

ELENA CARPI
Università di Pisa

Memoria de la lengua y memoria electrónica:
los analizadores morfosintácticos
y el estudio del español en Italia

1.

A diferencia de lo que acaecía hace unos años¹, la mayoría de los estudiosos de ciencias humanas ha llegado a aceptar hoy en día el hecho de que el encuentro entre diferentes enfoques metodológicos beneficia la elaboración de contribuciones originales, mucho más que limitar el ámbito de una investigación dentro de los confines de los métodos tradicionalmente pertenecientes a cada disciplina; es precisamente gracias a este cambio de punto de vista que la *lingüística aplicada* ha adquirido una acusada relevancia en campos como la lexicografía, los estudios de lingüística general y el tratamiento informático de grandes *corpora*. Muy importante en relación con la formación de *corpora* electrónicos se ha revelado el papel de la llamada lingüística *computacional*, que se diferencia de la *cognoscitiva* en cuanto no persigue la realización de prestaciones lingüísticas «inteligentes» – como el desarrollo de redes neuronales y de inteligencias artificiales – pero es capaz de proporcionar valiosas herramientas de recolección de informaciones de tipo léxico, gramatical, sintáctico y estilístico.

Conformemente al título del congreso, dedicado a la memoria, voy a recordar cuándo y cómo empezó en Italia la aplicación de los

¹ «Hay que admitir la existencia de opiniones opuestas acerca de las ventajas que el uso y el empleo de técnicas computativas electrónicas aporta al ámbito crítico literario. Al lado de algunos defensores no faltan los detractores quienes consideran estas metodologías auxiliares unos elementos de profanación en la evaluación estética tradicional». Rosselli (sd: 64)

métodos computacionales a la literatura y a la filología españolas, permitiendo la realización de estudios que, en períodos anteriores, habrían abarcado la vida entera de un hombre; presentaré asimismo ejemplos concretos de los datos que se pueden obtener y de las investigaciones que es posible llevar a cabo utilizando el último – en orden de tiempo – analizador morfosintáctico realizado en el Instituto de Lingüística Computacional del Consejo Nacional de Investigaciones (C.N.R.) de Pisa.

Sin embargo, antes de profundizar en el tema, me parece necesario resumir en breve la historia del interés hacia la dimensión cuantitativa textual² que, lejos de ser una invención del siglo de las computadoras, posee una tradición muy larga y se remonta a tiempos en los cuales los *chips* todavía no existían: entre los primeros en plantearse el problema de la frecuencia de las palabras de un texto se cuentan en efecto gramáticos y filólogos griegos y latinos, que ya formulaban consideraciones estilísticas contraponiendo los *hapax* a las formas de uso frecuente. La gramática de los alejandrinos se fundaba en el compromiso entre principios analogistas y anomalistas: la analogía indicaba la norma, mientras que la anomalía designaba la excepción, el uso más raro.

2.

Las primeras concordancias conocidas después de la época clásica datan de la Edad Media: son las de la Biblia, recopiladas con finalidades didáctico-religiosas, y abren camino a lo largo de los siglos siguientes a una amplia serie de libros que reúnen las concordancias de obras maestras de autores clásicos. Sin detenerme en los muchos ejemplos existentes, señalaré que los primeros listados de frecuencias se redactan en el siglo XIX en Berlín y en París, con el propósito de estudiar la frecuencia de grafemas, sílabas y palabras en alemán y francés, para mejorar el trabajo de los taquígrafos. A causa de la creciente atención por las técnicas de aprendizaje de los idiomas extranjeros, al comienzo del nuevo siglo se elaboran

² Cfr. Duro-Zampolli (1968) texto del cual he sacado algunos de los datos que presento.

diccionarios que toman como punto de partida la compilación – todavía manual – de listas de palabras fundamentales; por lo que corresponde al español, se publican valiosas investigaciones como la de Buchanan (1927) y de Keniston (1933), fruto del análisis estadístico aplicado a los textos más variados.

Pero es sólo después de 1950, con la llegada de los elaboradores electrónicos, que se abren nuevas posibilidades para el análisis de textos. Las concordancias de la Biblia, obtenidas en 1957 por John W. Ellison utilizando un elaborador que funcionaba con lámparas y fichas perforadas y tenía tiempos de cálculo muy largos, inauguran la era de la computación en lingüística.

En Italia la L.C. lleva el nombre de Padre Busa, que en 1949 empieza a analizar automáticamente el *Index Thomisticus*, mientras en 1964 la Academia de la Crusca da comienzo a los trabajos de su *Diccionario*. En 1968 se funda en Pisa, bajo la dirección del profesor Antonio Zampolli en aquel entonces coordinador de las actividades literarias IBM en Italia, la sección lingüística del Cnuce (Centro nacional universitario de cálculo electrónico), que llegará a ser un instituto autónomo del CNR (Consejo Nacional de Investigaciones) de Pisa. Es justamente dentro de dicho instituto que se ha formado el grupo de «hispanistas computacionales»³ que ha contribuido a la creación de las herramientas de análisis automático del español a las cuales me voy a referir.

3.

El primer analizador morfosintáctico italiano dedicado especialmente al español data de los años ochenta y se llamaba MORFSIN; anteriormente, la colaboración entre el grupo de técnicos e investigadores de la División Lingüística del CNUCE-CNR y los profesores Ferdinando Rosselli y Alessandro Finzi había desembocado en la publicación de las concordancias y frecuencias de la obra poética de Antonio Machado (1973), del léxico poético de Cesar Vallejo (s.d.) y de las obras dramáticas de Juan del Encina (1983), pero MORFSIN representó para esa época una herramienta de análisis

³ Reunidos por primera vez en Pisa en 1981.

automático de la lengua española totalmente nueva. Sin entrar en detalles técnicos diré que, utilizando el *Frequency Dictionary of Spanish Words* de Juilland y Chang Rodriguez (1964), ese analizador era capaz de lematizar y atribuir la categoría gramatical automáticamente al 80% de ocurrencias de un texto escrito en español moderno. Después de haber nombrado al prof. Rosselli, que en el campo de la crítica literaria española en Italia fue un pionero de esta tipología de investigaciones, me incumbe ahora hablar de otro estudioso que demostró tener una acusada confianza en la aplicación de la lingüística computacional a los estudios humanísticos, el prof. Guido Mancini. De su colaboración con el CNR y de la utilización de MORFSIN nacieron las *Concordanze del «Libro de la Vida» de Teresa de Avila* (Carpi Saba Sassi 1987) y las concordancias y frecuencias de las *Exclamaciones del alma a Dios* (Carpi Saba Sassi 1992) de la misma doctora abulense, así como la memorización en soporte electrónico del entero corpus de la Santa⁴. Para utilizar MORFSIN, pensado para la lengua moderna, con textos escritos en el español del siglo XVI, se modificaron las desinencias verbales y el diccionario de radicales y sufijos, añadiendo p.ej. la flexión – ava a la moderna – aba del imperfecto indicativo, el sufijo -eça a -eza y -ction a -ción, y fue necesario dar al programa informaciones como la inestabilidad gráfica de las sibilantes, como en el caso de la pareja conociessen/conociessen etc. pero, al final de la labor de adaptación técnica, MORFSIN era capaz de reconocer automáticamente hasta el 75,3% de las 112.200 ocurrencias de la autobiografía teresiana. MORFSIN proporcionó en aquel entonces una gran cantidad de datos, y permitió confirmar valiosas intuiciones críticas, como la validez de las reflexiones de Mancini (1955) sobre la elaboración estilística de la obra teresiana. El análisis de la frecuencia de los diminutivos teresianos (Carpi Saba Sassi 1996) con sus respectivos

⁴ El corpus completo llega a 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* (7357), *Constituciones* (7266), *Modo de visitar los conventos* (6171), *Vejamen* (863), *Respuesta a un desafío* (1380), *Poesías* (4705), *Apuntaciones* (4429), *Epistolario* (244.583).

contextos desmentía cuantos negaban la voluntad de estilo de la Santa (Menéndez Pidal 1947), demostrando que para describir la escritura de Teresa de Ávila no se podía hablar ni de casualidad ni de arbitrariedad. La más reciente investigación (Carpi 2000) llevada a cabo utilizando los resultados de MORFSIN concierne la biografía de Teresa de Ávila escrita por fray Luis de León – que la muerte del agustino dejó inacabada; las concordancias de la obra⁵ han representado una eficaz ayuda para cotejar el estilo de la última labor literaria del maestro de León con *De los Nombres de Cristo*, confirmando que la claridad conceptual deseada por el escritor salmantino y los recursos retóricos utilizados para llevarla a cabo se encuentran en su último trabajo magistralmente acabados.

MORFSIN utilizaba las herramientas informáticas del tiempo⁶ y – como es sabido – 10 años en informática corresponden a eras geológicas. La revolución representada por los PC ha llevado a la realización de un nuevo analizador para el español, AyDA⁷ (**A**nalizador y **D**esambiguador **A**utomático). AyDA es un producto versátil: para trabajar no necesita de un texto previamente «marcado» y se puede utilizar desde el ordenador de casa, puesto que ocupa muy poco espacio en el disco duro. El investigador puede crear un glosario personalizado, que participa en el reconocimiento del texto junto al diccionario principal del programa y puede ser aumentado indefinidamente, mientras que el de MORFSIN no podía superar las 500 palabras. La desambiguación de un texto se compone de distintas fases: durante la primera, AyDA identifica las palabras que no es capaz de reconocer. En un segundo momento, AyDA trabaja con el diccionario preparado manualmente por el usuario, que contiene lema y categoría gramatical de las palabras no reconocidas y

⁵ Cfr. Carpi y Mancho Duque (1999).

⁶ Morfsin funcionaba en *main frame* [VM-IBM] y era un sistema cerrado, porque el diccionario podía ser modificado, pero no aumentado indefinidamente.

⁷ AyDA ha sido realizado por Giuseppe Cappelli y Antonina Saba (Istituto di Linguistica Computazionale)[CNR-Pisa], María José Albalá (Consejo Superior de la Investigación Científica) [CSIC-Madrid], María Victoria Marrero (Departamento de Filología de la Universidad Nacional de Educación a Distancia) [UNED-Madrid], Ignacio Moreno Torres (Departamento de Filología de la Universidad de Málaga).

completa el análisis. La que presento es una muestra del *output* de AyDA, lo que se obtiene después de ejecutar el programa y antes de utilizar el *software* que genera las concordancias:

000001	Suma	N[	Suma
000002	de	N[	de
000003	Geographía	N[	Geografia
000004	que	PQ[	que
000005	trata	VOE1IP3	tratar
000006	de	E0[	de
000007	todas	TC[FP	todo
000008	las	RD[FP	el
000009	partidas	S0[FP	partida

La primera columna es el texto de la obra, en este caso la *Suma de Geographía* de Martín Fernández de Enciso, que se lee en vertical; la segunda corresponde a los códigos de la lematización gramatical – p.ej. “Suma de Geographía” está clasificado como nombre propio con la marca N, “que” como pronombre relativo con la marca PQ[–; la tercera es el lema que corresponde a cada forma. Utilizando unos sencillos mandos de búsqueda por lema, es posible agrupar todas las formas de una misma palabra, recuperando así aun las variantes gráficas.

Como ejemplo de las posibilidades de AyDA, voy a presentar unas muestras de su aplicación a la lengua del siglo XVI y al español moderno.

Como paso preliminar a la realización de un diccionario de la navegación del Siglo de Oro, el equipo al cual pertenezco, que integra investigadoras del CNR, de la UNED de Madrid y de la Universidad de Pisa, ha constituido un corpus de manuales de navegación del siglo XVI, denominado LÉNESO (**L**éxico **N**avegación **E**spañol **S**iglo de Oro).

LÉNESO comprende nueve textos entre los más destacados de ese período, por un total de más de 600.000 palabras⁸, los cuales

⁸ El corpus se compone de nueve tratados: *Arte del marear*, (Antonio de Guevara); *Breve compendio de la sphaera y de la arte de navegar*, (Martín Cortés y Albácar); *Compendio del arte de navegar*, (Rodrigo Zamorano); *Cuatritartu en Cosmographía Práctica*, Alonso de Chaves; *Instrucción náutica*, (Diego de García

han sido tratados con el analizador AyDA. Los porcentajes de reconocimiento de ocurrencias y formas de los seis textos ya completamente analizados superan el 80% y en algunos casos rozan el 90%:

textos	pág.	ocurr.	recon.	%	no recon.	formas	recon.	%	no recon.
1. <i>Guevara</i>		15474	13130	84%	2344	3373	2102	63%	1271
2. <i>Cortés</i>		48914	42713	87%	6201	7255	4785	65%	2470
3. <i>Zamorano</i>		22419	19812	88%	2607	2307	1674	64%	633
4. <i>Chaves</i>		145905	127436	86%	18554	6718	4195	62%	2523
5. <i>Palacio</i>	p. 321	58872	49497	84%	9375	7602	4621	61%	2981
6. <i>Enciso</i>	p. 150	78983	70271	88%	8712	8624	6125	71%	2499
	1.154	400.567	322.859		47.793	35.879	23.502		12.377

El diccionario del usuario fue constituido con todas las formas no reconocidas del primer libro memorizado, el *Cuatripartitu en Cosmographia Práctica* de Alonso de Chaves, y fue incrementado a medida que el trabajo avanzaba: debido a la tipología textual, estaba integrado sobre todo por tecnolectos y cultismos. Sometiendo los listados de AyDA al paso por el programa CLE (Concordancias Lematizadas del Español)⁹, se han obtenido concordancias lematizadas de todo el corpus, ordenadas alfabéticamente, que están proporcionando múltiples informaciones sobre el castellano de 1500. Uno de los obligados textos de referencia para cualquier estudioso del español del XVI, siglo que representa la transición entre castellano antiguo y moderno y en el cual llega a estabilizarse una norma gramatical y sintáctica que durante el periodo anterior había sido dudosa e incierta, es – sin lugar a duda – *The Syntax of Castilian Prose* de Hayward Keniston. Dicho ensayo es todavía más interesante para los especialistas de LC, porque el mismo autor lo presenta en la introducción como «*the first attempt in Spanish, and, so far as I know, in any language, to present a statistical analysis of grammatical usage*» (1933: xi); más adelante, el investigador estadounidense subraya a la vez la dificultad de su trabajo y la impo-

de Palacio); *Libro de las longitudes*, (Pedro de Medina), *Summa de Geographia*, (Martín Fernández de Enciso); *Tratado del Esphera y del arte del marear*; (Francisco Falero); *Ytinerario de navegación de los mares y tierras*, (Juan de Escalante de Mendoza).

⁹ Realizado por Rolando Bianchi-Bandinelli (Cnuce-CNR, Pisa).

sibilidad de clasificar todo el material. Por esta razón, aprovechando los datos proporcionados por AyDA, y puesto que los textos utilizados por Keniston se adscriben a los géneros más variados pero ninguno pertenece al sector técnico, me ha parecido interesante intentar comprobar con la ayuda de herramientas informáticas si, al cambiar tipología textual, algunos de los usos sintácticos evidenciados en *The Syntax of Castilian Prose* se confirmaban, o si al contrario existían variaciones de tipo diastrático. En este trabajo, sin la presunción de agotar un tema tan complejo, he fijando mi atención en el uso de los auxiliares.

Keniston transcribe numerosos contextos que permiten identificar en “haber” el auxiliar utilizado en el siglo XVI en la formación de los tiempos compuestos de los verbos transitivos; sin embargo, añade que es posible considerar auxiliares también otros sintagmas verbales, como “tener”, “dejar”, “traer” o “llevar”, cuyo uso está en relación con la significación del participio en épocas precedentes (1933: 449). Desafortunadamente el estudioso norteamericano no ofrece datos numéricos, así que resulta difícil comprender si se aprecian disparidades de frecuencia entre los «otros auxiliares» (1933: 449) o diferencias cuantitativas de estos últimos respecto a “haber”. El examen del corpus de la navegación permite detectar que la presencia de “dejar”, “traer” y “llevar” es en absoluto esporádica, y que no es posible compararla a la de “tener” y sobre todo de “haber”; respecto a los textos didascálicos-literarios, esta diversidad se puede explicar con la tendencia a la simplificación sintáctica que también actualmente caracteriza la prosa técnica.

	Dejar	Traer	Llevar	Tener	Haber	Total
Enciso	3	0	1	14	139	157
Chaves	1	0	0	28	220	249
Cortés	0	0	0	12	86	98
Guevara	1	0	0	7	33	41
Palacio	0	1	0	15	186	202
Siria	0	0	1	18	122	141
Zamorano	0	0	0	4	72	76
Total	5	1	2	98	858	964

El porcentaje de ocurrencias de “haber” confirma su predominio sobre “tener”: su frecuencia aumenta y se estabiliza hacia el final del siglo, como demuestran las obras de Zamorano (1581), Palacio (1587) y Siria (1602) si se cotejan con las de Enciso (1519), Chaves (1520-37) y Guevara (1539). Por lo que concierne la función de “tener”, en los contextos que presento se puede notar que, cuando no se utiliza como auxiliar en lugar de “haber”:

Ya tenemos provado como la marea creciente [...] (Chaves 48v.);

Pondrá vuestra merced las tres líneas que **tendrá hechas** en una de las partes [...] (Palacio 75v.);

[...] la altura del polo siempre el buen marinero la **tiene sabida** por donde navega. (Zamorano 57v.),

contribuye a la formación de unidades fraseológicas que se repiten de manera estereotipada en todos los manuales, llegando a adquirir casi el estatuto de frases hechas: así en Siria (148) se encuentra «como tengo dicho», en Palacio (7r) «Con las demás cosas que tengo dicho», en Cortés «Pues como tengamos dicho» (27r.), en Chaves «por algunas destas tres formas que tengo dicho» (19) etc. En muchos de los contextos, se nota también que las formaciones verbales con “tener” poseen un acusado valor temático, y se encuentran en frases que contienen datos conocidos; al afirmarse de dicho matiz semántico contribuye la frecuente presencia de adverbios de tiempo como “ya” y “antes”, y el hecho de que a menudo el sintagma verbal compuesto rige oraciones temporales que preceden la expresión de informaciones nuevas, y tienen la significación de condición preliminar de lo que va a pasar después:

Quando vuestra merced **tuviere hecha** la carta de marear en la forma dicha, / para guardarla, hará primeramente un repartimiento que llaman escala o tronco de leguas [...] (Palacio 73v).

Por lo que concierne el auxiliar de los verbos intransitivos,

excepto en tres ocasiones:

[...] **es amanecido** [...] (Chaves 66v.); [...] les puede dar aviso hasta **ser llegados** a tierra. (Chaves 22 v.), [...] si **es salido** a tierra [...](Guevara 27 r.),

en los manuales examinados se registra un uso distinto respecto a las obras literarias que componen el corpus de Keniston; en lugar de “ser” – que el estudioso estadounidense indica como la norma – todos los textos náuticos utilizan “haber”, como en los ejemplos siguientes:

[...] y después que **han llegado** a la Nueva España [...] (Chaves 46r.); 2v.); [...] que **avían allegado** allí. (Siria V); [...] y después de **aver salido** con virtud y fuerza de los rayos del Sol [...] (Siria 46).

Aunque los que presento son datos limitados a un caso concreto, dicha diversidad de usos sintácticos respecto al género literario permite suponer que la percepción de las diferencias, no sólo a nivel léxico, entre lenguaje técnico y lengua común, no es un fruto de la época moderna, y que la tendencia a la innovación que caracteriza el lenguaje técnico científico nace hace siglos.

Pasando al estudio del español moderno, AyDA puede ser una valiosa ayuda en detectar neologismos y préstamos: en una de las primeras fases de su ejecución, el programa elabora una lista de palabras no reconocidas, obtenida confrontando las ocurrencias del texto con las del diccionario principal. Una prueba hecha sobre un corpus de artículos económicos sacados de *ABC*, ha puesto en evidencia que las palabras no reconocidas eran préstamos, anglicismos y neologismos, o lemas que pertenecían al léxico especializado. Ha sido así posible recuperar sintagmas como *stock options*:

Repsol repartirá 1.406 millones entre 45 directivos por «stock options». (*ABC* 9/3/01),

y *holding*:

Auna, el holding de telecomunicaciones que agrupa las inversiones [...] (ABC 6/3/01).

Entre los neologismos destacan también palabras de reciente formación, como las creadas a partir del sufijo *euro-*:

[...] la difusión del código de buenas prácticas y la euroetiqueta [...] (ABC 7/3/01),

mientras al ámbito del léxico a fines específicos se adscriben palabras como “cotizar” o “accionarial”:

Esta posición accionarial sólo tendrá un objetivo financiero [...] (ABC 6/3/019).

Concluyendo, quiero expresar el deseo que mi contribución permita conocer mejor una herramienta pensada expresamente para el español, que proporciona datos de interés y ahorra a los investigadores tiempo precioso que puede ser destinado a otros estudios.

Bibliografía

- Buchanan, M.A., *A graded spanish word book*, Toronto, Ontario, 1927.
- Carpi, E., *Sobre el estilo de la biografía hagiográfica frayluisiana «De la vida, muerte, virtudes y milagros de la santa madre Teresa de Jesús»*, «Epos», XVI, 2000, pp. 105-132.
- Carpi, E., *El léxico del «Quatripartitu en Cosmographía Práctica» de Alonso de Chaves*, Estudios de la Uned, Madrid, Uned, 2001.
- Carpi E., Mancho Duque, M. J., *El léxico de «De la vida, muerte, virtudes y milagros de la santa madre Teresa de Jesús» de Fray Luis de León*, Estudios de la Uned, Madrid, Uned, 1999.
- Carpi, E., Saba, A., Sassi, M., *Concordanze del «Libro de la Vida» di Teresa de Ávila*, Pisa, DSU, 1987.
- Carpi, E., Saba, A., Sassi, M., *Exclamaciones del alma a Dios. Concordanze per lemma e repertorio delle forme enfatiche nell'opera omnia di Teresa de Ávila*, Pisa, DSU, 1992.
- Carpi, E., Saba, A., Sassi, M., *Los diminutivos orgánicos en la obra de Teresa de Ávila*, «Epos», vol. 12-1996, pp. 159-176.
- Duro, A., Zampolli, A., *Analisi lessicale mediante elaboratori elettronici*, Roma, Accademia Nazionale dei Lincei, 1968.
- Ellison, J.W. (ed.), *Nelson's Complete Concordance of the Revised Standard Version of the Bible*, New York, Thomas Nelson, 1957.
- Finzi, A., *Concordancias y frecuencias de uso en el léxico poético de Antonio Machado*, Pisa, Cattedra di Linguistica, Università di Pisa, Divisione Linguistica del CNUCE-CNR, 1973.
- García-Macho, M. L., Saba, A., *El léxico de la «Yliada de Homero en romance» traducida por Juan de Mena*, Estudios de la Uned, Madrid, Uned, 1998.
- Juilland, A., Chang-Rodriguez, E., *Frequency Dictionary of Spanish Words*, The Hague, Mouton, 1964.

- Keniston, H. A., *A basic list of spanish words and idioms*, Chicago, 1933.
- Keniston, H. A., *The syntax of castilian prose*, Chicago, 1933.
- Mancini, G., *Espressioni letterarie dell'insegnamento di Teresa de Ávila*, Modena, 1955.
- Menéndez Pidal, R., *La lengua de Cristobal Colón*, Buenos Aires, 1947.
- Rosselli, F., *Concordancias y Frecuencias de uso en el teatro de Juan del Encina*, Firenze, Università di Firenze, 1983.
- Rosselli, F., Finzi, A., *Diccionario de concordancias y frecuencias del léxico poético de Cesar Vallejo*, Istituto di Lingue Straniere, Economia e Commercio, Università di Firenze, Pisa, Divisione Linguistica del CNUCE-CNR di Pisa, s.d.
- Saba, A., *El léxico del «Breve compendio de la sphaera y de la arte de navegar» de Martín Cortés y Albácar*, Estudios de la UNED, Madrid, Uned, 2001.

