UNIVERSIDAD DEL ZULIA

Facultad Experimental de Ciencias

Laboratorio de Investigación de Tecnologías y Sistemas de Información (LITSI)

Enlace

Revista Venezolana de
Información, Tecnología y Conocimiento

año 12 Nº 3

Contenido Destacado

Año 12 / Nº 3 / Septiembre-Diciembre 2015

Dep. legal: ppi 201502ZU4693

Esta publicación científica en formato digital es continuidad de la revista impresa
ISSN: 1690-7515
Dep. legal: pp 200402ZU1624

Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento
Año 12 No. 3, Septiembre-Diciembre 2015, pp. 23-47.

Cómo citar el artículo (Normas APA):
Fernández, P. (2015). La creación de corpus lingüísticos diacrónicos: la perspectiva del transcriptor. Enl@ce Revista Venezolana de Información, Tecnología y Conocimiento, 12 (3), 23-47.

La creación de corpus lingüísticos diacrónicos: la perspectiva del transcriptor

Patricia Fernández Martín[1]

Resumen

El objetivo del artículo es reflexionar sobre un proceso de transcripción de textos del siglo XVIII que realizó la autora en 2008. Para ello, se dedica una primera parte a explicar el marco teórico, ejemplificado en la perspectiva del análisis del texto filológico, para facilitar la comprensión, en la segunda parte, de las fases en que dividimos el proceso de transcripción. En concreto, se defiende que las tres fases metodológicas esgrimidas que aparecen en las dos caras de la misma moneda (la del análisis-lector y la del transcripción-productor), hunden sus raíces en tres maneras diferentes de comprender la lingüística computacional, para lo que se diferencia entre i) la lingüística de corpus; ii) la lingüística de corpus computacional y iii) la lingüística computacional de corpus. Las principales conclusiones recogen la interrelación entre el proceso analítico y el proceso creador de los textos que contribuyen, poco a poco, a conformar un macrocorpus diacrónico digital.

Palabras clave: anotación de corpus; corpus lingüísticos; diseño de corpus electrónicos; lingüística de corpus

Recibido: 8/8/15 Devuelto para revisión: 13/10/15 Aceptado: 4/11/2015.

[1] Doctora en Lengua Española (UCM). Licenciada en Filología Hispánica (UCM), Lingüística (UAM) y Antropología Social y Cultural (UNED). Profesora de lengua española y de español para extranjeros.
Correo e-: patriciafernandezmartin@gmail.com

The Creation of Linguistic Corpora Diachronic: the Perspective of the Transcriber

Abstract

The objective of the article is to reflect on a transcription process of texts of the eighteenth century that made the author in 2008. To do this, devotes a first part to explain the theoretical framework, exemplified in the perspective of the analysis of the text philological, to facilitate the understanding, in the second part of the phases in which we divided the transcription process. In particular, it is argued that the three methodological stages that appear in the cited two sides of the same coin (the analysis-reader and the transcript-producer), had its roots in three different ways of understanding the computational linguistics, for what is difference between (i) the linguistic corpus; ii) the computational corpus linguistics and (iii) the computational linguistics of corpus. The main conclusions reflected the interrelationship between the analytical process and the process of creating the texts which contribute, little by little, to form a diachronic macrocorpus digital.

Key words: annotation of corpus; linguistic corpora; design of electronic; linguistic corpus.

1. Introducción

El reciente crecimiento del uso de las nuevas tecnologías de la información está obligando también a la filología y la lingüística a adaptarse con la nueva forma de hacer ciencia, enfocándolas naturalmente desde su propia perspectiva (Payrató 1998: 108 y ss; Luque Agulló 2004-2005; Moreno Sandoval 1998: 15, 27-29; Pöckl *et al.* 2004: 251-257; Fernández-Pampillón *et al.* 2010; Sevilla Muñoz *et al.* 2011; González Rey 2012), como el almacenamiento de datos (elaboración de bases de datos, creación de corpus) y su posterior análisis (análisis automatizados de corpus, programas de procesamiento de datos), la corrección ortográfica y ortotipográfica, así como de estilo y gramatical; la enseñanza de lenguas extranjeras (recursos multimedia y audiovisuales; aprendizaje asistido por ordenador; cuantificación de datos para priorizar los exponentes lingüísticos que enseñar); la investigación literaria (análisis cuantitativos de los textos; análisis métricos, sintácticos y semánticos sobre textos anotados manualmente; asignación de autoría y localización cronológica); la lexicografía (creación de bases de datos, ordenación alfabética automática, edición e impresión, lematización automática) o la simulación del procesamiento del lenguaje humano (generación de lenguas naturales, traducciones automáticas, sistemas de reconocimiento de habla).

Dentro de este contexto de relación entre la filología y las TIC, apoyado metodológicamente por las herramientas que la lingüística del corpus pone a su alcance (Parodi 2008; López Alonso 2014: 287 ss), se encuentra la creación, etiquetación y anotación de corpus diacrónicos, dado que el objetivo esencial de la presente investigación es relatar (y, a partir de ello, reflexionar) un proceso de transcripción de textos dieciochescos llevado a cabo por la autora[1], que tuvo lugar hace unos años gracias a una prórroga que

[1] Aunque en el texto se utilice el plural de modestia, típico del discurso humanístico, la autoría de la realización del análisis lingüístico y de las transcripciones relatadas pertenece a la autora del presente artículo que, por tanto, asume toda la responsabilidad sobre el mismo.

obtuvo el proyecto *Procesos de Gramaticalización del Español (II): formación de variedades (tipología, periodización, criollización)* del MCyT (REF. HUM2004-03610), dirigido por José Luis Girón Alconchel (Universidad Complutense de Madrid). Con el trabajo en referencia se prevé seguir ampliando el corpus lingüístico iniciado unos años atrás bajo versiones anteriores del mismo proyecto, para contribuir así a formar parte de la construcción de corpus lingüísticos digitales que venían creándose de forma más o menos simultánea (Fernández Moreno 2005: 98 ss; Enrique-Arias 2009b; Parodi 2010: 155-164).

Entre los corpus digitales dedicados a la lengua española[2] pueden encontrarse, con objetivos sincrónicos (Parodi 2008: 112 ss; López Alonso 2014: 291-293), el CREA y el CORPES de la Real Academia Española; el proyecto Val.Es.Co de la Universidad de Valencia, nacido para estudiar el español coloquial; el CEDEL2 de la Universidad Autónoma de Madrid y de la Universidad de Granada, cuyo objetivo es hacer un corpus escrito del español como segunda lengua; Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC), también de la Universidad Autónoma; los interuniversitarios C-ORAL-ROM, por un lado, en el que participan la mencionada Universidad Autónoma de Madrid y la Universitá di Firenze; y PRESEEA (Proyecto para el Estudio Sociolingüístico para el Español de España y de América), por otro, en el que toman parte los equipos de Alcalá de Henares, México, Madrid, San Juan de Puerto Rico, Las Palmas de Gran Canaria, Lérida, Cádiz, Zaragoza, Barranquilla, Cipolletti, Málaga, Culiacán, Guatemala, Miami, Bogotá, Pereira, Palma de Mallorca, Caracas, Granada; el Grial de la Universidad de Valparaíso dirigido por Giovanni Parodi (2006, 2010; López Alonso 2014: 290-291) o, finalmente, el llevado a cabo en la Brigham Young University por Davies (2009).

Con perspectivas diacrónicas están hechos, entre otros (Contreras Seitz 2009), el CORDE de la Real Academia Española; la red CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos), coordinada por Pedro Sánchez-Prieto Borja e integrada por grupos formados por la Universidad de Alcalá, Centro Superior de Investigaciones Científicas (España)", Universidad de Deusto, Universidad de Murcia, Universidad de Valladolid, Universidad de las Palmas de Gran Canaria, Universidad Complutense de Madrid, Universidad de Jaén, Universidad de Salamanca, Universidad de las Islas Baleares, Universidad de Padua, Universidad de Nêuchatel, Universidad de Gotemburgo, King's College, Munich y Los Andes (Venezuela), tal y como se indica en su página web; el corpus diacrónico del español del Reino de Granada (CORDEREGRA; cfr. Calderón Campos y García Godoy 2009), y el ya citado de Mark Davies en la Brigham Young University (Davies 2009), lo que deja entrever la enorme riqueza de perspectivas, que ofrece la creación de los corpus de que se dispone.

Una vez vistos algunos de los corpus lingüísticos ya creados,, la reflexión que se pretende efectuar a partir de la experiencia transcriptora parte de la forma en que se opera al analizar los textos con motivos filológicos, por lo que se dedica una primera parte a explicar el marco teórico, con el fin de comprender las fases que se desempeñan al hacerlo (cfr. § 2) para, en la segunda parte, exponer el proceso de

[2] Dejamos, pues, de lado otras lenguas en las que también se está trabajando (Torruella y Casañas 2009; Enrique-Arias 2009b; Palermo 2011; López Alonso 2014: 292).

transcripción de acuerdo a esas mismas fases (cfr. § 3), las cuales, como han de verse, hunden sus raíces en tres maneras diferentes de comprender la lingüística computacional (Moreno Sandoval 1998; Lee 2004; Kay 2006; Parodi 2010: 18-21).

2. Lingüística computacional: perspectiva de análisis

Puede decirse que la lingüística computacional (o procesamiento del lenguaje natural [PLN]) surge a mediados del siglo XX, en un contexto de interés teórico por ampliar los conocimientos sobre el lenguaje humano desde una perspectiva informática, cuyas aplicaciones (ingeniería lingüística, tecnologías lingüísticas, lingüística informática) llegarían sobre todo durante los años 90, con la propagación, primero, de los ordenadores personales y de los sistemas operativos gráficos (Windows) y, segundo, de Internet (Gazdar 1996; Moreno 1998: 24-29; Lee 2004).

La principal premisa metodológica de la lingüística computacional, se encuentra en la concepción matemática del lenguaje, de tal manera que si se consigue profundizar en el conocimiento de la primera se podrá llegar a comprender con mayor rigor el segundo, por lo que cabe entender, en este sentido, que la lingüística computacional no sea más que la herramienta experimental de la lingüística teórica (Moreno Sandoval 1998: 25). Las discrepancias llegarán entre los lingüistas computacionales a la hora de decidir cómo analizar el procesamiento del lenguaje. Así, por un lado, los modelos simbólicos, que hunden sus raíces en la lógica, entienden que los procesos mentales se basan principalmente en la manipulación de símbolos. Concretamente, parten de la base de que existe un lexicón cuyos componentes han de ser conectados mediante una serie de reglas para tener sentido, de donde se deduce que conocer las reglas que rigen la unión de los elementos que formen el léxico permite descifrar el significado de cualquier oración (Moreno Sandoval 1998: 47-49; Lee 2004: 114).

Asimismo, en los modelos simbólicos se trabaja en el grado de la competencia, por lo cual tienden a situarse en un lugar más abstracto de lo que permite esta, y consiguen con ello cierta "perfección" de los datos que se analizan (Moreno Sandoval 1998: 51). Esta perfección les permite hacer predicciones, lo cual, a su vez, facilita que las conclusiones válidas basadas en premisas verdaderas puedan generalizarse, garantizando así la posibilidad de aspiración a universal de estos modelos (Kay 2006).

Por otro lado, los modelos estadísticos o probabilísticos que, si bien surgen a mediados del siglo XX aunque injustamente ignorados (e incluso desprestigiados) durante décadas, no irrumpen con fuerza hasta los años 80, debido fundamentalmente a la importancia que el léxico adquirió durante este período, en el que se llegó casi a una obsesión por explicar todo como palabras más que como reglas (Lee 2004; Moreno Sandoval 1998: 153 ss; Civit Torruella 2003).

Esta necesidad de acudir a las construcciones léxicas y a conceptos como colocaciones, redes semánticas, herencia morfológica se debieron a las continuas irregularidades que se encontraban en la lengua y a otros inconvenientes hallados tanto en los modelos simbólicos, como la ambigüedad de los análisis sintácticos y semánticos; la falta de cobertura real de las reglas establecidas; la

cantidad de excepciones que aparecen al aplicar las reglas; la extensa combinatoria de elementos y, por tanto, posibles análisis de los que los hablantes no son conscientes en absoluto; y la inadecuación con la realidad simbolizada, es decir, la falta de coherencia existente entre los datos reales y el mismo sistema simbólico (Kay 2006; Moreno Sandoval 1998: 143-149, 164-166; Parodi 2008: 108 ss).

Asimismo, los modelos estadísticos se empleaban ya entonces en algunos campos de la lingüística aplicada, como la sociolingüística, la pragmática o la enseñanza de lenguas en los que siempre resultaba más útil utilizar un corpus basado en enunciados emitidos que emplear oraciones "de laboratorio" dependientes íntegramente de la competencia lingüística del investigador (o de sus a veces cuantitativamente escasos informantes) (Moreno Sandoval 1998: 164; Stubbs 2004; López Alonso 2014: 289).

De este modo, parece que cobra sentido incluir la lingüística de corpus dentro de la lingüística aplicada, tomada esta desde una perspectiva general (Stubbs 2004; Payrató 1998: 112; Luque Agulló 2004-2005: 168), ya que aquella constituiría un método, en lo que aquí ocupa, para el estudio de la historia de la lengua, tal vez el único campo de investigación lingüística que, intentando ser o aspirar a ser lo más teórico posible, siempre precisará de un corpus textual para poder existir (Kabatek 2004; Enrique-Arias 2009a; Fernández Martín 2012a; Clavería Nadal 2012).

Entonces, al considerar la extrema y absoluta importancia del empleo de corpus en cualquier investigación lingüística, especialmente la de corte diacrónico, cabe entender que prácticamente no haya a principios de los 90 una concepción de la lingüística radicalmente ajena al corpus (entendido, siguiendo a David Crystal, como «una colección de datos lingüísticos, ya sea de textos escritos o de transcripciones de habla grabada, que pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificación de hipótesis acerca de una lengua» Crystal, 1991: 32 *apud* Parodi, 2010: 21), independientemente de que se haya recopilado a mano o con medios informáticos (López Alonso 2014: 287-288). Por tanto, toda lingüística parece ser, actualmente, una lingüística de corpus (Caravedo 1999; Civit Torruella 2003; Stubbs 2004; Enrique-Arias 2009a: 13-15; Parodi 2010: 21; Fernández Martín 2012a, 2014), incluso aunque ese corpus se base en la competencia lingüística del investigador, sobre la que reflexione mediante la introspección (Moreno Sandoval 1998: 26; Pöckl *et al.* 2004: 251-256).

Esto implica que el interesado en los estudios diacrónicos ha de manejar primero los conocimientos (meta) lingüísticos necesarios para poder enfrentarse, en una segunda fase, al estándar computacional de trabajo. Este segundo estaría compuesto por la extracción de los datos a partir de un corpus digital y siguiendo ciertos principios metodológicos, lo que compondría la llamada lingüística de corpus computacional (Parodi 2008: 99, 2010: 18; Fernández Martín 2012a), que no deja de ser, en el fondo, una filología con carácter innovador por las nuevas tecnologías, pero que entiende que estas son una herramienta al servicio de su disciplina (Caravedo 1999; Civit Torruella 2003; Stubbs 2004; Enrique-Arias 2009a: 13-15; López Alonso 2014: 288).

En una tercera fase, el filólogo ha de aspirar a conocer, holísticamente, los sistemas lingüísticos

de estadios anteriores de la lengua, con el *ideal* para construir gramáticas que puedan utilizarse computacionalmente. Así, dentro de la lingüística aplicada, podría destacarse la creación y análisis de corpus; enfoques psicolingüísticos de análisis del habla; traducción automática; tratamiento de la información; corrección de textos ortográfica, ortotipográfica y de estilo; enseñanza de lenguas; y, quizá algo más alejados de la lingüística, se encontrarían otros usos posibles como la construcción de interfaces hombre-máquina o el reconocimiento de voz. De esta manera, se encuentra matizando, en esta especialidad, el campo de estudio de la considerada lingüística computacional de corpus (Parodi 2008: 99, 2010: 18), que no distaría demasiado, en realidad, de la lingüística computacional basada en modelos estadísticos y que se opondría, por tanto, a la lingüística computacional formal o simbólica (Gazdar 1996; Moreno Sandoval 1998; Payrató 1998: § 7.4; Luque Agulló 2004-2005; Stubbs 2004; Parodi 2008, 2010; Clark, Fox y Lappin, 2010).

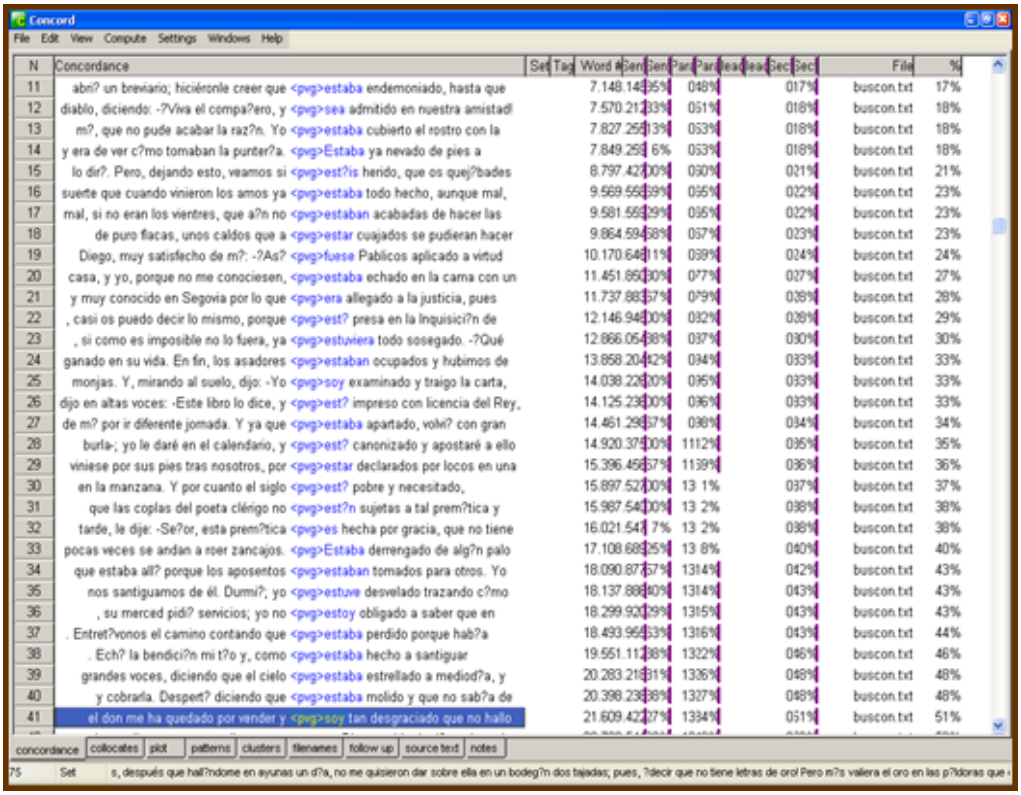
A continuación, se utiliza como marco teórico estas tres vertientes de la lingüística computacional (la lingüística [de corpus], la lingüística de corpus computacional y la lingüística computacional de corpus), considerándolas tres fases de un mismo método de investigación: el nuevo método filológico (Tagliavini 1973: 55-69; Garatea Grau 2005: 51-67). En concreto, se expone cómo se ha localizado y analizado ciertas perífrasis verbales en el *Buscón* de Francisco de Quevedo para luego realizar la misma operación desde la perspectiva del proceso transcriptor (cfr. § 3). Sea como fuere, no conviene olvidar que lo que aquí interesa, no es el resultado lingüístico (Fernández Martín 2014), sino el modo de proceder desde lo informático.

2.1. Fase I. Delimitación del objeto de estudio (lingüística [de corpus])

Las perífrasis verbales son unas construcciones lo suficientemente complejas (Olbertz 1998; Fernández de Castro 1999; Gómez Torrego 1999; Yllera 1999) como para necesitar una delimitación *previa* a su estudio. La paradoja de esta necesidad es que, entonces, este no se puede basar en un análisis de corpus digital ya hecho, por la sencilla razón, de que no se sabe qué se está buscando (Garachana y Artigas 2012). En otras palabras, la imposibilidad de localizar perífrasis verbales desde el significado y no desde la forma, debido a la ignorancia que *a priori* se va a tener sobre el concepto mismo de perífrasis (Fernández Martín 2014), parece obligar al investigador a crear un doble conjunto de corpus lingüísticos: el de las fuentes primarias, esto es, en nuestro ejemplo, el *Buscón*; y el de los estudios metalingüísticos formado, para lo que aquí interesa, por obras que versan sobre perífrasis verbales, como pudieran ser las ya mencionadas de Fernández de Castro (1999), Gómez Torrego (1999), Yllera (1999) u Olbertz (1998), entre otras.

Una vez seleccionados los textos de ambos grupos de corpus y dejando de lado ahora el corpus metalingüístico, de nuevo se plantea la creación de un corpus exclusivamente formado por aquellos ejemplos de interés, es decir, los que constan de perífrasis verbales. Para ello, se utiliza la versión en texto plano de la Biblioteca Virtual Miguel de Cervantes (es decir, la edición digital a partir del manuscrito de la obra *Manuscrito Bueno*, depositado en la Biblioteca de la Fundación Lázaro Galdiano [Madrid]) y se introdujo en el programa WordSmith (figura 1), de manera que permita obtener la mayor cantidad posible de ejemplos con posibles perífrasis verbales.

Figura 1
Estar + gerundio en el Buscón según el programa WordSmith



Fuente: elaboración propia, (2015)

2.2. Fase II. Interpretación de los ejemplos (lingüística de corpus computacional)

La segunda fase del proceso, consistiría en interpretar los ejemplos obtenidos en la fase previa, para lo cual, primero, convino clasificarlos y, después, analizarlos.

Para ello, se utilizó un libro de Excel en el que se incluían las distintas categorías semánticas que servían de partida, basadas esencialmente en la clasificación de Fernández de Castro (1999). Pese a haber realizado concienzudamente la fase de estudio metalingüístico, se dejó una pestaña llamada «Dudas» en el archivo, para registrar todos

aquellos ejemplos que iban a ser perifrásticos (o no) dependiendo de la definición que de perífrasis se hiciera finalmente (figura 2).

Figura 2.
Clasificación previa de las perífrasis en un documento de Excel

Fuente: elaboración propia, (2015)

En síntesis, pues, se entiende que para llevar a cabo el análisis minucioso de los ejemplos, el filólogo ha de recurrir nuevamente a la teoría previa (sus conceptos metalingüísticos, sus objetivos iniciales, sus premisas metodológicas), que le permita comprender la clasificación obtenida para ir exponiéndola en su primer borrador escrito. En este caso, el análisis se hizo a mano, ejemplo a ejemplo, como parece ser habitual (Stubbs 2004; Garachana y Artigas 2012), a no ser, naturalmente, que el corpus utilizado por el programa de análisis haya sido previamente anotado siguiendo los mismos fines que los del investigador (López Alonso 2014: 294; cfr. § 3.3).

2.3. Fase III. Aspiración universal (lingüística computacional de corpus)

En una última fase, se toma como premisa una actitud holística y extremadamente autoexigente, el estudioso de la historia de la lengua ha de atreverse a comparar las distintas ediciones que de un mismo texto puedan haber llegado, lo que resulta perfectamente factible, si se cuenta con versiones digitalizadas de las diversas publicaciones que haya tenido el texto.

Los avatares ecdóticos del *Buscón* son demasiado complejos para explicarlos aquí detalladamente (Ynduráin 1992; Rodríguez Mansilla 2004-2005), pero basta con indicar que se efectuó un análisis comparativo en dos niveles. Por un lado, se leyó la edición en papel de Ynduráin para localizar algún ejemplo concreto que pudiera resultar sugerente. Por otro lado, se utilizó otros motores de búsqueda como el CORDE (figura 3) y el Corpus del Español de Mark Davies (figura 4), en aquellos casos que se llamó la atención, para constatar que no hubo ninguna alteración de las búsquedas realizadas con el programa WordSmith. Se cree que los motores de búsqueda de ambos corpus virtuales, pese a sus respectivos inconvenientes (Davies 2009; Garachana y Artigas 2012; Fernández Martín 2012b), son complementarios porque el CORDE permite la búsqueda de ejemplos poco especificados en obras determinadas, mientras que en el Corpus del Español de Davies, se pueden

localizar colocaciones concretas en un siglo entero, sin indicar el texto específico que se quiere estudiar. En cualquiera de los casos, si no están previamente anotados o, estándolo, si el concepto metalingüístico de perífrasis del anotador no coincide con el del investigador (Schulte 2009; Garachana y Artigas 2012), hay que determinar manualmente su naturaleza perifrástica e, incluso, acabar eligiendo el empleo de la edición que se considera más fiel al texto original, y se descartan las demás: las diferencias entre ellas, plasmables en la aparición o ausencia del fenómeno lingüístico que interesa, pueden ser abismales (Fernández Martín 2014).

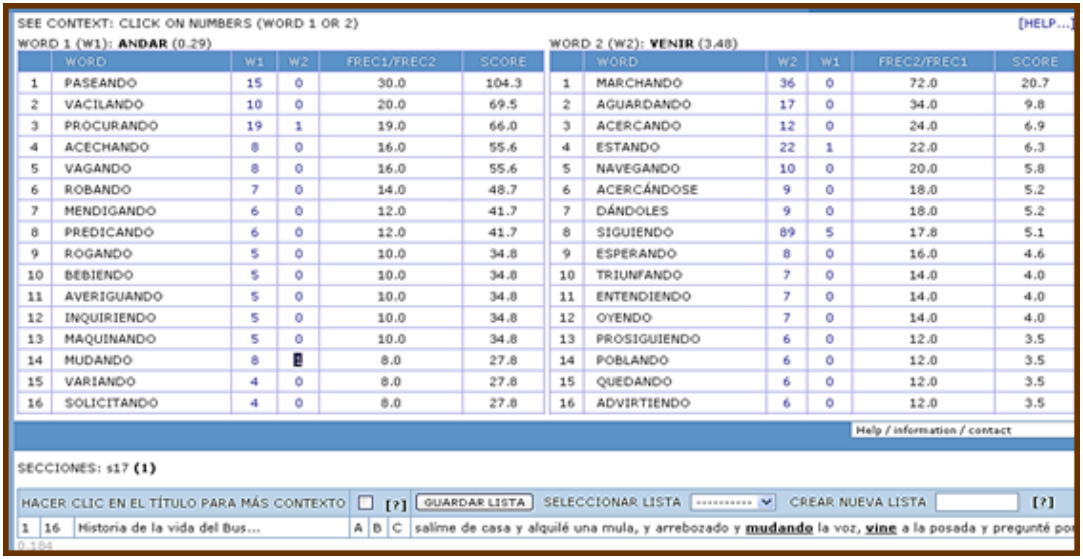
Por esta experiencia previa, pues, basada en la localización y análisis de las perífrasis verbales en el *Buscón*, se procedió a mantener una posición eminentemente holística cuando se propuso transcribir los textos del siglo XVIII de los que hablamos en la próxima sección.

Figura 3

Ejemplo de *debía* + infinitivo en el Buscón (CORDE)

Fuente: elaboración propia, (2015).

Figura 4.
Ocurrencias de {*andar/venir*} + gerundio en el siglo XVII
(www.corpusdelespanol.org)



SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2) [HELP...]

WORD 1 (W1): **ANDAR** (0.29)

	WORD	W1	W2	FREC1/FREC2	SCORE
1	PASEANDO	15	0	30.0	104.3
2	VACILANDO	10	0	20.0	69.5
3	PROCURANDO	19	1	19.0	66.0
4	ACECHANDO	8	0	16.0	55.6
5	VAGANDO	8	0	16.0	55.6
6	ROBANDO	7	0	14.0	48.7
7	MENDIGANDO	6	0	12.0	41.7
8	PREDICANDO	6	0	12.0	41.7
9	ROGANDO	5	0	10.0	34.8
10	BEBIENDO	5	0	10.0	34.8
11	AVERIGUANDO	5	0	10.0	34.8
12	INQUIRIENDO	5	0	10.0	34.8
13	MAQUINANDO	5	0	10.0	34.8
14	MUDANDO	8	8	8.0	27.8
15	VARIANDO	4	0	8.0	27.8
16	SOLICITANDO	4	0	8.0	27.8

WORD 2 (W2): **VENIR** (3.48)

	WORD	W2	W1	FREC2/FREC1	SCORE
1	MARCHANDO	36	0	72.0	20.7
2	AGUARDANDO	17	0	34.0	9.8
3	ACERCANDO	12	0	24.0	6.9
4	ESTANDO	22	1	22.0	6.3
5	NAVEGANDO	10	0	20.0	5.8
6	ACERCÁNDOSE	9	0	18.0	5.2
7	DÁNDOLES	9	0	18.0	5.2
8	SIGUIENDO	89	5	17.8	5.1
9	ESPERANDO	8	0	16.0	4.6
10	TRIUNFANDO	7	0	14.0	4.0
11	ENTENDIENDO	7	0	14.0	4.0
12	OYENDO	7	0	14.0	4.0
13	PROSIGUIENDO	6	0	12.0	3.5
14	POBLANDO	6	0	12.0	3.5
15	QUEDANDO	6	0	12.0	3.5
16	ADVIRTIENDO	6	0	12.0	3.5

Help / information / contact

SECCIONES: s17 (1)

HACER CLIC EN EL TÍTULO PARA MÁS CONTEXTO [?] GUARDAR LISTA SELECCIONAR LISTA ---------- CREAR NUEVA LISTA [?]

1 16 Historia de la vida del Bus... A B C salíme de casa y alquilé una mula, y arrebozado y **mudando** la voz, **vine** a la posada y pregunté por

Fuente: elaboración propia, (2015).

3. Lingüística computacional: perspectiva de transcripción

Como se ha dicho (cfr. § 1), la tarea que se expone a continuación seguía una línea de continuidad con respecto a trabajos anteriores realizados desde el año 2003 dentro del mismo proyecto y sus precedentes. La relativa prontitud en el empleo de las TIC aplicadas a la filología causó la necesidad de crear unos criterios de transcripción cuando aún eran escasos en España los ejemplos publicados de corpus diacrónicos del español. Este fue el motivo por que los mismos miembros del proyecto elaboraron una serie de criterios que permitieron al texto electrónico mantener la mayor fidelidad posible al original, como parecer ser deseable (Contreras Seitz 2009: § 2.2.2), tal y como se indica en la página web del PROGRAMES, dentro del apartado «Documentos»:

> Aparte de que ningún texto [de los seleccionados] ha sido incluido en ningún corpus lingüístico del español en red, los testimonios textuales elegidos dentro del proyecto carecen normalmente de edición moderna, o esta no sigue criterios de transcripción válidos para un estudio diacrónico, por haberse realizado en tal edición modernización gráfica, de puntuación e incluso de lengua. En cambio, nuestros criterios de transcripción *han sido de fidelidad absoluta al texto original* (la cursiva es nuestra).

Esta fidelidad al texto original, es la que provocó en la transcriptora la obsesión por dejar huella de la extensa heterogeneidad gráfica que mostraba especialmente uno de los textos (cfr. *infra*), lo que a su vez produjo la creación de nuevos criterios de transcripción y la consecuente reflexión sobre todo el proceso, con relación de su experiencia previa como analista (cfr. §§ 2, 3.3).

De la misma manera, el hecho de que se propusiera una transcripción paleográfica buscaba llegar al mayor número de investigadores posible, por ofrecer una serie de textos mínimamente editados que no partieran de objetivos de estudios previos, si bien, naturalmente, el proyecto se encontraba interesado en ciertos fenómenos lingüísticos cuya anotación se consideraba fundamental:

> Dados los objetivos de investigación del proyecto en cuanto a lingüística diacrónica (gramaticalización, morfología histórica, sintaxis histórica, semántica histórica), el equipo decidió formular una serie de criterios de transcripción de los textos acorde con tales objetivos: *se efectuará básicamente una transcripción paleográfica* (grafías, acentuación y puntuación); en cuanto a las abreviaturas, se conservarán las referidas a formas y fórmulas de tratamiento, mientras que se desarrollarán entre corchetes el resto. Se acordó también, no separar palabras por renglones, ni siquiera al final de página, ya que el programa WordSmith, no reconoce las palabras segmentadas y se trata de una información lingüística poco o nada relevante para el estudio (*la cursiva es nuestra*).

Asimismo, este interés por divulgar se plasma en la creación de un corpus cuyos textos han quedado preparados de dos formas diferentes, en un afán por construir un material útil al público, filológicamente riguroso y técnicamente procesable (Spence 2014): la versión etiquetada y anotada, válida para analizar con un programa de análisis de corpus (preferentemente se opta por WordSmith y, por ello, se ofrece en .doc, porque el servidor no permite ponerlo en .txt) y la versión para leer, etiquetada pero no anotada (Pagola *et alii* 2006: 1434-1440) y presentada en pdf. De este modo, el investigador tiene la opción de utilizar la lingüística de corpus siguiendo los cánones clásicos (Tagliavini 1973: 55-69; Garatea Grau 2005: 51-67) o los modernos (Schulte 2009; cfr. § 2), lo que permite dar una idea del potencial de facto de esta nueva herramienta metodológica (Caravedo 1999; Civit Torruella 2003; Stubbs 2004; Enrique-Arias 2009a: 13-15; López Alonso 2014: 288).

En cuanto a los textos con los cuales se trabajaron, el primero de los manuscritos transcrito, etiquetado y anotado, de género epistolar, fue el titulado *Papeles curiosos en prosa y verso de los años de 1710 y 1711*, correspondiente al MSS10907 de la Biblioteca Nacional de España. Constaba de

192 folios, de los que tan sólo fueron transcritos aquellos escritos en prosa (unos 50), aunque se mencionaba claramente en cada caso dónde se encontraban los poemas y cuánto ocupaban, dado que los textos poéticos no cumplen los requisitos lingüísticos adecuados a los objetivos propuestos en el proyecto. Desde la perspectiva de la transcripción, la letra humanística permitía que no fuera excesivamente complejo, salvo por el descifre de ciertas abreviaturas (Marín Martínez 1991).

El segundo texto, de género cronístico, del que se transcribieron los 243 primeros folios, era un manuscrito titulado: *Narraciones históricas desde el año 1700 hasta el año 1725*, escrito entre 1733 y 1742[3] (Castellví, 1997: 49-70; 1998, 1999), tal y como se indica en el folr7:

1. Por <ft> Don </ft> Francisco Castellvi, Natural del Principado de <sic> Cathaluna [*h* tachada] </sic>, nacido en la Real Villa de Montblanch Capital del Ducado de este nombre; con cuyo Título los Primogenitos de los Reÿes de Aragon condecoraban sus Personas, Colegial del Colegio de Cavalleros de la Purissima Concepcion de la Antiquissima Universidad de Lerida, Capitán que fuè en el Regimiento delos Ciudadanos de Barcelona llamado la Coronela en los años 1713 ÿ 1714.

La transcripción de este segundo texto, tampoco fue excesivamente complicada: la letra era igualmente humanística y eso facilitó enormemente la tarea. El trabajo de esta transcripción, fue la necesidad de ampliar el sistema de etiquetado que hasta ese momento se había empleado en el mismo proyecto, puesto que la riqueza de matices que aparecía en el texto era tan grande que se consideró conveniente dejar constancia de todas ellas, para posibles futuras investigaciones. Así, por ejemplo, al poco tiempo de empezar a leer el documento, el lector se encontraba con un párrafo lingüísticamente tan interesante como el siguiente (folr7):

2. En los Reÿnos de <superp> Castilla </ superp>, Aragon, Valencia, <tach> Serdeña </ tach> <superp> Cerdeña <superp>, Mallorca, Sicilia, Principado de Cathaluña [*h* tachada], ÿ sus Confines; que han sido el centro de la sangrienta Guerra que empezó en este Siglo. <tach> Con un Resumen en el fin de cada año, de los Sucesos que acaecieron, en todas partes en donde se guerrearia sobre la Sucesion de la España </tach>.

En efecto, el documento de Castellví se mostró desde la perspectiva caligráfica, ortográfica e incluso gramatical lo suficientemente complejo como para que se sintiera la necesidad de plasmar de algún modo tal amplitud de interpretaciones. Las dudas que surgieron entonces fueron: ¿cómo transcribir un texto tan heterogéneo como el que se tenía entre manos, sin dejar de ser fiel a las directrices marcadas desde el proyecto? ¿Cómo hacerlo sin entrar en la elaboración de una edición filológica? Y, quizá lo más preocupante, ¿cómo dejar huella en las transcripciones realizadas de toda la variedad lingüística que podía, tal vez, ser útil a otros estudiosos en el futuro?

Como se muestra a continuación, la solución apareció en la combinación de las perspectivas establecidas por la lingüística computacional (cfr. § 2), es decir, en escribir los textos como a la transcriptora le gustaría encontrárselos (cfr. *supra*, § 3), si precisara de su procesamiento informático para localizar algún fenómeno lingüístico concreto. No cesaba de actuar, de este

modo, como el lector ideal que todo escritor tiene en mente al producir un texto y que, en ocasiones, no es otro más que él mismo (Eco, 1993: 92 ss).

3.1. Fase I. Concienciación (meta) lingüística (lingüística [de corpus])

Una vez conseguidos los documentos en papel y utilizado el programa de análisis de datos Wordsmith para comprobar el correcto funcionamiento de las etiquetas y anotaciones, una conexión a la Internet para estar en constante comunicación con los miembros del proyecto durante el trabajo (Fernández Martín 2009), se procedió a aplicar todos los conocimientos lingüísticos para poder ir transcribiendo los documentos.

Años después, tras reflexionar sobre diversas aspectos metodológicas concernientes con el proceso filológico (Tagliavini 1973: 55-69; Garatea Grau 2005: 51-67), se entendió que esta era la primera fase del trabajo: la transcripción, no sólo exige conocimientos informáticos, los cuales son fundamentales para poder llevar a cabo el proceso de escritura con un cierto grado de habilidad, rapidez y eficacia, sino que también obliga a tener una alta competencia lingüística y metalingüística, imprescindible para poder comprender que el proceso comunicativo existente entre el escritor del documento y el lector-transcriptor (Nystrand 1986; Eco 1993) no se ciñe a la mera ortografía o a la relación letra leída-letra escrita, sino que en numerosas ocasiones hay que emplear a fondo las competencias más pragmáticas (discursiva, textual) para poder comprender lo que se está leyendo y, en consecuencia, lo que se va a transcribir.

Por tanto, se observa que en el proceso de análisis lingüístico tiene lugar a la hora de investigar el objeto de estudio para crear corpus lingüísticos acordes con él (cfr. § 2.1), en el de la transcripción ocurre en una introspección que ha de servirnos necesariamente para conocer cuáles son las limitaciones de nuestras competencias (meta)lingüísticas y, una vez descubiertas, pasar a la investigación (abreviaturas, tipografías, conocimientos históricos o biográficos del autor…) para rellenar esas lagunas. Es en este sentido, en que se defiende que la primera fase de cualquier proceso de transcripción (que es el principio de la creación de un corpus electrónico) pertenece a la lingüística de corpus, es decir, pertenece a la lingüística (Moreno Sandoval 1998: 26; Caravedo 1999; Civit Torruella 2003; Stubbs 2004; Enrique-Arias 2009a: 13-15; Parodi 2010: 21; cfr. § 2.1), dado que hay necesariamente una labor previa de documentación lingüística e introspección metalingüística, no necesariamente relacionada con el corpus que se pretende transcribir.

3.2. Fase II. Escritura del texto (lingüística de corpus computacional)

En esta segunda fase, simultánea en el tiempo a la anterior, pero separable de ella desde una perspectiva metodológica, se fue transcribiendo el texto, etiquetándolo estructuralmente y anotándolo con las categorías morfosintácticas oportunas, según las instrucciones recibidas y los criterios expuestos (cfr. *supra*).

Como se explica en la siguiente fase (cfr. § 3.3), la complejidad en el proceso de transcripción del segundo texto, el de Castellví, no se debía a las dificultades caligráficas (la letra humanística en la que se encontraba el documento no daba ningún problema de descifre), sino a las constantes enmiendas, tachaduras, correcciones o notas al margen que parecían avisar sobre el hecho de que fueron varios los momentos de redacción del texto

y, probablemente, fueron muchos los amanuenses que tomaron parte en el proceso (Castellví 1997: 41-44). El interés, por tanto, por las diferentes consciencias (meta) lingüísticas de la época hizo que el documento fuera tomado como una huella del paso del tiempo y de las implicaciones sociológicas que puede ir dejando la lengua en los escritores.

La anotación (Pagola *et alii* 2006: 1434-1440), por su parte, permitió señalar entre comillas latinas simples (<>) las categorías morfológicas o las funciones sintácticas relevantes para el proyecto de investigación en que dicho trabajo se desarrollaba, tal y como se explica en la web de PROGRAMES:

Etiquetas de fenómenos lingüísticos

<cdp>, </cdp>	Complemento directo *preposicional*
<conda>, </conda>	Condicional analítico
<dcl>, </dcl>	Duplicación clítica
<ft>, </ft>	Forma/fórmula de tra-*tamiento*
<futa>, </futa>	Futuro analítico
<lasm>, </lasm>	Laísmo
<lesm>, </lesm>	Leísmo
<losm>, </losm>	Loísmo
<marc>, </marc>	Marcador del discurso

Etiquetas de fenómenos lingüísticos solo empleadas en los textos transcritos 2003-2005

<cd> </cd>	Complemento directo
<ci> </ci>	Complemento indirecto
<ppas>, </ppas>	Participio de pasado
<ra> </ra>	Forma verbal en -ra
<reg> </reg>	Régimen preposicional *de los verbos*
<rel> </rel>	Relativo
<relart> </relart>	Artículo + relativo
<relcomp> </relcomp>	Relativo compuesto

También fue necesario, naturalmente, dejar constancia de los aspectos formales del texto, imprescindibles para que fuesen correctamente codificados por los programas de procesamiento y análisis textual que el interesado pretendiera utilizar (Pagola *et alii* 2006: 1434-1440):

Aspectos materiales del texto

<body>, </body>	Cuerpo del texto
<col#>, </col#>	Columna
<colf>, </colf>	Colofón
<curs>, </curs>	Cursiva
<folr#>, </folr#>	Folio (recto)
<folr[I]>, </folv[I]>	Folio I no numerado de preliminares (recto)
<folr[II]>, </folv[II]>	Folio II no numerado de preliminares (recto)
<folr[III]>, </folv[III]>	Folio III no numerado de preliminares (recto)
<folr[IV]>, </folv[IV]>	Folio IV no numerado de preliminares (recto)
<folv#>, </folv#>	Folio (vuelto)
<folv[I]>, </folv[I]>	Folio I no numerado de preliminares (vuelto)
<folv[II]>, </folv[II]>	Folio II no numerado de preliminares (vuelto)
<folv[III]>, </folv[III]>	Folio III no numerado de preliminares (vuelto)
<folv[IV]>, </folv[IV]>	Folio IV no numerado de preliminares (vuelto)
<foreign>, </foreign>	Fragmento o palabra en una lengua extranjera
<lim> </lim>	Liminares (índices, etc.)
<marg>, </marg>	Nota al margen
<ms>, </ms>	Notas manuscritas en *un impreso*
,	Párrafo
<pag#>, </pag#>	Página
<pie>, </pie>	Nota al pie de página
<port>, </port>	Portada
<prelim>, </prelim>	Preliminares
<sic>, </sic>	Errores

Por último, la interacción real que se dio entre esta segunda fase y la primera (cfr. § 3.1) facilitó, por un lado, el aprendizaje de nuevas perspectivas de análisis lingüístico y, por otro lado, el desarrollo de una mayor competencia (meta) lingüística basado en una estadio de lengua previo a la lengua hablada por la transcriptora.

Por ejemplo, uno de los aspectos que llamó sumamente la atención, estaba relacionada con el significado real de 'error' (Torijano Pérez 2004; Hoyos y Marrero 2008), especialmente desde un punto de vista diacrónico. ¿Algo que el mismo autor escribe de diversas maneras se puede considerar error? ¿Se considera error con respecto a la ortografía actual o con respecto a la que el mismo autor emplea? En este segundo caso, ¿cuál de las variantes utilizadas -por ejemplo, en los topónimos o antropónimos- es la errónea? ¿Se considera error un extranjerismo «mal» escrito? ¿Y si son criterios ortográficos conscientes -como poner casi siempre *francia* con minúsculas-? La decisión que se tomó, finalmente, para designar los <sic> fue partir del sistema ortográfico actual, expresando así, todos aquellos elementos distintos a los esperables según los cánones del siglo XXI (véase supra, ejemplo 1).

En síntesis, de manera semejante a como se veían en la segunda fase del análisis (cfr. § 2.2), que la linguísitica de corpus computacional permitía la extracción de los datos a partir de un corpus digital y su estudio según ciertos principios metodológicos (Parodi 2010: 18; Fernández Martín 2012a; López Alonso 2014: 288), desde la perspectiva de la transcripción se observa que paulatinamente se va creando un corpus electrónico, siguiendo ciertos principios metodológicos establecidos por el equipo de PROGRAMES (los criterios de etiquetación y anotación).

3.3. Fase III. Actitud holística interdisciplinar (lingüística computacional de corpus)

La fase más compleja, pero tal vez, la más interesante del proceso de transcripción que se define tuvo lugar cuando se tuvo que enfrentarnos a la pregunta, ya esbozada anteriormente, de cómo plasmar en las transcripciones realizadas toda la riqueza lingüística que podía, quizás ser útil a otros estudiosos en el futuro.

Dada la actitud holística, que asume, todo filólogo ha de mantener como premisa metodológica, se decide construir nuevos criterios de transcripción que permiten dar cuenta de las diferentes maneras de escritura que contenían los documentos. El objetivo al hacerlo era doble. Por un lado, se contaba con la pretensión de universalidad al abarcar interesantes fenómenos lingüísticos que podían ser analizados por investigadores de diversas teorías y escuelas metodológicas, por lo que se estarían así ampliando los conocimientos sobre la historia de la lengua española desde una enriquecedora manera interdisciplinar. Por otro lado, como se dice surgió la necesidad de contribuir a crear nuevos textos que puedan utilizarse computacionalmente dentro de los intereses de los estudios diacrónicos (o de otros cualesquiera), lo que, como se señala anteriormente (cfr. § 1), se considera el objeto principal de la lingüística computacional de corpus (Parodi 2010: 18).

Desde esta perspectiva, se encuentra que en esta etapa, los criterios generales de transcripción consistieron en (véanse ejemplos 1 y 2):

- respetar los rasgos tipográficos comunes (diéresis, mayúsculas, tildes agudas, graves o circunflejas...) siempre que fuera posible;

- especificar todas las correcciones aplicadas sobre el propio texto, sin diferenciar entre aquellas realizadas por la misma mano del cuerpo textual de otra mano diferente, porque no era posible discernirlo cuando no había corrección superpuesta o se limitaba el tachado a una simple raya;
- indicar igualmente cuándo un fragmento era ilegible (por una característica propia del texto, como que estuviera tachado o borrado, o por circunstancias externas, como roto, estropeado o, simplemente, este perdido), bien mediante la palabra [ilegible] entre corchetes, bien recomponiendo el fragmento del vocablo que faltara entre corchetes, como en *predica[ban]*[4], bien mediante la palabra [falta] o el símbolo [?] para indicar que no es materialmente posible reconstruir lo no visible, el primero, y para señalar que el texto que se lee, no es suficiente para elaborar el resto de la construcción, el segundo;
- mantener la ortografía original, aunque ello supusiera cometer faltas ortográficas desde la perspectiva actual y contar con diversas variantes de escritura de una misma palabra;
- mantener la puntuación (en general), con la intención de cuidar determinados elementos que quizá en un futuro pudieran convertirse en objetos de estudio. A este respecto, cuando los signos de puntuación aparecen tachados o corregidos, se ha optado por transcribir el que más claramente se perciba, ya que en muchas ocasiones era casi imposible saber a ciencia cierta cuál es el tachado y cuál es el superpuesto.

Esta complejidad de posibilidades, nos permitió ir creando diversas etiquetas necesarias para ofrecer una transcripción lo más completa posible, evitando de esta manera una interpretación del texto excesivamente cerrada. Este hecho ocasionó dos tipos de marcas: las comillas simples latinas utilizadas como el resto de etiquetas, empleadas de este modo para que en un futuro próximo puedan ser traducidas al lenguaje de marcación HTML y, por tanto, representadas gráficamente para ser legibles por humanos y, por otro lado, las que se ponían entre corchetes por ofrecer información que iba más allá de lo simplemente gráfico, ver tabla 1

Tabla 1
Marcas entre corchetes.
Leyenda de correcciones lingüísticas

Etiquetas. Aspectos materiales del texto

<tach>, </tach>	Tachado
<superp>, </superp>	Superpuesto (escrito encima, debajo o al lado de la línea iente)

[*a* tachada, *e* superpuesta]	Después de una palabra, cuando en ella aparece una letra tachada (*a*), y encima se escribe la letra correspondiente que trata de corregirla (*e*). (En el caso de que en una palabra haya varias letras iguales, la corrección afecta a la señalada mediante ordinales, leídas, pues, de izquierda a derecha; si no se especifica, afecta a todas por igual. Si son mayúsculas o minúsculas es irrelevante, a pesar de que en algunos casos se emplea esta distinción en lugar del orden en que aparece la letra.)

Fuente: elaboración propia, (2015)

Marcas entre corchetes. Leyenda de correcciones lingüísticas

[*e* superpuesta a *a*]	Después de una palabra, cuando sobre la misma letra (*a*) se escribe la que la corrige (*e*). (En el caso de que en una palabra haya varias letras iguales, la corrección afecta a la señalada mediante ordinales, leídas, pues, de izquierda a derecha; si no se especifica, afecta a todas por igual. Si son mayúsculas o minúsculas es irrelevante, a pesar de que en algunos casos se emplea esta distinción en lugar del orden en que aparece la letra.)
[símbolo] <marg> [símbolo] </marg>	Cuando en el cuerpo del texto, aparece un símbolo o llamada de atención, y luego vuelve a aparecer en el margen con alguna anotación que amplía o corrige lo dicho en el cuerpo textual.
[*m* sobre *n*]	Se ha colocado encima de la letra en cuestión (*n*) otra letra que la corrige (*m*) pero ni se ha tachado, ni se ha superpuesto a ella.
[corregido: *les*]	Lo que se corrige traspasa las fronteras de lo puramente fonético u ortográfico, y puede ser relevante para la morfosintaxis, pragmática, etc.
[*h* tachada]	Después de una palabra, cuando la letra mencionada está tachada. (En el caso de que en una palabra haya varias letras iguales, la corrección afecta a la señalada mediante ordinales, leídas, pues, de izquierda a derecha; si no se especifica, afecta a todas por igual. Si son mayúsculas o minúsculas es irrelevante, a pesar de que en algunos casos se emplea esta distinción en lugar del orden en que aparece la letra.)
d^{e}recho	La letra con formato 'superíndice' aparece superpuesta en el texto original, corrigiendo una ausencia.
[*cc* convertido en *x*]	Cuando en la palabra aparecen letras (*cc*) que han sido gráficamente transformadas para convertirse en otra (*x*), y así corregirla. (En el caso de que en una palabra haya varias letras iguales, la corrección afecta a la señalada mediante ordinales, leídas, pues, de izquierda a derecha; si no se especifica, afecta a todas por igual.)
[*Fadrique* superpuesto]	Cuando la corrección no se limita a escribir una letra sobre otra, sino que pretende superponer varias letras a una, o viceversa.
[*c* añadida]	La letra en cuestión se añade sin superponerse o tachar a otras, aprovechando un hueco entre las grafías.

Fuente: elaboración propia, (2015)

Esta forma de transcripción cuenta con la enorme ventaja de intentar incluir todos los datos sobre el escribir de la época, lo cual implica un interesante corpus que puede ser utilizado con intereses lingüísticos, desde la fonología (ortografía, ortoepía) y la morfología (laísmo, leísmo, loísmo) a la sociolingüística (influencia de lenguas extranjeras, conciencia lingüística) o pragmática (fórmulas de cortesía), pasando por la sintaxis (construcciones perifrásticas), el léxico

(extranjerismos, dialectalismos), la semántica (cambios de significado) o incluso la psicolingüística (procesamiento del lenguaje); o con intereses históricos (percepción que de la historia tienen los ilustrados, influencias de las ideas de la época en el autor, datos históricos mencionados, etc.).

Esta amplia gama de lectores objeto, permite tener una idea de lo global de nuestra pretensión, tanto en la selección del usuario tipo (Spence 2014), como en la posibilidad de que los textos se trabajen desde muy distintas disciplinas, como ya se ha dicho (cfr. § 2), bien según los métodos clásicos (Tagliavini 1973: 55-69; Garatea Grau 2005: 51-67), bien utilizando los modernos (Schulte 2009; López Alonso 2014: 287 s

No obstante, existen también algunos inconvenientes que pueden ser paliados con programas informáticos que dispongan de un potente motor de búsqueda (como Sketch Engine, por ejemplo). Dado que no se han desarrollado todas las abreviaturas, porque quizás estas puedan constituir objeto de estudio en algún momento para algún investigador[5], si se busca una palabra como *Barcelona,* deberá buscarse tal cual pero también como *Bar*^{na}, que es la abreviatura usual del mencionado texto. De la misma manera, todos aquellos términos en que aparezcan variaciones fonéticas reflejadas ortográficamente como *cesión*, será necesario buscarlos siguiendo las pautas de las variaciones (*sesion*, *session*, *cecion*), entre otros motivos, porque es muy probable que aparezca de diversas formas a lo largo del texto, y que todas ellas estén corregidas de una manera u otra, y no siempre siguiendo el mismo criterio de homogeneidad por parte del autor original.

Conviene señalar, que en los ejemplos en que aparece etiquetado cierto elemento lingüístico como <tach> (tachado), se ha eludido añadir <sic>,

[5] Hay algunas excepciones a esta regla general. En los textos de la Biblioteca Nacional, se ha desarrollado solamente las que refieren a palabras menos frecuentes, y se ha mantenido como abreviaturas las de las fórmulas de tratamiento (<ft>). En el documento de Castellví, sin embargo, se evita mantener las abreviaturas en tres casos: i) si la palabra abreviada es demasiado breve (por ejemplo, Rl, pa), en cuyo caso, sí han sido desarrolladas, aunque incluye en este desarrollo las letras que faltan entre corchetes, para indicar, precisamente, que en el manuscrito original aparecen como abreviaturas; ii) si toda la abreviatura se encuentra al mismo nivel dentro de la caja, es decir, no hay letras superpuestas (por ejemplo, ocbre, que aparece con una raya encima); o iii) si la abreviatura en cuestión es un tanto desconocida y se cree que al lector le costaría llegar a la palabra correspondiente. En cualquier caso, la mayoría de las abreviaturas aparecen desarrolladas y sin desarrollar en algún momento del texto, por mano del propio autor; de ahí que las más comunes no sean, en realidad, difíciles de deducir si se continúa con la lectura.

[6] Resulta extremadamente complejo distinguir la autoría de una simple raya puesta sobre el texto que se está analizando, como ocurre igualmente con las letras superpuestas a otras letras (imagínese una e inserta dentro de una a o a la inversa). Si a esa tachadura acompaña la palabra o letra que quiere sustituir a la percibida como errónea, se puede entonces analizar en caso de ser de la misma persona o no, de tratarse de un texto lo suficientemente extenso como para comparar grafías; si el único elemento superpuesto es una letra, como sucede en numerosas ocasiones, el saber si ha sido escrita por el mismo autor años después de su escritura o por otra mano resulta francamente complicado. Por lo general, se tiende a pensar que las notas al margen son de otra mano mientras que las tachaduras que carecen de elementos superpuestos (es decir, se dan en el mismo momento de la escritura, ya que se produce la corrección en la misma línea) son de la misma, pero no puedo afirmarlo con idéntica seguridad de las superposiciones con o sin tachaduras. Ante la duda, se optó por no indicarlo y dejar, en este trabajo, una explicación bien clara de la manera en que ha sido realizada la transcripción. Ruego se entienda la limitación a este respecto y se sugiere, por tanto, la consulta del manuscrito original para quien tenga mayor interés o para quien compruebe que los criterios de transcripción llevados a cabo aquí, no son todo lo útil que se desea.

fundamentalmente porque el hecho de que alguien lo tache implica que se es consciente del error que se ha cometido[6]:

> 1. <p> Fue este Reÿ en sus pocos años reflectivo, ÿ docil; ciñose à seguir los consejos de los sugetos de quienes su Padre se fió por mas rectos; en su adolecencia fuè osado, le costó quedar prisionero; ÿ en su detencion que fue poca, se mostró savio [b superpuesta a v] ÿ fuè fortunado porque en su prision aafirmó [primera a tachada] en sus sienes la Corona de Napoles, fue constante, valiente, premiador de servicios, ÿ magnanimo, perdonando con liberalidad los arrepentidos </folv220> <folr221> ÿ rendidos; à los primeros restituiendoles por entero sus bienes, <dcl> â los segundos dexandoles </dcl> <tach> el </tach> <superp> con </superp> que subsistir con esplendor; fuè de condicion constante, ÿ 10 años de <tach> duradera </tach> guerra en la Conquista de Napoles fue siempre <superp> sin intermision </superp> permanente <tach> en la perseverancia </tach> (Fransisco de Castellví, folv220-folr221).

En este caso, se han etiquetado la duplicación del complemento verbal (â los segundos *dexandoles*) y otros fenómenos que, como se ve, tienen unas características más cercanas a lo gráfico que a lo estrictamente lingüístico: el artículo *el* aparece tachado en el texto, y encima se presenta la preposición *con* que complementa a la subordinada que viene a continuación (*que subsistir con esplendor*); el adjetivo *duradera* es igualmente tachado, pero esta vez no hay elemento superpuesto, a diferencia de lo que sí ocurre con *sin intermision*, que aparece superpuesto al tachado *en la perseverancia*.

Veamos ahora otros fragmentos extraídos de los textos:

> 2. Passó luego el Reÿ <ft> Don </ft> Juan â Çaragoça [*z* superpuesta a ç] como Vicario general de la Corona, ÿ en 27 Agosto 1452, <tach> pr pro </tach> <sic> porrogó </sic> las Cortes ÿ empessaron [*z* superpuesta a *ss*] â 8 Nov[iem]bre (Francisco de Castellví, folv222).
>
> 3. El 7º [artículo perteneciente a la Divina Providencia es creer] que hande acabar las Guerras en nuestros Payses, y vendrà mediante Dios nro Señor <sic>) </sic> à juzgar <cdp> a los buenos </cdp>, y malos </folr53> <folv53> Vasallos [...] (MSS10907 de la Biblioteca Nacional).
>
> 4. El Reÿ quiso apartar el recurso, ÿ no dar lugar ala submision [*b* tachada] <tach> Yzo </tach> [símbolo] <marg> [símbolo] y hizo </marg> Leÿ de union; en 29 Marzo 1344 dio Privilegio de incorporacion à perpetuidad ala Corona de Aragon, Mallorca, Yslas, Ressellon [*o* superpuesta a la primera *e*], ÿ Cerdaña: Añadió que los Vassallos no estassen [corregido: *estuviessen*] obligados à obedecer los Successores, hasta aver jurado esta Leÿ: en 3 de Maÿo la firmaron los Catalanes (Francisco de Castellví, folr166).

En (4), además de la fórmula de tratamiento *Don* y de las dos superposiciones señaladas, cabe destacar la vacilación morfofonológica en el verbo *prorrogar*. En (5), aparte del complemento directo preposicional *a los buenos*, resulta llamativa la existencia de un paréntesis cuya apertura es inexistente. Finalmente, en (6) se pueden observar las constantes dudas en determinadas palabras (*submisión* y *sumisión*) y las diversas maneras de solucionarlas que el mismo amanuense emplea: tachar (*Yzo*), tachar y superponer (*Ressellon* y

Rossellon), llamar la atención con un símbolo y añadir la palabra corregida, seguida del símbolo, al margen (*Y hizo*). También se puede indicar lo interesante de la corrección *estassen* y *estuviessen* que traspasa las fronteras de lo puramente fonológico.

Así pues, se observa que la actitud holística buscada en el análisis filológico de los textos a través de la comparación interdiscursiva (cfr. § 2.3), en la otra cara de la moneda, la de la transcripción, se aprecia en el intento de dejar el texto cuan completo sea posible para que sirva de estudio al mayor número de expertos interesados en la historia externa de la lengua española.

5. Reflexiones finales

La percepción que de la lingüística de corpus se tiene, pues, en este artículo hace hincapié en la necesidad de entenderla como una herramienta metodológica al servicio de otros objetivos (meta) lingüísticos, lo que implica que no constituye ella sola una disciplina autónoma, si bien en determinados momentos puede convertirse, naturalmente, en objeto de estudio *per se*.

Al aplicar la que, por tanto, se considera la función primordial de la lingüística de corpus (herramienta al servicio de otros objetivos), se ha empleado tres vertientes de la lingüística informática (lingüística de corpus, lingüística de corpus computacional y lingüística computacional de corpus) como sendas fases del proceso de investigación filológica que, de la misma manera que se dan en el proceso analítico de los datos lingüísticos, necesariamente han de darse también en la primera fase de la creación de los corpus digitales (especialmente con intereses diacrónicos), esto es, la transcripción de los textos.

Se recuerda, por tanto, que la primera fase, compuesta por la adquisición de conocimientos lingüísticos (de corpus), se da, en el análisis, al determinar el objeto y los textos de estudio, mientras que, en la transcripción (considerada, como se dice, la fase esencial de la creación del corpus digital), tiene lugar a la hora de aplicar los conocimientos (meta) lingüísticos para descifrar los caracteres que se van a transcribir.

En la segunda fase, la lingüística de corpus computacional permite apoyar para interpretar los ejemplos de dicho corpus (clasificándolos, analizándolos), siguiendo un método de estudio determinado previamente. Desde la perspectiva de la transcripción, el método se dió de antemano por los intereses del proyecto en que se enmarcaba, pero de no ser así, el transcriptor habría tenido que crearlo.

La última fase, que ofrece aspiraciones universales como se entiende que corresponde a la lingüística computacional de corpus, permite la comparación anecdótica de los distintos textos analizados, desde la perspectiva del estudio, y obliga al transcriptor a mantener una actitud igualmente holística que manipule el texto lo menos posible, para facilitar así la perspectiva interdisciplinar de los estudios filológicos.

En la tabla 2 se ofrece de forma sintética estas ideas, consideradas esenciales para comprender la propuesta

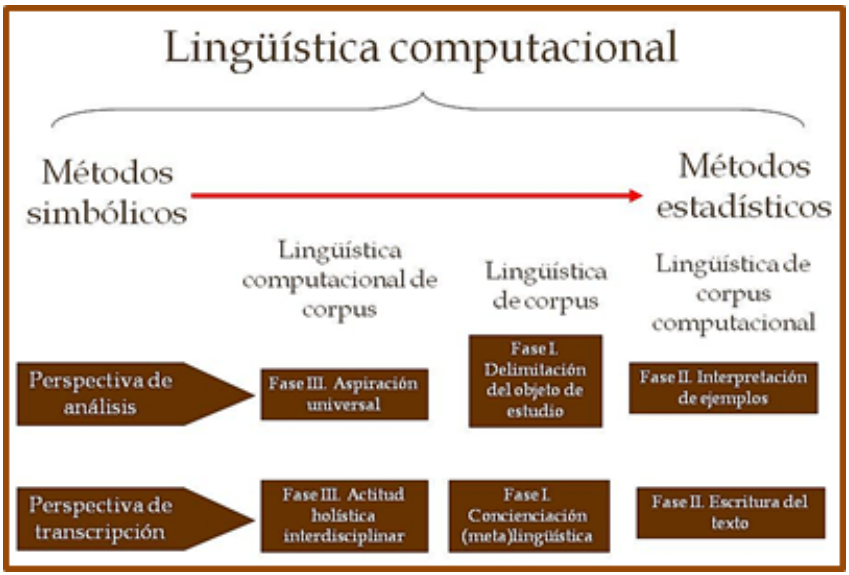
Tabla 2

	Perspectiva de análisis	*Perspectiva de transcripción*	*Vertiente de la lingüística computacional*
Fase I	Delimitación del objeto de estudio	Concienciación (meta)lingüística	Lingüística (de corpus)
Fase II	Interpretación de los ejemplos	Escritura del texto	Lingüística de corpus computacional
Fase III	Aspiración universal	Actitud holística interdisciplinar	Lingüística computacional de corpus

Fuente: elaboración propia, (2015)

Por otra parte, si se pudiera esbozar un *continuum* entre los modelos simbólicos y los modelos estadísticos que conforman, esencialmente, la lingüística computacional tal y como se comprende aquí, podría señalarse que la considerada primera fase del proceso (lingüística [de corpus]) se encontraría sumamente cerca de los modelos probabilísticos, pero no tanto, como la segunda

Figura 4.
Lingüística computacional y creación de corpus



Fuente: elaboración propia, (2015)

(lingüística de corpus computacional) que depende de ellos por completo: recuérdese que aún se puede hacer lingüística (de corpus), sin que esta sea computacional, es decir, aplicando métodos cualitativos o realizando los registros y el análisis de manera manual (incluso sobre el papel), sin dejar por ello de mantener el carácter universal de los modelos simbólicos.

La tercera fase (lingüística computacional de corpus), tal y como se ha concebido en este trabajo, esta mucho más cerca de los modelos simbólicos por su aspiración a la creación automática de gramáticas basadas en corpus, pero no llegaría a encontrarse en el extremo simbólico del *continuum*, porque este lugar estaría reservado a la lingüística más formal, ajena, en principio, a cualquier tipo de corpus.

En la (figura 4), se representa lo antes expuesto sobre estas líneas:

En consecuencia, se espera con estas reflexiones haber contribuido a situar metodológicamente el trabajo considerado como primordial en la creación de los corpus diacrónicos –esto es, la transcripción–, difícilmente separable de la tarea de análisis filológico que lo ha de preceder. Al fin y al cabo, durante la investigación lingüística todo filólogo es a la vez estudioso y creador de corpus, de la misma manera que todo escritor, durante el proceso de escritura es, simultáneamente, productor y receptor de lenguaje humano.

Referencias bibliográficas

Calderón, M. y García, Ma. T. (2009). «El Corpus Diacrónico del Español del Reino de Granada (CORDEREGRA)». En: A. Enrique-Arias (ed.). Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana Veurvert, pp. 229-249.

Caravedo, R. (1999). Lingüística del corpus. Cuestiones teórico-prácticas aplicadas al español. Salamanca: Ediciones Universidad de Salamanca.

Castellví, F. (1997) (eds. Mundet i Gifre, J. M. y Alsina Roca, J. M.). Narraciones históricas. Vol. I. Madrid: Fundación Francisco Elías de Tejada y Erasmo Pèrcopo.

Castellví, F. (1998) (eds. Mundet i Gifre, J. M. y Alsina Roca, J. M.). Narraciones históricas. Vol. II. Madrid: Fundación Francisco Elías de Tejada y Erasmo Pèrcopo.

Castellví, F. (1999) (eds. Mundet i Gifre, J. M. y Alsina Roca, J. M.). Narraciones históricas. Vol. III. Madrid: Fundación Francisco Elías de Tejada y Erasmo Pèrcopo.

Civit, M. (2003). Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. Alicante: Sociedad Española para el Procesamiento del Lenguaje Natural.

Clark, A.; Fox, C. y Lappin, S. (eds.) (2010). The Handbook of Computational Linguistics and Natural Language Processing. Oxford: Blackwell.

Clavería, G. (2012). «Corpus diacrónicos: nuevas perspectivas para el estudio de la historia de la lengua». En: E. Montero Cartelle y C. Manzano Rovira (coords.). Actas del VIII Congreso Internacional de Historia de la Lengua Española, Santiago de Compostela, 14-18 de septiembre de 2009. Santiago: Universidad de Santiago de Compostela, 405-429.

Contreras, M. (2009). «Hacia la constitución de un corpus diacrónico del español de Chile». RLA. Revista de Lingüística Teórica y Aplicada (Concepción, Chile), 47 (2), 111-134.

C-ORAL-ROM (Universidad Autónoma de Madrid). Recuperado el 1 de diciembre del 2014, del sitio web www.lllf.uam.es/c-oral-rom/index.html

C-ORAL-ROM (Universidad de Florencia). Recuperado el 1 de diciembre del 2014, del sitio web lablita.dit.unifi.it/coralrom/

Corpus de Referencia del Español Actual (CREA). Recuperado el 1 de diciembre del 2014, del sitio web corpus.rae.es/creanet.html -

Corpus del Español de la Brigham Young University, dirigido por Mark Davies. Recuperado el 1 de diciembre del 2014, del sitio web www.corpusdelespanol.org

Corpus del español del siglo XXI (CORPES). Recuperado el 1 de diciembre del 2014, del sitio web http://web.frl.es/CORPES/view/inicioExterno.view -

Corpus Diacrónico del Español (CORDE). Recuperado el 1 de diciembre del 2014, del sitio web corpus.rae.es/cordenet.html

Corpus diacrónico del español del Reino de Granada (CORDEREGRA). Recuperado el 1 de diciembre del 2014, del sitio web www.corderegra.es

Corpus Escrito del Español L2 (CEDEL2). Recuperado el 1 de diciembre del 2014, del sitio web www.uam.es/proyectosinv/woslac/cedel2.htm

Corpus hispánico y americano en la red: textos antiguos, coordinado por Borja Sánchez Prieto. Recuperado el 1 de diciembre del 2014, del sitio web www.charta.es.

Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC) Recuperado el 1 de diciembre del 2014, del sitio web http://www.lllf.uam.es/ESP/Corlec.html

Davies, M. (2009). «Creating useful historical corpora: A comparison of CORDE, the Corpus del español and the Corpus do português». En: A. Enrique-Arias (ed.). Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana Vervuert, pp. 137-166.

Eco, U. (1993). Lector in fabula. La cooperación interpretativa en el texto narrativo. Barcelona: Lumen.

El Grial. Interfaz de etiquetaje e interrogación de corpus textuales. Recuperado el 1 de diciembre del 2014, del sitio web www.elgrial.cl

Enrique-Arias, A. (2009a). «Introducción. Lingüística de corpus y diacronía de las lenguas iberorrománicas». En: Andrés Enrique-Arias (ed.). Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana Vervuert, 11-21.

Enrique-Arias, A. (ed.) (2009b). Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana Vervuert.

Fernández de Castro, F. (1999). Las perífrasis verbales en el español actual. Madrid: Gredos.

Fernández Martín, P. (2009). «El teletrabajo: reflexiones en torno a una experiencia filológica». En: P. Maya Álvarez y J. J. Caballero Trigo (dirs.). Internet como recurso para el empleo. Sevilla: Divulgación Dinámica, 40-46. Recuperado el 02 de diciembre del 2014, del sitio web http://www.gdel.es/pdf/Librosdigitales/InternetRecursoEmpleo.pdf

Fernández Martín, P. (2012a). «El estudio de la construcción pasiva en documentos del Archivo Municipal de Alcalá de Henares: reflexiones y ejemplos». En: Mª J. Torrens Álvarez y P. Sánchez-Prieto Borja (eds.). Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos. Berna: Peter Lang, 109-126.

Fernández Martín, P. (2012b). «Propuesta de un prototipo participial con base en cuatro perífrasis verbales". Boletín de Filología de la Universidad de Chile, 47:1, 33-68. Recuperado el 19 de diciembre del 2014, del sitio web http://www.boletinfilologia.uchile.cl/index.php/BDF/article/view/22220

Fernández Martín, P. (2014). «Cuestiones metodológicas en el estudio de las perífrasis verbales: interrelaciones entre sintaxis, semántica y pragmática». En: J. L. Girón Alconchel y D. M. Sáez Rivera (eds.). Procesos de gramaticalización en la historia del español. Madrid/Frankfurt: Iberoamericana/Vervuert, 119-158

Fernández, F. (2005). «Geografía lingüística de Hispanoamérica». En: J. M. Enguita; T. Buesa;

y Mª A. Martín Zorraquino (eds.). Jornadas Internacionales en memoria de Manuel Alvar. Zaragoza: Fundación Fernando el Católico/CSIC, 89-108.

Fernández-Pampillónros, A. Mª; GOICOECHEA de JORGE, Mª.; HERNÁNDEZ YÁÑEZ, L.; y LÓPEZ GARCÍA, D. (eds.) (2012). Filología y tecnología: introducción a la escritura, la informática, la información. 2ª edición revisada y ampliada. Madrid: Editorial UCM. Recuperado el 19 de diciembre del 2014, del sitio web http://eprints.ucm.es/23457/1/Filologia2completoB.pdf

Garachana, M. y Artigas, E. (2012). «Corpus digitalizados y palabras gramaticales», Scriptum Digital, 1, 37-65. Recuperado el 19 de diciembre del 2014, del sitio web http://diposit.ub.edu/dspace/bitstream/2445/35173/1/609211

Garatea, C. (2005). El problema del cambio lingüístico en Ramón Menéndez Pidal: el individuo, las tradiciones y la historia. Tübingen: Gunter Narr.

Gazdar, G. (1996). «Paradigm merger in natural language processing». En: Milner, R. y Wand, I. (eds.). Computing Tomorrow: Future Research Directions in Computer Science. Cambridge: Cambridge University Press, 88-109.

Gómez, L. (1999). «Los verbos auxiliares. Las perífrasis verbales de infinitivo». En: Demonte, V. y Bosque, I. (coords.). Gramática descriptiva de la lengua española (2). Las construcciones sintácticas fundamentales. Relaciones temporales, aspectuales y modales. Madrid: Espasa, 3323-3389.

González, Ma. I. (coord.) (2012). Unidades fraseológicas y TIC. Madrid: Instituto Cervantes, Biblioteca fraseológica y paremiológica, Serie «Monografías», nº 2. Recuperado el 19 de diciembre del 2014, del sitio web http://cvc.cervantes.es/lengua/biblioteca_fraseologica/n2_gonzalez/unidades_fraseologicas_y_tic.pdf

Hoyos, A. y Marrero, V. (2008). «Errores léxicos de habla: una perspectiva lingüística». En: Moreno Sandoval, A. (ed). Actas del VIII Congreso de Lingüística General, 25 al 28 de junio de 2008. Madrid: UAM, 1014-1025. Recuperado el 12 de diciembre del 2014, del sitio web elvira.lllf.uam.es/clg8/actas/index.html

Kay, M. (2006). «A life of language», Computational Linguistics. 31(4), 425-438.

Lee, L. (2004). «I'm sorry Dave, I'm afraid I can't do that: Linguistics, Statistics, and Natural Language Processing circa 2001». Computer Science: Reflections on the Field, Reflections from the Field, 111-118.

López Alonso, C. (2014). Análisis del Discurso. Madrid: Síntesis.

Luque, G. (2004-2005). «El dominio de la lingüística aplicada». Revista Española de Lingüística Aplicada, 157-173.

Marín, T. (1991). Paleografía y diplomática, 2 vols. Madrid: UNED.

Moreno, A. (1998). Lingüística computacional. Introducción a los modelos simbólicos, estadísticos y biológicos. Madrid: Síntesis.

Nystrand, M. (ed.) (1986). The structure of written communication. Studies in Reciprocity between Writers and Readers. Londres: Academic Press.

Olbertz, H. (1998). Verbal Periphrases in a Functional Grammar of Spanish. Berlín: Mouton de Gruyter.

Pagola, R. M.; Isasi, C.; Errasti, J. y Fernández, P. (2006). «Edición digital para el análisis lingüístico automático del corpus Bonaparte». En: Villayandre Llamazares, M. (ed.). Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística. León: Universidad de León, 1439-1441. Recuperado el 19 de diciembre del 2014, del sitio web http://fhyc.unileon.es/SEL/actas/Pagola_et_al.pdf

Palermo, M. (2011). «El CEOD: un archivo de la escritura epistolar italiana del siglo XIX». Manuscrits, 29, pp. 107-114. Recuperado el 19 de diciembre del 2014, del sitio web http://www.raco.cat/index.php/Manuscrits/article/viewFile/249948/334485

Parodi, G. (2006). «El Grial: interfaz computacional para anotación e interrogación de corpus en español».

RLA. Revista de Lingüística Teórica y Aplicada, 44 (2), II Sem., 91-115. Recuperado el 19 de diciembre del 2014, del sitio web en http://www.scielo.cl/pdf/rla/v44n2/art07.pdf

Parodi, G. (2008). «Lingüística de corpus: una introducción al ámbito». RLA. Revista de lingüística Teórica y Aplicada, Concepción (Chile), 46 (1), I Sem., 93-119. Recuperado el 19 de diciembre del 2014, del sitio web http://www.scielo.cl/pdf/rla/v46n1/art06.pdf

Parodi, G. (2010). Lingüística de corpus: de la teoría a la empiria. Madrid: Iberoamericana Vervuert.

Payrató, L. (1998). De profesión, lingüista. Panorama de la lingüística aplicada. Barcelona: Ariel.

Pöckl, W.; Rainer, F. y Pöll, B. (2004). Introducción a la Lingüística Románica. Madrid: Gredos.

Proyecto para el Estudio Sociolingüístico para el Español de España y de América. Recuperado el 1 de diciembre del 2014, del sitio web preseea.linguas.net/

Proyecto PROGRAMES. Recuperado el 1 de diciembre del 2014, del sitio web www.ucm.es/ocesosdegramaticalizacionenlahistoriadelespanol

Proyecto Val.Es.Co (Universidad de Valencia). Recuperado el 1 de diciembre del 2014, del sitio web www.valesco.es

Quevedo, F. (1622 [1992]) (ed. Domingo Ynduráin). La vida del Buscón llamado Don Pablos. Madrid: Cátedra.

Rodríguez Mansilla, F. (2004-2005). «'Émulo de Guzmán de Alfarache y tan agudo y gracioso como Don Quijote'. El lugar del buscón en la picaresca". Etiópicas, 1, 144-160. Recuperado el 19 de diciembre del 2014, del sitio web http://hdl.handle.net/10272/1602

Schulte, K. (2009). «"Using Non-Annotated Diachronic Corpora: Benefits, Methods and Limitations». En: A. Enrique-Arias (ed.). Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana Veuvert, 167-180.

Sevilla, J.; Fernández-Pampillón, A. y Poves, A. (eds.) (2011). El laboratorio de idiomas y la enseñanza-aprendizaje de lenguas. Madrid: Servicio de Publicaciones UCM. Disponible en http://eprints.ucm.es/23462/ [19/12/14]

Spence, P. (2014). «Siete retos en edición digital para las fuentes documentales». Scriptum Digital, Vol. 3 (2014), 153-181. Recuperado el 19 de diciembre del 2014, del sitio web http://scriptumdigital.org/documents/6-_Spence_-_CHARTA-III-.pdf

Stubbs, M. (2004). "Language Corpora". En: A. Davies y C. Elder (eds.). Handbook of Applied Linguistics, Oxford: Maxwell, 106-133.

Tagliavini, C. (1973). Orígenes de las lenguas neolatinas. Introducción a la filología romance. México: Fondo de Cultura Económica.

Torijano, J. A. (2004). Errores de aprendizaje, aprendizaje de errores. Madrid: Arco/Libros.

Torruella y Casañas, J. (2009). «Los ejes principales en el diseño de un corpus diacrónico: el caso del CICA». En: Cantos Gómez, P. y Sánchez Pérez, A. (eds.). A survey of corpus-based research. Murcia: Asociación española de Lingüística de Corpus, 21-36. Recuperado el 19 de diciembre del 2014, del sitio web http://www.um.es/lacell/aelinco/contenido/pdf/2.pdf

Yllera, A. (1999). «Las perífrasis verbales de gerundio y participio». En: Demonte, V. y Bosque, I. (coords.). Gramática descriptiva de la lengua española (2). Las construcciones sintácticas fundamentales. Relaciones temporales, aspectuales y modales. Madrid: Espasa, 3392-3439.

Ynduráin, D. (1992). «Introducción». En: Quevedo, F. (1626): La vida del Buscón llamado Don Pablos. Madrid: Cátedra, 13-87.

UNIVERSIDAD DEL ZULIA

Enlace

Revista Venezolana de Información, Tecnología y Conocimiento

Vol. 12, N°3_____

Esta revista fue editada en formato digital y publicada en diciembre de 2015, por el ***Fondo Editorial Serbiluz, Universidad del Zulia. Maracaibo-Venezuela***

www.luz.edu.ve
www.serbi.luz.edu.ve
produccioncientifica.luz.edu.ve