Parlaritaliano.it
Università degli Studi di Salerno, 26 de febrero de 2007

Los corpus como recurso compartido para la investigación lingüística

Joaquim Llisterri

Grup de Fonètica, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona

http://liceu.uab.cat/~joaquim

Los corpus como recurso compartido para la investigación lingüística

- ✓El papel de los corpus
- ✓El eterno problema de los recursos
- ✓Los recursos compartidos
- ✓El futuro de los recursos lingüísticos

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los corpus como recurso compartido para la investigación lingüística

- ✓El papel de los corpus
- ✓El eterno problema de los recursos
- ✓Los recursos compartidos
- ✓El futuro de los recursos lingüísticos

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus

- ✓El papel de los corpus en la lingüística
- ✓El papel de los corpus en las tecnologías lingüísticas

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus

- ✓ **El papel de los corpus en la lingüística**
- ✓ El papel de los corpus en las tecnologías lingüísticas

El papel de los corpus en la lingüística

FILLMORE, C. J. (1992) "'Corpus Linguistics' or 'Computer-aided armchair linguistics'", in SVARTVIK, J. (Ed.) *Directions in Corpus Linguistics. Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*. Berlin - New York: Mouton de Gruyter. pp. 35-66.

`http://linguistics.berkeley.edu/people/fac/fillmore.html`

El papel de los corpus en la lingüística

“Armchair linguistics does not have a good name in some linguistic circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn’t anybody exactly like this, but there are some approximations.)” (Fillmore 1992:35)

El papel de los corpus en la lingüística

“Corpus linguistics doesn’t have a good name in some linguistic circles. A caricature of the corpus linguist is something like this. He has all the primary facts that he needs, in the form of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence. (There isn’t anybody exactly like this, but there are some approximations.)” (Fillmore 1992:35)

El papel de los corpus en la lingüística

✓La lingüística de corpus

✓La lingüística con corpus

El papel de los corpus en la lingüística

✓La lingüística de corpus

✓La lingüística con corpus

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la lingüística
La lingüística de corpus

- La lingüística de corpus
  - Orígenes en la filología, la lexicografía, la sociolingüística, el análisis del discurso y, en general, la lingüística
  - Corpus textuales y corpus de lengua oral transcrita
  - Área de investigación consolidada y madura

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la investigación en lingüística
La lingüística de corpus

"A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research."

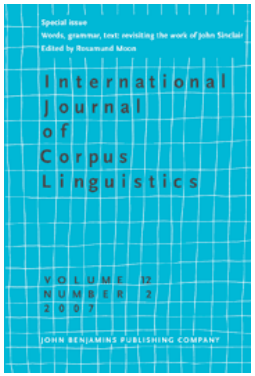
SINCLAIR, J. (2005) "Corpus and Text - Basic Principles", in WYNNE, M. (Ed.) (2005) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.
`http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm`

Universitat Autònoma de Barcelona

Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

La lingüística de corpus

- **Revistas**
  - ***International Journal of Corpus Linguistics*. Amsterdam: John Benjamins**
    `http://www.benjamins.com/cgi-bin/t_seriesview.cgi?series=IJCL`



El papel de los corpus en la investigación en lingüística

La lingüística de corpus

- **Revistas**
  - ***Corpus Linguistics & Linguistic Theory*. Berlin - New York: Mouton de Gruyter**
    `http://www.degruyter.de/rs/384_7546_ENU_h.htm`

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la investigación en lingüística

La lingüística de corpus

- **Revistas**
  - ***Corpora*. Edinburgh: Edinburgh University Press**
    `http://www.eup.ed.ac.uk/journals/content.aspx?pageId=1&journalId=12801`

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la investigación en lingüística

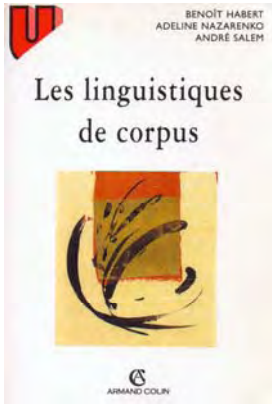
La lingüística de corpus

- **Manuales**
  - **McENERY, T.- WILSON, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics), 2nd edition, 2001**
    `http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm`

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

La lingüística de corpus

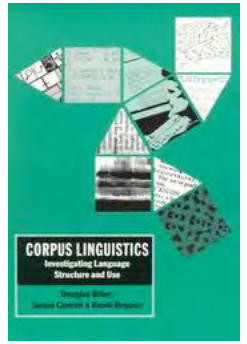
- **Manuales**
  - **HABERT, B.- NAZARENKO, A.- SALEM, A. (1997) *Les linguistiques de corpus.* Paris: Armand Colin (U Linguistique)**



El papel de los corpus en la investigación en lingüística

La lingüística de corpus

- **Manuales**
  - **BIBER, D. - CONRAD, S. - REPPEN, R. (1998) *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press (Cambridge Approaches to Linguistics)**

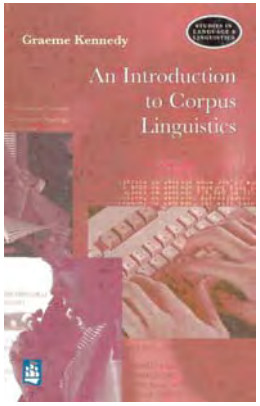


Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la investigación en lingüística

La lingüística de corpus

- **Manuales**
  - **KENNEDY, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman (Studies in Language and Linguistics)**

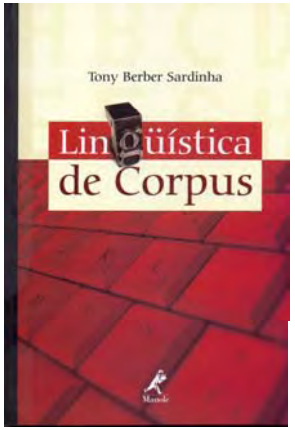


Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El papel de los corpus en la investigación en lingüística

La lingüística de corpus

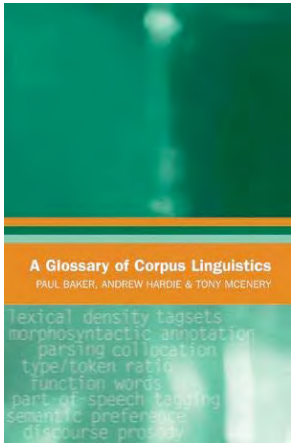
- **Manuales**
  - **BERBER SARDINHA, T. (2004) *Lingüística de Corpus*. Barueri, São Paulo: Editora Manole**



Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

La lingüística de corpus

- **Glosarios**
  - **BAKER, P. - HARDIE, A. - McENERY, T. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press (Glossaries in Linguistics)**



La lingüística de corpus

- **Antologías**
  - **SAMPSON, G.- McCARTHY, D. (Eds.) (2004) *Corpus Linguistics: readings in a widening discipline*. London - New York: Continuum International**

El papel de los corpus en la lingüística

✓La lingüística de corpus

✓La lingüística con corpus

La lingüística con corpus

- **La lingüística con corpus**
  - **Análisis de datos lingüísticos obtenidos de un corpus**
  - **Utilización de corpus en áreas de aplicación de la lingüística**

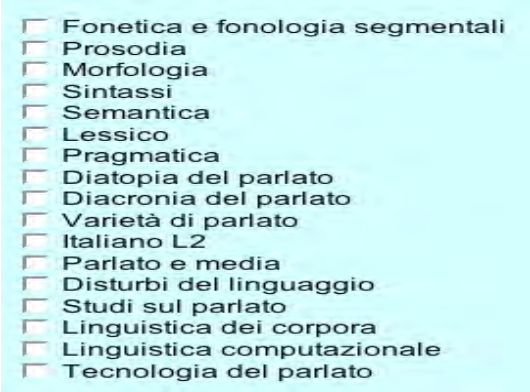
La lingüística con corpus

- **Teoría y descripción lingüística**
  - **Fonética**
  - **Fonología**
  - **Morfología**
  - **Lexicología**
  - **Sintaxis**
  - **Semántica**
  - **Pragmática**
  - **Análisis del discurso**
  - **Lingüística del texto**

La lingüística con corpus

- **Investigación en lingüística aplicada**
  - **Adquisición de segundas lenguas (L2)**
  - **Adquisición de la primera lengua (L1)**
  - **Patologías del lenguaje y del habla**
  - **Sociolingüística**
  - **Lingüística diacrónica**
  - **Lingüística contrastiva**
  - **Traductología**
  - **Lingüística judicial**
  - **Comunicación mediatizada por ordenador**
  - **Documentación de lenguas minorizadas**

La lingüística con corpus



`http://www.parlaritaliano.it/`

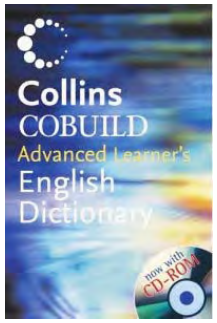
La lingüística con corpus

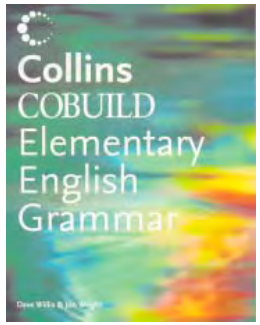
- **Productos en lingüística aplicada**
  - **Diccionarios de L2**
  - **Gramáticas de L2**
  - **Herramientas de enseñanza de lenguas asistida por ordenador**

La lingüística con corpus

Collins Cobuild

`http://www.collins.co.uk/books.aspx?group=151`





El papel de los corpus

✓El papel de los corpus en la lingüística

✓El papel de los corpus en las tecnologías lingüísticas

El papel de los corpus en las tecnologías lingüísticas

- Los recursos lingüísticos (LR, *Language Resources*)
  - Orígenes en la lingüística computacional (informática e ingeniería)
  - Auge debido a las técnicas de aprendizaje automático

El papel de los corpus en las tecnologías lingüísticas

- Los recursos lingüísticos
  - Corpus textuales, corpus orales con la señal sonora, colecciones de textos, colecciones de transcripciones, la web…
  - Corpus diseñados específicamente para el desarrollo de una aplicación

El papel de los corpus en las tecnologías lingüísticas

- Los recursos lingüísticos
  - Cuando existen herramientas, anotados de forma (semi)automática

El papel de los corpus en las tecnologías lingüísticas

- Desarrollo de tecnologías lingüísticas
  - Tecnologías del habla: conversión de texto en habla, reconocimiento automático del habla, sistemas de diálogo
  - Tecnologías del texto: analizadores morfológicos, sintácticos y semánticos, gramáticas computacionales, redes léxico-semánticas y ontologías

El papel de los corpus en las tecnologías lingüísticas

- Productos en tecnologías lingüísticas
  - Sistemas de dictado automático
  - Correctores ortográficos, gramaticales y de estilo
  - Sistemas de traducción automática y de traducción asistida
  - Herramientas para la recuperación de información

El papel de los corpus en las tecnologías lingüísticas

- Asociaciones
  - ELRA, European Language Resources Association `http://www.elra.info/`

El papel de los corpus en las tecnologías lingüísticas

- Congresos
  - LREC, International Conference on Language Resources and Evaluation
    - Granada (1998), Atenas (2000), Las Palmas (2002), Lisboa (2004), Génova (2006)

`http://www.lrec-conf.org/`

El papel de los corpus en las tecnologías lingüísticas

- Centros de distribución
  - ELDA, Evaluations and Language Resources Distribution Agency

`http://www.elda.org/`

El papel de los corpus en las tecnologías lingüísticas

- Centros de distribución
  - LDC, Linguistic Data Consortium

`http://www.ldc.upenn.edu/`

Los corpus como recurso compartido para la investigación lingüística

- ✓El papel de los corpus
- ✓**El eterno problema de los recursos**
- ✓Los recursos compartidos
- ✓El futuro de los recursos lingüísticos

El eterno problema de los recursos

✓La financiación

✓La accesibilidad

El eterno problema de los recursos

✓La financiación

✓La accesibilidad

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El eterno problema de los recursos
La financiación

- Los recursos (lingüísticos) requieren recursos (económicos)
- Recursos con financiación pública
  - ⧫ Universidades
- Recursos con financiación privada
  - ⧫ Empresas

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El eterno problema de los recursos
La financiación

LC-STAR Catalan phonetic lexicon

The LC-STAR Catalan phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission and the Spanish Government.

Production was performed at the Technologies and Applications of Language and Speech Center (TALP) of the Universitat Politècnica de Catalunya (UPC) (Spain). The owner of the database is UPC.

The lexicon comprises more than 100,000 words, including a set of more than 45,000 common words and a set of more than 45,000 proper names (including person names, family names, cities, streets, companies and brand names) with phonetic transcriptions in SAMPA.

The lexicon is provided in XML format. The database is stored on 1 CD.

ELRA Catalog Reference : S0207

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

El eterno problema de los recursos
La financiación

Members prices

- **Academic - Commercial 22.000 EUR**
- **Academic - Research 14.250 EUR**
- **Commercial - Commercial 22.00 EUR**
- **Commercial - Research 22.00 EUR**

Non-member prices

- **Academic - Commercial 29.250 EUR**
- **Academic - Research 38.700 EUR**
- **Commercial - Commercial 21.500 EUR**
- **Commercial - Research 29.250 EUR**

El eterno problema de los recursos
La financiación

VIT, Venice Italian Treebank

The VIT, Venice Italian Treebank is the effort of the collaboration of people working at the Laboratory of Computational Linguistics of the University of Venice in the years 1995-2005. It is partly the result of annotation carried out internally with no specific project in mind and no financial support. This work was partly related to the development of a lexicon, a morphological analyzer, a tagger, a deep parser of Italian. All these resources were finally ready at the beginning of the '90s when the LCL got involved in the first national projects.

The VIT contains about 272,000 words distributed over six different domains, and this is what makes it so relevant for the study of the structure of Italian language. In addition, some 60,000 tokens of spoken dialogues in different Italian varieties were annotated.

The annotation follows general X-bar criteria with 29 constituency labels and 102 PoS tags. VIT is also made available in a broad annotation version with 10 constituency labels and 22 PoS tags for machine learning purposes.

ELRA Catalog Reference: W0040

El eterno problema de los recursos
La financiación

Members prices

- **Academic - Commercial 7.000 EUR**
- **Academic - Research 3.000 EUR**
- **Commercial - Commercial 7.000 EUR**
- **Commercial - Research 7.000 EUR**

Non-member prices

- **Academic - Commercial 10.000 EUR**
- **Academic - Research 4.000 EUR**
- **Commercial - Commercial 10.000 EUR**
- **Commercial - Research 10.000 EUR**

El eterno problema de los recursos

✓La financiación

✓**La accesibilidad**

El eterno problema de los recursos
La accesibilidad

- **Para que un recurso pueda utilizarse debe conocerse su existencia y estar accesible**
- **Una historia española…**

El eterno problema de los recursos
La accesibilidad

Informe sobre recursos lingüísticos para el español (I): Corpus escritos y orales disponibles y en desarrollo en España. **Alcalá de Henares: Instituto Cervantes, 1994.**

Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en España. **Alcalá de Henares: Observatorio Español de Industrias de la Lengua, Instituto Cervantes, 1996.**

El eterno problema de los recursos
La accesibilidad

"Sería imprescindible disponer, en el marco nacional, de fuentes de información completas sobre los recursos y herramientas existentes y en desarrollo […] y estudiar la posibilidad de difusión de modo que se respetaran tanto los intereses de los centros que los han desarrollado como los requisitos que impone la financiación pública de los proyectos a través de los que se han conseguido fondos para su creación."

LLISTERRI, J.- GARRIDO, J.M. (1998) "La ingeniería lingüística en España", in *El español en el mundo. Anuario del Instituto Cervantes. 1998*, Madrid: Arco/Libros, 1998, pp. 299-391.
`http://cvc.cervantes.es/obref/anuario/anuario_98/llisterri/default.htm`

El eterno problema de los recursos
La accesibilidad

- **RILE, Servidor de Recursos para el desarrollo de la Ingeniería Lingüística en Español**
- **Ministerio de Industria y Energía, Programa de Fomento de la Tecnología Industrial, Iniciativa de Apoyo a la Tecnología, la Seguridad y la Calidad Industrial**
- **1.01.1999 - 31.03.2000**

El eterno problema de los recursos
La accesibilidad

"En este sentido, es urgente una acción encaminada a coordinar los corpora y las herramientas que existen en la actualidad y dedicar los recursos necesarios para dotar a las lenguas de España de dichos instrumentos, que tendrá una gran utilidad de investigación para diversas disciplinas, y estimulará además la investigación sobre dichas lenguas, no sólo por parte de grupos de España, sino del mundo entero."

***Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (2004-2007) Vol. II: Áreas prioritarias*. Presidencia de Gobierno, Oficina de Ciencia y Tecnología - Comisión Interministerial de Ciencia y Tecnología pp. 452-453.**

El eterno problema de los recursos
La accesibilidad

"En consecuencia, se propone inventariar y estudiar los corpora existentes, investigar en las bases metodológicas que permitan la compatibilidad entre ellos así como establecer criterios uniformes de etiquetamiento y configuración de sus contenidos textuales, no sólo en lo que atañe a los materiales ya existentes, sino también a los de los nuevos corpora cuya creación se vaya a promover. Finalmente se sugiere también facilitar el acceso general a estas importantes bases documentales, por ejemplo mediante la creación y mantenimiento de una página web desde donde se pueda realizar consultas a todos ellos."

***Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (2004-2007) Vol. II: Áreas prioritarias*. Presidencia de Gobierno, Oficina de Ciencia y Tecnología - Comisión Interministerial de Ciencia y Tecnología pp. 452-453.**

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los corpus como recurso compartido para la investigación lingüística

- ✓El papel de los corpus
- ✓El eterno problema de los recursos
- ✓**Los recursos compartidos**
- ✓El futuro de los recursos lingüísticos

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los recursos compartidos

- ✓**La creación de estándares y de recursos compartidos**
- ✓**Algunos estándares actuales**

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los recursos compartidos

✓La creación de estándares y la distribución de recursos compartidos

✓Algunos estándares actuales

- Para compartir, es necesario que los recursos estén estandarizados (o que, como mínimo, exista una compatibilidad entre estándares diferentes)

- Financiación de la Unión Europea
  - Creación y difusión de estándares
  - Creación y difusión de recursos estandarizados
  - Estudios y redes para la difusión de recursos compartidos

- 1991-1993 NERC, Network of European Reference Corpora
- 1993-1995 RELATOR, European Linguistic Resources Repository Network

Los recursos compartidos
La creación de estándares y la distribución de recursos compartidos

- 1993-1995 EAGLES, Expert Advisory Group on Language Engineering Standards
- 1996-1999 EAGLES II

`http://www.ilc.cnr.it/EAGLES96/home.html`

Los recursos compartidos
La creación de estándares y la distribución de recursos compartidos

- 1994-1995 PAROLE, Preparatory Action for Linguistic Resources Organization for Language Engineering
- 1996-1997 LE-PAROLE, Language Engineering - Preparatory Action for Linguistic Resources Organization for Language Engineering

Los recursos compartidos
La creación de estándares y la distribución de recursos compartidos

- 2000-2002 ISLE, International Standards for Language Engineering

`http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm`

Los recursos compartidos
La creación de estándares y la distribución de recursos compartidos

- 2001-2003 ENABLER, European National Activities for Basic Language Resources

`http://www.enabler-network.org/`

- 2003-2005 INTERA, Integrated European language data Repository Area

`http://www.mpi.nl/INTERA/`

Los recursos compartidos
La creación de estándares y la distribución de recursos compartidos

- 2005-2007 LIRICS, Linguistic Infrastructure for Interoperable Resources and Systems

`http://lirics.loria.fr/`

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los recursos compartidos
Algunos estándares actuales

✓La creación de estándares y la distribución de recursos compartidos

✓Algunos estándares actuales

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los recursos compartidos
Algunos estándares actuales

- Estándares de codificación textual
  - TEI, Text Encoding Initiative

  `http://www.tei-c.org/`

Universitat Autònoma de Barcelona
Joaquim Llisterri
Grup de Fonètica, Departament de Filologia Espanyola

Los recursos compartidos
Algunos estándares actuales

- Estándares de recogida de corpus orales en el ámbito de las tecnologías del habla
  - SpeechDat

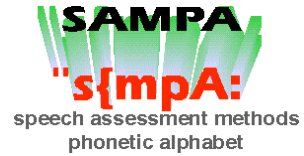
`http://www.speechdat.org/`

Los recursos compartidos
Algunos estándares actuales

- Transcripción fonética segmental
  - IPA, International Phonetic Association `http://www.arts.gla.ac.uk/IPA/ipa.html`

Los recursos compartidos
Algunos estándares actuales

- Transcripción fonética segmental
  - SAMPA, Computer Readable Phonetic Alphabet `http://www.phon.ucl.ac.uk/home/sampa/index.html`



Los recursos compartidos
Algunos estándares actuales

- Transcripción fonética suprasegmental
  - ToBI, Tones and Break Indices `http://www.ling.ohio-state.edu/~tobi/`
  - INTSINT, International Transcription System for Intonation `http://aune.lpl.univ-aix.fr/~hirst/home.html`

➧ SAVY R. - CROCCO, C. (2006) *Analisi prosodica: teorie, modelli e sistemi di annotazione, Atti del II convegno nazionale AISV*. Salerno, 30 novembre-2 dicembre 2005. Padova: EDK Editore. `http://www.parlaritaliano.it/aisv2005/home/proceedings.htm`

Los recursos compartidos
Algunos estándares actuales

- **Transcripción del discurso y de la conversación**
  - **EDWARDS, J. A.- LAMPERT, M. D. (Eds.) (1993) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates.**
  - **LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) (1995) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.**
  - **PraTiD: un sistema di annotazione pragmatica di dialoghi task-oriented**
    `http://www.parlaritaliano.it/`

Los recursos compartidos
Algunos estándares actuales

Du Bois et al 1990	Short pause	Longer pause	Timed pause
	..	...	... (1.5)
MacWhinney (1991)	Short pause	Longer pauses	Timed pause
	#	##, ###, #long	#1_5
Rosta (1990)	Short pause	Long pause	
	<,>	<,,>	
Svartvik and Quirk (1980)	Brief pause	Unit pause	Longer pauses
	.	_	_. __.

THOMPSON, P. (2005) "Spoken Language Corpora", in WYNNE, M. (Ed.) Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books: 59-70. `http://ahds.ac.uk/guides/linguistic-corpora/chapter5.htm`

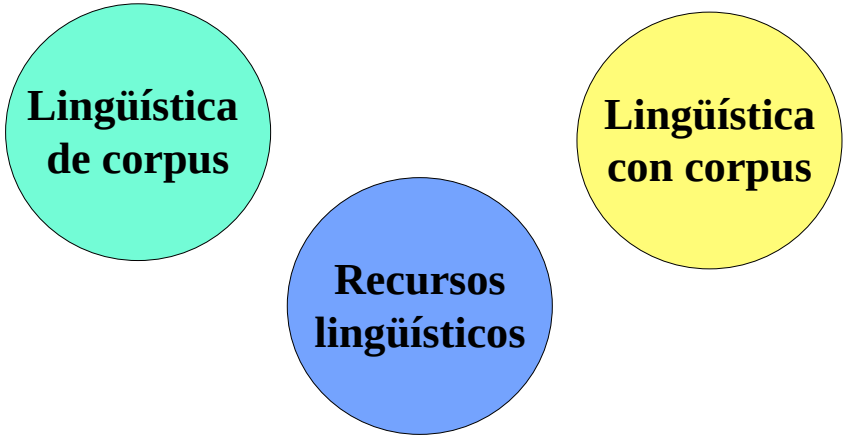
Los recursos compartidos
Algunos estándares actuales

- **Anotación sintáctica**
  - **LEECH, G.- BARNETT, R.- KAHREL, P. (1996) *Preliminary Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-SASG1.8, March 1996. `http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html`**
  - **…**
  - **ANANAS: annotazione e analisi sintattica**
    `http://www.parlaritaliano.it/`

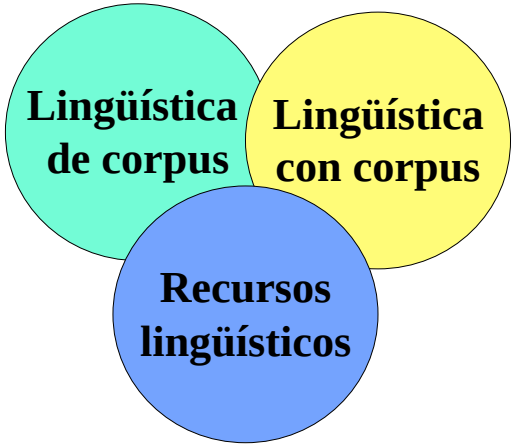
Los corpus como recurso compartido para la investigación lingüística

- ✓El papel de los corpus
- ✓El eterno problema de los recursos
- ✓Los recursos compartidos
- ✓**El futuro de los recursos lingüísticos**

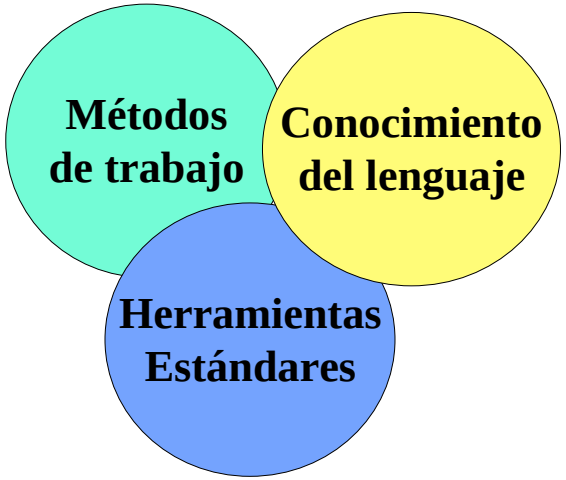
El futuro de los recursos lingüísticos



El futuro de los recursos lingüísticos



El futuro de los recursos lingüísticos



El futuro de los recursos lingüísticos

- Las herramientas de anotación requieren conocimiento lingüístico
- La anotación permite "trabajar" realmente con un corpus desde una perspectiva lingüística

El futuro de los recursos lingüísticos

- **Recursos con formato, codificación y anotación/etiquetado estandarizados o compatibles**
- **Recursos en red preparados para realizar consultas**
- **Recursos locales para el desarrollo de tecnologías o el estudio detallado**

Los corpus como recurso compartido para la investigación lingüística

✓El papel de los corpus
✓El eterno problema de los recursos
✓Los recursos compartidos
✓El futuro de los recursos lingüísticos

Los corpus como recurso compartido para la investigación lingüística

BERMEJO, I.- CARRERAS, X.- CASTELL, N.- CASTELLÓN, I.- COELLO, E.- GONZALO, J.- KALFON, N. - MARTÍ, M.A.- RODRÍGUEZ, S.- PADRÓ, L.- PEÑAS, A.- READ, T.- VERDEJO, M.F. (2000) "RILE: Servidor de Recursos para el desarrollo de la Ingeniería Lingüística en Español", *Procesamiento del Lenguaje Natural*, 26: 141-142. http://www.sepln.org/revistaSEPLN/revista/26/bermejo.pdf

EDWARDS, J. A.- LAMPERT, M. D. (Eds.) (1993) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

FILLMORE, C. J. (1992) "'Corpus Linguistics' or 'Computer-aided armchair linguistics'", in SVARTVIK, J. (Ed.) *Directions in Corpus Linguistics. Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*. Berlin - New York: Mouton de Gruyter. pp. 35-66.

LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) (1995) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

LLISTERRI, J.- GARRIDO, J.M. (1998) "La ingeniería lingüística en España", in *El español en el mundo. Anuario del Instituto Cervantes. 1998*, Madrid: Arco/Libros, 1998, pp. 299-391. http://cvc.cervantes.es/obref/anuario/anuario_98/llisterri/default.htm

SAVY R. - CROCCO, C. (2006) *Analisi prosodica: teorie, modelli e sistemi di annotazione, Atti del II convegno nazionale AISV*. Salerno, 30 novembre-2 dicembre 2005. Padova: EDK Editore. http://www.parlaritaliano.it/aisv2005/home/proceedings.htm

SINCLAIR, J. (2005) "Corpus and Text - Basic Principles", in WYNNE, M. (Ed.) (2005) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

THOMPSON, P. (2005) "Spoken Language Corpora", in WYNNE, M. (Ed.) Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books: 59-70. http://ahds.ac.uk/guides/linguistic-corpora/chapter5.htm

Parlaritaliano.it
Università degli Studi di Salerno, 26 de febrero de 2007

Los corpus como recurso compartido para la investigación lingüística

Joaquim Llisterri

Grup de Fonètica, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona

http://liceu.uab.cat/~joaquim