Los corpus electrónicos y la investigación lingüística

FAPESP *Week* Montevideo

Virginia Bertolotti
Facultad de Información y Comunicación

Corpus lingüísticos electrónicos

- Conjuntos de datos auténticos seleccionados de acuerdo con objetivos de investigación determinados y luego informatizados

- Su confección y su empleo han llevado a un **“giro informático**” en buena parte de los estudios sobre el lenguaje: sobre las disciplinas y sobre los investigadores

  Impacto en la disciplina:
  a. Nuevas formas de edición
  b. Necesidad de categorizar

  Impacto en los investigadores:
  a. Cantidad y disponibilidad
  b. Trabajo colaborativo
  c. Internacionalización de la investigación

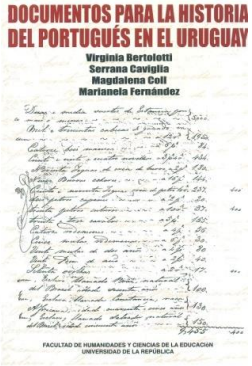
- Su confección: elegir qué datos seleccionar, qué variables hacer procesables y la creación de categorías constituyen una actividad investigativa en sí misma.

Corpus informatizados del español

- *Corpus Diacrónico del Español (CORDE)*
  Real Academia Española, coordinado por Guillermo Rojo
- *Corpus Hispánico y Americano en la Red: Textos Antiguos* (*CHARTA*)
  Universidad de Alcalá de Henares, coordinado por Pedro Sánchez-Prieto Borja
- *Corpus del español* de Mark Davies
- *Corpus del español del siglo XXI*, Real Academia Española, coordinado por Guillermo Rojo
- ***Corpus diatópico y diacrónico del español en América* (CORDIAM)**

Antes: corpus en papel

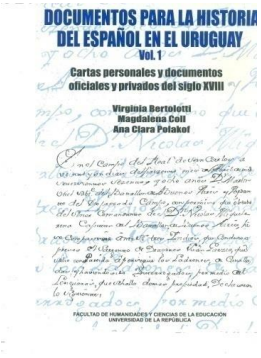
- Ej. Corpus de documentos para el estudio de la historia del portugués en el Uruguay

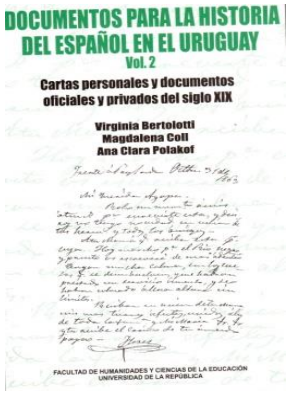


- Ej. Corpus de documentos para el estudio de la historia del español en el Uruguay:

  Vol. 1 – Siglo XVIII

  Vol. 2 - Siglo XIX





Antes

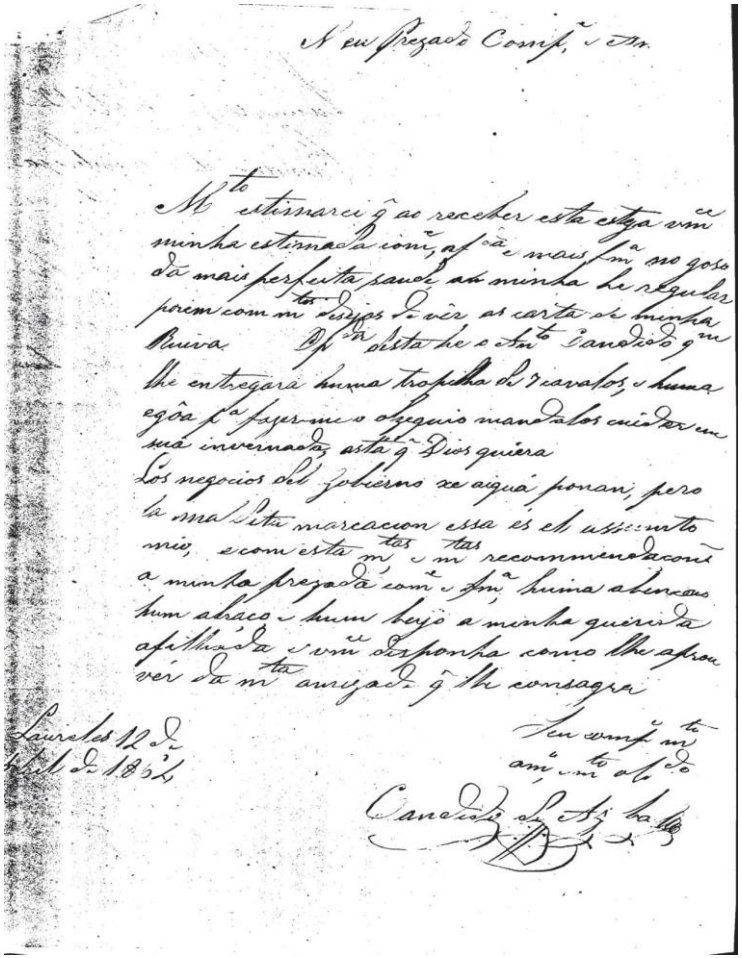
Documento 14: Carta con pasajes en español y en portugués . Escrita por Cándido Azambuya desde Laureles. El destinatario es el Sr. Constantino García da Rosa en Cuaró.
Siglo: XIX **Año: 1864**
Trans: LO Rev: MC

[fol. 1r]

Meu Prezado Comp.º e Amº

Mto estimarei qe ao receber esta esteja vmce
minha estimada come, of da a mais fma no goso
Da mais perfeita saude a minha he regular
porem com mtos desejos de ver as carta de minha
Ruiva Opdes desta he o Anto Candido qm
lhe entregara huma tropilha de 7 cavalos, e huma
Egôa pa fazerme o obzequio mandalos cuidar em
Sua invernada, astá qe Dios quiera
Los negocios del Gobierno de aiguá penan, pero
la maldita marcacion essa és el assumto
mio, e com esta mtas mtas recommendaçõns
a minha prezada come e Ama huma abencao
hum abraço e hum beijo a minha querida
afilhada e vmce disponha como lhe aprou
ver da mta amizade qe lhe consagra

Seu compe mto
Amº e mto Atto
Candido de
Azambuya {RUB



Un ejemplo de un corpus histórico electrónico:

Corpus Diacrónico y Diatópico del Español de América

CORDIAM

www.cordiam.org

1. Presentación de *CORDIAM*

- Corpus de corpus
- Corpus de documentos manuscritos, extraídos directamente de los archivos
- Manuscritos transcriptos y pasados a soporte digital
- Incluidos actualmente en un motor de búsqueda *ad-hoc*

2. Prestaciones del *CORDIAM*

- **Búsquedas simples**
- **Búsquedas sintácticas complejas y búsquedas discontinuas**, con más de un término **lematizado**.
- **Información cuantitativa**: número de casos en número de documentos, en universo de palabras. Facilita el análisis de fenómenos lingüísticos.
- **Plantilla de 17 metadatos asociados** a cada documento, de relevancia lingüística, histórica y social.
- **Accesibilidad inmediata a los metadatos**. Calidad y cantidad de los parámetros de búsqueda en la plantilla = **7 parámetros de búsqueda**.

2. Prestaciones del *CORDIAM*

- **Visualización simultánea** de la concordancia y de su contexto mayor.
- **Visualización simultánea** de la concordancia y de sus metadatos.
- **Búsquedas con diversos tipos de ordenamientos**: siglo-país-tipo textual / tipo textual-país-siglo, etc.
- **Ordenación alfabética** de concordancias.
- **Guardado automático en base de datos** de todas las concordancias seleccionadas.
- **Acceso al documento completo**, guardado e impresión.

3. Plantilla de metadatos

1. **Nombre**: I
2. **Siglo: 16**
3. **Año: 1546**
4. **Autor (datos étnicos): s/d**
5. **Autor (hombre o mujer): hombre**
6. **Autógrafo:** no
7. **País actual:** PER
8. **Topónimo actual**: Lima
9. **Topónimo histórico**: Lima
10. **Adscripción histórica**: Audiencia de Lima, Virreinato del Perú
11. **Tipo textual: Documentos jurídicos**
12. **Archivo**: Biblioteca Nacional de Perú, A-208. Expediente del pleito seguido por Francisco, indio natural de Nueva España.
13. **Número de folios**: 1
14. **Número de palabras aproximado**: 310
15. **Créditos**: José Luis Rivarola, *Documentos lingüísticos del Perú. Siglos XVI y XVII. Edición y comentario*, anejo de la *Revista de Filología Española*, 103, Madrid: Consejo Superior de Investigaciones Científicas, 2009.
16. **Facsimilar** disponible: no
17. **Síntesis**: Declaración de Martín Fernández, zapatero.

4. Géneros textuales existentes

49 géneros con superposiciones textuales

- Cartas entre particulares
- Cartas oficiales
- Descripciones geográficas
- Relaciones de sucesos
- Actas de cabildo
- Probanzas de méritos
- Juicios civiles
- Juicios de residencia
- Capitulaciones
- Informes
- Memoriales de méritos
- Requerimientos
- Inventarios de barcos
- Inventarios de bienes de difunto
- Notas, billetitos y recados
- Testamentos
- Testimonios en juicios
- Docs. probatorios en juicios
- Bandos
- Decretos
- Denuncias / Querellas
- Sentencias
- Nombramientos
- Peticiones
- etc., etc., etc. etcétera

5. Tipos textuales del *CORDIAM* (*Creación de categorías*)

Requisitos deseables

- Pocos agrupamientos.
- Agrupamientos grandes con “homogeneidad” interna.
- Etiquetas generales que respeten las diferentes rutinas comunicativas escritas durante esos cuatro siglos.

- **Ejes taxonómicos divisorios:** tipo de circulación del documento + cómo llega a un archivo.

- **Más privado > Más público**

6. Muestra de tipos textuales

1. Documentos entre particulares

Mi mas estimada y q*ue*rida esposa de / mi corazon me alegrare q*ue* al rresibo desta / te alles con la salu q*ue* yo para mi deseo / en conpañia, de mis dos amadas ijas de mi corason / y de tu familia y mia / la q*ue* yo difruto es buena para que me mandes / q*ue* lo are como me toca de obligasion

2. Cronísticos

Y / así diçen que los vnos salieron de qüebas, los otros de çerros, /[25] y otros de fuentes, y otros de lagunas y otros de pies de árboles, / y otros desatinos desta manera; y que por auer salido y enpeçado / a muntiplicar destos lugares y auer sido de allí el prinçipio / de su linaje, hizieron guacas y adoratorios estos lugares / en memoria del primero de su linaje que de allí proçedió; /[30] y así cada nación se uiste y trae el traje con que a su guaca / uestían.

7. Muestra de tipos textuales

3. Jurídicos

yo Jose Candido Baes besino de el pueblo de Antim*an*o y residente de la Ciudad de S*a*n Felipe / Ante V paresco y digo que el rreo nombrado ylario Silba es un honbre que me a sentensiado a muerte con una lansa que a sacado en mi misma casa y p*o*r no aber tenido los testigo (sic) no me presente ante V y de contra a una mujer que tengo en mi Casa a sacado un puñal para matarla en la casa de el S*eñ*or Miguel Bara p*o*r un pique que tie <inter: ne> con hella p*o*r una mujer que el tenia y llo la hise salir de el Sitio de Carapa y p*o*r Cullo moti*vo* Cuantas beses pasa p*o*r mi casa a dis*tintas* horas de la noche se benga Con pegar un astaso a las tijas de mi Corredor que estan a la bista las tijas quebradas en dicho Coredor

4. Administrativos

resçivymos las dichas escryturas e fees de almonedas en n*ues*tro poder porq*ue* no se perdjesen, e res*e*rvamos n*ues*tro d*ine*ro a salbo p*ar*a q*ue* por raçon de lo susod*ic*ho no nos paren p*er*juiçio si las dichas devdas saljesen inçiertas {f.3} (...) A v*uestra sacra* mag*estad* supljcamos vmyllmente sobre todo lo que a V*uestra sacra* mag*estad* escrevymos nos ynvye a mandar lo que fuere s*er*vydo

8. Necesidad de sistematización (revisión de prácticas de edición)

- Un corpus electrónico NO es un texto:
  - Lectura fragmentaria / Lectura integral
  - Lectura vertical / Lectura horizontal
  - Búsquedas / Contenidos

- La transcripción para un corpus electrónico no puede ser una reproducción fotográfica del original
- Una transcripción exacta es poco funcional y es confusa para búsquedas electrónicas: contiene símbolos que «representan» palabras.
- Necesario mantener palabras completas para hacer búsquedas y evitar pérdida léxica

9. Criterios de sistematización

No alterar. Sí sistematizar para permitir búsquedas

- Desatado de abreviaturas en cursivas; actualización de grafías antiguas; no ruptura de palabras por cambio de línea / folio

- y **supl^{co}** a su md y a los demas mis señores relyxosos → supl*i*co
- **abnque** → aunque **bna** → una
- Salud te desea tu **Desgra / ciada** mama → Desgraciada / mama
- bien sabeys de mi condiçion q p^{a} **yr/^{10} me** con poca posiblidad alla → q*ue* p*ara* yrme /^{10} con
- que no se **be [*fol.1v] ra** ýjos como bosotro → bera {f.1v} ýjos

Síntesis

El giro informático en la Historia de lengua y en la Lingüística Histórica:

- - nos ha hecho repensar las categorías de la disciplina
- -nos ha hecho crear nuevas categorías
- -nos ha hecho repensar las prácticas (ecdóticas) de la disciplina

Crear corpus es mejorar las condiciones para la investigación pero también hacer investigación.