Scriptum digital
Vol. 12 (2023), pp. 39-53

Del corpus VAL.ES.CO. 3.0 a los futuros corpus diacrónicos orales: perspectivas de futuro

SALVADOR PONS BORDERÍA
Universidad de Valencia/ Grupo Val.Es.Co.
Salvador.pons@uv.es
ORCID-ID: https://orcid.org/0000-0001-5788-5506

RESUMEN
Este artículo presenta, de forma programática, las características que deben cumplir los futuros corpus diacrónicos orales, un nuevo tipo de corpus que debe combinar los requisitos de los actuales corpus orales con los requerimientos de los corpus diacrónicos. Sobre la base de la versión actual del corpus oral Val.Es.Co., se establecen las modificaciones que se deben aplicar a un corpus oral para que pueda abarcar su desarrollo microdiacrónico.

PALABRAS CLAVE: Lingüística de corpus, Val.Es.Co. 3.0, corpus orales, español coloquial, corpus diacrónicos, Pragmática

From the VAL.ES.CO. 3.0 corpus to the forthcoming oral diachronic corpora: looking forward

ABSTRACT
This article presents, in a programmatic way, the features to be met by future oral diachronic corpora. Oral diachronic corpora are a new type of corpus that combines the requirements of current oral corpora with the requirements of diachronic corpora. This paper establishes the kind of changes to be applied to current oral corpora in order to become (micro)diachronic corpora. The 3.0 Val.Es.Co. oral corpus is taken as an exemplification.

KEY WORDS: Corpus linguistics, Val.Es.Co., oral corpora, spoken Spanish, Pragmatics, diachronic corpora.

1. INTRODUCCIÓN[1]

La lingüística española ha desarrollado, en los últimos cincuenta años, una brillante tradición de lingüística de corpus: si hablamos de los corpus específicamente orales, desde los documentos pioneros del PILEI (Lope Blanch 1971) [2], que introdujeron las primeras

[1] Este trabajo ha sido posible gracias al proyecto CIPROM/2021/038 *Hacia la caracterización diacrónica del siglo XX* (DIA20), del proyecto PROMETEO de la Generalitat Valenciana, y al proyecto de I+D+I PID2021-125222NB-I00 *Aportaciones para una caracterizacion diacronica del siglo XX*, financiado por MCIN/ AEI /10.13039/501100011033/ y por FEDER Una manera de hacer Europa.

[2] El *Proyecto para el estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica* fue el primer estudio conjunto panhispánico que se planteó la creación de corpus orales paralelos que permitieran la comparación entre el español hablado de las diferentes áreas dialectales del diasistema español. Dirigido por el profesor Lope Blanch desde México, puso

Fecha de recepción: 06/02/2023
Fecha de aceptación: 24/03/20223
ISSN: 2014-640X

grabaciones orales y pusieron las bases del trabajo conjunto en grupos con una metodología única, hasta la pléyade de corpus particulares desarrollados por investigadores con fines de investigación muy concretos, los desarrollos se han caracterizado por una cuidadosa atención a las cuestiones metodológicas, por una alta capacidad de colaboración entre grupos afines y por el desarrollo de soluciones informáticas que favorezcan las búsquedas.

Aunque en sus inicios los corpus hispánicos se caracterizaron por ser sincrónicos y por registrar material oral, el desarrollo de corpus diacrónicos es, hoy en día, similar en cantidad y, en muchos aspectos, superior en calidad a sus gemelos sincrónicos. Esencial en este proceso global de desarrollo ha sido la apuesta decidida desde las instituciones académicas por la creación de grandes bases de datos, tanto sincrónicas (CREA) como diacrónicas (CORDE) que, continuadas en el tiempo, dieron lugar a su continuación (CORPES) y a su corrección (CDH) o ampliación (CORDIAM)[3]. Asimismo, el trabajo continuado de grupos de investigación aislados, como Val.Es.Co. o ESLORA, así como la coordinación de grupos en proyectos como PRESEEA o CHARTA[4], ha producido corpus específicos (https://eslora.usc.es), reflexiones teóricas (Pons Bordería 2022) y experiencia conjunta de trabajo, un intangible no por inmaterial menos importante.

Los cincuenta años transcurridos desde la publicación de los primeros materiales orales, tanto desde la tradición sociolingüística (Alvar 1972) como desde la propia del español coloquial (Lope Blanch 1977), han añadido historicidad a la práctica científica. Así, los corpus sincrónicos orales se han ido convirtiendo, lenta pero inexorablemente, en corpus diacrónicos orales. Llegados a este punto, las dos tradiciones en la lingüística de corpus hispánica comienzan a converger: corpus sincrónicos orales, por un lado, y corpus diacrónicos escritos, por el otro, creando un híbrido que, no por esperable, resulta menos problemático. Se trata de los *corpus diacrónicos orales*, una nueva categoría que, compartiendo las características de las ramas que la conforman, presenta, sin embargo, rasgos que le son propios.

Al no haberse desarrollado esta categoría en la lingüística hispánica, creemos conveniente abordar este tema en un artículo prospectivo y programático que sirva para establecer algunas bases teóricas de los desarrollos prácticos que van a tener que aplicar los corpus sincrónicos orales que, por su duración en el tiempo, hayan traspasado las barreras de la sincronía más estrecha. En este sentido, a los corpus orales les espera una etapa en la que van a tener que *crecer diacrónicamente* para abordar con coherencia el paso del tiempo.

El presente artículo partirá de un caso concreto, el del corpus Val.Es.Co. 3.0[5], para abordar los problemas que plantea su estructura y su adecuación a un estudio diacrónico. Sobre la base de este caso particular, se establecerá una serie de consideraciones que puedan ser de validez para otros corpus orales en su camino a la diacronía, referentes a la periodización de los datos, al tratamiento de las muestras, y a la tipología de los materiales.

las bases para la cooperación iberoamericana en creación de corpus y fue la semilla de macroproyectos actuales como el conjunto de corpus PRESEEA. Su importancia para la lingüística hispánica todavía no está lo suficientemente reconocida.

[3] CREA: http://corpus.rae.es/creanet.html. CORDE: https://corpus.rae.es/cordenet.html. CDH: https://apps.rae.es/CNDHE/view/inicioExterno.view;jsessionid=B70EB8E6890D2FCC9C5A54AE95DB10F1. CORDIAM: https://www.cordiam.org/.

[4] Proyecto PRESEEA: https://preseea.uah.es. CHARTA: https://www.redcharta.es.

[5] www.valesco.es

2. CARACTERÍSTICAS DEL CORPUS VAL.ES.CO. 3.0

A finales de 2022 se lanzó la versión 3.0 del corpus Val.Es.Co, un resultado de investigación producido por el grupo de investigación homónimo que se ha desarrollado sin interrupción desde 1995. El corpus Val.Es.Co es un corpus oral que recoge setenta y dos conversaciones coloquiales grabadas mediante el método de la observación participante y transcritas inicialmente según el llamado *método jeffersoniano*, desarrollado por el Análisis Conversacional americano (Jefferson, Sacks, Schegloff 1974).

Desde 1995, se pueden distinguir cuatro etapas en el corpus. Las dos primeras corresponden a la publicación del corpus en papel en 1995 y en 2002 (Briz et al. 1995; Briz y Grupo Val.Es.Co. 2002). La tercera es el corpus Val.Es.Co. 2.0 (Cabedo y Pons Bordería 2013), primera versión digital del corpus, y la última es la versión 3.0 (Pons Bordería 2022), también digital. La andadura del corpus a lo largo de un cuarto de siglo refleja bien las acomodaciones necesarias para adaptar este producto de investigación a los tiempos: así, el salto al formato digital supuso la necesidad de cambiar el sistema de transcripción al formato *.xml* a través de un programa de transcripción específico (en nuestro caso, ELAN). Asimismo, las conversaciones se sincronizaron con una barra de tiempo, lo que permite una medida exacta de la duración de turnos y de silencios. Esta alineación permite extraer muestras para análisis fonéticos o prosódicos mediante programas especializados, como PRAAT.

El corpus 3.0 ha desarrollado una serie de mejoras sobre la primera versión en línea que han supuesto un desafío teórico y práctico para el que han sido necesarios tres años de trabajo. Enumeramos algunas de las novedades:

a) La distribución de los hablantes se ha ajustado a nuevas franjas de edad. La división tradicional en tres grupos de edad (18-25, 26-55, +55) se ha ajustado para permitir su comparación con otros corpus, como el conjunto de corpus PRESEEA a (18-34, 35-55, +55), y se ha propuesto una segmentación alternativa, que desplaza el inicio de la tercera generación, más acorde con la pirámide poblacional actual (18-34, 35-65, +65). En efecto, ajustar el inicio de la generación III a la edad de jubilación es coherente con la evolución de la pirámide poblacional española y con la mejorada esperanza de vida de las últimas décadas. El motor de búsqueda permite seleccionar la información con ambas distribuciones de edad.

b) Todo el sistema de transcripción se ha adaptado al eje temporal. Mientras que el corpus 2.1 alineaba la transcripción por grupos entonativos, el corpus 3.0 alinea audio y transcripción en modo karaoke. En este nuevo formato, la transcripción tiene tantas líneas como hablantes y se puede comparar a un papiro que se desdoblara en el sentido de la progresión temporal. Si la transcripción en papel, limitada a la estructura de la página, respondía al modelo *de pergamino*, la transcripción 3.0, que se despliega en el tiempo, responde al modelo *de papiro* (Fig. 1):

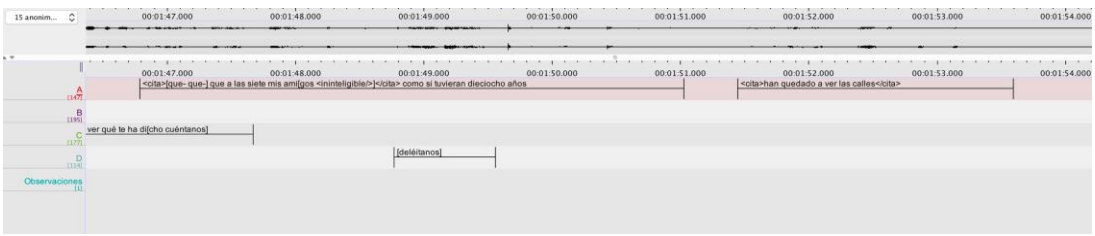

Figura 1: Transcripción en modo papiro

c) Aunque la transcripción se realiza en el programa ELAN con un sistema de etiquetado *.xml*, lo que permite las búsquedas automatizadas, en la interfaz se restituyen los signos originales de la transcripción jeffersoniana mediante un sistema de equivalencias, que se reproduce parcialmente en la Figura 2:

Fenómeno	**Signo/Símbolo**	**Etiqueta**
Tonema ascendente	↑	**<ta/>**
Tonema descendente	↓	**<td/>**
Tonema suspendido	→	**<ts/>**
Tonema circunflejo	(no existía)	**<tc/>**
Anonimización	(no existía)	**<an>**Nombre/Apodo**< /an>**
Fenómenos paralingüísticos (risas, toses, gritos…)	(RISAS), (TOS), (GRITOS)	**<risas/>, <tos/>, <gritos/>**
Entre risas	Se marca como nota al pie	**<e_risas>**texto**</e_risas>**
Estilo directo	Y me dijo *tía qué fuerte* (cursiva)	me dijo **<cita>**tía qué fuerte**</cita>**

Figura 2: relación entre etiquetado TEI y signos de transcripción

Esto facilita la lectura de las conversaciones como texto, ya que muchos estudios de tipo pragmático (cortesía, humor, etc.) necesitan leer y entender fragmentos amplios de texto. En pantalla, el texto de la Figura 1 aparece como sigue:

(1) A: yy también esta tarde → (CHASQUIDO) me flip- flipando ¡Pepa la Charanga! me dice →
C: a ver qué te ha di[cho cuéntanos]
A: [*que- que-*] *que a las siete mis ami*[*gos* (( ))]=
D: [deléitanos]
A: =como si tuvieran dieciocho años *han quedado a ver las calles*

d) Se ha estandarizado el proceso que lleva de una grabación a una conversación transcrita en la red mediante un protocolo de veintidós tareas sucesivas.

e) El corpus ha sido sometido a una tokenización inicial mediante el etiquetador XIADA, desarrollado por el grupo ESLORA[6].

f) Una parte del corpus, que asciende a 50 000 palabras (dieciséis conversaciones), ha sido segmentado en unidades y subunidades discursivas siguiendo el modelo de segmentación discursiva desarrollado por el grupo Val.Es.Co (Pons Bordería 2022). Este proceso ha durado dos años en su fase previa y un año de segmentación intensiva en el

[6] https://eslora.usc.es/

Scriptum digital
Vol. 12 (20223), pp. 39-53

que un equipo de doce personas, dividido en cuatro grupos de tres (dos segmentadores y un validador), ha dividido el corpus en unidades inferiores a la intervención (residuo, subacto, acto e intervención) y en unidades superiores a la intervención (turno, diálogo y discurso).

g) Este análisis ha sido incorporado al corpus, de modo que un investigador puede elegir, para este subcorpus, entre recuperar la transcripción desnuda o añadir a la transcripción capas de análisis (p. ej., transcripción más segmentación en subactos, transcripción más segmentación en subactos y actos, etc.).

h) Lograr esto ha supuesto un gran reto informático, que incluye la creación de un motor de búsqueda capaz de recuperar la información sobre unidades discursivas, así como la implementación de una web de administración completa e intuitiva.

i) Por último, se ha creado un nuevo *frontend* con un sistema de visualización de la información que reproduce los movimientos del ojo al leer: así, la información se despliega de arriba a abajo para la recuperación de ejemplos y de izquierda a derecha para la ampliación de información:

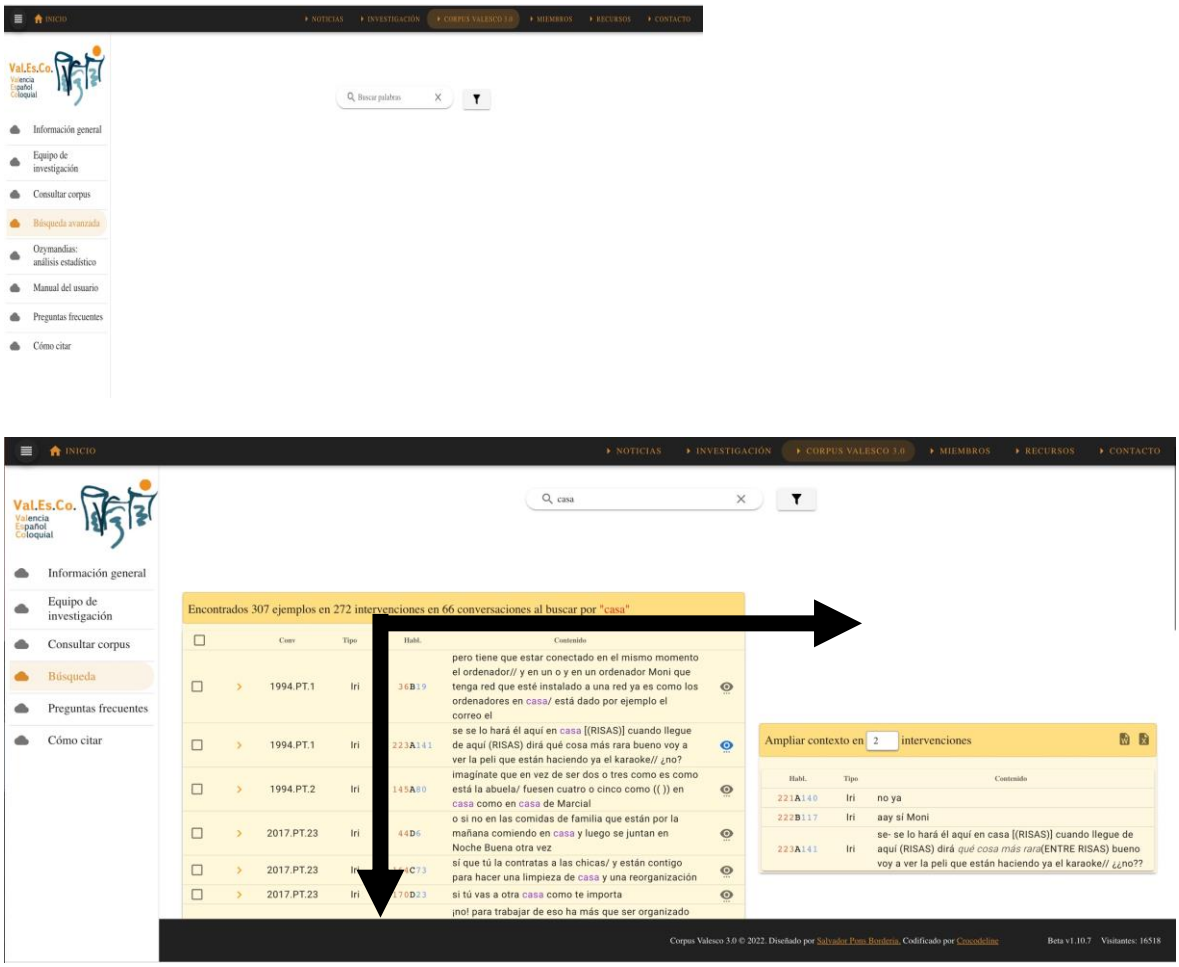

Figura 3: disposición inicial de la página de búsqueda y expansión posterior

Este es, en síntesis, el punto de llegada de un proceso de investigación que se ha desarrollado de forma ininterrumpida durante los últimos treinta años (las primeras conversaciones fueron grabadas en 1989, durante el último año de estudios de los primeros miembros del grupo). Recuperar la historicidad de dicho corpus supone no solo

unir todos los materiales producidos durante este tiempo, sino encontrar un mínimo común múltiplo en el que puedan convivir, y compararse, transcripciones en distintos formatos, muestras desiguales y grupos desigualmente representados.

Enumeramos algunos de los problemas que se plantean para convertir el corpus Val.Es.Co 3.0 en un corpus diacrónico, caso particular este sobre el que plantearemos una generalización en la sección 3.

a) Cuestiones de formato (1). Las conversaciones grabadas hasta 2002 disponen de una transcripción en formato *.docx*, frente a las posteriores, para las que se utilizó ELAN. El segundo grupo se puede exportar fácilmente a *.docx*, pero el primero no se puede convertir a ELAN, así que, para unificar los formatos, se hace necesario transponer las transcripciones en *.docx* a ELAN cortando y pegando fragmentos de unos dos segundos de transcripción. Como, además, en las conversaciones intervienen varios hablantes simultáneamente, ese copia-pega no puede ser una tarea automática, sino un auténtico trabajo de orfebre para el que se necesita tiempo, investigadores y tenacidad.

b) Cuestiones de formato (2). La situación de los archivos de sonido es más problemática aún. Algunos de estos archivos se han perdido y otros presentan una calidad de grabación tan precaria que difícilmente se entenderían en formato .mp3. En el mejor de los casos, podrán presentarse como material adicional, pero no alineados con la transcripción. Es necesario asumir que no todas las conversaciones van a ir acompañadas de su archivo de sonido y, por tanto, no van a poderse utilizar para investigaciones fónicas o prosódicas, más allá de los signos de transcripción que indiquen dichos fenómenos.

c) La muestra. De los treinta años cubiertos por el corpus, disponemos actualmente en la web de conversaciones grabadas en los años 1994, 1995, 1996, 2011 y 2016 a 2021. Añadiendo las publicadas en papel, la muestra se ampliaría a los años 1989, 1991, 1992, 1993, 1999 y 2000. Cubren, en total, dieciséis años distintos, pero agrupados en dos extremos: la década de los 90, por un lado, y la de 2010, por otro. El hueco intermedio se puede rellenar parcialmente con algunas conversaciones que formaron parte de proyectos que no se continuaron (un corpus de conversaciones de deportistas), pero hay que asumir esta heterogeneidad como limitación insoslayable del corpus. La división en microdiacronías está condicionada por los huecos del sistema, algo que resulta habitual para los diacronistas.

d) Los criterios de agrupación. La primera versión del corpus (1995) estructuró la muestra siguiendo el criterio conversaciones prototípicas ~ conversaciones periféricas[7]. La segunda (2002), por su parte, adoptó el criterio sociolingüístico por niveles socioculturales (alto / medio / bajo). Las versiones en línea (2.0 y 3.0) unen ambos criterios. Esto quiere decir que una estructuración conjunta del corpus deberá presentar una serie coherente de

[7] En 1995 se introdujo la distinción entre registro *formal* e *informal* como dos prototipos de límites abiertos, en cuya intersección se daban los registros intermedios: semiformal y semiinformal. De este modo, cada acontecimiento comunicativo podía situarse en un punto de este continuo. En el caso de las conversaciones, se observó que podían desplazarse a distintos puntos de esta escala en función de la marcha del intercambio: así, lo que empieza como una entrevista transaccional puede derivar en una conversación coloquial. Para dar cuenta de este fenómeno se acuñó el término *coloquialización* y se definió dicho proceso a partir de la acción conjunta de cuatro variables: relación de igualdad entre los participantes, relación vivencial de proximidad, marco de interacción familiar y temática no especializada.
La organización del corpus de 1995 se estructuró de acuerdo con esta distinción teórica. Para comprobar que dicha distinción era más predictiva para determinar la coloquialidad que la división por rasgos sociolingüísticos, el corpus de 2002 se articuló en torno a los niveles socioculturales de los hablantes.

criterios de agrupación del corpus para permitir su comparabilidad, por un lado, y para ser fiel a su historicidad, por el otro.

La situación descrita arriba no es exclusiva de un grupo de investigación. *Mutatis mutandis*, puede considerarse que la situación de los corpus mantenidos en el tiempo (excepción hecha de los corpus académicos) presenta problemas similares, debidos, en su mayor parte, a las fluctuaciones en los equipos de investigación y al carácter azaroso de las ayudas públicas. Crear y, sobre todo, mantener un corpus requiere una cantidad de horas muy considerable y unos recursos humanos que no siempre están al alcance ni siquiera de los departamentos más numerosos. Lo normal es que las plantillas de los corpus consten de un IP con vinculación permanente y un cuerpo flotante de estudiantes, doctorandos y profesores de distintas categorías que ven su paso por el corpus como una etapa puntual. Así, sin una plantilla estable, los corpus crecen a velocidad irregular. Añádase a esto que las horas dedicadas a revisar transcripciones, subir a la red o detectar problemas con el *software* no cuentan como méritos puntuables en concursos ni sirven para obtener sexenios; es decir, son una inversión muy poco rentable en términos de currículo.

Pero si, siguiendo a Labov (1972) la investigación histórica consiste en "hacer un buen uso de malos datos", será necesario ir, más allá de las características accidentales que determinan el ser, lo actual, a los requisitos ideales que conforman el deber ser, lo obligatorio dentro de lo posible. Así pues, sobre la base de estos datos, la pregunta pertinente es la de qué características debería presentar un corpus diacrónico oral.

3. CARACTERÍSTICAS DE UN CORPUS DIACRÓNICO ORAL

Dado el carácter programático de este trabajo, vamos a plantear tales características como un conjunto de requisitos anidados que están determinados por las tres palabras que componen la expresión *corpus diacrónico oral*: qué implica ser un corpus, qué añade el apellido diacrónico, y qué supone el que, además, sea de procedencia oral.

No todo conjunto de materiales presentado bajo un mismo título es un corpus. Siguiendo a Sinclair (1996, 5), resulta útil distinguir entre *corpus*, *archivos de texto* y *colecciones de ejemplos*. Los primeros son:

> A machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of texts is compiled with the intention (1) to be representative and balanced with respect to a particular variey or register or genre and (2) to be analyzed linguistically (Gries 2009, 7)

Esto significa que todo corpus se compila hoy en día para su procesamiento electrónico y que debería cumplir los criterios de representatividad y equilibrio; es decir, que la relación entre *población* y *muestra* (en el sentido estadístico de estos términos) sea proporcionada.

Frente a los corpus, los *archivos de texto* (Gries 2009) son una base de datos que no ha sido construida para ser analizada lingüísticamente (por ejemplo, el archivo de las sesiones del Parlamento español en sus diferentes legislaturas) y que, por esta razón, no presenta equilibrio en las variables que la componen.

Por último, una *colección de ejemplos* es el nombre con el que se designa un conjunto de ocurrencias compiladas para un fin específico (por ejemplo, para una tesis doctoral).

Estas definiciones ponen una primera restricción a la noción de corpus: para ser considerado como tal, un proyecto debe partir de un criterio homogéneo y organizador en su base. No todos los corpus llamados corpus, siguiendo esta definición, lo serían.

Ser *diacrónico* implica una serie de restricciones, que se añaden a las anteriores. La más evidente es la temporal: un corpus diacrónico debe presentar, al menos, dos cortes sincrónicos que permitan establecer una ruta evolutiva en la que el estado de lengua A se pueda comparar con el estado de lengua B. Los corpus diacrónicos tradicionales han operado, o bien sobre toda la historia del español (como en el caso de la red CHARTA o los corpus académicos CORDE o CDH), o bien sobre un corte sincrónico determinado (como el corpus Biblia Medieval); en estos casos, la comparación con la sincronía actual se da, de forma más o menos implícita, bien contra el conocimiento agente del investigador (López Serena 2007), bien contra el sistema actual del español.

El reflejo de estados de lengua pretéritos impone una serie de restricciones sobre el material disponible. Dado que los corpus diacrónicos, hasta el momento, han trabajado con materiales escritos, las restricciones propias del método filológico se han aplicado a la construcción de corpus. En este sentido, la creación del CDH académico a partir del CORDE resulta un caso paradigmático, como queda reflejado en las críticas realizadas por diferentes autores (Garachana y Artigas 2012, Enrique Arias 2018, Molina y Octavio de Toledo 2017, entre otros). De la crítica de estos y otros autores al corpus académico sobresalen cuestiones como las siguientes (Molina y Octavio de Toledo 2017, 9-37):

a) la selección de las ediciones más adecuadas para cada obra, con indicación de originales y copias, lo que debería excluir copias tardías o poco fiables (ver para este punto el *Cordemáforo* de Molina y Octavio de Toledo 2017).

b) la correcta representación de la escritura de cada época, mediante el uso de ediciones críticas con grafías no modernizadas.

c) la correcta datación y catalogación de cada obra, siguiendo las investigaciones filológicas más fiables.

d) la proporción de obras seleccionadas para cada periodo histórico, que debería producir muestras representativas y equilibradas de la producción textual de cada época.

e) la proporción de obras para cada área geográfica, lo que es especialmente relevante cuando se aborda la diatopía del español. El desequilibrio entre los textos peninsulares y los del español de América ha llevado a compilar corpus específicos de las variedades americanas, como el excelente CORDIAM.

Además de estas cuestiones, altamente problemáticas, el material de los corpus diacrónicos impone otras restricciones que se deben tratar en el buscador, como son:

a) La variación gráfica, que afecta a la relación entre formas y lemas. En este sentido, una correcta lematización deberá tener en cuenta todas las variantes gráficas (usuales o no) halladas en los textos del corpus.

b) La autoría del documento, ya sea original o copia.

c) El adecuado equilibrio de géneros textuales en el corpus por periodos históricos.

d) La selección representativa de autores para cada periodo considerado.

El añadido del término *oral* al de *corpus diacrónico* no produce una expresión composicional; es decir, las restricciones que se aplicaban a los corpus diacrónicos no se aplican en bloque a los corpus diacrónicos orales. Por el contrario, el término *oral* libera de

alguna de las servidumbres de un corpus diacrónico para imponer, como corresponde a un nuevo objeto de estudio, otras nuevas.

Comenzando por los requisitos diacrónicos que no se dan en un corpus diacrónico oral, las muestras orales quedan libres de muchas cuestiones filológicas, como las variaciones gráficas, las cuestiones de autoría (de momento) o la selección de autores, ya que los documentos orales tienen una gestión mucho más directa en este sentido. Asimismo, las cuestiones diatópicas están especificadas de partida en los corpus ya existentes, donde se establece un ámbito claro de acción, ya sea un punto geográfico concreto (corpus Val.Es.Co.), varios puntos comparables (corpus PRESEEA), o varios puntos comparables –áreas rurales– dentro de un límite espacial –España– (corpus COSER). Estos límites establecen la frontera que dichos corpus no van a sobrepasar (por ejemplo, si el corpus COSER incluyera ciudades en su estudio sería más bien un nuevo corpus que una extensión del ya existente). Esto no quita para que los corpus diacrónicos orales de nueva planta no especifiquen el ámbito de variación que van a cubrir.

Lo que los corpus diacrónicos orales deberán especificar en su macroestructura son los cortes temporales que se pretenden documentar y los criterios de organización para cada corte temporal. Estas decisiones concretas, propias de cada corpus, son, a su vez, dependientes de dos decisiones generales: la periodización del español oral en el siglo XX y la distinción de los géneros textuales orales propios que deberían cubrir los distintos corpus. Desarrollamos a continuación estas cuestiones, comenzando por las más generales:

3.1. Límites temporales y aspectos de periodización

Si los corpus diacrónicos orales se plantean como corpus monitores, que se van a seguir alimentando en el tiempo, dichos corpus van a tener un punto de inicio concreto (el de la primera grabación oral de que se disponga) y un punto de llegada abierto; es decir, se asemejarán a un intervalo con límite cerrado a su izquierda y límite abierto a su derecha ([…)). Dado que la grabación de voz más antigua registrada hoy en día se encuentra en la Biblioteca Nacional y data de 1897[8], los límites del campo están bien especificados. A partir de este punto los distintos corpus tienen que trazar su campo de acción en función de sus especificidades. En el caso del corpus Val.Es.Co, por ejemplo, su ámbito de acción diacrónico queda limitado por su primera grabación, que es de 1989. Así, cada corpus oral deberá establecer su rango de una forma coherente con su función y objetivos.

La periodización de los géneros orales del español del siglo XX es una tarea por desarrollar, ya que depende de investigaciones futuras, específicas sobre el desarrollo de cada uno de estos géneros. El caso de Salameh (en prensa) sobre las narraciones deportivas es un ejemplo programático del tipo de estudios que van a ser necesarios para cubrir este hueco descriptivo. Sin embargo, es útil comenzar por algún tipo de distinción básica y creemos que, como punto de partida, resulta pertinente la distinción establecida por Cortés Rodríguez (1994) entre estudios de análisis del discurso *premagnetofónicos* y estudios *magnetofónicos*. Siguiendo a dicho autor, podemos plantear una primera división, de tipo técnico, entre el periodo previo a la existencia del magnetófono y el periodo posterior, lo que divide el siglo XX en dos etapas: la etapa inicial, entre finales del siglo XIX

[8] http://bit.ly/3e2JR82 (Consulta realizada el 21/01/2023).

y 1960, y el periodo posterior a 1960. En la primera de estas etapas, el número de grabaciones es escaso y los corpus deben hacer frente a la *falta* de muestras. Por el contrario, en la segunda etapa los datos se multiplican y los corpus deben enfrentarse al *exceso* de muestras. Desarrollamos esta idea en el apartado 4 de este trabajo.

3.2. Cuestiones de catalogación, conservación y digitalización

Al hablar de los archivos sonoros del siglo XX pudiera pensarse en un cuerpo cerrado de documentos bien catalogado, pero lo cierto es que, en buena medida, desconocemos de cuántos archivos consta el conjunto total de grabaciones de, por ejemplo, la primera mitad del siglo XX. No se trata de que los lingüistas desconozcan los catálogos existentes, sino de que existe en los distintos archivos material todavía sin catalogar y de que, además, los distintos centros que los hospedan no conocen los archivos de centros similares. Por tanto, la tarea de ordenación y catalogación del material sonoro es un trabajo primordial y necesario.

Por otro lado, esos archivos, cuyo número se desconoce, distan de ser un conjunto ordenado de soportes físicos: Cilindros de cera, discos de pizarra, perforados o de vinilo; cintas abiertas, de casete o DAT, más los formatos digitales de las últimas décadas, en distintos grados de conservación, que deben ser volcados a un formato único, de tipo digital. Las grabaciones más antiguas y delicadas pueden estar en un estado muy precario y deben ser tratadas por personal especializado de los archivos correspondientes antes de llegar a manos de los lingüistas; la colaboración con los conservadores de dichos archivos se hace, por tanto, esencial.

Por último, la cuestión de los derechos de reproducción y la necesaria buena voluntad de las instituciones implicadas para su publicación en corpus públicos es quizá el problema que se nos antoja más grave, y para cuya resolución harán falta horas de reuniones con distintas instancias, negociaciones nada fáciles y trabas burocráticas sin cuento. Nada, por otra parte, a lo que no estén acostumbrados los investigadores.

4. TIPOLOGÍA PARA UN FUTURO CORPUS DIACRÓNICO ORAL

Tomando la división establecida en la Sección 3 entre el *periodo de la falta de muestras* y *periodo del exceso de muestras* como base para la organización de un corpus diacrónico oral, dicha distinción deberá someterse a subdivisiones más finas en función de la cantidad de materiales que se consiga recuperar.

El periodo de la *falta de muestras* requiere de una labor previa de *arqueología sonora* a la búsqueda de materiales orales, especialmente los anteriores a la Guerra Civil. En este sentido, será necesario realizar un catálogo de todos los archivos que cuenten con muestras sonoras (museos, archivos, fundaciones públicas y privadas, sociedades e incluso archivos personales); comprobar si están disponibles en línea, si están catalogadas y si son accesibles, así como negociar con sus responsables las condiciones de una posible cesión para usos de investigación o, en una situación ideal, para su consulta en abierto en un corpus lingüístico. Esta última posibilidad resulta altamente problemática porque las fuentes sonoras pueden estar protegidas por derechos de autor y, además, estar sujetas a

las condiciones de la institución que las aloja. Este proceso de recuperación del patrimonio sonoro español se nos antoja largo y tedioso, pero absolutamente necesario para hacer accesible un material tan digno de protección como un incunable; de hecho, consideramos que las primeras grabaciones constituyen verdaderos *incunables de la oralidad*, y como tales deberían tratarse.

De entre los materiales por rescatar en esta primera etapa se cuentan las primeras grabaciones radiofónicas (desde 1924, aunque prácticamente perdidas), los documentos producidos durante la Guerra Civil, o las crónicas deportivas de los años cincuenta. Asimismo, la existencia de grabaciones privadas (previas al advenimiento del magnetófono o coetáneas) puede completar los registros existentes.

Para el segundo periodo, por el contrario, hay que lidiar con el exceso de datos: la mejora en los procedimientos para el archivo de material sonoro, la existencia de más emisoras de radio, la aparición de la televisión, y la irrupción del magnetófono, que permite al ciudadano medio hacer grabaciones de su propia habla o de la de su entorno, multiplica las posibilidades de documentación. Por esta razón, puede resultar útil establecer una distinción teórica entre los tipos de archivos que se pueden encontrar los investigadores. De forma programática, estableceremos una división provisional en *corpus lingüísticos*, *archivos orales no lingüísticos*, *materiales orales* y *materiales de la oralidad*. Estas cuatro clases representan una gradación entre las muestras más fieles al acto comunicativo y las que, siendo orales, tienen más de mímesis que de oralidad.

En el apartado de *corpus lingüísticos*, disponemos de las primeras entrevistas grabadas (los materiales del PILEI) y de las conversaciones de Criado de Val (1966), así como los materiales de los distintos corpus orales que se han recopilado en España y en Hispanoamérica desde finales de los ochenta (Esgueva y Cantarero 1981, Carbonero Cano 1985, Briz et. al. 1995, Lope Blanch 1996, López Barrios y Mendoza 1997, Gómez Molina 2001, etc.), desiguales en método, extensión, duración y objetivos, pero una base que no habría que desdeñar.

Pero, además de materiales grabados por lingüistas para uso lingüístico, disponemos de *archivos orales no lingüísticos*, que guardan muestras orales sin pretensión de organización lingüística. Destaca por encima de todos ellos el Archivo de la Palabra de la Biblioteca Nacional, los archivos de las emisoras de radio y televisión más antiguas y el Archivo del Congreso de los Diputados (https://www.congreso.es/archivo-audiovisual), que conserva muestras orales desde la X Legislatura (2011). Además de estos, existe un número indeterminado de archivos sonoros dispersos por distintos centros de investigación, tanto nacionales como regionales, que resulta necesario consultar y evaluar. Este es uno de los apartados que más sorpresas puede deparar.

Una tercera fuente para este mapa sonoro del siglo XX la constituyen los *materiales orales*, muestras orales de hablantes, anónimos o no, grabados en entrevistas o en programas de radio y televisión. Con motivo de catástrofes naturales o acontecimientos sociales, los micrófonos se desplazaban al lugar de los hechos y recababan la opinión de las personas que allí se encontraban. Asimismo, entrevistas a héroes anónimos, o a deportistas de origen popular, como ciclistas, boxeadores, toreros o futbolistas, constituyen muestras diafásicas (principalmente en registros semiformal y semiinformal) que informan, a su vez, de la variación diatópica en el continuo diacrónico.

Por último, y ya dentro de las muestras miméticas, podemos distinguir los *materiales de la oralidad*, reconstrucciones orales de la oralización en obras de teatro, películas, series

de televisión, etc. Su carácter hablado las distinguirá de los corpus de la oralidad, que desarrollaremos en el siguiente apartado. En este grupo se incluyen obras artísticas que tienen entre sus objetivos reproducir unos usos orales determinados, ya sea en obras de teatro (*Los quinquis de Madriz*), en películas (el cine quinqui de finales de los setenta, las comedias de la Movida), en series de televisión (*Siete Vidas*, *Los Serrano*), en chistes o en canciones coplas, cuplés, números de revista y, más tarde, en canciones humorísticas (*Autotango del cantautor*, *¿Qué pasa contigo, tío?, Saca el güisqui, cheli*, etc.) o no (como en Serrat o en Estopa). La frontera entre este material y el que se incluye en los corpus de la oralidad es sutil y está sujeta a estudio más detallado: por ejemplo, las obras de teatro de Arniches pueden considerarse tanto material de la oralidad como elementos de un corpus de la oralidad.

5. MÁS ALLÁ: CORPUS ORALES Y CORPUS DE LA ORALIDAD

Como se ha indicado en la sección anterior, tres quintos del siglo XX quedan comprendidos en la etapa de la falta de muestras. Por esta razón, y junto a la tarea de arqueología sonora, se hace necesario completar este periodo con un material complementario, al que damos en llamar *corpus de la oralidad*. Se agrupan aquí aquellos materiales escritos que reflejan elementos del español coloquial de la época en que se produjeron, como novelas, prensa o revistas y sirven para triangular los resultados encontrados en los corpus orales. En este material, se pueden distinguir los producidos por un único autor (caso de las novelas o el teatro) de las creaciones colectivas, como las cabeceras de prensa:

El listado de las obras relevantes para los corpus de la oralidad cuenta desde hace años con algunas de las investigaciones pioneras sobre español coloquial (Arniches –Seco 1972– o Sánchez Ferlosio –Hernando Cuadrado 1988–), pero han quedado relegadas las de autores como Jardiel Poncela, Arturo Barea, Francisco Umbral, Alfonso Sastre, José Luis Alonso de Santos o José Ángel Mañas, que merecerían estudios diacrónicos específicos.

En el apartado que podría denominarse colectivo destacan las publicaciones satíricas, ya que este género, debido a su finalidad humorística, dispone de una libertad para reproducir usos orales que podían estar vedados en otro tipo de géneros discursivos. Al existir prensa satírica a lo largo de todo el siglo XX (desde *La Traca* hasta *El Jueves*) se dispone de un género discursivo único que dota de coherencia al material de los corpus de la oralidad.

Las publicaciones contraculturales, que se inician a principios de los años setenta y sobreviven hasta los noventa (Llopis Cardona 2023), tienen en el uso coloquial del lenguaje una herramienta de reivindicación política y social. Fanzines (*El Rrollo Enmascarado*), cómics (*Makoki*, *El Víbora*) y revistas como *Ajoblanco* ocupan un lugar central en la comprensión del español de los primeros procesos de coloquialización (Pons Bordería en prensa).

Las fotonovelas y las revistas juveniles son una fuente de gran valor para el estudio del lenguaje juvenil entre los años 70 y 2000, especialmente en la documentación de modas lingüísticas. Este material permite triangular los resultados obtenidos de la consulta de los corpus de lenguaje adolescente (corpus COLA y ALCORE/ COVJA), pero plantea un importante problema técnico, ya que la disposición gráfica de texto e imágenes en dichas

publicaciones, así como la profusión de tipos de letra variados para responder a la imagen juvenil y cercana de estas revistas hace muy difícil su tratamiento informatizado (más allá del mero escaneado en OCR) con la tecnología actual.

En la tipología de acabamos de presentar existen problemas de límites, especialmente entre los materiales de la oralidad y los corpus de la oralidad: en efecto, el estudio de la oralización de una obra no esconde que dicha obra haya sido compuesta por un autor, de modo que situar ciertas muestras, como canciones u obras de teatro, en uno u otro apartado, requiere de criterios adicionales que deberán especificarse de forma más detallada.

Más allá de estos materiales lingüísticos, y ya fuera de la consideración de corpus, se encuentran los *estudios metalingüísticos* de los primeros investigadores sobre el lenguaje hablado, que se pueden considerar investigaciones sobre microdiacronías pretéritas. El trabajo pionero de Beinhauer (1963:1929), por ejemplo, no se debe considerar ya una fuente del español hablado actual, sino una completísima descripción histórica del español hablado en los años diez y veinte; de igual modo, el estudio de Vigara Tauste (1980) refleja la lengua hablada a finales de los setenta. Lo mismo se puede afirmar de los trabajos de Seco (1970) sobre el habla de Arniches o del de Hernando Cuadrado (1988) sobre *El Jarama*. Es cierto que se da un alto grado de continuidad entre las descripciones de estos trabajos y las características del español hablado actual; por este motivo, un objetivo importante de la investigación consistirá en deslindar la variación diacrónica de su continuidad en el tiempo.

La clasificación establecida en estas páginas puede resumirse en el siguiente esquema, que es también una hoja de ruta:

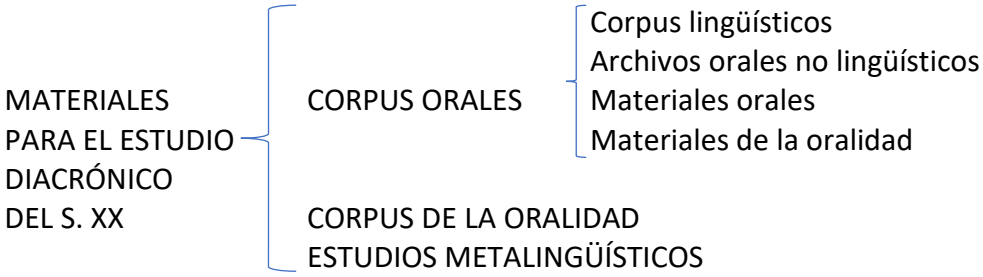

Figura 4: Materiales para el estudio diacrónico del siglo XX

5. CONCLUSIONES

En estas líneas hemos pretendido establecer una hoja de ruta programática para el diseño y estructuración de los futuros corpus diacrónicos orales que, por primera vez, se ocuparán de la diacronía de la palabra en español. Como corresponde a un nuevo objeto de estudio, se han abordado dos tipos de cuestiones: el diseño de la planta del corpus (en el que se incluyen los problemas de periodización, tipo de material, géneros discursivos y proporción entre sus partes) y los problemas de localización y tratamiento de las muestras (lo que implica el hallazgo y catalogación de los materiales existentes, su tratamiento e informatización y los posibles problemas legales y burocráticos vinculados a su aprovechamiento lingüístico). Asimismo, hemos establecido una tipología inicial en la que

cada unidad de estudio (ya sea archivo sonoro, obra literaria, material escrito o estudio lingüístico) ocupa su lugar en un continuo.

Esperamos que estas consideraciones sirvan de guía a los investigadores que, en un futuro muy cercano, se adentren en esta nueva y apasionante página de la historia de la Filología española.

BIBLIOGRAFÍA

Academia Mexicana de la Lengua, *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) <www.cordiam.org>

Alvar, Manuel (1972): *Niveles socio-culturales en el habla de Las Palmas de Gran Canaria*. Las Palmas, Cabildo Insular.

Beinhauer (1963:1929): *El español coloquial*. Madrid, Gredos.

Briz, Antonio et al. (1995): *La conversación coloquial. Materiales para su estudio*. Vol. XV. València: Universidad.

Briz, Antonio, y Grupo Val.Es.Co. (2002): *Corpus de conversaciones coloquiales*. Anejo I de *Oralia*. Almería, Universidad.

Cabedo, Adrián y Salvador Pons Bordería (dirs.): Corpus Val.Es.Co. 2.1.

Carbonero Cano, Pedro (1985): Sociolingüística andaluza. Sevilla, Universidad.

CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos): [en línea] www.corpuscharta.es.

Cortés Rodríguez (1994): Tendencias actuales en el estudio el español hablado. Almería, Universidad.

Criado de Val (1966): "Esquema de una estructura coloquial". Español Actual 8, p. 9.

Enrique Arias (2018): "Some methodological issues in the corpus-based study of morphosyntactic variation". En Richard J. Whitt (ed.): Diachronic Corpora, Genre and Language Change. Amsterdam, John Benjamins, 261-280.

Esgueva, Manuel, y Margarita Cantarero. 1981. *El habla de la ciudad de Madrid: Materiales para su estudio*. Madrid: CSIC.

ESLORA: Corpus para el estudio del español oral <http://eslora.usc.es>, versión 2.1 de junio de 2022, ISSN: 2444-1430

Garachana, Mar y Artigas, Esther (2012): "Corpus digitales y palabras gramaticales". *Scriptum Digital* 1, 37-65.

Gómez Molina (2001): El español hablado de Valencia, I: Materiales para su estudio. Valencia, Cuadernos de Filología.

Gries, Stefan (2009): *Quantitative Corpus Linguistics with R: A Practical Introduction*. Londres, Routledge.

Hernando Cuadrado (1988): El español coloquial en El Jarama. Madrid, Playor.

Inés Fernández-Ordóñez (dir.) (2005-): Corpus Oral y Sonoro del Español Rural, <http://www.corpusrural.es>.

Labov, William (1972): «Some Principles of Linguistic Methodology». *Language in Society* 1 (1): 97-120.

Llopis Cardona, Ana (en prensa): Factores socioculturales y fases en los procesos de difusión del español coloquial. Spanish in Context 20.

Lope Blanch, Juan M. (1971): El habla de la ciudad de México. Materiales para su estudio. México, UNAM.

Lope Blanch, Juan M. (ed.). 1996. *Memoria de la V reunión de trabajo de la Comisión Ejecutiva del estudio del español hablado culto*. México: U.N.A.M.

López Barrios y E. Mendoza (1997): *El habla de Sinaloa. Materiales para su estudio*. Culiacán, Universidad Autónoma de Sinaloa.

López Serena, Araceli. 2007. *Oralidad y escrituralidd en la recreación literaria del español coloquial*. Madrid, Gredos

Pons Bordería, Salvador (2022): *Creación y análisis de corpus orales. Saberes prácticos y reflexiones teóricas*. Berna, Peter Lang.

Pons Bordería, Salvador (dir.): *Corpus Val.Es.Co 3.0.* <http://www.valesco.es>

PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CDH) [en línea]. *Corpus del Diccionario histórico de la lengua española.* <http://www.rae.es>

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es>

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es>

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es>.

Rodríguez Molina, Javier y Álvaro Octavio de Toledo 2017): La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística. Scriptum Digital 6, 5-68.

Rojo, Guillermo (2010): "Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA". *Lingüística*, 24, 11-50.

". En Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*,¡. Berlin, de Gruyter, 197-212.

Rojo, Guillermo (2016): "Citius, maius, melius: del CREA al CORPES XXI". En Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*,¡. Berlin, de Gruyter, 197-212.

Harvey Sacks, Emanuel A. Schegloff y Gail Jefferson (1974): "A Simplest Systematics for the Organization of Turn-Taking for Conversation". Language 50, 4, 696-735.

Seco, Manuel (1970): Arniches y el habla de Madrid. Madrid, Alfaguara.

Sinclair, John. 1996. «Preliminary recommendations on Corpus Typology». Accedido 2 de junio de 2020. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

Vigara Tauste (1980): Aspectos del español habladoaportaciones al estudio del español coloquial. Madrid, SGEL.