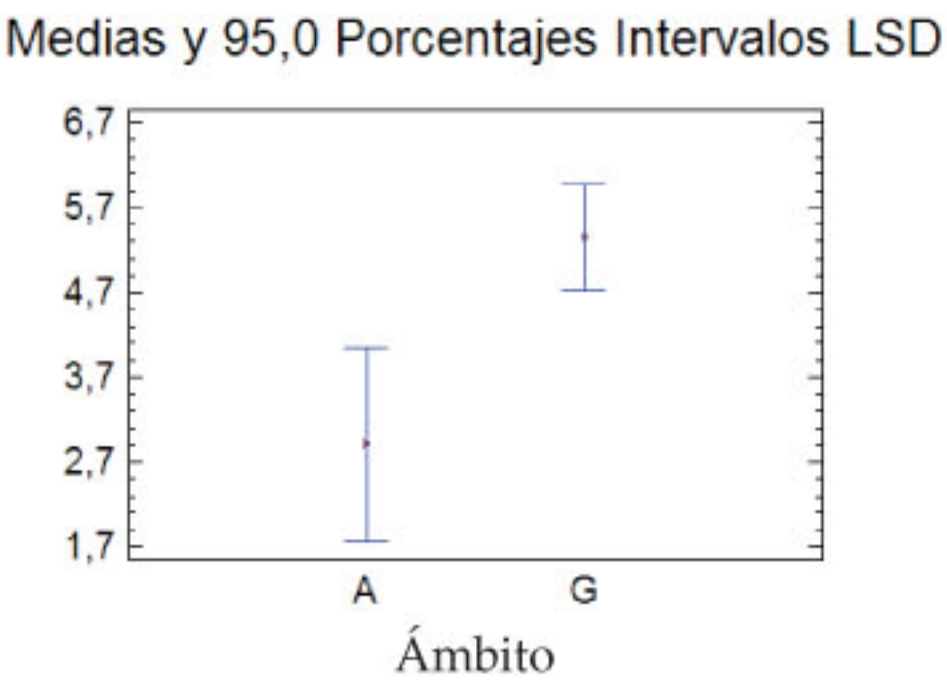


Gráfico 3. Número de siglas mixtas o impropias

- 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.

4 Algunos autores también denominan a las siglas mixtas siglas impropias o sigloides (cf. Martínez de Sousa (1984), Mestres (1996) y Casado Velarde (1985)).

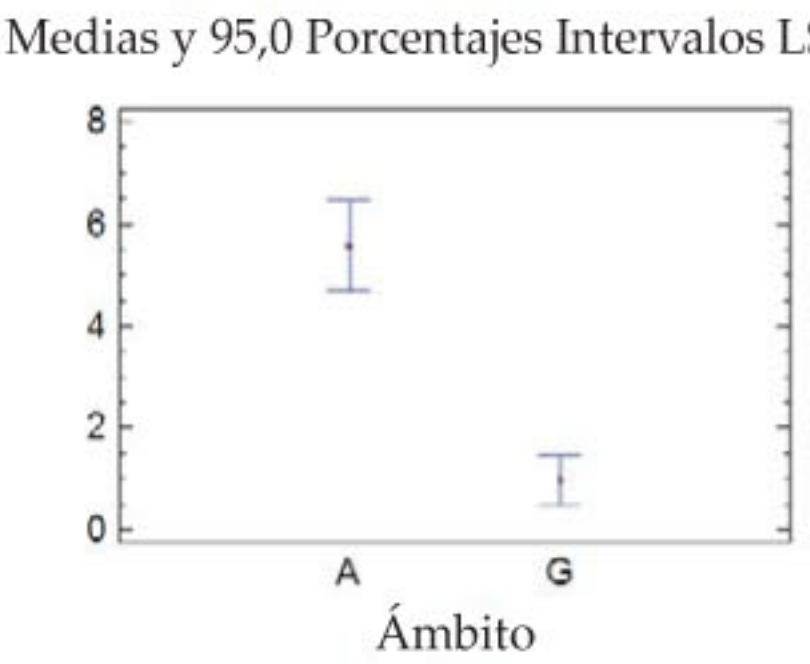


Gráfico 4. Número de siglas generales (nombres de instituciones)

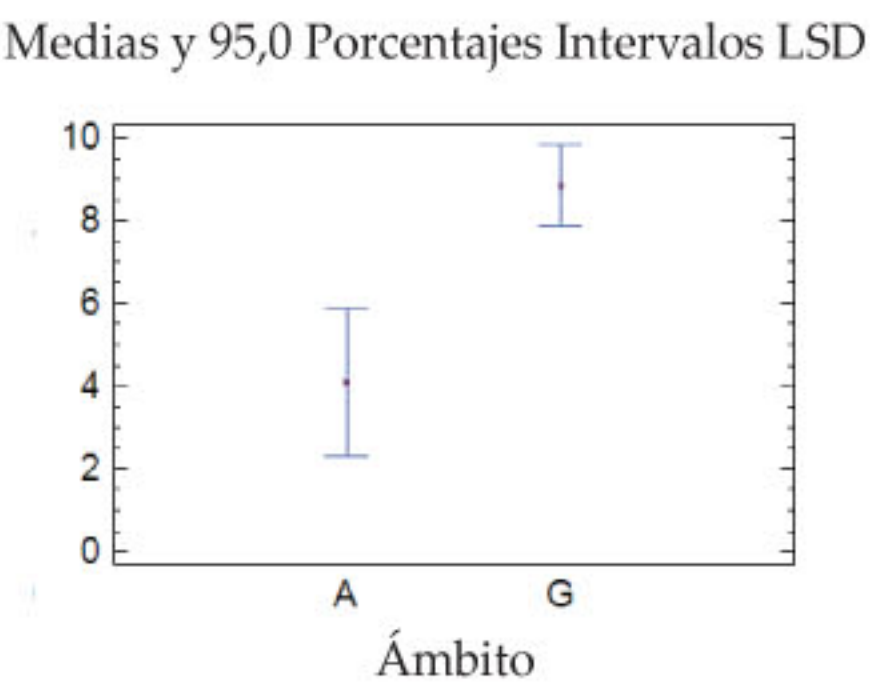


Gráfico 5. Número de siglas del discurso especializado

- 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 0.0013; 0.0000 y 0.0000, respectivamente.

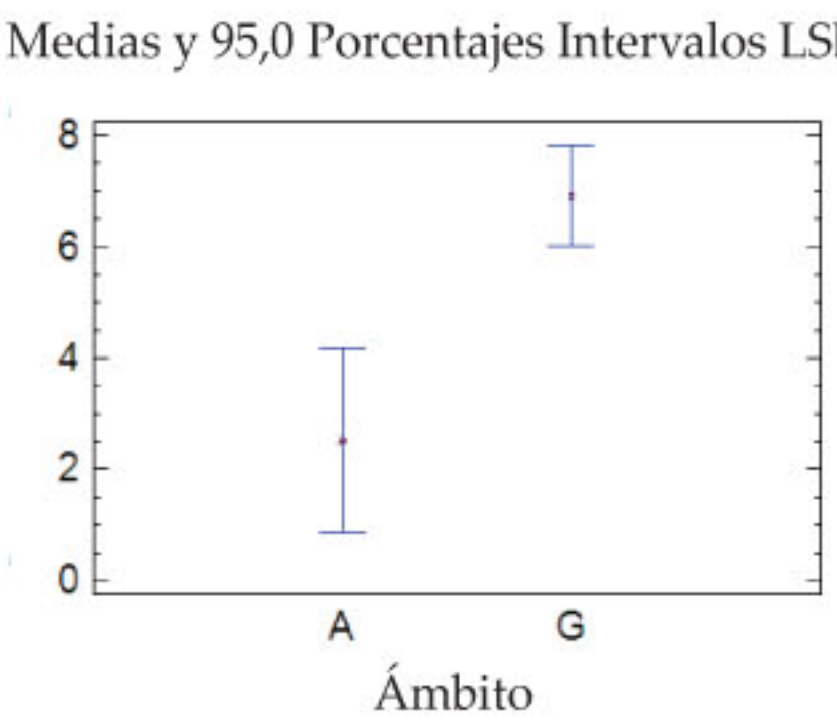


Gráfico 6. Número de siglas creadas en inglés

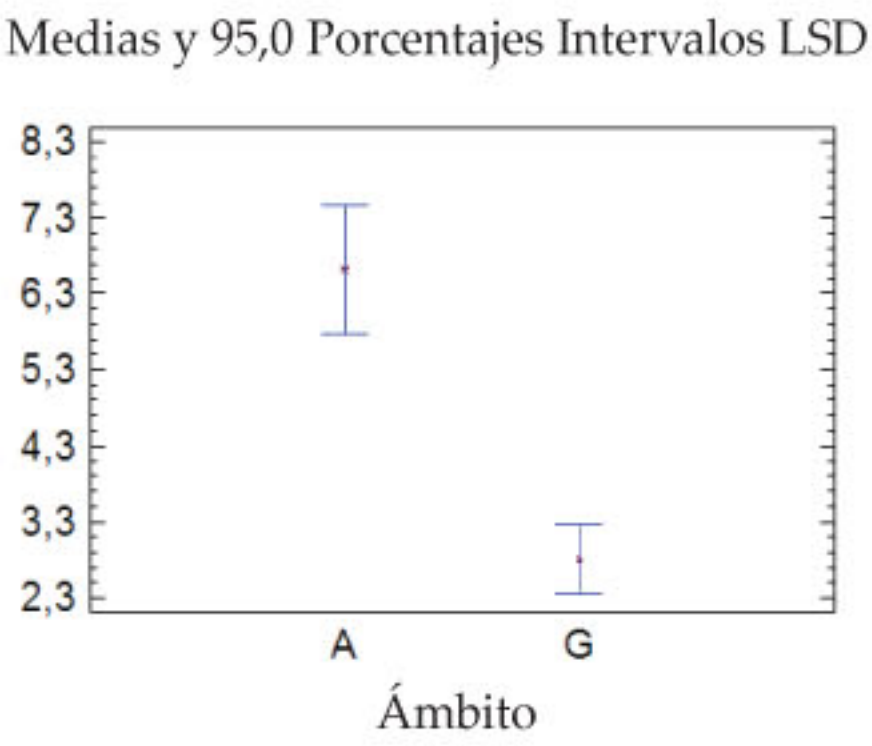


Gráfico 7. Número de siglas creadas en español

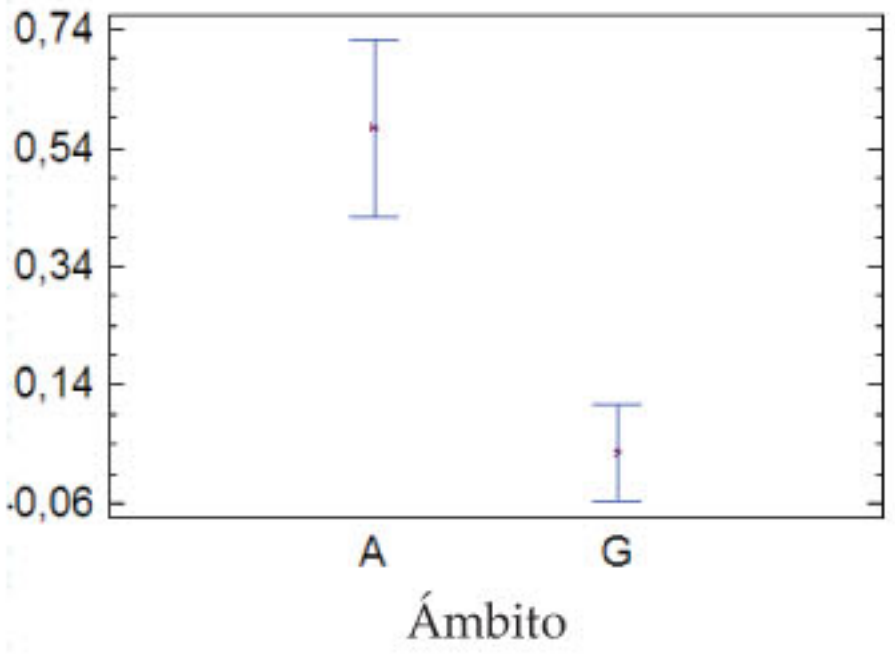


Gráfico 8. Número de siglas creadas en otras lenguas

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

Los dos análisis efectuados anteriormente han 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 o más caracteres), propias, generales y españolas.

Como hemos explicado al principio, se ha efectuado un análisis de componentes principales (ACP), 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, que 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.

Tabla 4. Análisis de componentes principales

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

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

Tabla 5. Pesos de los componentes principales 1 y 2

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

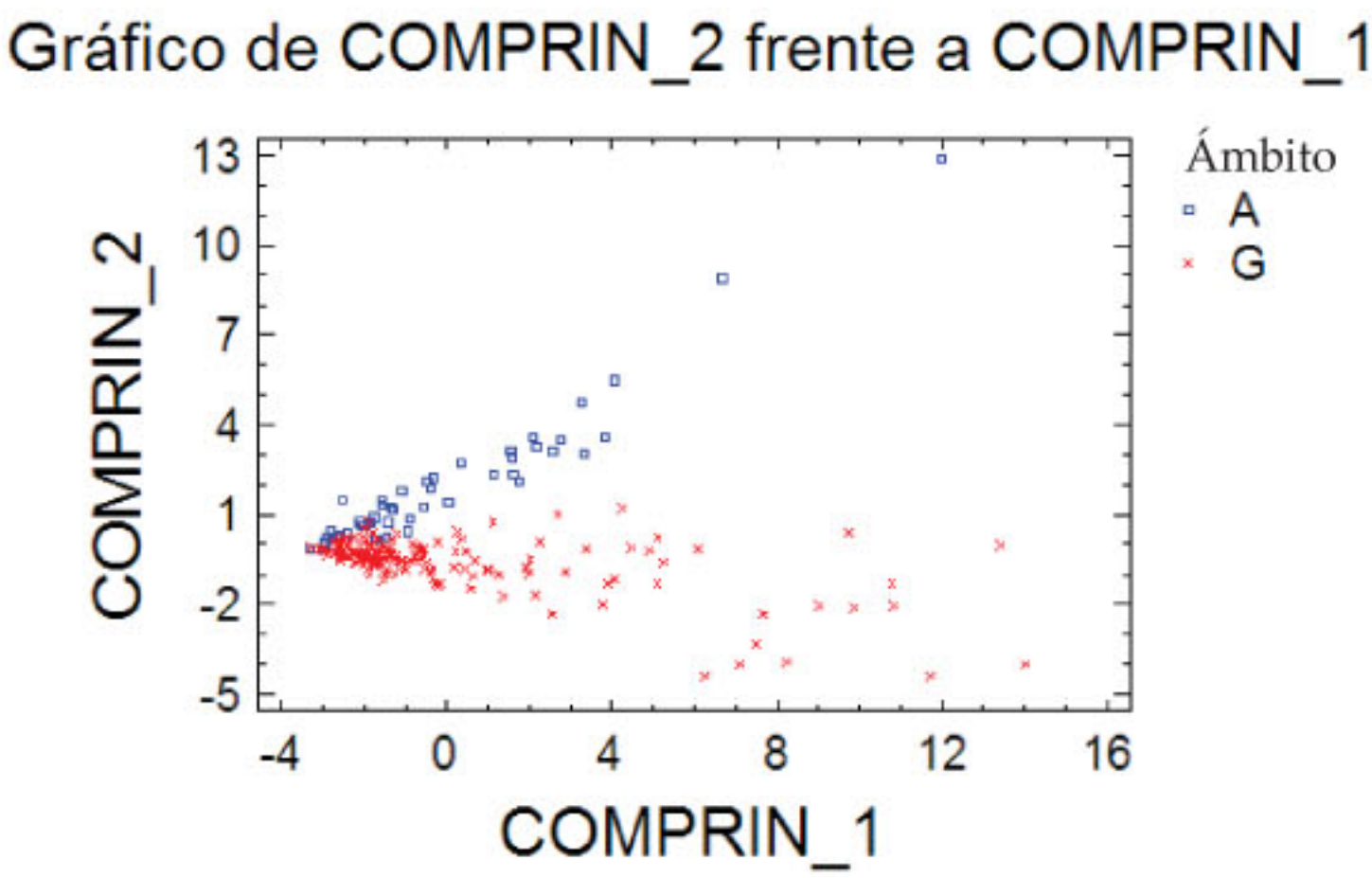
La interpretación de los componentes muestra que el componente principal 1 (COMPRIN1) presenta una alta correlación positiva con todas las variables, hecho que permite asimilarlas al ámbito de GH.

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 por traducción, 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 negro y a GH con gris. 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 9. Comparación de los componentes principales 1 y 2



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 editoriales, mnemotécnicas, estilísticas o de economía lingüística 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 (Languages for Specific Purposes) o la lingüística computacional. En el presente artículo nos hemos propuesto observar las siglas y sus características lingüísticas y estadísticas. En lo concerniente a la descripción lingüística 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 definitiva, 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.

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.

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.

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 han confirmado los indicios de que la aparición de las siglas varía de un ámbito de especialidad a otro.

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 (2000), 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 dos 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.

Por último, puede decirse que la siglación es un fenómeno de frecuencia variable según los ámbitos de especialidad; que las siglas son unidades representativas de términos y de palabras, por tanto, están presentes tanto en el discurso especializado como en el general; que son unidades que corresponden a nombres comunes y, a veces, a nombres propios; y que son unidades que tienden a aparecer sin su forma desarrollada cuando están fijadas en el discurso (el caso de ADN, ARN, ONU, etc.);

REFERENCIAS

Alvar, M. & Miró, A. (1983). *Diccionario de siglas y abreviaturas*. Madrid: Alhambra.

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.

Bloom, D. A. (2000). Acronyms, abbreviations and initialisms. *BJU International, 86*, 1-6.

Cabré, Mª T. (1993). *La terminología: teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.

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.

Casado Velarde, M. (1985). *Tendencias en el léxico español actual*. Madrid: Coloquio.

Fijo, Mª I. (2003). *Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas*. Tesis doctoral inédita, Universidad Pablo de Olavide, Sevilla, España.

Giraldo, J.J. (2010). Hacia una revisión del concepto de siglación. *Panace@, 11 (31)*, 70-76. Consultado el 8 de mayo de 2012 en http://www.medtrad.org/panacea/IndiceGeneral/n31_tribuna_Ortiz.pdf

Liu, H.; Lussier, Y.; Friedman, C. (2001). (2001). *A Study of Abbreviations in the UMLS*. Consultado el 1 de octubre de 2012 en http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243414/pdf/procamiasymp00002-0432.pdf

Liu, H.; Aronson, A.; Friedman, C. (2002). *A Study of Abbreviations in MEDLINE* Abstracts. Consultado el 8 de mayo de 2012 en http://www.ncbi.nlm.nih.gov/pubmed/12463867

Martínez de Sousa, J. (1984). *Diccionario internacional de siglas y acrónimos*. Madrid: Pirámide.

Mestres, J. Mª (1996). La problemàtica de les abreviacions i els diccionaris. *Revista de Llengua i Dret, 26*, 9-28.

Nakos, D. (1990). Sigles et noms propres. *Meta 35 (2)*, 407-413.

Rodríguez, F. (1981). *Análisis lingüístico de las siglas: especial referencia al español e inglés*. Tesis doctoral inédita, Universidad de Salamanca, Salamanca, España.

SOBRE EL AUTOR:

John Jairo Giraldo Ortiz

Doctor en Lingüística aplicada por la Universidad Pompeu Fabra (Barcelona, España). En la actualidad es docente Asistente de la Escuela de Idiomas de la Universidad de Antioquia (Medellín, Colombia). Áreas de docencia e investigación: terminología, traducción y neología.

Correo electrónico: johnjairo.giraldo@gmail.com

Fecha de recepción: 25-10-2011

Fecha de aceptación: 06-09-2012