**RLA**. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 47 (2), II Sem. 2009, pp. 111-134.
CL ISSN 0033 - 698X

# HACIA LA CONSTITUCIÓN DE UN CORPUS DIACRÓNICO DEL ESPAÑOL DE CHILE*

## TOWARDS CHILEAN SPANISH LANGUAGE DIACHRONIC CORPUS

MANUEL CONTRERAS SEITZ
Universidad Austral de Chile, Chile
manuelcontreras@uach.cl

**RESUMEN**

Este artículo describe y analiza la constitución de corpus diacrónicos hispánicos (españoles e hispanoamericanos) con el objeto de situar el marco de referencia del CorDECh (Corpus Diacrónico del Español de Chile). El trabajo presenta una intersección entre disciplinas tradicionales, como la filología y la paleografía, con especialidades lingüísticas que se han revitalizado en los últimos años, como es el caso de la lingüística del corpus, en virtud de la conformación de un corpus diacrónico del español de Chile. Además, en este texto se da cuenta de los criterios de constitución de este corpus (cronológico, de selección y transcripción documental, diatópico y de variedad de registros); asimismo, se discuten aspectos de la automatización del corpus y las implicancias y proyecciones para el análisis del mismo. Finalmente, se repasan los criterios de etiquetación del corpus, de acuerdo con los aspectos de la Text Encoding Initiative (TEI), adoptados por la RAE en el *Prontuario de Marcación SGML*, utilizado para el CORDE (Corpus Diacrónico del Español), junto con presentar algunos ejemplos de estos aspectos, para lo cual se utilizó el editor XML/ XSLT Cooktop 2.5.

*Palabras clave*: Lingüística del corpus, corpus diacrónico, Español de Chile.

**ABSTRACT**

The conformation of Hispanic diachronic corpora (Spanish and Hispano-American) in order to place the framework of the CorDECh (Corpus Diacrónico del Español de Chile) is described and analyzed in this paper. The work shows an intersection between traditional disciplines, such as the philolology and the paleography, with linguistic disciplines that

* Trabajo derivado del proyecto Fondecyt Nº 1040072. Para esta presentación, en sus aspectos estructurales y con fines comparativos, se ha tenido como referencia el artículo de Carrera y Herrán (2006).

have been revitalized in the last years, such as corpus linguistics, in order to shape a Chilean Spanish diachronic corpus. In addition, the criteria of corpus formation is presented in this text (chronological, documentary selection and transcription, diatopics and register varieties); in the same way, some aspects of corpora automatization and the implications and projections for its analysis is discussed here. Finally, the criteria for corpus labelling are reviewed, according to the Text Encoding Initiative (TEI), adopted for the RAE in the *Prontuario de Marcación SGML*, used for the CORDE (Spanish Language Diachronic Corpus), together with showing some examples of these aspects, for which the Cooktop 2.5 XML/XSLT editor was used.

*Keywords*: Corpus linguistics, diachronic corpus, Chilean Spanish language.

*Recibido: 14-01-2009. Aceptado: 10-08-2009*.

## 1. ANTECEDENTES TEÓRICOS Y DE CORPUS DIACRÓNICOS ESPECÍFICOS

Al igual que los *Atlas Lingüísticos*, cuando se estudia la variación sincrónica del lenguaje, la constitución de un corpus diacrónico de la lengua es la base fundamental para el estudio histórico de los diversos aspectos de nuestro idioma, pues ya no sólo se aprecia la óptica interna del estudio lingüístico, sino que se va entretejiendo la red contextual que circunda y "envuelve" el texto que se está analizando[1].

En este sentido, la vinculación de la filología con disciplinas complementarias, como la paleografía y la diplomática, no sólo está por sobre la mera auxiliaridad, sino que más bien constituyen el núcleo fundamental de la misma, ya que la propia filología se caracteriza por ser un área multidisciplinaria. Es más, comporta una relación con la conciencia lingüística y escrituraria del hablante, a quien las distinciones de este último tipo le eran más propicias en la medida en que el sistema fonológico que poseía resguardaba determinadas oposiciones del mismo. De este modo, el lingüista que precise el expurgo de documentos de archivo para llevar a cabo su labor, encontrará que necesariamente deberá recurrir al método de la paleografía de lectura.

En un sentido extrapolable a los estudios filológicos y lingüísticos, señala Jara (1996) lo siguiente en relación con las fuentes documentales:

[1] O tal como lo expresa Jucker y Jacob (1995) respecto de la pragmafilología: "Traditionally, historical linguists have spent most of their efforts on sound changes and on the phonology and morphology of historical texts. Syntax and semantics have always been less popular among the language historians. Pragmaphilology goes one step further and describes the contextual aspects of historical texts, including the addressers and addressees, their social and personal relationship, the physical and social setting of text production and text reception, and the goal(s) of the text".

> Sin fuentes primarias no es posible la creación histórica, el planteamiento de nuevos problemas /…/. Algunas de esas fuentes primarias constituyen para la Nueva Historia, canteras llenas de riquezas reveladoras y altamente indispensables a una ciencia que ha cambiado sus objetivos y que ha desarrollado nuevas técnicas, aptas para el tratamiento de sus materiales y de su documentación.

Este ámbito propiamente filológico se ve actualizado y complementado en su metodología por medio de la *lingüística del corpus*, de larga data, pero que en tiempo reciente se ha redefinido y ha irrumpido en el ámbito hispánico de la investigación histórica, con sus aportes específicos en la constitución de corpus diacrónicos. Con anterioridad, en Finlandia, la Universidad de Helsinki fue una de las pioneras en el desarrollo de lo que podría llamarse un 'megacorpus' diacrónico, cuando en 1987 ya Ossi Ihalainen, Merja Kytö y Matti Rissanen (1987) daban cuenta de los avances del *Helsinki Corpus of English Texts*; el proyecto se desarrolló entre 1984 y 1991, contando con un corpus total de 1.572.800 palabras, con aproximadamente 450 textos (de entre 5.000 y 10.000 palabras) que abarcaron el milenio comprendido entre 710 y 1710.

A partir de este momento no sólo se generaliza el término lingüística del corpus, sino que hay un nuevo impulso a esta disciplina, producto del auge de la lingüística computacional, la mayor disponibilidad de corpus electrónicos y el desarrollo de tecnologías de reconocimiento óptico de caracteres (OCR) que facilita el ingreso de textos en las bases de datos. De hecho, el trabajo con corpus informatizados v/s el realizado con corpus impreso (en libros, esencialmente) permite otorgar una serie de ventajas que hasta ahora no habían sido consideradas en el ámbito de la historia del español de Chile.

Evidentemente, las más claras dicen relación con la rapidez, consistencia y precisión en el procesamiento de las bases de datos, así como la facilidad para el acceso y manipulación de los textos. Asimismo, una vez que se disponga de un corpus marcado, permitirá la automatización de largas labores de análisis gramatical y sintáctico, especialmente, que antes debían efectuarse manualmente. Además, sin dejar de lado el análisis cualitativo de los datos, permite introducir un factor cuantitativo en los estudios de diacronía de la lengua vinculados con los factores incidentes en la variación de la misma, a partir de la disponibilidad de numerosos recursos y aplicaciones que permiten correlacionar una gran cantidad de datos.

### 1.1. Qué es, qué no es y qué aspira a ser nuestro corpus

En cuanto a sus características constitutivas, la tipología de los corpus se estructura de acuerdo a los siguientes criterios de ordenamiento y definiciones (STEL 2007):

**a) Según la modalidad de la lengua:** *Textuales* o escritos y *orales* o muestras de lengua hablada.

**b) Según el número de lenguas:** *Monolingüe*, el que da cuenta de una lengua o variedad lingüística y *bilingües* o *multilingües*, como muestra dos o más lenguas, cuyos corpus pueden ser comparables[2], paralelos *o* alineados.

**c) Según la cantidad y distribución de los textos:** *Grandes*, esto es, que no tienen límites de palabras o es muy elevado, no siguen criterios de representatividad o equilibrio; *equilibrados*, es decir, cuentan con la misma proporción de diferentes tipos de textos; *piramidales*, sus textos están distribuidos en estratos o niveles, de manera que en cada nivel hay más variedad y menos textos; *cerrados*, los cuales están constituidos por un número predeterminado de palabras y *abiertos* o *monitor*, en constante crecimiento, gracias a la introducción periódica de nuevos textos según proporciones definidas, incluyendo, por lo general, textos completos en vez de muestras.

**d) Según la especificidad de los textos:** *Generales*, que pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible; *especializados*, cuyos textos pueden aportar datos para la descripción de un tipo particular de lengua; *genéricos* o textos pertenecientes a un único género, cuyo objetivo es caracterizar dicho género; *canónicos* que representan la obra completa de un autor; *cronológicos* o textos de una época concreta que tienen por objetivo estudiar la lengua producida durante ese período; *diacrónicos* o *históricos* cuyos textos de diferentes etapas temporales permiten observar la evolución de la lengua y *sincrónicos*, en los que su finalidad es estudiar una o más variedades lingüísticas en un momento determinado, generalmente para establecer comparaciones entre variedades o dialectos.

**e) Según la representatividad del corpus:** *Textuales*, esto es, formados por textos enteros; *de referencia*, formados por fragmentos, porque interesa más el nivel de lengua, el equilibrio y la representatividad que el texto en sí; *léxicos* o *sample corpus*, formados por fragmentos muy breves de textos, de una longitud constante.

**f) Según el proceso al que se someta el corpus:** *Simples*, de datos brutos, no anotados, no codificados, es decir, son textos guardados sin formato alguno y sin añadir ninguna información adicional; *verticales* que son el resultado de disponer en forma de columna las palabras de un texto ordenadas según criterios alfabéticos o frecuenciales; *codificados* o *anotados*, formados por textos a los que se han añadido, de forma manual o automática, determinadas informaciones referidas a la estructura de textos o aspectos puramente lingüísticos. Estos últimos pueden ser

[2] *Corpus comparables*: textos similares en cuanto a sus características y criterios de selección que se usan para comparar variedades de la lengua; *corpus paralelos*: es el mismo texto traducido a una o más lenguas, usado en traducción automática y en entornos bilingües o multilingües; y *corpus alineados*: son aquellos corpus paralelos en los que, para facilitar su explotación, los textos están dispuestos uno al lado de otro en párrafos o frases, de manera que sea fácil extraer las equivalencias (se usan como entrenamiento en traducción automática).

"analizados morfológicamente", "parentizados" o "treebanks"[3].

Según la clasificación y definiciones precedentes el CorDECh, al menos en una primera etapa, se conformará como un corpus de documentos de carácter ***monolingüe, grande, diacrónico, textual*** y ***simple***. Será un corpus simple, por cuanto la etapa de trabajo que se ha fijado en este momento apunta a la constitución de un volumen importante de textos, a fin de hacer 'visible' una parte fundamental del período histórico de los siglos XVI y XVII, donde se pondrá especial énfasis, dado que es la etapa más compleja de leer en la documentación, debido al tipo de letra que la caracteriza como es la procesal encadenada. Para estimar una cifra, se supone que el CorDECh, bajo las condiciones que luego señalaremos, estaría conformado por alrededor de 60.000 imágenes (1 imagen = 1 página de texto).

En una segunda etapa se pretende conformar un corpus *codificado* y *anotado*. Esto es, se aspira a que el CorDECh reciba tanto marcas lingüísticas como textuales, a fin de que pueda ser utilizado de la manera más provechosa posible por investigadores de áreas disciplinarias afines, para lo cual se cuenta con la asesoría del equipo de trabajo del Corpus Diacrónico del Español (CORDE) de la RAE.

### 1.2. Corpus diacrónicos hispánicos

En el ámbito de los estudios diacrónicos del español, claramente en Latinoamérica existe un trabajo por realizar, sobre todo a partir de la necesidad, ya formulada por Guillermo Guitarte en los años 60[4] de recurrir a los archivos para trazar la historia del español americano. Esto no quiere decir que en el ámbito hispánico no existan corpus diacrónicos de larga data ya. Sólo por citar algunos casos, se mencionan los más amplios:

***CORDE:*** Real Academia Española (Corpus Diacrónico del Español), http://

[3] *Analizados morfológicamente*: anotados con información morfológica (categorías morfosintácticas, con mayor o menor detalle); *parentizados*: anotados con información sintáctica superficial, marcada con paréntesis o corchetes; *analizados* o *treebanks*: el texto está procesado sintácticamente de manera completa, con un análisis exhaustivo.

[4] Guillermo Guitarte señala: "Me refiero al estudio de la lengua en los documentos guardados en los archivos de América. Poco se ha hecho a este respecto, y debemos confesar que, en este punto, la filología se halla en retraso frente a la historia, que hace tiempo utiliza las fuentes documentales para sus investigaciones. Para trazar la historia del español de América, en cambio, las fuentes han sido sobre todo los cronistas e historiadores de Indias; de manera adicional se han empleado textos literarios y gramáticas y, ocasionalmente, algunos documentos. Esto fue lo que ya hizo Cuervo, pero si en su tiempo era un procedimiento aceptable y representó un gran progreso, creo que hoy podemos ampliar la base de nuestros conocimientos" (1968: 159). En 1992, Alvar señala: "Porque el conocimiento de la realidad americana desde comienzos del siglo XVI tendrá que hacerse por transcripción rigurosa de los textos y el conocimiento de la procedencia de los colonizadores. El primer motivo debe llevarnos a la formación de unas colecciones documentales tan rigurosas en sus lecturas como las que en su día hicieron Staaff, Menéndez Pidal y Navarro Tomás con referencia a los dialectos peninsulares".

www.rae.es. Cuenta con cerca de 300 millones de palabras (España: 79,16%; Hispanoamérica: 19,48%; Otras zonas: 1,36%). Inicios del español hasta 1975.

***ARTHUS:*** Universidad de Santiago de Compostela (Archivo de Textos Hispánicos de la Universidad de Santiago), http://gramatica.usc.es/EspWelcome.html. Contiene corpus textuales y orales de diferentes épocas de la historia española en España e Hispanoamérica. Está etiquetado sintácticamente.

***Base de dades morfosintàctica de l'espanyol medieval:*** Grupo de Lexicografía y Diacronía (Seminario de Filología e Informática, Universidad Autónoma de Barcelona), http://seneca.uab.es/sfi. Corpus fragmentario pero representativo del español medieval, con características diastráticas y diatópicas. Su finalidad es estudiar los cambios gramaticales del español preclásico.

***Corpus del Español:*** Prof. Mark Davies (Brigham Young University), http://www.corpusdelespanol.org/. Es un corpus de 100 millones de palabras (20 millones entre 1200-1400 y 40 millones para cada uno de los períodos entre 1500-1700 y 1800-1900).

Junto con ellos existen iniciativas relacionadas con la constitución de diccionarios históricos, que también tienen como base un determinado corpus diacrónico, aunque no siempre éste sea público o conocido. Por ejemplo, el *Léxico Hispanoamericano* de Peter Boyd-Bowman es una obra de gran ayuda para este ámbito, pero de la cual se carece de acceso real a sus fuentes. En los últimos tiempos, en todo caso, muchos de estos proyectos han estado publicando referencias parciales o algunas de sus fuentes, con el fin de que lleguen a una mayor cantidad de investigadores e interesados en la materia. En tal sentido se encuentran los *Documentos para la Historia Lingüística de Hispanoamérica* (ALFAL-RAE, 1993), los *Documentos Lingüísticos de la Nueva España* (Company, 1994) o el volumen *Ilegibilidad y Cotidianidad* que recoge 102 documentos chilenos del período colonial.

### 1.2.1. El *CHEM*

En el ámbito hispanoamericano destaca la implementación del ***CHEM*** (Corpus Histórico del Español de México) a cargo de la UNAM. Este corpus se ha puesto en marcha a partir del 2005, mediante un proyecto coordinado por el Instituto de Ingeniería[5]. Considera la adición de materiales aportados por lingüistas, filólogos e historiadores, estimados fundamentales para la representación del español de México entre los siglos XVI y XIX. La primera versión de este proyecto incluye la informatización de los *Documentos Lingüísticos de la Nueva España* (1994), 320 textos transcritos por Concepción Company, que abarcan todo el período colo-

[5] Incluye la participación de la UNAM (a través del Instituto de Ingeniería, Instituto de Investigaciones Filológicas, Facultad de Ingeniería y Facultad de Filosofía y Letras), del Instituto Mora y de la Brigham Young University. El patrocinio es de la DGAPA (Dirección General de Apoyo al Personal Académico) por medio del proyecto IN400905, 2005-2007.

nial, desde 1525 a 1816. El corpus contempla calas históricas[6], una delimitación geográfica a la zona del Altiplano Central, así como criterios temáticos, donde se ha tratado de privilegiar el carácter coloquial de los textos. La autora recoge *cartas* –de emigrantes, misioneros, al Rey, al Consejo y, de menor frecuencia, las más coloquiales–, *denuncias y testimonios en juicios* –particularmente los más "populares", como aquellos por asesinato, brujería, blasfemia, ultrajes y despojos, además de juicios de residencia–, *inventarios y testamentos* –particularmente valiosos para el estudio léxico de los ámbitos de la vida cotidiana– y *peticiones e informes* –por la presencia de diversas copias y cómo aquello podría reflejar etapas del cambio fonológico, además de que en los primeros, sobre todo, puede encontrarse el escaso material relacionado con el español aprendido por indígenas–.

Por otra parte, Company incluye un criterio de "origen del autor", esto es, descontando el período en que la Nueva España es territorio poblado exclusivamente por españoles (los primeros 50-60 años), procura que el autor del documento haya nacido en suelo mexicano. Junto con ello, trata de incluir una representación "racial y social" que dé cuenta de la conformación de la sociedad mexicana de la Colonia (castizos, criollos, españoles, indios, mestizos, mulatos, negros, portugueses), aunque también incluye una buena cantidad de documentos de cuyo autor no se tiene referencia (44,7%).

Sin dejar de resaltar la importancia de este trabajo, no es aún un corpus marcado, sino que, tal como el que se propone, es un corpus *simple*.

### 1.2.2. El *CorDECh*

En este marco es que se plantea desarrollar el ***Corpus Diacrónico del Español de Chile*** (*CorDECh*) que se inscribe dentro del marco más general del CORDE. De todas maneras es necesario considerar que la constitución de un corpus de esta naturaleza presenta no pocas implicancias teóricas y metodológicas. Esto, unido a la falta de experiencias sistemáticas, desde la perspectiva de la constitución de un corpus diacrónico para el español de Chile, hace que la tarea deba abordarse con especial cuidado para iniciarla. Es cierto que ya estaba disponible el capítulo referido a "Santiago de Chile", en la colección de *Documentos para la historia lingüística de Hispanoamérica* (1993), coordinada por Mª Beatriz Fontanella de Weinberg, pero esta iniciativa no prosperó como punto de partida del trabajo permanente de un equipo y quedó como un capítulo de libro publicado como anejo del BRAE LIII (47 textos, desde 1565 a 1795, entre las páginas 163-260).

[6] Señala Company (1994) al respecto que "los materiales fueron seleccionados eligiendo un grupo considerable de documentos cada cincuenta años aproximadamente, en el supuesto de que en ese lapso, unas dos generaciones, los cambios pueden hacerse más fácilmente perceptibles en lengua escrita. Para cada corte cronológico consideré un margen de flexibilidad de búsqueda de unos quince años, y así se establecieron siete etapas que corresponden a seis cortes cronológicos, aproximadamente: 1525-1540, 1570-1585, 1620-1635, 1680-1695, 1735-1750, 1780-1795 y 1805-1820".

## 2. FIJACIÓN DEL CORPUS

En esta primera etapa de conformación del CorDECh se ha fijado para éste algunos criterios en la recolección documental, de acuerdo a lo ya señalado con anterioridad, esto es, que nuestro corpus será *monolingüe*, *grande*, *diacrónico*, *textual* y *simple*. Dichos criterios no son arbitrarios, sino que se fundamentan en la experiencia anterior acumulada de equipos de trabajo europeos e hispanoamericanos, los que se señalan a continuación.

### 2.1. Criterio cronológico

Al tratarse de un corpus *diacrónico* este factor resulta casi redundante, por cuanto lo que se pretende con esta iniciativa es dar cuenta de los cambios lingüísticos en el período que comprende los tres siglos coloniales. El corpus se inicia, hasta este momento, con un texto de Luis de Cartagena, de 25 de septiembre de 1548 y finaliza con uno de Agustín Díaz, fechado el 13 de febrero de 1798. La extensión de los documentos es desigual, desde un folio hasta diez o doce. La distribución de documentos, de acuerdo al período tratado, es la siguiente:

**Siglo XVI** (1548-1599): 51 documentos
**Siglo XVII:** (1600-1650) 50 documentos – (1651-1699) 38 documentos.
**Siglo XVIII:** (1700-1750) 46 documentos – (1751-1799) 36 documentos

### 2.2. Criterio de selección y transcripción del corpus

#### 2.2.1. Selección

En este caso, se trata de un corpus cuya principal variable de selección documental es el tiempo. Teniendo esto presente, se ingresan al menos dos criterios adicionales en la selección de volúmenes, en relación con la difusión y los beneficiarios / audiencia: *pertinencia* y *relevancia*. ***Pertinencia***, entendida como la adscripción cronológica del documento al período en estudio (Colonia, siglos XVI al XVIII), siguiendo las pautas de constitución antes descritas y ***relevancia***, como criterio cuantitativo, esto es, se seleccionarán los volúmenes que contengan el mayor porcentaje de documentos *pertinentes*, de acuerdo con los catálogos de los Fondos Históricos. El tamaño final de la muestra pretende abarcar, en un primer momento, 80 volúmenes de dichos fondos, con un promedio de 400 folios, recto y verso, por volumen.

La digitalización del corpus se efectuará por fotograma, a razón de 600 dpi, con un zoom de 900%, lo que entrega un promedio de "peso" de imagen de 5 Mb,

en formato *tiff*. La imagen se escaneará como "película positiva b/n", para que una vez digitalizado el fotograma se invierta la imagen a "negativo" y se visualice como 'positiva', dado que al hacerlo directamente como "película negativa b/n" se pierde parte importante del documento. Una vez digitalizadas, las imágenes se procesarán con Photoshop, a fin de corregir brillo y contraste, fundamentalmente, para visualizarlas con óptimas características para su lectura y transcripción. Los volúmenes serán transcritos y archivados en un DVD, así como en disco duro, debido a la permanencia del material digitalizado, cuya durabilidad es mayor en este tipo de unidad que en el soporte portátil. Se calcula que, como mínimo, los archivos digitales de los 80 volúmenes ocuparían un espacio de alrededor de 320 GB.

### 2.2.2. Transcripción

El objetivo de la transcripción paleográfica, según Cencetti (1978), se puede resumir en que: "Studio del loro contenuto e, su un piano più ampio, alla storia della cultura in genere. Il suo studio comprende pertanto: /…/quello della storia della scritura alfabetica (paleografia in senso stretto); quello dei segni accesori della scritura alfabetica (interpunzione, numerali, segni ortografici e critici, ecc.) /…/.".

Desde el punto de vista metodológico, se han empleado tres formas para la transcripción paleográfica documental: las vinculadas con la historia, la filología y la lingüística. La primera, dice relación con el método empleado por historiadores para transmitir el acervo documental, donde se moderniza todo el documento (grafías, puntuación, léxico, etc.). El caso más representativo es, quizá, la *Colección de Documentos Inéditos para la Historia de Chile*, de José Toribio Medina. Respecto de las normas de transcripción paleográfica, de acuerdo con esta óptica, dice Jara (1996) que se abocan a "modernizar la ortografía, pero conservando el sonido original, para mantener el sabor arcaico. En consecuencia, esta publicación no es útil para filólogos, salvo en lo que respecta a saber el lugar en que está ubicado el documento. Es una advertencia que debe ser tenida en cuenta".

En el segundo caso se responde, como señala Kordić (2005), a "las normas textológicas de la serie Biblioteca Antigua Chilena (BACh)", establecidas por Mario Ferreccio, que han sido pilar de numerosas ediciones críticas generadas por el Seminario de Filología Hispánica de la Universidad de Chile, como por ejemplo la edición del *Cautiverio Feliz* de Pineda y Bascuñán. Estas normas las sintetiza, más adelante, como sigue:

> La reducción fonografemática realizada contempla el principio básico del respeto y conservación de todo rasgo gráfico que implique efectiva o eventualmente la representación de un rasgo fónico diferencial, significativo; todo aquel recurso que, tras el examen del comportamiento gráfico del escriba, demuestre ser inoperante, se moderniza, con el fin de evitar en

> el texto editado la presencia de inútiles grafías exóticas que sobrecarguen visualmente el texto y confundan al lector

En este sentido se simplifican las geminadas *ff*, *ss*, *rr*; se simplifican usos como *qu-cu* (quales > cuales), uso de nasales según uso actual (*enpezar* > *empezar*), restitución de vocal protética (*scribano* > *escribano*); se aplican normas actuales de puntuación, acentuación y uso de mayúsculas y minúsculas.

Finalmente, con el tercer criterio, se pretende entregar un reflejo lo más fiel posible del texto, ya que como señala Alvar y Alvar (1981):

> La edición paleográfica de un texto tiene sus propias peculiaridades: trata de hacer asequible con signos actuales lo que resultaría de otro modo de penosa o imposible lectura para quien no tenga cierto tipo de conocimientos. Pero, por otra parte, trata de presentar ese material de la manera más fiel con respecto al original que transcribe. No es –como se ha dicho erróneamente– algo que pueda suplir a la fotografía, sino lo que la fotografía no puede dar: la sencillez, sin transgredir en nada de lo que consta en el original.

La mayor parte de este corpus tiene un carácter inédito, ya que los textos se han transcrito directamente de los archivos, o bien, teniendo a mano una fotocopia o edición digitalizada de los mismos. En algunos casos, que se especifican, se han incorporado los textos que forman parte de los *Documentos para la Historia Lingüística de Hispanoamérica*, los cuales han sido editados teniendo a la vista la fotocopia de los originales, desplegando las abreviaturas y, en algunas ocasiones, corrigiendo las transcripciones efectuadas en su momento.

Para este trabajo se han adoptado los criterios del grupo de trabajo de ALFAL; las transcripciones son de carácter literal estricto, ajustándose a las normas siguientes:

1. Se respetará en todo la grafía original del texto.
2. Sólo se apartará del mismo en cuanto contemplará la separación gramatical de las palabras.
3. Esto regirá no sólo en cuanto se separarán las palabras unidas, sino también en cuanto se unirán las letras de una palabra que estén separadas.
4. Se conservarán las abreviaturas.
5. La *s* larga y la *s* de doble curva (redonda) se transcribirán con s redonda (la usual en la grafía moderna).
6. La *c* con cedilla se transcribirá literalmente: ç.
7. Se respetará el uso de *i* e *y*, ya sea como vocales o consonantes según el texto original: myll, maior, vyo.
8. La *r* mayúscula con valor fonético de *rr*, se mantendrá: Río, Rodrigo, coRe.

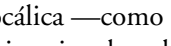
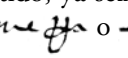
9. Se respetará la duplicación de letras: cossa, cappitan, ottra, ffecha.
10. Se conservarán las contracciones: del, della, desta, ques.
11. Se respetará el signo copulativo *t.*
12. Se conservará la puntuación del original.
13. Se conservará el uso de mayúsculas y minúsculas del original.
14. Se respetará la acentuación (o su ausencia) del original.

En estricto rigor, la transcripción que se sigue en el CorDECh se aparta de estos parámetros en los siguientes casos:

a) cuando hay contracciones donde una letra es utilizada para indicar dos términos, sin que el final de una sea el inicio de la otra, se realiza la separación (v.gr.: êl "e***n e***l", D$^{or}$on "D***oct***or ***D***on");

b) las abreviaturas se despliegan para facilitar la lectura del texto, lo cual se realiza, cuando es posible verificarlo, conforme a las pautas internas del texto o del autor. De no ser así, según la forma actualmente aceptada. El desarrollo se indica con letra cursiva, mientras que el texto original se mantiene con letra común.

c) Se conserva la ese alta (ſ) y la sigmática (σ) para adecuarse a las pautas internas del grafismo de los autores y para acercar la transcripción paleográfica al reflejo más cercano del documento. En una edición crítica, evidentemente, este criterio debiera modificarse.

d) Los casos de R mayúscula en posición interna y en posición inicial, siempre que no correspondan a nombre propio, se transcriben como ***rr***. Es un problema frecuente en las transcripciones paleográficas la interpretación de la grafía correspondiente a la vibrante múltiple, ya que ésta puede figurar con carácter simple mayúsculo, incluso en posición intervocálica —como en  — o con un tipo invertido, ya sea inicial o interior de palabra, como en los casos de  ,  o  .

Por el momento, el método de transcripción ha sido paleográfico estricto, con despliegue de las abreviaturas, ya que el carácter que ha tomado este trabajo, marcadamente individual, no ha dejado tiempo como para realizar una edición crítica, filológica, de estas transcripciones, lo que permitiría indicar al lector común las pausas, oraciones y párrafos contenidos, actualizando las normas de puntuación y ortografía. Carrera y Herrán (2006) señalan, además, al respecto que: "Este último modo de transcribir, sin entorpecer posibles trabajos de grafémica o fonético-fonológicos, facilita los análisis lingüísticos de niveles como el sintáctico o pragmalingüístico. Y también se muestra más útil para la futura edición electrónica...". Desde este punto de vista, las observaciones hechas por Kordić (2005) son más que pertinentes y deben ser las que se adopten en un futuro trabajo de edición filológica, crítica, de los documentos.

Se muestra, en la Tabla I, las dos versiones de un mismo texto, paleográfica y filológica, para que se comprenda mejor lo expresado.

**Tabla I**: Transcripciones documentales (ANS, Escribanos de Santiago 8, foja 34v; 13 de abril de 1592).

| Transcripción Paleográfica | Transcripción Filológica |
|---|---|
| ✠<br>e*n* la muy noble y leal çiudad de ſant*iag*o rrey**[no]** / [2] de chille A treçe dias del mes de abril de mjl**[l]** / [3] y qui*nient*os y nouenta y dos a*ñ*os ante mj gines de t**[oro]** / [4] maçote Escriu*an*o rr*ea*l pu*blic*o y de cab*ild*o de Eſta d*ic*ha çiud**[ad]** / [5] y de los t*estig*oσ aqui contenidos pareçio pres*en*te don / [6] fran*çisc*o de gaete v*ezi*no de osorno. rresidente e*neσ*ta d*ic*ha / [7] çiudad Enfermo e*n* vna cama y dixo que por / [8] quanto. El a fe*c*ho y otorgado su testam*en*to ant**[e]** / [9] mi El pres*en*te Escriu*an*o. e*ne*l qual dejo nonbrado sepol / [10] tura albaçea y Eredero. y quiere mudar la sep**[ol]** / [11] tura que dejando e*n* ſu fuerça y bigor El tes**[ta]** / [12] m*en*to que tiene otorgado saluo e*n* q*uan*to a lo que / [13] toca a la sepoltura su boluntad Es de / [14] que le Entierren e*ne*l monest*eri*o. de s*eñ*or sant**[o]** / [15] domjngo de Esta çiudad por los frajles del / [16] e*n* la capilla del general ju*an* jufre su suegr**[o]** / [17] y se le diga la misa cantada y bígília po**[r]** / [18] los frailes del d*ic*ho conbento. y con esto como / [19] d*ic*ho Es En todo lo demas deja El d*ic*ho teσta / [20] mento) e*n* ſu fuerça y bigor y asi lo dixo y / [21] otorgo siendo presentes por t*estig*oσ pablo flor**[es]** / [22] y alonso gonçalez de medina y antoni**[o]** / [23] morales de albornoz y El otorg*an*te desta / [24] a quien doy fee que conozco no firmo por no / [25] poder por la grauedad de ſu Enfermedad / [26] rrogo al d*ic*ho pablo flores t*eσtigo* lo firme por e**[l]** / [27] de su n*ombr*e __ / [28] a rruego y por t*estig*o *pablo* / [29] *flore*ſ **[firmado]** / [30] paso ante mj gines de toro maçote **[firmado]** / [31] scriu*an*o rr*ea*l pu*blic*o y de cab*ild*o | ✠<br>En la muy noble y leal ciudad de Santiago, Reino de Chile, a trece días del mes de abril de mil y quinientos y noventa y dos años, ante mí, Ginés de Toro Mazote, escribano real, público y de cabildo de esta dicha ciudad, y de los testigos aquí contenidos, pareció presente Don Francisco de Gaete, vecino de Osorno, residente en esta dicha ciudad, enfermo en una cama, y dijo que por cuanto él ha fecho y otorgado su testamento ante mí, el presente escribano, en el cual dejó nombrado sepoltura, albacea y heredero, y quiere mudar la sepoltura.<br>Que, dejando en su fuerza y vigor el testamento que tiene otorgado, salvo en cuanto a lo que toca a la sepoltura, su voluntad es de que le entierren en el monesterio de Señor Santo Domingo de esta ciudad, por los frailes de él, en la capilla del General Juan Jufré, su suegro, y se le diga la misa cantada y vigilia por los frailes del dicho convento.<br>Y con esto como dicho es, en todo lo demás deja el dicho testamento en su fuerza y vigor, y así lo dijo y otorgó, siendo presentes por testigos: Pablo Flores y Alonso González de Medina y Antonio Morales de Albornoz. Y el otorgante de ésta, a quien doy fe que conozco, no firmó por no poder, por la gravedad de su enfermedad. Rogó al dicho Pablo Flores, testigo, lo firme por él, de su nombre.<br>A ruego y por testigo, Pablo Flores **[firmado]**.<br>Pasó ante mí<br>Ginés de Toro Mazote **[firmado]**<br>escribano real, público y de cabildo. |

En todo caso, la base del CorDECh será, de todas formas, la transcripción paleográfica estricta, dado que este tipo de transcripción sirve de sustento para cualquier tipo de estudio lingüístico, histórico o sincrónico, permite la proyección hacia diccionarios históricos, además de servir como fuente para la difusión del material transcrito, reeditándolo para alcanzar públicos diversos. Por otra parte, la digitalización del documento presenta invaluables ventajas a la hora de transcribir y de conformar los 'abecedarios' internos de cada documento y de cada autor. Este trabajo puede tener proyecciones, inclusive, para el desarrollo de software de reconocimiento óptico de caracteres manuscritos, en la medida en que, habiendo recopilado un corpus documental amplio, se puedan establecer los rasgos fundamentales de cada grafema –así como de su conjunto de alógrafos–, independientemente del individuo que los traza, por medio de las teorías específicas dentro del campo del modelamiento matemático.

### 2.3. Criterio diatópico

En cuanto a la delimitación geográfica de los documentos, se atenderá a la denominación conocida como *Reino de Chile*, sobre la cual hay que hacer algunos alcances. Si bien la Corona había delimitado claramente sus posesiones en América, correspondiendo a este territorio la signatura de *Capitanía General de Chile*, sin embargo, quien escribe no se identifica con esta caracterización oficial, es más, en la primera mención que se encuentra en los documentos se hace referencia a estas tierras como los *reinos de la Nueva Extremadura*. Villalobos (1986) señala que la denominación de "reino" era característica de estas provincias de ultramar, careciendo de una connotación definida, jurídicamente hablando, y podía designar a cualquier territorio más o menos extenso que se encontrase delimitado administrativamente. En la práctica, se utilizó esta denominación para Nueva Granada, Quito y Chile.

La conformación primera de las cuatro gobernaciones en las que se repartiría el territorio americano fue una solución sin conocer el terreno, por lo que luego se privilegió las conquistas efectivas, fijándose los límites *de facto* más que siguiendo las cédulas otorgadas. Pedro de Valdivia, entonces, se encarga de fijar los márgenes del nuevo "reino" americano desde el valle de Copiapó (paralelo 27º) hasta las cercanías del canal de Chacao (paralelo 41º). Los límites hacia "lo ancho" conservaban el espíritu de las primeras capitulaciones, es decir, otorgaban a Chile cien leguas (unos 634 km), lo cual dejaba dentro de la jurisdicción territorial parte de Tucumán, Cuyo y la Patagonia. Debido al distinto desarrollo histórico y lingüístico que luego tomaron estas regiones, se considerará para este caso sólo los territorios "aquende los Andes".

### 2.4. Variedad de registros según variedad de textos

La discusión respecto de la categorización de los documentos, como *públicos* o *privados*, está hoy día aún vigente. Heredia (1985) establece una tipología bastante simple y práctica, en la que intervienen dos variables (emisor-destinatario y tipo de carta), para establecer las relaciones entre ellos, determinando los tipos de cartas que pueden encontrarse.

**Tabla II:** Tipología de cartas.

| Emisor-Destinatario | Tipo de Carta |
|---|---|
| Autoridad soberana-autoridades delegadas | Carta real |
| Autoridades delegadas-autoridad soberana | Carta oficial |
| Particular-autoridad constituida | Particular |
| Particular-particular | Privada |

Sin embargo, el ámbito de lo público o lo privado es más amplio. Duranti (1996), comentando precisamente la discusión que ha tenido en el ámbito de la diplomática la definición de estos conceptos, señala:

> La conclusión de todo esto es que la definición de la naturaleza de un documento que sea la más aceptable para propósitos diplomáticos debe colocar al documento en relación con su autor. De acuerdo con esto, *un documento es público si es creado por una persona pública o bajo su dirección o en su nombre*, es decir, si la voluntad que determina la creación del documento es pública por naturaleza. Una persona pública es una persona jurídica que desempeña funciones consideradas públicas por el sistema jurídico en el que la persona actúa y, mientras las desempeña, está revestido de algún poder soberano para ejercerla. Por contraste, *un documento es privado si es creado por una persona privada o bajo su dirección o en su nombre*; es decir por una persona que desempeña funciones consideradas privadas por el sistema jurídico en el que la persona actúa.

Considerando este factor, realizaremos una clasificación primera de los documentos, separando los ámbitos *públicos* de los *privados*. A partir de esta primera subdivisión, se establecerá la distinción categorial de los textos, de acuerdo a referencias tipológicas de los géneros en los cuales pueden ser adscritos. En este sentido no es posible realizar una taxonomía *a priori*, ya que dependerá de la constitución del corpus y de la variedad estilística que se recoja en estos textos indianos; sin embargo, hasta ahora la predominancia de la *carta* como tipo textual es lo que se muestra en los materiales recogidos en diversas instancias de compilación. No ha-

brá, en el CorDECh, una selección previa de acuerdo al tipo de documento, sino más bien una descripción y análisis de los tipos hallados, en virtud de que se trata de conformar un corpus *diacrónico textual grande*, lo que implica que se recogerán textos completos, de longitud diversa y estilos varios, de todo el período colonial.

## 3. ASPECTOS DE LA AUTOMATIZACIÓN DEL CORPUS

Este apartado tratará de algunos referentes vinculados con la automatización de corpus y, específicamente, en lo relacionado con la lexicografía diacrónica.

De manera general, es posible señalar que el uso más inmediato de cualquier corpus está en el plano de la *lexicografía*, como elemento determinante de los artículos que deben hallarse en el repertorio léxico, acepciones, construcción y régimen que debe considerarse, entre otras posibilidades. Lo anterior conlleva un amplio repertorio en análisis léxico y semántico en cuanto a formación de palabras, cambios de significado, evolución del léxico, por citar algunos. En este sentido, López Morales (2005) señala la importancia de contar con un Diccionario Histórico de la Lengua Española, "instrumento lexicográfico tan característico de toda gran lengua de cultura". La *lingüística computacional* también se beneficia de la formación de corpus, especialmente para la construcción de 'diccionarios-máquina', diccionarios electrónicos, tratamiento automático del lenguaje, por ejemplo. Asimismo, y relacionada con la anterior, la *estadística lingüística* puede establecer índices de frecuencia de formas, determinando tendencias de uso, ámbitos (general, técnico, neologismo, etc.) y, evidentemente, para obtener información sobre diversos niveles y estilos de lengua, de acuerdo con los parámetros de la sociolingüística y el análisis del discurso. Por último, es posible obtener materiales valiosos tanto para la *enseñanza de la lengua*, en la medida en que estos corpus sirvan para la confección de gramáticas de las lenguas descritas, en los niveles correspondientes (morfosintáctico, léxico-semántico y pragmático), como para el ámbito de las *industrias de la lengua*, donde cada día más se requiere de aplicaciones lingüísticas sustentadas en estos corpus[7].

### 3.1. Antecedentes en el ámbito hispanoamericano

El primer gran trabajo relacionado con la lexicografía histórica ha sido sin duda el

[7] Kytö y Rissanen (1997) señalan: "A historical linguist must therefore rely on a corpus, either in the old sense of the word, that is a text or a collection of texts yielding linguistic evidence on the phenomenon studied, or in the new sense of the word, that is the computerized version of the same. And if (s)he wishes to avoid the study of first-hand textual evidence, (s)he has to rely on another scholar's earlier corpus work, or on some more refined outcome of it, such as dictionaries or concordances".

de Peter Boyd-Bowman, quien a través del *Hispanic Seminary of Medieval Studies*[8] produjo una copiosa recopilación de información que se materializó en el *Léxico Hispanoamericano*, el cual abarcó todo el período colonial. Como formalización en un grupo de estudio más amplio, se discutieron las bases de una investigación en lexicología histórica en el seno del *Proyecto de Estudio Histórico del Español de América, Canarias y Andalucía*, de la ALFAL, coordinado en primera instancia por Mª Beatriz Fontanella de Weinberg, y luego, por Elena Rojas.

Este grupo de trabajo produjo dos compilaciones de textos coloniales, publicadas por la RAE (1993, 2000) las cuales constituyeron una primera fuente documental con criterios estandarizados de transcripción paleográfica. En un momento posterior, se elaboró una plantilla de recogida de datos léxicos, con el fin de poner en común la instancia de investigación diacrónica, visualizando la posibilidad de reunir en un solo conglomerado las investigaciones de los diversos países hispanoamericanos y de las respectivas regiones españolas.

Hasta el momento se tiene noticia de cuatro investigaciones sistemáticas al respecto, en este marco, la del español de México, a cargo de Chantal Melis y Concepción Company[9]; la de Canarias, que tiene como responsables a Dolores Corbella y Cristóbal Corrales[10]; la de Santo Domingo, al frente de Micaela Carrera de la Red[11] y la de Venezuela, recientemente, al frente de María J. Tejera[12]. Esto no quiere decir que no existan trabajos de lexicografía anteriores o posteriores a éstos, pero sí debe destacarse de ellos su base documental original, lo que contribuye a darle un carácter más fidedigno a los datos estudiados, así como a la publicación

[8] No hay que dejar de señalar, en todo caso, que las conexiones con proyectos traen, a su vez, otras iniciativas. De esta manera, relacionados con el Seminario de Madison, encontramos el *Diccionario del Español Medieval*, de la Universidad de Heidelberg y el *Diccionario Español de Textos Médicos Antiguos (DETEMA)*, coordinado por María Teresa Herrera, basado en 33 textos medievales, dedicados al estudio de la anatomía, higiene y patología humanas y conservados en bibliotecas españolas.

[9] *Léxico histórico del español de México. Régimen, clases funcionales, usos sintácticos, frecuencias y variación gráfica*, México, UNAM, 2 vols. [1998-2002]. Es un trabajo de 1.000 páginas, cuya base se encuentra en los 320 documentos publicados por Company (1994).

[10] Parte de estos trabajos se han visto reflejados en el *Diccionario Histórico del Español de Canarias (DHECan)*, Instituto de Estudios Canarios, La Laguna, 2001; 1.622 páginas.

[11] A partir del trabajo con Francisco José Zamora, coordinados por Germán de Granda, en la publicación de los documentos de *Santo Domingo*, 30 textos coloniales seleccionados y transcritos paleográficamente y de los documentos del *Reino de Nueva Granada*, en los ***DHLH***, ha desarrollado diversos proyectos y publicaciones en este ámbito. Actualmente, trabaja de manera conjunta con el Departamento de Lingüística de la Universidad de Los Andes (Mérida, Venezuela) en la tarea de recogida y análisis documental de la antigua Provincia de Mérida, así como en un proyecto de estudio de identidad y lengua en los Andes colombiano-venezolanos.

[12] El equipo de trabajo venezolano, a partir de la documentación publicada en los ***DHLH***, ha aumentado el corpus a 111 textos de los tres siglos coloniales (39 para el XVI, 34 en el XVII y 38 pertenecientes al siglo XVIII), correspondientes a las primeras ciudades fundadas en el territorio. La región andina estaba subrepresentada, lo cual se solucionó en parte con la incorporación del Dr. Enrique Obediente (Universidad de Los Andes) al equipo, quien publicó, en el 2003, 37 textos correspondientes a esa zona, entre 1564 y 1657.

de las fuentes de las cuales se extrae la información, lo que las hace verificables y contrastables[13].

## 4. EL ETIQUETADO ELECTRÓNICO DEL CORDECH

En la actualidad es indispensable pensar que, en una creciente sociedad globalizada, el intercambio de información se hace cada vez más necesario, por lo que es de necesidad insoslayable plantearse la codificación del material transcrito a fin de considerar dos criterios: la *intercambiabilidad* de manera independiente de los recursos tecnológicos (software o hardware) y la *conservación* de la información, es decir, que producto de esta codificación no haya pérdidas o cambios en los datos.

A fin de poder lograr ambas metas, se recurrirá al metalenguaje SGML, en versión del *Prontuario de Marcación SGML* que la Real Academia Española ha adaptado para la informatización del CORDE, ya que se consideran allí las marcas para las diversas tipologías diacrónicas de los textos que pudieran encontrarse, tanto en prosa como en verso.

Algunas marcas de este tipo son de carácter ***intratextual***: prosa ("la marca <p> señala, en prosa, el comienzo de un párrafo nuevo en el texto; los textos en prosa de los corpus se segmentan en unidades <s>, que se corresponden aproximadamente con los enunciados y con las oraciones ortográficas".), verso ("la marca <verse> se utilizará para iniciar una tirada de líneas dispuestas a modo de versos. Dichas líneas, que coincidirán casi siempre con lo que entendemos por "verso" irán precedidas de la marca <lb>, debemos cerrarlo al finalizar una tirada de versos; la marca <lb> se utiliza para señalar el comienzo de cada línea de una composición en verso"), números de página ("la marca <pb> indica el comienzo de una nueva página en la edición que se toma como fuente para el texto electrónico del corpus. Su atributo *n* tiene como valor el número de la nueva página. El elemento <pb> debe aparecer justo antes de la nueva página, no al final, independientemente del lugar donde se encuentre el número en la página de la edición original. La marca <pb> debe estar separada del texto precedente y siguiente por espacios en blanco. Si el cambio de página no coincide con salto de párrafo, la marca <pb> se deja en la misma línea"), marcación de texto resaltado ("se utiliza la marca <hi> para señalar fragmentos de texto resaltado, ya sea entrecomillado, en cursiva, en negrita, en versalitas, en mayúsculas o subrayado. La marca <hi> debe rodear el texto tipográficamente resaltado. La marca que indica el final de texto resaltado

[13] Sin duda que no puede dejar de mencionarse, aun cuando no pertenezca al ámbito "hispánico", el trabajo desarrollado en el marco del *Diccionario del Español Medieval*, del Instituto de Filología Romance de la Universidad de Heidelberg. Según los propios autores "es la primera clasificación sistemática del español medieval (que va del siglo X hasta principios del siglo XV), acompañada de un análisis lexicológico y lexicográfico, y de comentarios lingüísticos basados en la valoración de las fuentes".

es </hi>. Para señalar las diferencias entre tipos de resalte tipográfico se utiliza el atributo *rend* dentro del elemento <hi>"), marcación de las citas del texto ("el elemento <quote> se emplea para marcar citas. La finalidad de la incorporación de esta marca al esquema de codificación del CORDE es que no se mezclen, en la fase de recuperación de la información, fragmentos de texto de épocas o autores diferentes"), expresiones no castellanas ("la marca <foreign> sirve para diferenciar del resto del discurso las palabras, expresiones o frases en lengua no castellana, con la finalidad de constituir un subgrupo dentro del proceso de lematización que permita la recuperación de las palabras del corpus en otros idiomas"), glosas del texto ("el elemento <gloss> sirve para marcar las anotaciones que aparezcan en los márgenes de las ediciones de cita, ya sean del propio autor o de autoría diferente a la del texto principal"), cambios de autoría ("el elemento <change> sirve para marcar todo fragmento distinto de una cita que corresponda a una autoría diferente de la principal").

Otras marcas dicen relación con la ***estructura del corpus***: "La estructura global de cada corpus una vez codificado consta de dos grandes partes: el *prólogo* y el *corpus textual*, articuladas a su vez también de manera bipartita. El prólogo reúne todo un conjunto de declaraciones (declaración SGML, DTD) que informan sobre el tipo de etiquetas empleadas en la codificación del texto. El corpus, segunda parte del documento SGML, es el documento textual completo. No incluye ningún tipo de declaración, sino texto, codificación y referencias de entidad. Se divide en dos grandes partes: la cabecera del corpus y la serie de elementos <TEI.2>, esto es, textos codificados con sus cabeceras correspondientes"[14].

Esta es una propuesta que será adoptada para el trabajo con el CorDECh, ya que hasta el momento ha existido una concentración en la transcripción documental, por cuanto la constitución de un equipo de trabajo integral que aborde todos los aspectos del proyecto ha sido, cuando menos, dificultosa.

A continuación, en la Tabla III se observa la plantilla de la cabecera del proyecto CorDECh. Se ha utilizado el software *Cooktop 2.5*, que es un editor de XML/XSLT de distribución libre, desarrollado por Víctor Pavlov (http://xmlcooktop.com). Para verificar la sintaxis de la codificación, el programa cuenta con un validador que permite detectar las "partidas falsas" o aquellos elementos que no cuentan con un cierre en su descripción.

[14] SGML: Standard Generalized Markup Language; TEI: Text Encoding Initiative; DTD: Document Type Definition.

**Tabla III:** Cabecera del proyecto.

```
<?xml version=”1.0” encoding=”UTF-8”?>
<cooktop 2.5 RNGSCHEMA=http://www.tei-c.org/release/xml/tei/custom/schema/
relaxng/tei_corpus.rng type=”xml”?>
<teiCorpus xmlns=http://www.tei-c.org/ns/1.0>
    <TEI xmlns=”http:// www.tei-c.org/ns/1.0>
        <teiHeader>
            <fileDesc>
                <titleStmt>
                    <title>CORPUS DIACRÓNICO DEL ESPAÑOL DE CHILE
(CorDECh)</title>
                    <author>Instituto de Lingüística y Literatura</author>
                    <sponsor>Universidad Austral de Chile</sponsor>
                    <principal>Manuel Contreras Seitz</principal>
                </titleStmt>
                <editionStmt>
                    <edition>Transcripción paleográfica de documentos digitalizados por
el Archivo Nacional de Chile</edition>
                    <source>Fondos Históricos del Archivo Nacional de Chile</source>
                </editionStmt>
            </fileDesc>
        </teiHeader>
    </TEI>
</teiCorpus>
```

La gramática o DTD (*Document Type Definition*) es un lenguaje mediante el cual se definen con precisión aquellos elementos que son necesarios en la elaboración de un documento o un grupo de documentos estructurados de manera similar. Estas “declaraciones de elementos” aportan el nombre oficial de las etiquetas que aparecerán dentro de los delimitadores (por ejemplo, <change>), y describen lo que cada elemento puede contener (modelo de contenido). Para este caso, la DTD considerará elementos como los que se muestran en la Tabla IV. Esta estructura corresponde a un texto de Luis de Cartagena, de 25 de septiembre de 1548, donde se otorga un traslado de la merced hecha por Pedro de Valdivia a Inés Suárez.

**Tabla IV:** Cabecera de un texto de 1548.

```
<TEI xml:id=”MyTextNumber1”>
    <teiHeader>
        <fileDesc>
            <subjectStmt>
                <subject sameAs=“NomeclColec”>Chile-16-1548</subject>
                <subject type=“resumen”>Traslado de merced de don Pedro de Valdivia a
Doña Inés Suárez</subject>
                <author xml:id=“autógrafa”>Luis de Cartagena>/author>
            </subjectStmt>
            <editionStmt>
            <history><p>Documento en el que se describen los trabajos pasados por Inés
Suárez en la batalla de Santiago, por cuyos méritos se le concede merced de tierras
e indios</p></history>
            <origin><p>Escrito en la ciudad de Santiago de Chile</p></origin>
            <doc>1548-09-25</doc>
            <recordHist>
                <source><p>Catalogado directamente desde el manuscrito original</p></
                source>
            </recordHist>
        </fileDesc>
    </teiHeader>
</TEI>
```

## 5. CONSIDERACIONES FINALES

Varias son las características que distinguen a este corpus de los que aquí se han mencionado y que, en definitiva, constituyen el aporte fundamental de éste.

En primer lugar, y lo que más destaca en una primera aproximación, es la extensión del corpus: 60.000 folios. Si se considera un promedio de 600 palabras por folio, fácilmente se alcanza la cifra de 36.000.000 de palabras para la totalidad de éste, es decir, una cantidad similar a todo el corpus hispanoamericano del CORDE.

En segundo lugar, el contar con dos versiones simultáneas: una transcripción paleográfica estricta, así como una edición filológica de los textos, permitirá abarcar una variedad de campos de investigación que hoy se veían limitados por las características de una u otra, así como difundir ampliamente una documentación que hoy se ve restringida al espacio académico.

El tercer aspecto destacable del CorDECh es, precisamente, su disponibilidad. No sólo se contará con la publicación escrita del corpus, sino que la digitalización documental y las respectivas transcripciones también podrán ser consultadas en

línea, siguiendo con esto la dirección que hoy tienen los archivos nacionales e internacionales[15].

Asimismo, las características de su digitalización no sólo permiten que los documentos sean legibles y manipulables a través de los softwares computacionales pertinentes, sino que posibilitará el trabajo en el desarrollo de un reconocedor óptico de caracteres manuscritos especializado para los textos hispánicos coloniales, lo que permitiría hablar de una verdadera 'revolución' en la disciplina paleográfica y en el ámbito de la filología hispánica. En este sentido, se tiene presente la investigación llevada a cabo, entre otros, por Li, Tan, Ding y Liu (2004), Bertolami, Zimmermann y Bunke (2006), Basu, Chaudhuri, Kundu, Nasipuri y Basu (2007), Tracy, Papaodysseus, Roussopoulos, Panagopoulos, Fragoulis, Dafi y Panagopoulos (2007), Zand, Naghsh Nilchi y Amirhassan Monadjemi (2008), Louloudis, Gatos, Pratikakis y Halatsis (2008) y Liu, Wu, Zha y Liu (2008).

Para finalizar, sólo dos cosas que señalar respecto del trabajo con el CorDECh, al menos en su primera etapa.

En primer término, los textos seleccionados tratarán de ser una muestra documental proporcionada del discurso informal, tanto oficial como no-oficial, pues sobre todo en este último tipo de discursos el redactor tiende a preocuparse más por el contenido de tales documentos y mucho menos de la forma. Dentro de los textos transcritos, es probable que la mayor parte de ellos se caracterice por su naturaleza más bien oficial que no-oficial, ya que la experiencia señala que siguen siendo muy pocos los encontrados que respondan al requisito de ser no-oficiales o privados, debido a que la documentación que da testimonio de las diversas actividades de la vida pública nacional sólo da cuenta de la vida privada y, junto con ello, del modo familiar o cotidiano de la expresión, cuando la emisión o recepción de ellos podía estar a cargo de algún personaje connotado o trascendente en el quehacer de la comunidad de la época. Sin embargo, para suplir estas posibles deficiencias, se hará acopio de documentos lo suficientemente valiosos, aun cuando no se trate de correspondencia privada para los fines de este estudio.

En segundo lugar, cabe señalar que, si bien es cierto el software que se utiliza para el análisis demuestra potencialidades en cuanto al volumen de corpus por analizar y el tiempo en que puede efectuarse el trabajo, es sólo un instrumento metodológico más. Debe tenerse presente que la marcación del corpus –lo cual entrega todas las posibilidades futuras para el estudio del mismo– debe ser rigurosa, con

[15] El *Archivo Nacional de Chile* y la *DIBAM* han desarrollado, en este sentido, el portal **Memoria Chilena**. Por otro lado, es posible hallar en el Portal de Archivos Españoles (PARES - http://pares.mcu.es/) una gran cantidad de documentos digitalizados de diversa procedencia. Asimismo, la Biblioteca Virtual del Patrimonio Bibliográfico (http://bvpb.mcu.es/es/estaticos/contenido.cmd?pagina=estaticos/presentacion) es un importante fondo digital de documentos hispanos que abarcan desde el siglo X al XXI, esto es, desde la *Concordia regularum patrum videlicet beati benedicti* (s.X) hasta *Del "Consolat de mar" al "Libro llamado Consulado de mar": aproximación histórica* (2003).

indicadores claros y estandarizados. Con esa finalidad es que se trabajará siguiendo los parámetros establecidos en el formato del CORDE. De todos modos, hay que tener en consideración algunas otras propuestas, como las señaladas por Carrera y Herrán (2006), en las que plantean algunos aspectos textuales y discursivos para tener en cuenta en un análisis pragmático del documento:

- <documento>: identificación del país, siglo y año.
- <definición diplomática>: adscripción textual del documento.
- <autor id>: identificación de quien escribe y su calidad de autógrafa o heterógrafa.
- <emisor id>: calidad de quien escribe o suscribe el documento.
- <receptor id>: destinatario del documento.
- <tratamiento>: fórmulas de tratamiento del texto dirigidas al destinatario.
- <dirección id>: fórmula de intitulación.
- <ámbito discursivo de producción>: carácter institucional/personal, público o privado del documento.
- <registro>: tipo de discurso (epistolar, jurídico, etc.).
- <modo discursivo>: narrativo, argumentativo, expositivo, etc.
- <acto comunicativo>: estructura del evento (dialógico/respuesta, etc.)
- <acto de habla perlocutivo>: marcación en el documento de esta estructura (v.gr.: "dar quenta a v.s. de todo lo que ha pasado").

## REFERENCIAS

Alvar, Manuel. 1992. "La investigación del español de América: Proyectos inmediatos". Ponencia plenaria. Sevilla: Actas del Congreso de la Lengua Española. [En línea]. Disponible en http://cvc.cervantes.es/obref/congresos/sevilla/plenarias/ponenc_alvar.htm. Consulta: 01/07/2007.

Alvar, Manuel y Elena Alvar. 1981. *Cancionero de Estúñiga*. Edición paleográfica. Zaragoza: Institución "Fernando el Católico".

Basu, S.; C. Chaudhuri, M. Kundu, M. Nasipuri y D.K. Basu. 2007. "Text line extraction from multi-skewed handwritten documents", en *Pattern Recognition,* 1825-1839.

Bertolami, Roman; Matthias Zimmermann y Horst Bunke. 2006. "Rejection strategies for offline handwritten text line recognition", en *Pattern Recognition Letters*, Elsevier; pp. 2005-2012.

Carrera de la Red, Micaela y Andrea Herrán Santiago. 2006. "Apuntes sobre la elaboración de un corpus electrónico de documentos del español de América", en Actas del XXV Simposio Internacional de la Sociedad Española de Lingüística, Milka Villayandre Llamazares (ed.), Universidad de León, Dpto. de Filología

Clásica; pp. 263-287. [En línea] Disponible en: http://www3.unileon.es/dp/dfh/SEL/actas.htm.

Cencetti, Giorgio. 1978. *Paleografía Latina*. Roma: Jouvence.

Company, Concepción. 1994. *Documentos Lingüísticos de la Nueva España. Altiplano-Central*. México: Instituto de Investigaciones Filológicas, Centro de Lingüística Hispánica, UNAM.

Duranti, Luciana. 1996. *Diplomática. Usos nuevos para una antigua ciencia*. Colección Biblioteca Archivística. Carmona: S&C ediciones.

Guitarte, Guillermo. 1968. "Para una historia del español de América basada en documentos: El seseo en el Nuevo Reino de Granada (1550-1650)", en Actas de la 5ª Asamblea Interuniversitaria de Filología y Literatura Hispánicas. Bahía Blanca, pp. 158-165.

Heredia Herrera, Antonia. 1985. *Recopilación de estudios de diplomática indiana*. Sevilla: Diputación Provincial.

Jara, Álvaro. 1996. *Protocolos de los Escribanos de Santiago*, transcripción paleográfica de Álvaro Jara y Rolando Mellafe, Centro de Investigaciones Diego Barros Arana, Archivo Nacional, Fuentes para el Estudio de la Colonia. Santiago: Ediciones DIBAM.

Jucker, Andreas y Andreas Jacob. 1995. "The historical perspective in pragmatics", en *Historical Pragmatics.* Andreas Jucker (ed.) Amsterdam: John Benjamins. pp. 3-33.

Kordić, Raïssa. 2005. *Testamentos coloniales chilenos*, estudio preliminar de Cedomil Goić, Biblioteca Indiana, Centro de Estudios Indianos (CEI). Madrid: Universidad de Navarra, Iberoamericana/Vervuert.

Kytö, Merja y Matti Rissanen. 1997. "Language analysis and diachronic corpora", en *Tracing the trail of time. Proceedings from the Second Diachronic Corpora Workshop*, Raymond Hickey, Merja Kytö, Ian Lancashire y Matti Rissanen (eds.). Amsterdam: Rodopi, pp. 9-22.

Li, Yuan-Xiang; Chew Lim Tan, Xiaoqing Ding y Changsong Liu. 2004. "Contextual post-processing based on the confusion matrix in offline handwritten Chinese script recognition", en *Pattern Recognition* 37, pp. 1901-1912.

Liu, Hong; Qi Wu, Hongbin Zha y Xueping Liu. 2008. "Skew detection for complex document images using robust borderlines in both text and non-text regions", en *Pattern Recognition Letters* 29, pp. 1893-1900.

López Morales, Humberto. 2005. "La actuación de las Academias en la historia del idioma", en *Historia de la lengua española*, 2ª ed. actualizada, Rafael Cano (ed.). Barcelona: Ariel.

Louloudis, G.; B. Gatos; I. Pratikakis y C. Halatsis. 2008. "Text line detection in handwritten documents", en *Pattern Recognition* 41, pp. 3758-3772.

Ossi Ihalainen, Merja Kytö y Matti Rissanen. 1987. "The Helsinki Corpus of English Texts: Diachronic and dialectal report on work in progress", en *Corpus*

*Linguistics and Beyond*, *Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, edited by Willem Meijs, Costerus, New Series, Volume 59. Amsterdam: Rodopi, pp. 21-32. Para mayor información puede consultarse: http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

Real Academia Española. 1993. Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII, BRAE, Anejo LIII, ALFAL, Comisión de Estudio Histórico del Español de América, Mª Beatriz Fontanella de Weinberg (comp.). Madrid: Espasa-Calpe.

Real Academia Española. 2001. *Corpus diacrónico del español. Prontuario de marcación SGML*, Madrid.

Servei de Tecnología Lingüística (STEL). 2007. *Introducción a la lingüística del corpus*. [En línea] Disponible en: http://www.ub.edu/stel/esp_support.htm. Consulta: 01/07/2007.

Tracy, S.V., C. Papaodysseus, P. Roussopoulos, M. Panagopoulos, D. Fragoulis, D. Dafi y Th. Panagopoulos. 2007. "Identifying hands on ancient athenian inscriptions: First steps towards a digital approach", en *Archaeometry* 49, Vol. 4, pp. 749-764.

Villalobos, Sergio. 1986. *Historia del pueblo chileno*. Santiago: Andrés Bello.

Zand, Mohsen; Ahmadreza Naghsh Nilchi y S. Amirhassan Monadjemi. 2008. "Recognition-based segmentation in Persian charachter recognition", en *Proceedings of World Academy of Science, Engineering and Technology* 28, pp. 183-187.