Corpus
linguistics
19

BIBLIOTECA DE CATALUNYA. DADES CIP

Jornada de Corpus Lingüístics *(4a : 1996 : Barcelona)*
IV i V Jornades de Corpus Lingüístics, 1996-1997
Text en anglès, introducció en català. – Bibliografia
ISBN 84-477-0649-8
I. Cabré, M. Teresa, dir. II. Lorente, Mercè, dir.
III. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada IV. Jornada de Corpus Lingüístics (5a : 1997 : Barcelona) V. Títol: IV i V Jornades de Corpus Lingüístics, 1996-1997
1. Lingüística computacional – Congressos
801.3:681.3(061.3)

Direcció: M. Teresa Cabré
Coordinació: Mercè Lorente
Organització IV jornada: M. Teresa Turell
Organització V jornada: Enric Vallduví
Documentació: Mireia Ribera
Responsable de publicacions: Josep M. Fontana
Edició: M. Rosa Bayà, Helena Pàmpols, Montserrat Pibernat i Elisabet Ricart

Primera edició: setembre de 1998
© els autors
© Institut Universitari de Lingüística Aplicada
La Rambla, 30-32
08002 Barcelona

Disseny de la coberta: Cass
Impressió: PPU
Dipòsit legal: L-748-1998
ISBN: 84-477-0649-8

ÍNDEX

INTRODUCCIÓ

Aquest volum presenta les ponències incloses en les dues jornades sobre corpus lingüístics (quarta i cinquena edicions) organitzades per la Unitat de Recerca de Variació Lingüística de l'Institut Universitari de Lingüística Aplicada (IULA). Cal destacar el caràcter generalista, teòric i aplicat, d'ambdues jornades. El propòsit fou oferir una visió àmplia de la recerca que es fa actualment en àmbits lingüístics que són objecte d'estudi a la Unitat de Variació i que també són d'interès per a les altres unitats de recerca de l'IULA i per a especialistes en lingüística de corpus de fora de la Universitat Pompeu Fabra. Les Jornades de corpus lingüístics s'emmarquen dins de les activitats de la Xarxa Temàtica de Lingüística Aplicada, formada per grups de recerca dedicats a la constitució de corpus lingüístics i de bases de dades lèxiques.

La IV **Jornada de corpus lingüístics**, celebrada el 28 de maig de 1996, va versar sobre la constitució, l'etiquetatge i l'explotació dels corpus lingüístics des d'una perspectiva lingüística variacionista. La perspectiva de variació lingüística que va inspirar aquesta jornada implica, d'una banda, la consideració de la variació tant interna com externa i, de l'altra, un tractament de la variabilitat inherent del llenguatge natural que va més enllà d'una simple descripció, en el sentit que intenta establir les conseqüències d'aquesta variació en la modelació lingüística, millorar la nostra comprensió analítica del llenguatge, i al mateix temps contribuir al desenvolupament de noves eines de descripció i modelació del català i de les altres llengües que són objecte d'estudi a l'IULA.

La primera i segona ponències, *The HCRC Map Task Corpus: A design for many applications* i *The HCRC Map Task: Some results*, a càrrec d'Ellen Bard i Cathy Sotillo, respectivament, de la Universitat d'Edimburg, tenen a veure amb la constitució i explotació d'un corpus de diàleg a partir del *Map task*, un instrument desenvolupat al Human Communication Research Center d'aquesta universitat. Aquest corpus de diàleg, que forma part d'una base de dades més àmplia, la *Human Communication Dialogue Database*, té possibilitats d'explotació tant estructural, com pragmàtica i sociolingüística i, per tant, pot ser de gran interés per al *Corpus Tècnic* de l'IULA durant l'etapa d'etiquetatge i explotació dels nivells lingüístics esmentats en últim terme. La tercera ponència, *Theoretical and practical aspects of multilingual corpus construction: The MLCC experience*, a càrrec de Henry Thompson, també de la Universitat d'Edimburg, tracta dels *Multilingual Corpora for Cooperation* (MLCC) —projecte dut a terme pel Language Technology Group, també del Human Communication

Research Center—, que recull textos periodístics financers en sis llengües europees. Val a dir que aquesta ponència planteja sobretot qüestions tècniques i logístiques relacionades amb la gestió dels corpus lingüístics, sense menysprear la rellevància que aquesta experiència també pot representar en el vessant més científic del procés de constitució del *Corpus Tècnic* de l'IULA. La versió que s'inclou aquí d'aquestes tres primeres contribucions és una transcripció editada de la presentació oral feta per cada un dels ponents.

La quarta ponència, presentada per Susan Pintzuk de la Universitat de York, versa sobre el *Helsinki Corpus of Old English Texts* (anglès antic), conegut també com a *Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus*, per les universitats que han participat en la seva constitució i explotació. I finalment, l'última ponència, a càrrec d'Anthony Kroch i Ann Taylor de la Universitat de Pennsilvània, tracta el *Penn-Helsinki Parsed Corpus of Middle English* (anglès medieval). En el cas d'aquestes dues últimes ponències interessa remarcar les pautes innovadores que introdueixen pel que fa al disseny de corpus, els resultats de la recerca i les perspectives futures d'explotació.

La **V Jornada de Corpus Lingüístics** (10 d'abril de 1997) va tenir per tema la utilització dels corpus en la recerca semàntica i pragmàtica. Les contribucions incloses en aquesta jornada van oferir una visió completa i contrastada de l'estat de la qüestió en la recerca en etiquetatge i extracció d'informació pragmaticosemàntica: van cobrir tant l'espectre metodològic (codificació i etiquetatge, extracció, utilització teòrica i aplicada) com l'espectre teòric (des de la semàntica oracional a l'articulació retòrica dels textos).

Jacob Hoeksema, de la Universitat de Groningen, és un dels pioners en la utilització de corpus en recerca semàntica i és un expert de primer ordre en negació i ítems de polaritat. La seva ponència, *Corpus study of negative polarity items*, va posar en evidència les dificultats intrínseques que hi ha en l'etiquetatge semàntic dels corpus. Harald Baayen, investigador del prestigiós Institut Max Planck de Nijmegen, és un especialista en semàntica lèxica i morfosemàntica i té una sòlida experiència en la lingüística de corpus. La seva contribució, *Lexis, word frequencies, and text types*, va il·lustrar el bon ús dels corpus en la recerca semàntica però també va alertar sobre el mal ús i l'abús que se'n pot fer. Ellen Prince, directora del Departament de Lingüística a la Universitat de Pennsilvània i investigadora del Institute for Research for Cognitive Science de la mateixa universitat, és especialista en pragmàtica i es dedica des de fa anys a l'etiquetatge pragmàtic de corpus amb la finalitat d'extraure'n informació amb mètodes quantitatius. La seva ponència, *On the nature of metalinguistic intuitions and its methodological implications for pragmatic research*, va detallar la seva experiència en aquest camp i va posar en

relleu la importància dels corpus en la recerca pragmàtica. Els dos últims ponents, Julia Lavid i John Bateman, posseeixen una sòlida trajectòria de recerca en projectes sobre explotació pragmàtica de corpus finançats per la Unió Europea, com DANDELION (ESPRIT) i GIST (LRE). Les seves ponències, *The relevance of corpus-based research for contrastive linguistic and computational studies: thematization as an example* i *Using corpora for uncovering text organization: goals, requirements and methodologies*, van versar sobre la utilitat dels corpus en la recerca sobre organització textual i l'estructura retòrica dels textos.

Finalment, volem agrair la col·laboració de les següents institucions, que han fet possible l'organització d'aquestes dues jornades: la Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia (ajuts CO94-0722 i CO94-1165); el Comissionat per a Universitats i Recerca de la Generalitat de Catalunya (ajuts ARCS96-33 i ARCS97-46), i el Vicerectorat de Política Científica i Tercer Cicle de la Universitat Pompeu Fabra (ajuts P796.008 i P797.008).

M. Teresa Turell i Enric Vallduví
Organitzadors de les jornades