María Lourdes García-Macho
Manuela Sassi

EL LÉXICO DE *GENERACIONES Y SEMBLANZAS* DE FERNÁN PÉREZ DE GUZMÁN

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ÍNDICE

INTRODUCCIÓN

1. Antecedentes

El trabajo que presentamos deriva de la colaboración[1] empezada hace cuatro años entre la Universidad de Pisa, el Instituto de Lingüística Computacional del Consejo Nacional de Investigaciones de Pisa (C.N.R.), y la Universidad Nacional de Educación a Distancia (UNED) de Madrid. Objetivo de esta colaboración es impulsar el intercambio de recursos y conocimientos entre estas instituciones para evitar en lo posible pérdidas de tiempo y duplicaciones[2].

En un primer momento fijamos cuál sería la contribución de cada uno de los participantes, teniendo en cuenta los medios de que disponía cada persona, de sus conocimientos y del tiempo disponible. A continuación realizamos un estudio sobre la manera de llevar a cabo el trabajo con el fin de fijar los objetivos del proyecto, evaluar las posibilidades técnicas ya existentes en el ILC para el análisis y la lematización de los textos, y establecer cuánto tiempo requeriría su eventual adaptación a nuevas exigencias. Para este fin analizamos algunas obras y evaluamos el comportamiento de los recursos informáticos disponibles.

Empezamos con el análisis de tres obras: la *Ylíada en Romance,* de Juan de Mena, *Generaciones y Semblanzas,* de Fernán Pérez de Guzmán y *De la Vida, Muerte, Virtudes y Milagros de la Santa Madre Teresa de Jesús,* de Fray Luis de León, que constituyen un corpus de unas cuarenta mil ocurrencias. El trabajo sobre estas obras resultó muy útil porque, a partir de una experiencia concreta, pudimos aislar los problemas que planteaba el tratamiento de textos antiguos, y por lo tanto pudimos perfeccionar nuestros recursos técnicos.

Los resultados obtenidos en estos dos años han sido: mejorar las posibilidades técnicas con vistas a una reducción del trabajo manual en la revisión de los textos lematizados, la definición más segura de una metodología para el análisis de textos de los siglos XV y XVI, y la producción de las concordancias lematizadas de las tres obras citadas. Una vez acabada esta fase que consideramos preparatoria, hemos emprendido el análisis de textos del siglo XVI empezando por los *Abecedarios* de Francisco de Osuna.

Hablaremos más abajo de la metodología y de los recursos adoptados; para más detalles véase la bibliografía al final.

---

[1] Esta colaboración nació en 1994 con una Acción integrada financiada parcialmente por el Ministerio de Asuntos Exteriores y de Educación de España y por el MURST (Ministero dell'Università e della Ricerca Scientifica) de Italia, siendo coordinada por María Lourdes García-Macho y por Blanca Periñán.

[2] Queremos agradecer a la Dra. Blanca Periñán el enorme interés que ha mostrado siempre por nuestro trabajo y los importantes consejos que nos ha ofrecido, del mismo modo agradecemos al Dr. José A. Pascual su apoyo y la ayuda recibida para la publicación de esta obra, así como a la Dra. Julia Butiñá y al equipo de publicaciones de la UNED.

2. Instrumentos informáticos del ILC

Para nuestro trabajo se necesitan medios informáticos capaces de tratar el material lingüístico y sistemas de análisis para la lematización automática ya que, debido a la gran cantidad de datos, es importante reducir al máximo la intervención manual.

En estos ultimos años, en el ILC se han realizado investigaciones informáticas y muchos programas que podrán ser utilizados para este proyecto. Algunos están ya disponibles, mientras que otros están a punto de ser acabados, como el analizador morfosintáctico de la lengua española moderna[3]. Entre los ya disponibles elegimos los que mejor se prestaban a nuestro objetivo como el Morfsin[4], el DBT y la Estación Lexicográfica, que constituyen una ayuda indispensable en el campo lexicográfico para el análisis, la gestión y el tratamiento de datos lingüísticos. A continuación ofrecemos una síntesis de las prestaciones de estos tres medios.

El Morfsin es un analizador morfosintáctico de la lengua española moderna, pero también fue utilizado con éxito para analizar textos antiguos[5]. Su realización remonta a los años ochenta cuando los *P.C.* todavía no tenían ni la difusión ni las capacidades de hoy, y por estas razones funciona en *main frame* [VM-IBM]. Es un sistema cerrado, porque el diccionario puede ser modificado pero no aumentado indefinidamente. Con algunas modificaciones [cfr. 3.1] fue posible adaptarlo para nuestros fines, con mucho éxito en los resultados.

El DBT es una sistema de interrogación de bases de datos textuales que, generando automáticamente los índices, brinda la posibilidad de obtener las frecuencias y las concordancias tanto de un texto aislado como de un corpus entero. Además ofrece la posibilidad de consultar los archivos para comprobar las co-ocurrencias a una distancia que varía según las exigencias del lexicógrafo. Entre las posibilidades que brinda este sistema está la de imprimir los resultados de las interrogaciones en una forma tipográfica listos para ser editados.

La Estación Lexicográfica es un sistema que se utiliza para realizar archivos de datos lingüísticos, como diccionarios monolingües o bilingües, glosarios, bibliografías, etc. Es un recurso versátil que permite insertar datos y contextos textuales de manera fácil y

---

[3] Este Analizador que funciona sobre *personal computer*, es modular y puede facilitar, junto al análisis morfológico, algunos datos estadísticos relativos al sistema. Nacido de la experiencia del trabajo con Morfsin, ha sido realizado por el ILC, en colaboración con el Consejo Superior de Investigaciones Científicas y con la Universidad de Málaga. El sistema se encuentra todavía en fase de experimentación, pero se estima que pueda estar acabado en 1998.

[4] MORFSIN (Analizador Morfosintáctico de textos en Lengua Española) realizado en 1986 por Antonina Saba, Daniela Ratti, Maria Novella Catarsi, Giuseppe Cappelli, en el Instituto de Lingüística Computacional del Consejo Nacional de Investigaciones de Pisa (C.N.R.).

[5] Se ha sometido a análisis la obra completa de Teresa de Ávila en edición gráfica moderna. Se trata de un corpus de 632.463 ocurrencias así distribuidas: *Libro de la vida* (112.200), *Camino de perfección* (52.038), *Libro de las fundaciones* (74.752), *Moradas del castillo interior* (59.124), *Relaciones* (41.605), *Conceptos del amor de Dios* (15.990), *Exclamaciones* (7.357), *Constituciones* (7.266), *Modo de visitar los conventos* (6.171), *Vejamen* (863), *Respuesta a un desafío* (1.380), *Poesías* (4.705), *Apuntaciones* (4.429), *Epistolario* (244.583).

estructurada, y puede acompañar al usuario desde las primeras fases del tratamiento de datos lingüísticos hasta llegar a la preparación de los mismos para la impresión que se hubiera propuesto realizar.

El DBT y la Estación Lexicográfica, que son independientes de la lengua y se apoyan en el idioma, no necesitan modificación ninguna, mientras que el MORFSIN, para los textos antiguos con grafías originales, necesita algunos cambios en los diccionarios y en las listas de desinencias y afijos.

3. Metodología adoptada para el análisis de *Generaciones y Semblanzas*

Para comprender mejor los cambios que hemos aportado en el Morfsin, vamos a dar a continuación una sumaria descripción de su manera de proceder. Para más detalles véase la bibliografía al final.

3.1. Descripción del Morfsin

Se empieza con un pre-procesador que analiza y lematiza, según las categorías tradicionales, las palabras gramaticales, los nombres propios y las locuciones. El procesador morfológico analiza las demás formas verbales y nominales de la siguiente manera: cada forma del texto se segmenta a partir de la derecha para identificar las posibles desinencias y los sufijos, y a partir de la izquierda para detectar los prefijos. En cada paso de la segmentación se busca en el diccionario el hipotético radical aislado; si existe, se verifica la compatibilidad de los segmentos obtenidos con los encontrados en el diccionario. Si la búsqueda tiene éxito, se obtiene el análisis morfológico completo y el lema de la forma examinada. Esta operación continúa hasta encontrar todas las homografías posibles. Las formas se consideran sólo una vez, y el análisis que se obtiene se asigna a todas las ocurrencias de esas formas: de esta manera se mejora el procedimiento del análisis.

Después de haber analizado completamente un texto, el analizador sintagmático local pasa a la desambiguación de las formas ambiguas, mediante una serie de reglas morfosintácticas; las ambigüedades de tipo funcional se resuelven según la presencia o la ausencia de determinadas categorías que se hallan en el contexto inmediato de la forma examinada.

3.2. Datos lingüísticos y adaptaciones añadidas para el análisis de *Generaciones y Semblanzas*

Como decíamos, para obtener un análisis completo de *Generaciones y Semblanzas* se han modificado: el diccionario de radicales, las listas de desinencias verbales y la de los sufijos. Antes de describir estos cambios, nos parece importante poner de relieve cómo se han estructurado los datos del Morfsin, a partir del diccionario de radicales.

El diccionario del Morfsin contiene los lemas más frecuentes del *Frequency Dictionary of Spanish Words* más las eventuales homografías[6]; cada uno de estos lemas tiene un código que identifica la forma verbal o nominal y brinda la información necesaria para clasificar e identificar el lema. Con un diccionario de alrededor de 7000 entradas, el sistema puede reconocer más de 70.000 formas.

En el esquema 1 se da un ejemplo de dos radicales *am* y *anim* tal como los hemos introducido en el diccionario: las líneas indican las entradas de los lemas *amo*, *amar*, *reamar, animar*, *animal* adj., *animal* sust., *animación, reanimación, reanimar.*

Los códigos a la derecha remiten a la lista de las desinencias [cfr. esquemas 2 y 3]. La letra "b" delante al código S24 indica que el sufijo *-ción* cuando es plural no se acentúa.

Prefijo	Radical	Infijo	Sufijo	N. paradigma
	am-			S01
	am-			V1
re-	-am-			V1
	anim-			V1
	anim-		-al-	A23
	anim-		-al-	S34
	anim-	-a-	-ción-	b S24
re-	-anim-	-a-	-ción-	b S24
re-	-anim-			V1

Esquema 1: Ejemplo de las entradas del Diccionario de Radicales

	MS	FS	MP	FP
A23	*0[7]	0	es	es
S01	*o	a	os	as
S16	-	*a	-	as
S24	-	*0	-	es
S34	*0	-	es	-

Esquema 2 : Lista de las desinencias nominales

Los códigos **MS** y **MP** significan masculino singular y plural; **FS** y **FP** femenino singular y plural; **A** adjetivo; **S** sustantivo; **V** verbo; **1 ps, 2 ps, 3 ps**, primera, segunda, tercera, etc. persona del singular; **1 pp, 2 pp, 3 pp**, primera, segunda, tercera persona del plural; los números **1, 2** y **3** asociados al código del verbo indican la conjugación. Las líneas con puntos indican que se ha omitido una parte de los datos.

---

[6] Como referencia para las homografías se ha utilizado el *Diccionario ideólogico de la lengua española,* de Julio Casares y el *Diccionario de uso del español,* de María Moliner.

[7] El asterisco indentifica la desinencia del lema y el 0 (cero) significa que no hay desinencia.