CorpusRedEs

Proyecto de creación y anotación de un corpus de comunicación mediada por ordenador en español[1]

Ana Pano Alamán°, Patricio Moya Muñoz*
°Università di Bologna, *Universitat Politècnica de València

This paper presents the project CorpusRedEs, which aims at building an annotated corpus of Computer-Mediated Communication in Spanish. The corpus will gather texts from different *cybergenres* or socio-technical *modes* of CMC, including the diatopic varieties of Spanish as well as several domains. The annotation of the macrostructure of texts is based on the TEI-XML standard adapted to CMC, in order to favor the interoperability between platforms and the easy recovery of data by users. In this sense, we suggest that the *posting* element considered in other projects for the segmentation of CMC interaction units, may be enriched with further elements and attributes used for the annotation of spoken language corpora, for an accurate description of the interactional dynamics that take place within these texts.

Keywords: Computer-Mediated Communication, corpus, TEI, XML, Spanish varieties.

1. Antecedentes y justificación

Actualmente, hay un creciente interés por la creación de corpus de *Comunicación mediada por ordenador* o *computadora* (CMO/CMC)[2] con estrictos estándares de anotación, que faciliten la interoperabilidad entre las

[1] Este artículo es fruto de la continua colaboración entre ambos autores, no obstante, los epígrafes 1, 4 y 5 han sido redactados por Ana Pano Alamán, mientras que 2 y 3 han corrido a cargo de Patricio Moya Muñoz. El trabajo ha sido parcialmente financiado por el Programa de Formación de Capital Humano Avanzado del CONICYT - Ministerio de Educación de Chile.

[2] Conjunto de modalidades de interacción surgidas de la aplicación de las nuevas tecnologías a la comunicación pública e interpersonal. El *Análisis del discurso mediado por ordenador* es, en cambio, el acercamiento a la comunicación en la Red y a través de móvil desde la metodología del Análisis del discurso.

plataformas y que aseguren su aplicación dentro de la comunidad académica. Construir un corpus de interacciones comunicativas que se originan en Internet puede parecer una tarea sencilla, en la medida en que el discurso ya se encuentra digitalizado, es accesible fácilmente en la Red y no requiere, en principio, transcribirse para su posterior análisis (Mancera Rueda & Pano Alamán 2014).

No obstante, hay una serie de elementos que pueden ser problemáticos si no se toman las decisiones adecuadas. Por ejemplo, cuando se elabora un corpus de este tipo, enseguida surgen incompatibilidades entre los métodos asentados de la Lingüística de corpus y los datos recogidos en los distintos canales de CMO. En la recolección de datos, además de tener en cuenta aspectos como la extensión del corpus o su representatividad, Beiβwenger & Storrer (2008) y Androutsopoulos & Beiβwenger (2008) advierten de que es necesario distinguir los datos que genera el sistema –por ejemplo, las líneas de entrada y salida de un canal de chat, o la fecha de envío de un *tuit*– de los que genera el usuario, como texto, enlaces, emoticones e imágenes. Por otra parte, King (2009) enumera una serie de aspectos problemáticos que conciernen, entre otros, la ubicación geográfica de los hablantes, la validez de algunas interacciones de cara al análisis lingüístico o sociolingüístico, e incluso la peculiar ortografía que se observa en algunos tipos de CMO. Según el autor, es necesario interrogarse sobre si cabe mantener esa ortografía o adaptarla a la norma para análisis *PosTag*; si se deben considerar los *nicknames* o desecharlos por irrelevantes para el análisis; y, en el caso de interacciones de carácter privado, cómo solicitar la autorización de los participantes para tratar y publicar los datos. La mayor parte de los corpus existentes aborda estas cuestiones aportando distintas soluciones en función de los objetivos de análisis o de los proyectos en los que se enmarcan. Entre estos, cabe destacar: Sms4science, corpus multilingüe de SMS; NUSSMSCorpus, con SMS en inglés y en chino; DeRiK y CoMeRe, corpus de referencia de CMO para el alemán y el francés, respectivamente; y Web2Corpus_it, con textos de interacciones mediadas por ordenador en italiano.

Por otro lado, muchos de los esfuerzos dedicados a la constitución de estos corpus se han centrado en complementar y mejorar el estándar de etiquetado XML-TEI, que hasta el momento carece de elementos específicos para la anotación de textos de CMO. Por ejemplo, para el alemán, se ha llevado a cabo un trabajo que descansa sobre una rigurosa metodología de almacenamiento (Beißwenger & Storrer 2008; Beißwenger *et al*. 2012; Beißwenger *et al.* 2013). En ella, los investigadores distinguen entre un etiquetado de la macroestructura de la CMO, en relación con los distintos "modos sociotécnicos" (Herring 2002, 2007) de comunicar en la Red, y los fenómenos lingüístico-pragmáticos que serían específicos de cada uno. Para el italiano, se han llevado a cabo iniciativas

similares, en concreto, se han adoptado algunas de las propuestas planteadas ya por Beißwenger *et al*. (2012), aunque con diferencias significativas. Así, tal y como señalan Chiari & Canzonetti (2012: 601), el objetivo de Web2Corpus_it "non è fornire uno standard per le tipologie CMC, ma uno schema che renda conto nella metadatazione e nell'annotazione delle principali caratteristiche di interesse per un'indagine (socio)linguistica in senso lato". Por último, para el francés, Chanier *et al.* (2014: 5) han diseñado un modelo de corpus de CMO a partir del concepto de *interaction space*, "[…] encompassing, from the outset, the richest and the more complex CMC genres and situations. […] our goal is to release guidelines for all CMC documents and not for each CMC genre".

Para el español, no se ha elaborado aún un corpus en el que también se consideren estos aspectos. Hasta ahora, las investigaciones[3] han descansado sobre corpus diseñados y constituidos en relación con análisis específicos, lo que no permite aprehender la diversidad de modos de la CMO ni cubrir la variedad diatópica, entre otras, de este tipo de interacciones en el mundo hispanohablante[4]. A pesar de lo anterior, existen algunas aportaciones que van en este sentido.

Vela Delfa y Cantamutto (2015) consideran ampliamente la cuestión de la recolección de datos en este ámbito y plantean la creación de un repositorio abierto y colaborativo de interacciones comunicativas digitales en español (Proyecto CODICE). El repositorio constaría de una serie de muestras aportadas por otros investigadores y por nuevos datos recogidos en base a una plantilla que incluiría los parámetros tecnológicos y de situación de enunciación de la comunicación en entornos digitales. Esta idea supone un gran avance, puesto que se pueden compartir muestras que antes no se encontraban disponibles y porque permite reflexionar y aportar soluciones a los límites de los métodos de recogida adoptados hasta ahora. No obstante, cuando se aborda la cuestión de la "transcripción" de los datos dentro de este proyecto, no se especifica qué sistema se adoptaría. Aunque la propuesta se relaciona con el antecedente Text Bank CHILDES[5], basado en el sistema de transcripción de conversaciones CHAT convertible a XML, no se detalla cómo se codificarían los datos macro y

[3] En https://discursoenlared.wordpress.com/bibliografia-lengua-espanola/ es posible consultar una bibliografía sobre los estudios de CMO en español (accessed December 10, 2015).

[4] A. Pano Alamán & P. Moya Muñoz, *Una aproximación a los estudios sobre el discurso mediado por ordenador en lengua española* (trabajo inédito).

[5] C. Vela Delfa & L. Cantamutto, *Al abordaje de la comunicación digital: elaboración de un repositorio del español*. Comunicación *7 Congreso Internacional de Lingüística de Corpus* (CILC2015), Valladolid, 5-7 marzo 2015. http://bit.ly/1OZ2mXM (accessed December 10, 2015).

micro textuales o qué tipo de metadatos, además de los relativos a la situación de enunciación, deberían incluirse para permitir su consulta e interrogación, considerando también que los corpus que se incluirían en el repositorio se han elaborado y, en algunos casos, anotado, en función de objetivos de análisis lingüísticos, sociolingüísticos o pragmáticos diferentes. En este punto coincidimos con distintos analistas en que la constitución de un corpus depende de los usos que se le quieran dar al mismo (Beißwenger & Storrer 2008; King 2009; Borghetti, Castagnoli & Brunello 2011; Chiari & Canzonetti 2012; Real Academia Española 2013).

Cabe mencionar también el CORPES XXI (Corpus del Español del Siglo XXI), de la Real Academia Española (2013), que contiene datos obtenidos de la Web y en cuya elaboración y anotación, mediante el sistema XML, se han considerado ciertos rasgos que permiten delinear la diversidad que presenta una lengua, como son, por ejemplo, el medio (oral, escrito, electrónico) y la zona geográfica en la que se habla. Para este corpus se ha utilizado un sistema de codificación basado en el esquema TEI, diseñado especialmente para este tipo de datos. Sin embargo, aunque incluye documentos provenientes de Internet, su inclusión corresponde a un porcentaje menor en comparación con el resto de documentos (solo el 7,5% de los textos escritos recogidos) y no se especifica qué tipo de documentos son.

Se confirma por tanto la necesidad de constituir corpus de comunicación mediada por ordenador en español, que consideren tanto los aspectos de recogida de los datos, como las cuestiones de anotación de los mismos de cara a su posterior análisis.

2. Objetivos y fases del proyecto

El objetivo principal de CorpusRedEs es diseñar y construir un corpus de CMO en español que: a) sea representativo del tipo de interacciones que tienen lugar en los distintos modos sociotécnicos que van surgiendo en la Red; b) contemple las principales variedades diatópicas del español; y c) cubra distintos dominios o temáticas.

En la primera fase del proyecto, se está llevando a cabo, por un lado, una recuperación selectiva de textos pertenecientes a distintos modos de CMO, suficientemente relevantes para dar cuenta de la variabilidad de interacciones en español que se producen en este contexto: chat, correo electrónico, foro, lista de distribución, comentario a noticia, blog, red social, wasap, microblog y wiki. Por otro lado, desde una perspectiva sincrónica y diatópica, y para cada modo de

CMO, se procede a la selección de textos pertenecientes a variedades del español, por "zonas lingüísticas habituales", en la línea del CORPES XXI.[6] De esta forma, se evita considerar la comunicación digital como un todo unitario y sin matices o tratar solamente determinadas variedades del español. Así, en la primera etapa del proyecto, el corpus estará constituido por alrededor de 1.000.000 de tokens, para tener una muestra inicial de textos con los que empezar a trabajar, de forma similar a como está planteado Web2Corpus_it (Chiari & Canzonetti 2012: 601), por modos y, en nuestro caso, variedades geográficas. Una de las cuestiones que hemos considerado a la hora de diseñar el corpus es la de la representatividad. Un corpus puede ofrecer información detallada acerca de una lengua particular, pero es imposible constituir uno que abarque toda una lengua. Por tanto, cabe buscar soluciones para establecer una proporcionalidad adecuada del corpus que conduzca a ciertas proyecciones. En este sentido, nuestro objetivo, en este estadio inicial, ha sido el de constituir lo que Parodi llama un "pre-corpus", "con el fin de proponer hipótesis de trabajo y de explorar ciertas características o categorías para una posterior recolección más amplia y robusta" (Parodi 2010: 27).

En la segunda fase del proyecto, se elaborará un esquema o plantilla en el que se identifiquen, para cada modo sociotécnico de CMO seleccionado, los parámetros tecnológicos relativos al medio y los parámetros relacionados con el contexto de enunciación o sociosituacionales (Herring 2007). En la tercera etapa se elaborará un documento en XML que operará como modelo para la anotación de corpus representativos de las tipologías de CMO consideradas. En este sentido, CorpusRedEs pretende contribuir a la reflexión iniciada por el TEI Special Interest Group sobre *Computer-Mediated Communication*, centrado en "modelling user contributions (*posts*) to written CMC dialogues; modelling CMC document structures (*macrostructures* in forum threads, chat logfiles, Twitter timelines); developing perspectives for the representation of discourse in multimodal CMC"[7].

La última fase del proyecto comprende la puesta a disposición en línea del corpus anotado y de los modelos de base, para su consulta y mejora por parte de investigadores interesados en utilizar el corpus o colaborar en el proyecto. Los textos se publicarán en una plataforma web, que incluirá información de carácter bibliográfico sobre teoría, metodología y aplicaciones del análisis del discurso mediado por ordenador en lengua española.

[6] http://www.rae.es/sites/default/files/Zonas_linguisticas_habituales._CORPES_XXI.pdf (accessed December 10, 2015).

[7] http://www.tei-c.org/Activities/SIG/CMC/ (accessed December 10, 2015).

3. Metodología

CorpusRedEs se plantea como un corpus para uso general, es decir, los textos no se han almacenado bajo el alero de otro proyecto de investigación que establezca algún tipo de delimitación externa para su recogida. La recolección se ha llevado a cabo de manera manual, teniendo en cuenta aspectos éticos como la privacidad de los autores y la autoría de los textos. En relación con el primer aspecto, hemos eliminado los datos de los autores cuando la comunicación se produce entre particulares, tanto en interacciones privadas (e.g. correo electrónico, wasap), como públicas (e.g. foros de debate, microblog). Mantenemos en cambio los datos sobre la identidad de los hablantes cuando se trata de personas que desempeñan una actividad pública y cuyo conocimiento puede resultar relevante para el análisis pragmalingüístico; en cuanto a la autoría, se solicita por escrito a los autores o a los propietarios de la plataforma donde se publican los textos (e.g. medio digital donde se publican los comentarios a una noticia o red social), la autorización a utilizar los datos para fines de investigación.

Siguiendo a Beißwenger *et al.* (2012: 6), los textos recopilados cumplen los siguientes criterios:

> [they are] (i) based on the TCP/IP protocol suite for data exchange, (ii) dialogic (with all participating users being able to switch between the role of a recipient/reader and the role of a producer/author of messages), and (iii) based on writing as the main encoding medium for the users' dialogue contributions (that is, the verbal parts of the contributions must be encoded using writing, though they may also include graphics, embedded audio, or video files).

En este sentido, diferenciamos el texto plano etiquetado en XML, que puede contener enlaces hipertextuales (URL, *hashtag*), de los archivos relativos a vídeos, audios, imágenes y otros recursos multimedia incrustados en el documento. A partir de los criterios mencionados, el corpus no incluiría SMS, ya que no están basados en el protocolo TCP/IP; páginas web estáticas, en las que la comunicación es esencialmente unidireccional; o mensajería de voz (i.e. Skype), al no llevarse a cabo fundamentalmente a través del código escrito.

El etiquetado, realizado con el editor XML *Oxygen*, frecuentemente utilizado en proyectos de Humanidades Digitales Hispánicas (i.e. Remetca) y que incluye ya las etiquetas y plantillas de TEI P5, se centrará en los aspectos macroestructurales de los textos escritos almacenados. No se considerarán, para esta etapa del proyecto, aquellos elementos relacionados con la multimedialidad (Bateman, Delin & Henschel 2004). No obstante, esto no quiere decir que este

tipo de datos no jueguen un papel esencial en la CMO, por lo que, en un primer momento, sin descargar las imágenes u otros elementos multimedia y archivarlos en una base de datos, nos limitaremos a recuperar el enlace en el que se encuentra alojado el contenido multimedia referenciado, para poder recuperar y trabajar sobre ese contenido en fases sucesivas del proyecto.

Tampoco se etiquetarán por ahora los rasgos microtextuales, propiamente lingüísticos, de los textos recogidos, y que algunos investigadores consideran como específicos de la CMO, como son, por ejemplo para el alemán, las interjecciones (*ach*) y los llamados *interaction words* (*grins, lol*), *addressing terms* (@zora), *interaction templates* (figuras en wasap) y *responsives* (*okay*) (Beißwenger *et al.* 2012). Nos alejamos de esta anotación microlingüística puesto que, para nosotros, no es posible hablar de un lenguaje típico de la CMO. La vasta diversidad de escenarios y propósitos de uso de las diferentes formas de comunicación en la Red sobrepasan cualquier caracterización lingüística *a priori* (Androutsopoulos 2006), aunque los textos presenten rasgos comunes, como la tendencia a la coloquialización en los niveles ortográfico, morfosintáctico y léxico, principalmente (Mancera Rueda & Pano Alamán 2013). Seguimos, en este punto, la propuesta de Chiari y Canzonetti (2012: 602), quienes para su corpus llevan a cabo un etiquetado del documento en sí, esto es, de su macroestructura, sin describir o analizar elementos lingüísticos.

Como apuntábamos, resulta imprescindible distinguir las características tecnológicas y situacionales para cada variedad de CMO, en la medida en que dicha distinción facilitará la identificación de los rasgos macroestructurales que deben ser identificados y almacenados en el corpus de base. Para ello, se elaborará una plantilla con indicación de los parámetros tecnológicos relativos al medio y de los parámetros sociosituacionales relacionados con el contexto comunicativo. En la Tabla 1, se recogen estos factores, fundamentales para llevar a cabo cualquier investigación en CMO.

Tabla 1. Factores clasificación modos CMO (Moya Muñoz 2015, adapt. de Herring 2007).

Tipo de factores	**Característica**	**Definición**
Factores del medio	Sincronicidad	Conexión al mismo tiempo entre los usuarios
	Transmisión del mensaje	La transmisión se realiza mensaje por mensaje o signo por signo
	Persistencia	El tiempo que se mantienen los mensajes almacenados
	Tamaño del mensaje	Número de caracteres que el sistema permite incluir en un único mensaje.
	Canales de comunicación	Texto, audio, vídeo, imágenes, etc.
	Mensajes anónimos	Autoría desconocida
	Mensajes privados	Un usuario puede contactar en privado con otro
	Filtración	Hay administradores que filtran los mensajes
	Cita	Los mensajes se pueden incrustar en otros mensajes para facilitar la interacción
	Formato de los mensajes	Determina la presentación visual de los mensajes
Factores situacionales	Estructura de la participación	Número de participantes
	Características de los participantes	Características ideológicas, demográficas, culturales, etc.
	Propósito	Metas de la interacción
	Tópico	Tema
	Tono	Formal o informal
	Actividad	Acción llevada a cabo
	Normas	Prácticas establecidas aceptadas por el grupo
	Código (lengua y/o variedades de lengua)	Variedades del español

4. Propuestas para la anotación de los textos

Para Beißwenger *et al*. (2012: 8), la unidad mínima para el etiquetado de CMO, basándose en el sistema de la TEI, es el *posting*, que es posible considerar como parte de un *thread*. Para estos autores:

> In CMC documents it [posting] represents the largest structural unit that can be assigned to one author and one point in time. The category *posting* is defined as a content unit that

> has been sent to the server 'en bloc'. Its function is to make a (written) contribution to the ongoing dialogue.

Desde otra perspectiva, Chiari & Canzonetti (2012: 602) consideran para su análisis la unidad *thread*, aunque cabe decir que se trata de un término problemático porque puede confundirse con los *thread* de los foros de debate. Para estos autores, es una unidad más amplia que el *post*: "Il *thread* appartiene di volta in volta ad una specifica tipologia ed è costituito da una sequenza di *post*. In alcuni casi il primo *post* ha uno status particolare, da noi denominato *trigger*, poiché corrisponde all'origine e alla ragione del *thread*". Concordamos con estos investigadores en que, para la anotación de este tipo de textos, es necesario considerar la unidad más amplia *thread* o *hilo* en español dado que permite identificar los *post* o mensajes como distintas intervenciones en una misma interacción. Por otra parte, la unidad *post*, asociada a cada intervención, resulta adecuada para una anotación en profundidad. De hecho, tanto Beißwenger *et al*. (2012: 8-13), para etiquetar interacciones en Wikipedia en alemán, como Chiari & Canzonetti (2012: 602-603), para su corpus de distintas tipologías de CMO en italiano, proponen diversos elementos y atributos para etiquetar el *post*, como son, respectivamente, @who, @synch o @revisedby y *sender*, *time* o *quoting*.

Aunque coincidimos ampliamente con estas propuestas, consideramos que habría que: a) ampliar la definición de *thread* o *hilo*, en español, como conjunto de intercambios de distintas intervenciones, relacionadas, por ejemplo, por una misma temática u objetivo comunicativo, en un específico modo de CMO; y b) considerar las relaciones que pueden establecerse entre los *post* o mensajes que se insertan en un determinado *thread* o hilo de discurso. Como bien señalan los investigadores italianos, cabe tener en cuenta el estatus particular de cada *post*: este puede ser aislado y dar origen o no a un *thread*, esto es, a una suma de mensajes que corresponderían a las contribuciones de los hablantes en una interacción; pero también puede constituir una respuesta a un *post* precedente, como réplica o respuesta a una contribución ajena, o incluso abrir nuevos hilos en el mismo contexto comunicativo o en otros, como, por ejemplo, en los comentarios a una misma noticia en un diario digital o en una red social.

Pensando en la estructura esencialmente dialógica de la CMO (Pano Alamán 2008), cabe tener pensar que un *post* de inicio (e.g. intervención en chat, entrada en un blog, noticia en un diario digital, artículo en Facebook), puede ir seguido de comentarios, respuestas, réplicas a esas respuestas, citas, repeticiones, reacciones como "me gusta" o "compartir", entre otras. En este sentido, la etiqueta *webaction* ("pulsanti e elementi che permettono di interagire – invia, modifica, rispondi") propuesta por Chiari & Canzonetti (2012: 603) da

respuesta, en parte, a la necesidad de describir las modalidades de interacción en un *thread*, aunque está más centrada en las posibilidades del dispositivo tecnológico que en la interacción propiamente dicha y las relaciones entre las intervenciones iniciativas y reactivas que constituyen esos intercambios.

Por estos motivos, parece necesario considerar unidades y elementos de anotación que puedan combinarse con los elementos *post* y *thread* mencionados. En TEI P5, el elemento <p> permite identificar párrafos (*paragraph*) en textos escritos, y <u>, los enunciados o fragmentos de texto hablado (*utterance*), precedidos y seguidos normalmente de un silencio o de un cambio de hablante. Según Beißwenger *et al.* (2012: 3), ninguna de estas etiquetas es válida para anotar textos de CMO. El párrafo se entiende en TEI como unidad organizativa del texto que señala las decisiones del autor o del editor en la composición del mismo; en la CMO, en cambio, la organización del texto está determinada no solo por el autor de un *post*, quien establece el inicio y el final del mismo, sino también por la tecnología (i.e. rutina del servidor) (Beißwenger *et al.* 2012: 10), como los ficheros de registro de chats o el sistema automatizado de publicación de los comentarios en un foro. Como afirman estos investigadores, "a paragraph is a holistic unit determined by (one author's) *global* text coherence, whereas a posting in CMC is an atomic constituent of a written dialogue determined by the ongoing dialogue's *local* coherence" (cursiva de los autores).

Se asume aquí, como en muchos otros estudios, que la CMO es un *diálogo* o *conversación*, esto es, una actividad comunicativa, que es prototípicamente oral, pero que aquí se realiza mediante el código gráfico, en la que dos o más hablantes se alternan los papeles de emisor y receptor y negocian el sentido de los enunciados. Desde esta perspectiva, el elemento <u> que, de acuerdo con la TEI, "contains a stretch of speech usually preceded and followed by silence or by a change of speaker", puede resultar potencialmente útil para anotar las contribuciones de cada hablante en el contexto de la CMO, de modo similar a como se ha aplicado en el corpus de lengua oral C-ORAL-ROM, diseñado para distintas lenguas romances (cfr. "Entities for dialogue representation", Moneglia 2005: 32).

No obstante, debemos interrogarnos sobre si es posible aplicar tal cual el elemento <u> al tipo de intervenciones que se dan en la CMO, si se tiene en cuenta, por ejemplo, que la pausa o el silencio en los intercambios digitales pueden presentar distintos valores respecto a los que se producen en la conversación oral prototípica, o que en este contexto no es posible hablar, por ejemplo, de solapamientos entre intervenciones. En la CMO, las pausas/silencios pueden estar originidos en factores que muchas veces desconocemos: puede haber factores tecnológicos que hagan que el mensaje tarde en llegar, lo que

puede ser erróneamente interpretado como un silencio por parte del destinatario; o puede haber una intención precisa por parte de los hablantes y, en ese caso, la pausa tiene un significado pragmático que cabrá interpretar con la ayuda de otros elementos del contexto gráfico (puntos suspensivos, emoticones, etc.). Así, parece necesario considerar dos tipos de anotación de la macroestructura de estos textos, en función de los condicionantes propiamente tecnológicos del medio, por un lado, y de los condicionantes propios del tipo de discurso mediado por ordenador, esencialmente dialógico, por otro. En relación con estos últimos, creemos que el elemento <u> no debe ser desechado *a priori* para la anotación de textos de CMO, puesto que permite identificar, por ejemplo, el tipo de transiciones o pausas que pueden darse en ese contexto.

En todo caso, volviendo al elemento *post*, englobado en un *thread* o hilo, y propuesto específicamente para la comunicación electrónica, creemos que representa una válida opción para la anotación formal de los mensajes, aunque es necesario enriquecerla, como reconocen implícitamente Beißwenger *et al.* (2012: 13), cuando afirman que los *post* se estructuran en dos dimensiones: "the above/below dimension, which usually stands for a *temporal before/after relation*; the left/right dimension, in which one can use indentation to emphasize the *topical affiliation of one message to a previous message*" (cursiva nuestra). Nuestra propuesta prevé, pues, en relación con este elemento, la aplicación de atributos relativos no solamente al autor (@who), sino también al tipo de relaciones temporales o espaciales (@next o @prev), entre los mensajes, entre otros aspectos. En este punto, resulta fundamental apurar si y de qué manera los atributos y miembros del elemento <u> (enunciado en la lengua oral) permitirían afinar la descripción del nuevo elemento <post> (mensaje en la CMO). En relación con este elemento, puede resultar útil considerar las categorías de no simultaneidad y de multimodialidad planteadas por distintos autores (Herring 1999; Alcántara Plá 2014), en su caracterización de la interacción mediada por las nuevas tecnologías.

5. Resultados y futuros desarrollos

Uno de los resultados que se espera conseguir es el de obtener una serie de documentos en XML-TEI, adaptado a la comunicación digital, que operarán como modelos de anotación para los modos considerados de CMO en español, con factores tecnológicos, contextuales y situacionales distintos.

Asimismo, se espera facilitar la interoperabilidad entre corpus de distintos modos sociotécnicos y de lenguas, así como entre los sistemas de interrogación

de los mismos. En relación con este punto y de acuerdo con lo señalado en el epígrafe anterior, se pretende contribuir a la reflexión sobre los esquemas de anotación de este tipo de textos dentro del TEI SIG para la comunicación en la Red. En particular, parece necesario profundizar sobre la cuestión de la validez del elemento <u>, aplicado ya a la anotación de corpus orales del español, y sobre su relación con el elemento <post>, para una anotación adecuada de las unidades propias del discurso mediado por ordenador. Junto a este, otros aspectos requieren mayor reflexión y serán abordados en futuros desarrollos del proyecto. Se trata, en concreto, de la representatividad de las muestras, la problemática definición de los géneros (o *modos*) de CMO y la necesidad de mejorar los instrumentos de descripción de estos textos.

Referencias

Alcántara Plá, M. 2014. Las unidades discursivas en los mensajes instantáneos de wasap. *Estudios de Lingüística del Español* 35(1): 214-233.

Androutsopoulos, J. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10(4): 419-438.

Androutsopoulos, J. & Beiβwenger, M. 2008. Introduction. Data and Methods in Computer-Mediated Discourse Analysis. *Language@Internet* 5(9). http://www.languageatinternet.org/articles/2008/1609/introduction.pdf (accessed December 10, 2015).

Bateman, J., Delin, J. & Henschel, R. 2004. Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making. In E. Ventola, C. Charles & M. Kaltenbacher (eds), *Perspectives on Multimodality*. Amsterdam/Philadelphia: John Benjamins.

Beißwenger, M. & Storrer, A. 2008. Corpora of Computer-Mediated Communication. In A. Lüdeling & M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Berlin: De Gruyter, 292-308.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. 2012. A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative*, Issue 3. http://jtei.revues.org/476 (accessed December 10, 2015).

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and linguistic computing* 28(4): 531-537.

Borghetti, C., Castagnoli, S. & Brunello, M. 2011. I testi del web: una proposta di classificazione sulla base del corpus PAISA. In M. Cerruti, E. Corino & C. Onesti (eds), *Scritto e parlato, formale e informale: La comunicazione mediata dalla rete*. Roma: Carocci, 147-170.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J. & Seddah, D. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics* 29 (2): 1-30.

Chiari, I. & Canzonetti, A. 2012. La comunicazione mediata dal computer: la questione del genere e il problema dell'annotazione. In E. Garavelli & E. Suomela-Härmä (eds), *Atti del*

XII Congresso SILFI Società Internazionale di Linguistica e Filologia Italiana. Cesati: Firenze, 595-606.

Herring, S.C. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4(4). http://jcmc.indiana.edu/vol4/issue4/herring.html (accessed December 10, 2015).

Herring, S.C. 2002. Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology* 36: 109-168.

Herring, S.C. 2007. A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet* 4(1). http://www.languageatinternet.org/articles/2007/761 (accessed December 10, 2015).

King, B. 2009. Building and analysing corpora of computer-mediated communication. In P. Baker (ed.), *Contemporary Corpus Linguistics*. London: Continuum, 301-320.

Mancera Rueda, A. & Pano Alamán, A. 2013. *El español coloquial en las redes sociales*. Madrid: Arco Libros.

Mancera Rueda, A. & Pano Alamán, A. 2014. Las redes sociales como corpus de estudio para el Análisis del discurso mediado por ordenador. In S. López Poza & N. Pena Sueiro (eds), *Humanidades digitales: desafíos, logros y perspectivas de futuro* (*Janus*: Anejo 1). A Coruña: Universidade da Coruña 305-315. http://bit.ly/1gYmw1f (accessed December 10, 2015).

Moneglia, M. 2005. The C-ORAL-ROM resource. In E. Cresti & M. Moneglia (eds), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins Publishing, 1-70.

Moya Muñoz, P. 2015. La influencia de la Web 2.0 en la comunicación: una aproximación desde la Comunicación Mediada por Ordenador. *Lengcom* 4(3). http://bit.ly/1Q9ReIO (accessed April 9, 2015).

Pano Alamán, A. 2008. *Dialogar en la Red. La lengua española en chats, e-mails, foros y blogs*. Bern: Peter Lang, European University Studies.

Parodi, G. 2010. *Lingüística de corpus: de la teoría a la empiria*. Frankfurt: Iberoamericana Vervuert.

Real Academia Española 2013. *Corpus del Español del Siglo XXI*. http://www.rae.es/recursos/banco-de-datos/corpes-xxi (accessed December 10, 2015).

TEI. http://www.tei-c.org/index.xml (accessed December 10, 2015).

Vela Delfa, C. & Cantamutto, L. 2015. Methodological Approach to the Design of Digital Discourse Corpora in Spanish. Proposal of the CÓDICE Project. In P. Fuertes, E. Álvarez, R. Fernández, P. Garcés, B. López, M. Niño, I. Pizarro, A. Sáez, M. Sastre & M. Velasco (eds), *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics*. Elsevier, 494-499.