Institut Universitari de Lingüística Aplicada
Laboratori de Tecnologies Lingüístiques
Universitat **Pompeu Fabra**
iulalatel@upf.edu
935-421-344

Lingüística de Corpus

Guia docent

versió provisional núm 0.1

Professors: Lluís de Yzaguirre i Maura
lluis.de.yzaguirre@gmail.com
despatx 300B de Rambla (només en hores de visita)

Araceli Alonso Campo
araceli.alonso@upf.edu
despatx 300A de Rambla (només en hores de visita)

NB en condicions normals, els professors de l'assignatura rebrien fora de les hores concertades, però la situació interina de l'IULA al Rectorat, abans del trasllat a ca l'Aranyó fa que no siguin a Rambla més que per les hores estrictes de visita i que al Rectorat no tinguin condicions per atendre consultes o tutories. Doncs, aquells alumnes que necessitin excepcionalment alguna tutoria caldrà que la concertin per correu electrònic.

Continguts de l'assignatura

Teoria

- Lingüistica de corpus: entre la teoria, la pràctica i les aplicacions
  - exemples
- Tipus de corpus i estàndards
- Utilització de corpus
  - REGEX
- Creació de corpus i estratègies de processament
  - marcatge estructural
  - enriquiment
  - alienació

Pràctiques en cinc blocs de dues setmanes

- corpus orals
- corpus escrits + REGEX + familiarització amb els principals corpus
- obtenció de textos: WGET + P. Gutenberg + cercadors
- wordsketch (inclòs bootcat)
- SCP

Habilitats i competències

Aquesta assignatura, amb les limitacions d'un programa de deu setmanes, es proposa mostrar als estudiants principalment tres perfils relacionats amb els corpus, bo i introduint les habilitats i les competències que necessiten adquirir:

- investigador que s'incorpora a un equip de recerca que treballa en la constitució de corpus, sigui per detectar i incorporar textos, sigui per transformar-los als estàndards adoptats, sigui per enriquir-los amb els marcatges triats; en cas d'interessar-se per la programació informàtica, pot especialitzar-se en l'adaptació o fins i tot en la creació de les aplicacions d'enriquiment (lematitzadors, desambiguadors, *parsers*...);

- investigador que s'afegeix a un grup de lingüistes convencionals interessats a ampliar la seva recerca amb evidències i dades obtingudes de corpus, amb la funció d'intermediari, capaç de determinar quins corpus satisfan millor els requeriments de l'equip, com s'han d'interrogar i com s'han de transferir els resultats per integrar-los amb les dades de l'equip o en les publicacions de resultats;

- investigador individual que necessita constituir un corpus personal per a la seva recerca personal (tesi doctoral, alguna mena de peritatge...) i gestionar-lo amb eines de domini públic;

evidentment queden altres perfils d'usuaris potencials de corpus (professor de cursos d'idiomes instrumentals, lexicògraf, terminòleg...) i, en el cas que hi hagi estudiants que hi tinguin interès en especial, cercarem la manera de donar-los satisfacció.

Avaluació

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 medecina 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.

Bibliografia recomanada

ALONSO, A.; DE YZAGUIRRE, L.; FOLGUERÀ, R. , TEBÉ, C. (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, ps. 123-136. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ISBN: 84-477-0831-4.

ALONSO, A.; CABRÉ, T.; DE YZAGUIRRE, L.; TEBÉ, C. (2002) " La utilización de corpus paralelos alineados en la docencia de la traducción y de los lenguajes de especialidad ". En: Iglesias, L.; Doval, S. (ed.) Studies in Contrastive Linguistics. Proceedings of the Second International Contrastive Linguistics Conference, ps. 71-82. Santiago de Compostela: Publicacións de la Universidade de Santiago de Compostela. ISBN: 84-9750-027-X.

BARBERA, Manuel; CORINO, Elisa; ONESTI, C. (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, LL., A. MATAMALA, T. CABRÉ, " El lematizador 'PALIC' del IULA (UPF)" dins Trabajos en lingüística aplicada, AESLA, Barcelona, 2001 (ISBN 84-477-0733-4)

DE YZAGUIRRE, LL., C. BACH, A. MATAMALA, N. CASTILLO, E. USTRELL, " AMBILIC, el desambiguador lingüístico del Corpus del IULA" dins Trabajos en lingüística aplicada, AESLA, Barcelona, 2001 (ISBN 84-477-0733-4)

DE YZAGUIRRE, LL., S. TORNER, A. MATAMALA, " El tratamiento automático de las ambigüedades segmentales del castellano " dins Trabajos en lingüística aplicada, AESLA, Barcelona, 2001 (ISBN 84-477-0733-4)

DE YZAGUIRRE, LL.; RIBAS, M.; VIVALDI, J. I M.T. CABRÉ (2000) " Some Technical Aspects About Aligning Near Languages", comunicació presentada a la LREC-2000, Atenes, 31 maig-2 juny 2000. Publicat a Gabrilidou, M. *et al* (ed.) *Second International Conference on Language Resources and Evaluation. Proceedings*, vol I, Atenes: National Technical University of Athens Press, p. 545-548."

DE YZAGUIRRE, LL.; TEBÉ, C.; ALONSO, A. I R. FOLGUERÀ: " El seguimiento de la implantación de términos vía Internet: estrategias de cálculo y control ", en Correia, Margarita, "Terminologia e Indústrias da Língua", Lisboa, 2001, pàgs. 323-336, ISBN 972-9051-48-8

KILGARRIFF, Adam; RYCHLY, Pavel; TUGWELL, David (2004). "The Sketch Engine". En Williams, G. y 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.lexmasterclass.com/people/Publications/2003-KilgGrefenstette-WACIntro.pdf]

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

REPPEN, R.; FITZMAURICE, S.; BIBER, D. (ed.) (2002) *Studies in corpus linguistics, 9*. Ámsterdam: 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]

TORRUELLA, Joan y LLISTERRI, Joaquim (1999). "Diseño de corpus textuales y orales". En 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]

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/]

Bibliografia per al seminari

Per al seminari de lectures està previst treballar els capítols següents d'alguna de les publicacions precedents. S'hi podria arribar a incloure algun altre capítol de les mateixes o alguna altra publicacions o el text íntegre d'algun altre dels articles recollits a la bibliografia recomanada.

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

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.

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.

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; 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.

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

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]

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/]