Al abordaje de la comunicación digital: elaboración de un repositorio del español

Cristina Vela Delfa
UNIVERSIDAD DE VALLADOLID
Lucia Cantamutto
UNIVERSIDAD NACIONAL DEL SUR -CONICET

7º Congreso Internacional de Lingüística del Corpus
Panel: *Diseño, elaboración y tipología de corpus*

La presente investigación persigue responder a dos preguntas fundamentales:

i) ¿en qué situación metodológica se encuentran las investigaciones sobre la comunicación digital?

ii) ¿es posible diseñar un corpus abierto y colaborativo (REPOSITORIO) de comunicaciones digitales en nuestra lengua?

Objetivos

Colecciones de datos lingüísticos: el caso de la comunicación digital

1. **Corpus derivados de proyectos de investigación:** compilados en la elaboración de datos para proyectos de investigación particulares a partir de preguntas de investigación.
2. **Corpus para uso general:** no se integran con alguna pregunta de investigación particulares y sirven para distintas hipótesis.

---

3. **Corpus de datos sin procesar o simples: se accede a los datos tal como fueron recolectados inicialmente.**
4. **Corpus anotados:** los datos están anotados ya sea de manera manual o a través de algún software específico.

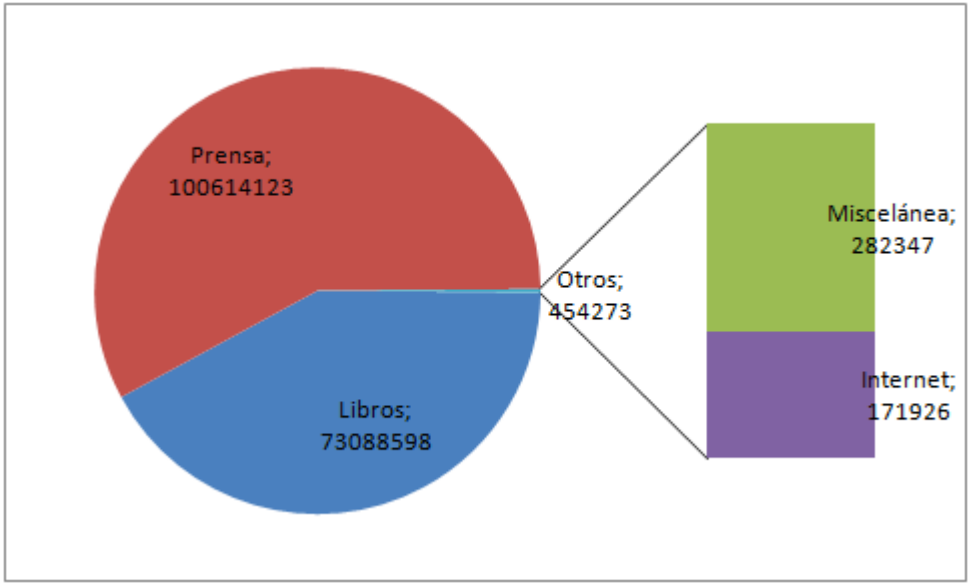
Corpus de comunicaciones digitales

Nombre del corpus	Idioma/s	Tipo de corpus	Descripción	Enlace o Bibliografía
CoSy:50 Corpus	Inglés	Corpus simple	50 presentaciones de 152 conferencias de informática	Yates, 1996.
German-Swedish IRC-Corpus	Sueco y alemán	Corpus simple	Chats	http://www.linguistik-online.de/15_03/pankow.pdf
SpamAssassin Public Corpus	Inglés	Corpus simple para uso general	> 6000 mensajes de correo electrónico spam	http:// spamassassin.apache.org /publiccorpus/
E-Mail corpus from the*COSMA* project	Alemán	Corpus de proyectos de investigación	160 mails	http://www.coli.uni-saarland.de/publikationen/softcopies/Declerck:1997:EKE.pdf
Dortmund Chat Corpus	Alemán	Corpus anotados para uso general	> 500 chats	http:// www.chatkorpus.tu-

Tabla 1: Elaboración propia en base a la lista provista por CMC-Corpora disponible en **http://www.cmc-corpora.de/** y Beißwenger, M., & Storrer, A. (2008)

Representatividad de comunicaciones digitales en corpus generales del español: CORPES

Gráfico 1: Distribución de formas por soporte. Elaboración propia (datos extraídos de http://web.frl.es/CORPES/org/publico/pages/ayuda/informacion.view, consulta: noviembre de 2014).



Corpus de comunicaciones digitales en español

Tipo de interacción	Carácter	Dominio	País	Descripción del corpus	Referencia
Correo electrónico	Monolingüe	Privado	España	>1800 mails recogidos entre 2001 y 2004, 1350 entre 2011-2014	Vela Delfa, 2005.
Chat	Bilingüe	Privado	España	100.000 palabras de chat recogidas en 2004	Mariottini, 2006.
Chat	Monolingüe	Semi-público	España	55 chats de entre 15 minutos y 2 horas entre 2004 y 2007.	Alvarez Martínez, 2008.
Chat	Monolingüe	Privado y público	Argentina	20 chats grupales y de persona a persona de Messenger y ICQ, entre 2001 y 2002	Noblia, 2009
Redes sociales	Monolingüe	Semi-público	Argentina	>70.000 palabras de 1897 comentarios recogidos en 2013	Kaul-Marlangeon y Cordisco, 2014
SMS	Monolingüe	Privado	Argentina	>3000 SMS de diferentes grupos etarios recogidos entre 2011 y 2014	Cantamutto, 2012 y Cantamutto, 2014.

CODICE: Comunicación Digital: Corpus del Español

Antecedentes de repositorios digitales abiertos y colaborativos de datos lingüísticos:

1. **THE TALK BANK**

   →CHILDES

CODICE: objetivos

1. Creación de un repositorio de comunicaciones digitales en español, a partir de las aportaciones de los trabajos parciales de investigadores de este campo de estudio.
2. Optimización de los recursos invertidos en la recopilación de muestras de lenguas
3. Disposición tanto datos de fuentes primarias como trabajos que aborden aspectos teórico y metodológicos sobre la comunicación digital.
4. Así mismo, se plantea como objetivo complementario la creación de unos estándares comunes en la recogida de los datos, en lo que concierne principalmente a los factores contextuales y situacionales, a fin de facilitar los análisis sociopragmáticos.

CODICE: etapas

1. REFLEXIÓN METODOLÓGICA
2. DISEÑO DEL RESPOSITORIO
3. IMPLEMENTACIÓN DEL RESPOSITORIO
4. DIFUSIÓN
5. ORAGANIZACIÓN DE DATOS

Comunicación mediada por ordenador vs. Discurso Digital

CMO (sigla adaptada de su equivalente en inglés CMC - Computer-Mediated-Communication): aquella comunicación producida cuando dos o más personas interactúan transmitiendo mensajes a través de un ordenador o de otro dispositivo tecnológico (Herring, 2001, 612).

La noción de Discurso Digital supone un constructo que, más allá de servir para conformar una clase de elementos, adquiere un trasfondo metodológico importante.

Los datos del discurso digital

- Multimodalidad (Herring, 1996)
- Multisimultaneidad (Alcántara Plá, 2014)
- Hiperpersonalidad (Whalter, 1996)

Multimodalidad

“Todo discurso es esencialmente multimodal y dicha multimodalidad no afecta únicamente al flujo de códigos semióticos, sino que incide en los mecanismos comunicativos de producción y comprensión que interviene en la construcción de los esquemas de significado (Obando, 2012: 881)”

EL DISCURSO DIGITAL ES CADA VEZ MÁS MULTIMODAL:

-Plataformas Multimodales Interactivas
- Plataformas NO Multimodales Interactivas

Multisimultaneidad

Los interlocutores se involucren de forma simultánea en varios intercambios paralelos constituye una condición intrínseca del medio digital, que es posible gracias la confluencia de varios factores:

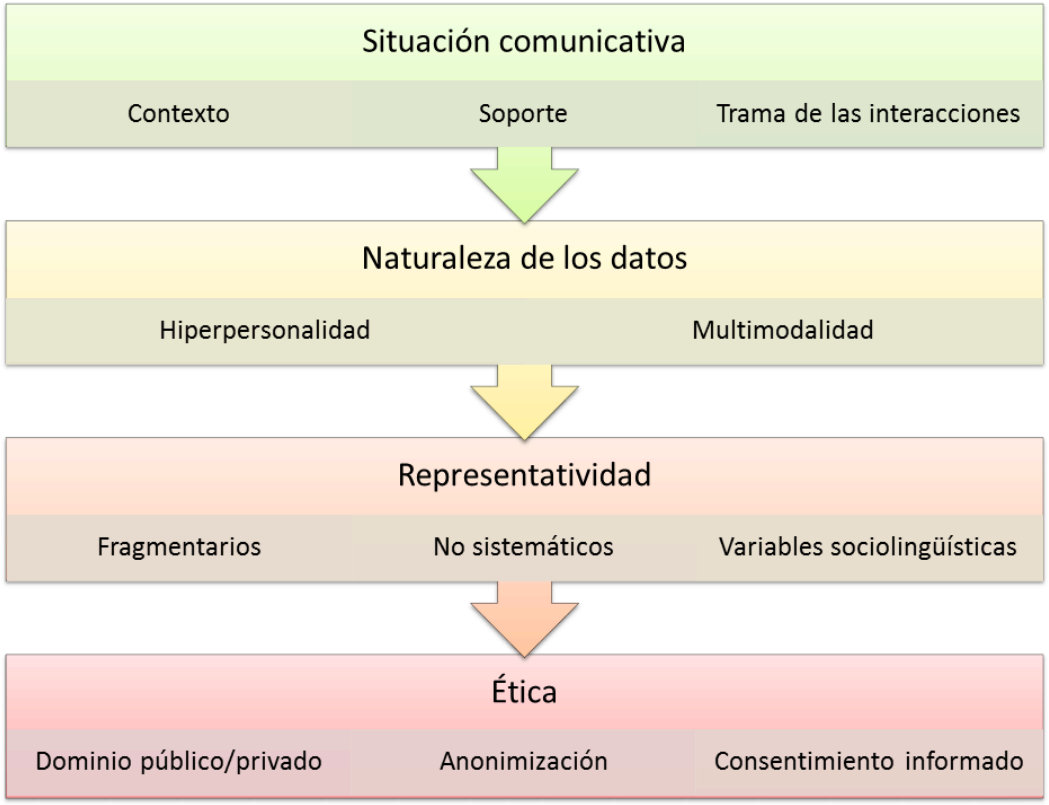
1. la persistencia textual
2. el carácter diferido del intercambio
3. la ausencia de copresencia física

Hiperpersonalidad

- Otorga un mayor control en la construcción de nuestra imagen personal: presentación selectiva de su imagen personal a través de los mensajes que crean y envían.

- **Relaciones con un grado de intimidad** muy elevado: emoción afectividad, etc.

Condicionantes, características y respuestas



La plantilla de trascripción

https://docs.google.com/forms/d/1TodONrQYbvmJBrXkPOHTUi_McmtBrJdy95zpIL8Ip8Q/edit

Hemos discriminado cuatro núcleos principales a los cuales debemos atender a la hora de desarrollar los metadatos:

a) la situación comunicativa,
b) la naturaleza de los datos,
c) la representatividad y
d) cuestiones éticas.

La plantilla de trascripción

LA SITUACIÓN DE COMUNICACIÓN:

1. Información sobre el dispositivo
2. Descripción icónica del paratexto
3. Descripción situación de enunciación
4. Recogida de datos de retroalimentación

DISTINGUIMOS DOS NIVELES

1) **Contextualización básica:** con datos de descripción del soporte y de identificación de la situación de comunicación, información que se incluiría en una suerte de metadatos que acompañarían al archivo de la trascripción principal
2) **Contextualización enriquecida**: con archivos complementarios de diversa naturaleza que se anexarían al archivo de la trascripción principal

La plantilla de trascripción

LA NATURALEZA DE LOS DATOS:

Se fijaran dos niveles de datos (Herring, 2014):

1) el texto limpio o plano (que puede estar enriquecido con etiquetamiento en html http://goo.gl/forms/O86VWMP8JC): criterio de segmentación de unidades, anclaje de archivos complementarios

2) otros archivos complementarios, tantos como se hayan podido recoger, con videos, audios, capturas de pantalla y otras formas de *fijar* las manifestaciones multimediales.

Las datos textuales serán obligatorios, pero los datos multimodales se incluirán en la medida en que hayan sido cedidos. Por ellos, resultarán menos comunes en las muestras elicitadas, pero serán considerados muy recomendables en las muestras de introspección.

La plantilla de trascripción

LA REPRESENTATIVIDAD:

A fin de asegurar la representatividad, incluso a partir de cruce de muestras con datos parciales, se incluirá un apartado con metadatos, que acompañarán a la trascripción, relativos a información sociolingüíistica del informalmente. Se atenderán aspecto como el nivel de formación, el grado de familiaridad con los medios digitales, frecuencia de uso del medio digital en general y de la aplicación en particular, edad, sexo y otras informaciones relevantes. Se completará con la inclusión de datos relativos al nivel pragmático, como la relación entre los interlocutores (distancia social, poder relativo), el situación de comunicación (registro, tono).

http://goo.gl/forms/O86VWMP8JC

La plantilla de trascripción

CUESTIONES ÉTICAS:

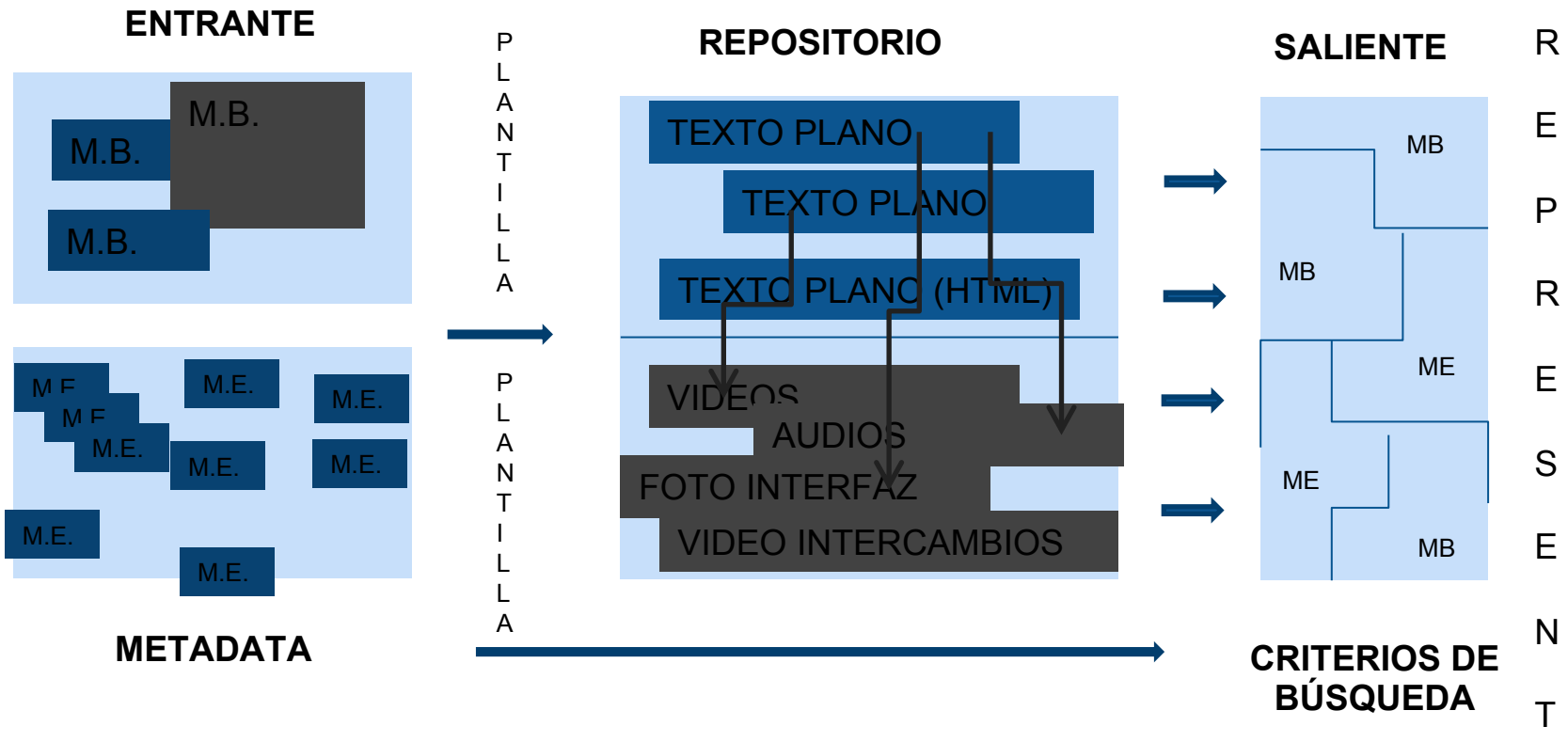
- Reflexión sobre el origen de los datos: ¿datos públicos o privados?

  → DATOS DISPONIBLES
  → DATOA ELICITADOS

- En el caso de datos elicitados, se pedirá que estos se acompañen de los correspondientes consentimientos informados.

- Las muestras deben enmascarar los detalles que hicieran posible la identificación personal de los informantes:

El repositorio CODICE: diseño



conclusiones

1. VENTAJA DE COMPARTIR MUESTRAS: COMPELMENTARIEDAD DE DATOS PARCIALES

2. DOS PERFILES DE MUESTRAS:
    1. generales→ representatividad
    2. instrospección → riqueza metadata

3. ESTANDARIZACIÓN DE MUESTRAS

4. REPRESENTATIVIDAD

5. ESTABILIZACIÓN METODOLÓGICA

https://sites.google.com/site/proyectocodice/

cveladelfa@gmail.com
luciacantamutto@gmail.com

Gracias