

 <b>U</b> UNIVERSITAT DE BARCELONA <b>B</b>	<b>Pla docent de l'assignatura</b> <b><i>Pendent de revisió lingüística</i></b>
---	--

### Dades generals de l'assignatura

**Nom de l'assignatura:** Lingüística de corpus  
**Codi de l'assignatura:** 564884  
**Curs acadèmic:** 2008-2009  
**Coordinació:** MIGUEL ANGEL CENTELLES VELILLA  
**Departament:** Facultat de Biblioteconomia i Documentació  
**Departament:** Dept. Biblioteconomia i Documentació

### Hores estimades de dedicació a l'assignatura

<b>Factor hores/ECTS</b>	<input type="text" value="25"/>
Hores de treball dirigit	<input type="text"/>
Hores d'aprenentatge autònom	<input type="text"/>
Hores presencials	<input type="text"/>
<b>Hores totals de treball de l'alumnat</b>	<input type="text" value="0"/>
<b>Crèdits ECTS</b>	<input type="text" value="0"/>

### Competències que es desenvolupen en l'assignatura

- avaluació de l'adequació d'un corpus determinat a les necessitats d'una recerca
- localització de casuística en un corpus en base a

1. familiarització amb el seu etiquetari o
2. ús d'expressions regulars (corpus no enriquits)

- obtenció de textos per a la creació, increment o actualització de corpus
- creació, gestió i explotació d'un corpus personal per a la recerca o la docència (fins a 0,5Mw)

- transferència de formats propietaris a estàndards públics dels textos a incorporar i viceversa amb els resultats obtinguts
- avaluació dels beneficis eventuais d'una migració entre estàndards d'etiquetatge
- ús i manteniment d'eines d'enriquiment

### Blocs temàtics de l'assignatura

1. corpus orals
2. corpus escrits + REGEX + vincles
3. obtenció de textos: WGET + P. Gutenberg + cercadors
4. wordsketch (inclòs bootcat)
5. SCP

### Avaluació acreditativa dels aprenentatges de l'assignatura

L'avaluació està basada en tres components, que són

- treballs dels seminaris de pràctiques
- examen sobre les lectures amb valoració de l'aportació personal als seminaris de lectures
- projecte personal de corpus per a la recerca pròpia.

L'estudiant podrà optar per ser avaluat en totes tres modalitats o només en una o dues; si ho fa només en una, cal que la seva aportació estigui perfecta per arribar a 5 sobre 10; si ho fa en dues, pot arribar a un 8 com a màxim; si ho fa en totes tres, el seu sostre de qualificació és el 10 amb opció a matrícula. Doncs, aquesta proposta d'avaluació està dissenyada pensant en un prototip d'estudiant que treballara en dues de les tres línies proposades.

Els estudiants que vulguin aprofundir en la matèria en la línia Bolonya participant de projectes de l'IULA tenen diverses opcions. En algun cas, fins i tot es pot pensar en organitzar el pràcticum (aquells que n'hagin de fer i encara el tinguin pendent) en algun dels projectes següents:

- constitució d'un corpus oral orientat a la recerca en ortologia catalana
- constitució d'un corpus alienat sobre medicina de l'esport cat/cast
- explotació d'un corpus per alimentar un diccionari d'aprenentatge quant als noms de massa.

Els estudiants que tinguin altres inquietuds d'aquest estil, poden demanar hora per tal que analitzem conjuntament si hi ha alguna possibilitat de trobar-los encaix en algun projecte real que s'ajusti a les seves motivacions.

<b>Fonts d'informació bàsiques de l'assignatura</b>
---

**Llibre**

ALONSO, Araceli; DE YZAGUIRRE, Lluís; FOLGUERÀ, Rosanna, TEBÉ, Carles. (2002). "La mesura de la implantació terminològica: dades, variables i resultats". Actes de la I Jornada sobre Terminologia i Serveis Lingüístics. Barcelona, 18 de maig de 2001. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 123-136.

ALONSO, Araceli; CABRÉ, Teresa; DE YZAGUIRRE, Lluís; TEBÉ, Carles (2002). "La utilización de corpus paralelos alineados en la docencia de la traducción y de los lenguajes de especialidad ". Dins Iglesias, L.; Doval, S. (ed.) Studies in Contrastive Linguistics. Proceedings of the Second International Contrastive Linguistics Conference. Santiago de Compostela: Publicacions de la Universidade de Santiago de Compostela. 71-82.

BARBERA, Manuel; CORINO, Elisa; ONESTI, Cristina. (2007). Corpora e linguistica in rete. 1a. edic. Perugia: Guerra Edizioni.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). Corpus linguistics: investigating language structure and use. 1a. edic. Cambridge: Cambridge University Press. [signature: P128.C68 B53 1998]

DE YZAGUIRRE, Lluís; MATAMALA, Anna; CABRÉ, Teresa (2001). " El lematizador 'PALIC' del IULA (UPF)". Trabajos en lingüística aplicada, AESLA. Barcelona.

DE YZAGUIRRE, Lluís, BACH, Carme; MATAMALA, Anna; CASTILLO, Núria; USTRELL, Eugènia (2001). " AMBILIC, el desambiguador lingüístico del Corpus del IULA". Trabajos en lingüística aplicada, AESLA. Barcelona.

DE YZAGUIRRE, Lluís; TORNER, Sergi; MATAMALA, Anna (2001). " El tratamiento automático de las ambigüedades segmentales del castellano". Trabajos en lingüística aplicada, AESLA. Barcelona.

DE YZAGUIRRE, Lluís; RIBAS, Marta; VIVALDI, Jorge; CABRÉ, M.<sup>a</sup> Teresa (2000). "Some Technical Aspects About Aligning Near Languages". Dins Gabrilidou, M. et al (ed.) (2000). Second International Conference on Language Resources and Evaluation LREC-2000. Proceedings, vol I. Atenes, 31 maig-2 juny 2000. Atenes: National Technical University of Athens Press. 545-548.

DE YZAGUIRRE, Lluís; TEBÉ, Carles; ALONSO, Araceli; FOLGUERÀ, Rosanna (2001). "El seguimiento de la implantación de términos vía Internet: estrategias de cálculo y control". Dins Correia, M. Terminologia e Indústrias da Língua. Lisboa: ILTEC, União Latina, Fundação Calouste Gulbenkein. 323-336.

DE YZAGUIRRE, Lluís; TEBÉ, Carles; ALONSO, Araceli; FOLGUERÀ, Rosanna (2002). "L'anàlisi lexicomètrica dels textos especialitzats: un terreny d'intersecció entre terminologia, documentació i traducció". Dins Cabré, T. i Domènech, M. (eds.) (2002).

Terminologia i Serveis Lingüístics. I Jornada de Terminologia i Serveis Lingüístics (18 de maig de 2001). Barcelona: Institut Universitari de Lingüística Aplicada-Universitat Pompeu Fabra. 151-161.

KILGARRIFF, Adam; RYCHLY, Pavel; TUGWELL, David (2004). "The Sketch Engine". Dins Williams, G. i Vessier, S. (ed.) (2004). Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Lorient: Université de Bretagne. 105-116. [<http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engineelx04.pdf?format=raw>]

KILGARRIFF, Adam; GREFENSTETTE, Gregory (2003). "Introduction to the Special Issue on Web as Corpus." Computational Linguistics 29 (3). (Also guest editors for the Special Issue). Reprinted in Practical Lexicography: a Reader. Fontenelle, editor. Oxford University Press. 2008. [<http://www.lexmasterclass.com/people/Publications/2003-KilgGrefenstette-WACIntro.pdf>]

MCENERY, Tony; WILSON, Andrew (2001). Corpus linguistics: An Introduction. 2a. edic. Edinburgh: Edinburgh University Press. [signatura: P302.3 .M345 2001] [<http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>]

REPPEN, Randi; FITZMAURICE, Susan; BIBER, Douglas (ed.) (2002). Studies in corpus linguistics, 9. Amsterdam: John Benjamins, cop. [signatura: P120.V37 U75 2002]

SINCLAIR, John (1991). Corpus, concordance, collocation. Oxford: Oxford University Press. [signatura: PE1611 .S46 1991]

STUART, Keith (2005). "New Perspectives on Corpus Linguistics". RAEL: Revista electrónica de lingüística aplicada, Nº 4. 180-191. [[http://dialnet.unirioja.es/servlet/revista?tipo\\_busqueda=CODIGO&clave\\_revista=6978](http://dialnet.unirioja.es/servlet/revista?tipo_busqueda=CODIGO&clave_revista=6978)]

TOGNINI-BONELLI, Elena (2001). "Corpus Linguistics at Work". Studies in Corpus Linguistics 6. Amsterdam: John Benjamins Publishing Co. [signature: P128.C68 T64]

TORRUELLA, Joan; LLISTERRI, Joaquim (1999). "Diseño de corpus textuales y orales". Dins Blecua, J. M.; Clavería, G.; Sánchez, C.; Torruella, J. (1999). Filología e informática. Nuevas tecnologías en los estudios filológicos. Barcelona: Seminario de Filología, Departamento de Filología Española, Universidad Autónoma de Barcelona, Editorial Milenio. 45-77. [[http://liceu.uab.es/~joaquim/publicacions/Torruella\\_Llisterri\\_99.pdf](http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf)]

WYNNE, Martin (ed.) (2005). Developing Linguistic Corpora: A Guide to Good Practice. AHDS Literature, Languages and Linguistics, Oxford: Oxbow Books. [<http://ahds.ac.uk/linguistic-corpora/>]

BARBERA, Manuel; CORINO, Elisa; ONESTI, Cristina. (2007). "Cosa è un corpus?". Dins BARBERA, M.; CORINO, E.; ONESTI, C. (ed.) (2007). Corpora e linguistica in rete. 1a. edic. Perugia: Guerra Edizioni.

BARNBROOK, Geoff. (1996). "First capture your data". *Language and Computers. A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics). 23-41.

BERBER, Tony (2000). "Lingüística de corpus: histórico e problemática". *D.E.L.T.A. Vol. 16, nº 2*. 323-367.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). "Methodology Box 1. Issues in corpus design". *Corpus linguistics: investigating language structure and use*. 1a. edic. Cambridge: Cambridge University Press. 246-250. [signature: P128.C68 B53 1998]

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). "Methodology Box 7. Statistical measures of lexical associations". *Corpus linguistics: investigating language structure and use*. 1a. edic. Cambridge: Cambridge University Press. 265-268. [signature: P128.C68 B53 1998]

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). "Methodology Box 8. The unit of analysis in corpus-based studies". *Corpus linguistics: investigating language structure and use*. 1a. edic. Cambridge: Cambridge University Press. 269-274. [signature: P128.C68 B53 1998]

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). "Investigating the use of language features". *Corpus linguistics: investigating language structure and use*. 1a. edic. Cambridge: Cambridge University Press. 19-131. [signature: P128.C68 B53 1998]

GRAEME, Kennedy (1998). "Variation in the distribution of modal verbs in the British National Corpus". En Reppen, R.; Fitzmaurice, S.; Biber, D. (ed.) (2002). *Studies in Corpus Linguistics, 9: Using Corpora to Explore Linguistic Variation*. Ámsterdam: John Benjamins Publishing Co. 73-90. [signatura: P120.V37 U75 2002]

HUNSTON, Susan (1998). "Pattern grammar, language teaching, and linguistic variation". En Reppen, R.; Fitzmaurice, S.; Biber, D. (ed.) (2002). *Studies in Corpus Linguistics, 9: Using Corpora to Explore Linguistic Variation*. Ámsterdam: John Benjamins Publishing Co. 167-183. [signatura: P120.V37 U75 2002]

KILGARRIFF, Adam; RYCHLY, Pavel; TUGWELL, David (2004). "The Sketch Engine". Dins Williams, G. i Vessier, S. (ed.) (2004). *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France, 6-10 July 2004*. Université de Bretagne Sud. Lorient: Université de Bretagne. 105-116. [<http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw>]

KILGARRIFF, Adam; GREFENSTETTE, Gregory (2003). "Introduction to the Special Issue on Web as Corpus". *Computational Linguistics* 29 (3). (Also guest editors for the Special Issue). Reprinted in *Practical Lexicography: a Reader*. Fontenelle, editor. Oxford University Press. 2008. [<http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf>]

McENERY, Tony; WILSON, Andrew (1996). "What is a corpus and what is in it?". *Corpus Linguistics*. 2a. edic. Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics). 2001. 29-32.

REPPEN, Randi (2002). "Using corpora to explore linguistic variation". Dins Reppen, R.; Fitzmaurice, S.; Biber, D. (ed.) (2002). *Studies in corpus linguistics*, 9. Amsterdam: John Benjamins, cop. [signatura: P120.V37 U75 2002]

SINCLAIR, John (1991). "Creación de corpus". Dins Vidal, J. (coord.) (1991). *Las industrias de la lengua*. Ávila: Fundación Sánchez-Albornoz. 95-107.

STUART, Keith (2005). "New Perspectives on Corpus Linguistics". *RAEL: Revista electrónica de lingüística aplicada*, Nº 4. 180-191.  
[[http://dialnet.unirioja.es/servlet/revista?tipo\\_busqueda=CODIGO&clave\\_revista=6978](http://dialnet.unirioja.es/servlet/revista?tipo_busqueda=CODIGO&clave_revista=6978)]

TORRUELLA, Joan; LLISTERRI, Joaquim (1999). "Diseño de corpus textuales y orales". Dins Blecua, J. M.; Clavería, G.; Sánchez, C.; Torruella, J. (1999). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología, Departamento de Filología Española, Universidad Autónoma de Barcelona, Editorial Milenio. 45-77.  
[[http://liceu.uab.es/~joaquim/publicacions/Torruella\\_Llisterri\\_99.pdf](http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf)]

VOLK, Martin (2002). "Using the web as Corpus for Linguistic Research". Dins Pajulasu, R. i Hennoste, T. (eds.) (2002). *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*. Tartu: Publications of the Department of General Linguistics 3. University of Tartu.

WYNNE, Martin (ed.) (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. AHDS Literature, Languages and Linguistics. Oxford: Oxbow Books.  
[<http://ahds.ac.uk/linguistic-corpora/>]