Corpus, corpora

a cura de

Lluís Payrató

Emili Boix

M. Rosa Lloret

Mercè Lorente

PPU

CORPUS, CORPORA

ACTES DEL 1r i 2n COL·LOQUIS LINGÜÍSTICS DE LA UNIVERSITAT DE BARCELONA (CLUB -1, CLUB -2)

CORPUS, CORPORA

ACTES DEL 1r i 2n COL·LOQUIS LINGÜÍSTICS DE LA UNIVERSITAT DE BARCELONA (CLUB -1, CLUB -2)

A cura de
Lluís Payrató, Emili Boix, M. Rosa Lloret i Mercè Lorente

PPU

Secció de Lingüística Catalana
Departament de Filologia Catalana
Universitat de Barcelona

Barcelona, 1996

Primera edició, 1996

PPU, S.A.
Promociones y Publicaciones Universitarias, S.A.
Muntaner, 200, entl. 2a 08036 Barcelona
Tel. (93) 209 53 40 Fax (93) 209 55 41

ISBN: 978-84-9168-680-4

ÍNDEX

Preàmbul

Els treballs aplegats en aquest volum recullen les contribucions fetes als dos primers col·loquis lingüístics de la Universitat de Barcelona, celebrats respectivament el 20 de desembre de 1993 i el 19 de desembre de 1994, i organitzats per la Secció de Lingüística Catalana del Departament de Filologia Catalana.[1]

El primer d'aquests dos col·loquis, amb el títol de "Corpus, corpora", va incloure els articles reunits en la primera part d'aquesta obra, i la seva presentació anava encapçalada pel text següent, que il·lustra prou bé les finalitats de la trobada:

> "La descripció d'una llengua no hauria de comptar exclusivament amb la intuïció (masculina o femenina) dels lingüistes. Ni tampoc basar-se només en enregistraments de parlants de valls d'aigües regalades o de companys de departaments universitaris. Els treballs de descripció i de prescripció lingüístiques han d'incorporar també materials escrits i orals que representin bona part de la llengua (la de la vida de cada dia i dels mitjans de comunicació de masses).
> La descripció del català contemporani a partir de materials representatius del seu repertori de varietats geogràfiques, històriques, socials i funcionals
>
> *és complexa* perquè exigeix una inversió pressupostària de llarg abast i un esforç constant de coordinació entre investigadors de diferents ciències del llenguatge, però sobretot
>
> *és possible* tècnicament, informàticament i humanament, i

[1] L'organització d'aquests col·loquis i la seva publicació s'han vist afavorits pel projecte de recerca PB90-0505, i també han comptat amb el suport de la Facultat de Filologia i de la Divisió de Ciències Humanes i Socials de la Universitat de Barcelona, i de la CIRIT de la Generalitat de Catalunya.
Els editors volen fer constar el seu agraïment a Carme Bach per la seva contribució en el procés d'edició d'aquest volum.

> *és necessària* per descriure l'estructura i el canvi lingüístics més adequadament, per avançar teòricament i per presentar propostes prescriptives al màxim raonades.
>
> Les experiències de corpus ja constituïts, com, per exemple, el British National Corpus i el Corpus du Français Parlé de Montréal, mostren que projectes similars, adaptats a les nostres necessitats, són factibles."

Les ponències de José Manuel Blecua (sobre els corpus, en general, i en particular sobre els escrits) i de Joaquim Llisterri (sobre els orals) ofereixen un estat de la qüestió i un conjunt de reflexions que serveixen per *situar-se* en una temàtica diversificada i desenvolupada amb ímpetu durant els últims anys. Joaquim Rafel, per una banda, i Emili Boix, M. Teresa Cabré, Mercè Lorente i Lluís de Yzaguirre, per una altra, completaven aquest panorama amb l'exposició de dos corpus concrets: el de l'Institut d'Estudis Catalans i el de la Universitat de Barcelona, respectivament. Una taula rodona sobre recerques i aplicacions a partir de corpus lingüístics cloïa el col·loqui, i es publica transcrita com a final de la primera part del llibre.

El CLUB - 2, amb el títol de "Metodologia per a l'anàlisi de corpus lingüístics", es va plantejar com la continuació lògica del col·loqui anterior. Tal com es feia constar també en el text que li servia de presentació, volia ser un fòrum de discussió dels múltiples problemes pràctics amb que s'enfronten els *treballadors* dels corpus:

> "Com s'han de *recollir* i *seleccionar* les dades perquè el corpus sigui representatiu?
> Com s'han de *transcriure* i *representar* de manera homogènia els materials obtinguts?
> Com s'ha de *codificar* la informació lèxica, gramatical i discursiva de les dades seleccionades?
> Com s'ha d'*automatitzar* el tractament de les dades per a les diferents orientacions de la recerca?"

Així doncs, les ponències de la segona part del present volum recullen les aportacions de diferents especialistes en relació a temes específics dels corpus: la selecció de les dades (Cristina Sánchez), la transcripció (Lluís Payrató), els aspectes computacionals (Toni

Badia) i els informàtics (Lluís de Yzaguirre). Igualment, seguint la línia encetada en el primer col·loqui, es va presentar un exemple concret de constitució d'un corpus, en aquest cas el de la Universitat de València, en la ponència a càrrec d'Antonio Briz.

Confiem ara que la publicació conjunta de tots aquests textos referma l'objectiu de tots dos col·loquis: potenciar un camp de recerca fructífer i que té nombroses aplicacions al cas particular de la llengua catalana.

Lluís Payrató, Emili Boix, M. Rosa Lloret i Mercè Lorente

PRIMERA PART

Reflexiones al margen de los corpus escritos*

José Manuel Blecua

(Universitat Autònoma de Barcelona)

En la historia de la lingüística aparece como constante la relación entre la teoría y los datos, apasionante relación que se manifiesta de manera abierta o velada y constituye, íntimamente unida a la posibilidad de construcción de modelos teóricos, el aspecto más fructífero de las distintas perspectivas del siglo XX. Una nueva visión de los problemas lingüísticos aparece en la década de los años cincuenta cuando se añade una posibilidad más a la construcción de modelos con la aparición del concepto de *corpus*, concepto que posteriormente, gracias a las nuevas tecnologías aplicadas a la investigación lingüística, permitirá trabajar con formidables cantidades de datos reales. No es el momento aquí de tratar con detalle la historia del concepto; en el precioso libro editado por Karin Aijmer y Bengt Altenberg, *English Corpus Linguistics*, aparece un excelente trabajo de Geoffrey Leech, *The state of the art in corpus linguistics*, en el que se traza la historia del concepto y las polémicas que ha suscitado desde que se inició el trabajo de Randolph Quirk, en 1959, al que siguieron las primeras labores de informatización en el corpus de la Universidad de Brown (1961-64) de un millón de palabras, y en el de Londres-Lund, de 500.000 palabras del inglés hablado,[1] grupo que constituye la primera generación, según la clasificación de Leech; en la segunda época, aparecen tipos más complejos, como el *corpus* de Birmingham, dotado ya de 20 millones de ocurrencias, en el que se combinan

* El autor agradece vivamente la ayuda y los consejos prestados por Gerardo Arrarte, Joaquim Llisterri y Joan Torruella. La paciencia y bondad de Lluís Payrató han hecho possible que estas líneas vean finalmente la luz.

[1] Vid. el completo trabajo de J. Llisterri, *Els corpus lingüístics orals*, en este mismo volumen.

materiales procedentes de la lengua escrita y de la lengua hablada (Renoulf, 1987); como es perfectamente conocido, el interés por la lengua hablada en el *corpus* es muy posterior al de la lengua escrita *(NERC-1*, 1994, pág. 75-109). La tercera generación estará formada por el tipo de *corpus* de cientos de millones de palabras (Leech, 1991, pág. 10, Leech y Fligelstone, 1992), en el que se utilizan medios tecnológicos mucho más complejos (Burland, 1992).

Como ha escrito M. Alvar (1991), "Un *corpus* es un conjunto homogéneo de documentos de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) que se toman como modelo de un estado o nivel de lengua predeterminado, al cual representan." Como advierte el investigador citado, el *corpus* ideal sería el que contuviera todos los enunciados de la lengua, pero como ello no es posible, no queda más remedio que concebirlo como limitado, como modelo representativo; este modelo puede estar formado por documentos completos (*corpus textual*) o por fragmentos de textos (*corpus de referencia*); el planteamiento y distribución de los elementos que componen un *corpus* es un problema sumamente complejo y, desde luego, supone en gran medida los posibles resultados (vid. por ejemplo el trabajo de Lara y Ham Chande, 1974, y el capítulo sobre la materia en *NERC-1*, pág. 59-74). El corpus permite trabajar con datos reales, es un apoyo realista a la investigación y como consecuencia se pueden emprender nuevos trabajos descriptivos, tanto en el léxico, en la sintaxis, en el discurso o en la prosodia; se pueden comparar variedades de la misma época; como ejercicio, se puede comparar el léxico de la poesía de Garcilaso de la Vega con el diametralmente diferente que aparece, a la muerte del poeta, en los inventarios de su casa. El *corpus* presenta inigualables posibilidades en el desarrollo y en la verificación de analizadores morfológicos y sintácticos, lo mismo que puede ocurrir desde el punto de vista de los generadores. Los materiales de un *corpus* realizado y anotado con criterios científicos se pueden convertir con facilidad en elementos útiles para distintos campos de investigación: la enseñanza de la lengua materna, la enseñanza a extranjeros (como ha realizado J. de Kock), la lexicografía o la estadística, por ejemplo. En su explotación, se pueden plantear muchos casos de relación entre el hombre y la máquina, además de los problemas teóricos y aplicados que presenta la construcción de las muy diversas "herramientas" informáticas que se requieren para su correcto uso y su mejor aprovechamiento.

Son muchos y muy variados los problemas que plantea la realización y la explotación juiciosa de un *corpus*; algunos de ellos son comunes a toda investigación lingüística y afectan tanto a *corpus* de la lengua escrita como de la lengua hablada, como los problemas de la reducción de *variantes* a *invariantes*, mientras que otros nacen de la propia naturaleza de las máquinas o programas que originan una nueva metodología de investigación.

Uno de los primeros problemas que se le planteaba, ya hace muchos años, al lingüista tradicional, y que se le sigue planteando hoy al investigador que trabaja en un *corpus*, es el de la reducción de *variantes* a *invariantes*. Este problema, que parece inocente, es -sin duda alguna- el primero de todos: todas las lenguas naturales presentan en su realización un conjunto más o menos amplio, según las zonas o niveles, de realizaciones *fonéticas*, *morfológicas*, *sintácticas* o *léxicas* que es necesario reducir a unas unidades básicas, de carácter abstracto, creadas artificialmente por los científicos: fonemas, unidades morfológicas, oraciones (como resultado abstracto de plantear la relación entre *enunciados* que se producen en las realizaciones del lenguaje y su estructura abstracta). Dejo aparte, por ser de sobras conocido, el problema de la reducción de *ocurrencias* léxicas al concepto, sumamente abstracto, de *palabra*: formas de los verbos reducidos al infinitivo o formas del nombre al masculino singular, por ejemplo.

Este problema de la reducción de *variantes* a *invariantes* cobra hoy toda su actualidad y aparece reflejado en muy diversos aspectos de la investigación realizada con la ayuda de las nuevas tecnologías; aparecerá con toda su pujanza en las cuestiones relativas al análisis y la síntesis del habla; en los problemas que plantean las confecciones de concordancias, en la lematización de los diccionarios electrónicos, igual que ocurría con los diccionarios y glosarios tradicionales. Problema que se agrava todavía más en el tratamiento informático de los textos antiguos, tratamiento en el que varios especialistas sumamente competentes han propuesto que la única manera rápida de resolver este problema es la lematización por las formas canónicas actuales con un envío en la base de datos léxica a un campo de grafía en el que se introduce la forma concreta. Intimamente unido al problema anterior, de reducción de *variantes* a *invariantes*, incluso forma muchas veces parte de él, es el complejo problema de la segmentación; ya difícil en la lengua escrita, recuérdense ciertas críticas de la aplicación de la informática a los primeros diccionarios de frecuencias, que se agrava todavía más en el análisis de la lengua hablada.

Si es muy complejo plantear segmentaciones en frases dotadas de aparente ambigüedad en sus características lineales o de constituyentes: "gorras para niños de hule" / "Se hacen felpudos a medida de coco", todavía lo es más en el caso de los enunciados de la lengua hablada, ya que la coarticulación convierte a las lenguas naturales en instrumentos extraños en el mundo de la información.

Las lenguas naturales viven en la *variación*; reconocer esta realidad es extraordinariamente duro y afecta a los problemas que traen los tratamientos informáticos. Los lingüistas tradicionales, desde Saussure a Chomsky, han concebido las lenguas naturales despreciando variaciones que dependen de factores muy diversos: el espacio, el tiempo, el sexo, la edad, el oficio o profesión, el hablante, el oyente, la situación, el registro o el tema de que se trata. Hoy sabemos que existen correlaciones entre factores lingüísticos y factores sociales, y que estas correlaciones se manifiestan en aspectos tan diversos como la fonología (el seseo y el yeísmo) o el diverso tratamiento del léxico y de la gramática en los sublenguajes, dejando aparte la cuestión más compleja de los lenguajes reducidos, ferrocarriles, aviación, marina o técnica aeroespacial, por no citar más que algunos de los muy actuales.

Además, pues, del problema ya citado de la reducción de *variantes* a *invariantes*, y el del reconocimiento de la *variación* como una realidad intrínseca de las lenguas, aparece un desplazamiento en el estudio lingüístico actual que se manifiesta en el interés por la lengua hablada, además del tradicional que siempre ha despertado la lengua escrita. La historia de la autoridad textual y gramatical pedía que los estudios lingüísticos se realizasen siempre sobre ejemplos de los escritores reconocidos como "autoridades", tal como manifiesta todavía hoy el *Esbozo* académico, aunque se amplíe la nómina y los escritores sean contemporáneos (vid. pág. 281, nota 36, sobre la preferencia de *hendir* sobre *hender*, en la que se cita a como autoridad a Laín Entralgo.) A nadie se le oculta la enorme dificultad de trabajar con la lengua hablada, como consecuencia de una falta de tradición, de medios, de trabajos previos, incluso de la llamada "paradoja del investigador"; no se trata únicamente de incorporar las variedades de la lengua, contextuales, espaciales o sociales, sino de incorporar y anotar con anotaciones inequívocas, también, el diálogo y la conversación (como ejemplo de dificultades y de fronteras borrosas en el marco de la enunciación, de la representación escrita de un fragmento oral, en el entremés cervantino de *La guarda cuidadosa*, vid. Huyn-Armanet y

Pineira-Tresmontant, 1989). A estos problemas sumamente generales (reducción de variantes a invariantes, variación lingüística, presencia de la lengua hablada como parte importante del modelo de representatividad), habría que añadir cuestiones muy específicas, propias sobre todo de la *lengua literaria*, que hacen que el concepto de *texto* que habitualmente utilizan los lingüistas quede superado en diferentes direcciones por múltiples problemas de los que citaré únicamente tres tipos:

a) presencia del *multilingüismo* en la lengua cotidiana y, fundamentalmente, en la obra literaria, fenómeno antiguo y de extraordinaria complejidad; como ejemplo, véase el clásico trabajo de Jörder sobre esta cuestión en los sonetos de Lope de Vega (Jörder, 1936, pág. 236 y 238-267);

b) la naturaleza de ciertos textos que no obedecen al concepto tradicional del carácter lineal de la escritura (cf. el excelente trabajo de C. M. Sperberg-McQueen, 1991, para la Edad Media, pero válido para tantas cuestiones actuales de la "escritura en libertad".) Problemas extraordinariamente atractivos para los amantes de determinadas obras literarias y que el especialista en *corpus* que contenga documentos literarios no puede obviar, so capa de simplificar hasta extremos que pueden falsear totalmente en su esquematismo períodos literarios como el Barroco o las vanguardias contemporáneas (baste recordar los emblemas y los elementos de arquitectura efímera, como certámenes, arcos o túmulos). Puesto que me propongo estudiar por extenso este tratamiento en un *corpus* en lugar más apropiado, remito ahora al curioso trabajo de J. Romera (1980) sobre el monje riojano Vigilán y al estudio general de Víctor Infantes de Miguel en la misma publicación. Aurora Egido, con su penetrante erudición, se ha planteado en varios lugares estas cuestiones; el lector encontrará abundantes luces teóricas en su interesante trabajo *La página y el lienzo: Sobre las relaciones entre poesía y pintura* (Egido, 1990, pág. 164-197). Por último,

c) el texto y las variantes de copias manuscritas o de ediciones impresas, sobre lo que no insistiré por ser conocido desde la antigüedad, tratado tanto desde el punto de vista de la ecdótica como de las distintintas posibilidades a la hora de construir un *corpus*.

A pesar que siempre se maneja como criterio definitivo, por su espectacularidad, el criterio del tamaño de un *corpus*, no es lo único importante, pues como ha advertido G. Leech (1991, pág. 10) una colección de textos informatizados no constituye un *corpus*, observación aparentemente elemental, pero que debería estar escrita con letras de oro en

los lugares donde se pretende iniciar investigaciones de este tipo. Es básico que los textos que constituyen el *corpus* posean la necesaria representatividad, de aquí que parezca menos eficaz un corpus estático que un corpus dinámico, el donominado *corpus monitor*, "que no tenga límites de extensión, porque, como la misma lengua, esté siempre desarrollándose." [...] "Un *corpus* así es lo que necesita toda lengua de rango internacional" (Sinclair, 1991, pág. 103-104). El *corpus*, ha advertido Leech, debe poseer una deseable proporción entre lengua hablada y lengua escrita; desde sus primeros pasos, tiene que contar con un planteamiento claro de los problemas legales, cuestión olvidada demasiadas veces, y además, debe desarrollar un conjunto de programas informáticos que permitan el uso riguroso y rápido del volumen de datos contenido en el *corpus* (Leech, 1991, pág. 12).

El rigor científico exige que un *corpus* esté debidamente documentado, anotado y etiquetado; en general, la anotación de un corpus supone la presencia de una teoría lingüística como base de las distintas clasificaciones de categorías y subcategorías y supone, sobre todo, el carácter normalizado de las distintas anotaciones y etiquetas que permite un uso racional de los materiales y un óptimo aprovechamiento de los recursos utilizados. El *corpus* debe poseer una documentación que especifique de manera normalizada el tipo de documento, anotaciones que organicen las particularidades de cada texto y las etiquetas para las distintas clasificaciones lingüísticas. La historia de este proceso de normalización es muy interesante (Smith, 1987; Bernard et al., 1988, Sperberg-McQueen, 1991, Goldfarb, 1993, Sperberg-McQueen y Burnard, 1994, *NERC-1*, 1994) y culminará en determinados estándares cuando se acaben las labores de grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*). Los procesos de normalización en la anotación de un *corpus* son básicos para que este tipo de recursos lingüísticos sea compatible con otras investigaciones similares, en la misma lengua o en otras, economizan grandes cantidades de trabajo y permiten un uso racional en la investigación de una lengua. Sin embargo, se trata de tareas importantes, sumamente complejas y laboriosas; la lectura y meditación del trabajo de Sinclair (1992) sobre anotación automática de *corpus* es obligada para toda persona que pretenda opinar sobre estas cuestiones o que tenga una fe ciega en la capacidad analítica de las máquinas. Los proyectos actuales de investigación de *corpus* están muy unidos a los proyectos de investigación léxica, como sucede con el francés *Tresor de la langue française*, cercano a

los doscientos millones de palabras, el interesante *Corpus General de Referència de la Llengua Catalana* para la lengua catalana o el que existe para el español de México, bajo la dirección de L. F. Lara, aunque un *corpus* bien concebido y anotado puede tener diferentes usos y posibilitar estudios de naturaleza muy variada, como ha demostrado G. Rojo al utilizar un *corpus* realizado con fines gramaticales para estudiar la frecuencia de fonemas (Rojo, 1991).

En el terreno de los recursos lingüísticos, junto a los materiales que constituyen el género denominado *corpus*, aparecen los *archivos digitales* (imagen y texto): ADMYTE (Archivo Digital de Manuscritos y Textos Españoles) (Marcos Marín, 1991b, 1994) o conceptos próximos como los anunciados por J. Torruella para la literatura catalana (1992, 1993), con el excelente paralelo de la literatura italiana en la obra *LIZ* (Pasquale Stoppelli y Eugenio Picchi, 1993). Nos encontramos ante materiales que pueden constituir *corpus* o pasar, con algún tramiento, a la categoría de *colecciones*; el futuro, gracias a la normalización de las anotaciones, va a permitir distintas posibilidades de usos y la aparición de materiales muy ricos, a la vez, para la investigación lingüística y también para la literaria; en este prometedor y cercano futuro el *hipertexto* va a ocupar un lugar de honor (Faulhaber, 1991a, con excelente bibliografía en pág. 205-237).

Estas palabras finales llenas de absoluta confianza en la aplicación de las tecnologías informáticas al análisis del lenguaje no pueden ocultar, sin embargo, la clara advertencia a los que creen que una vez concebido y realizado un *corpus* con todos los requisitos científicos, aparecen, mágicamente, como consecuencia de esta labor los diccionarios y las gramáticas de la lengua en cuestión. Como gusta de explicar José Antonio Pascual, los *corpus* no son más que complementos del trabajo con los que nunca se puede suplir las tareas que han de realizar lexicógrafos y gramáticos.

Bibliografia

AIJMER, K. y ALTENBERG, B. (1991), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Longman, Londres.

ALTENBERG, B. (1991), *A bibliography of publications relating to English computer corpora*, en S. Johansson y A. Stenström, *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín, pág. 355-396.

ALVAR EZQUERRA, M. en colaboración con BLANCO, M. J. (1991), *Creación de un corpus textual del español*, Simposio de la lengua española. Ciencia y tecnología.

ALVAR EZQUERRA, M. y VILLENA PONSODA J. A. (coords.) (1994), *Estudios para un corpus del español*, "Anejo 7 Analecta Malacitana".

ARRARTE, G. y LLISTERRI, J. (1994), *Informe sobre recursos lingüísticos para el español (I): Corpus escritos y orales disponibles y en desarrollo en España*, Instituto Cervantes, Madrid.

BARNARD, D. T., FRASER C. A. y LOGAN, G. M. (1988), *Generalized Markup for Literary Texts*, "Literary and Linguistic Computing", 3, 1, pág. 26-31.

BURNARD, L. (1992), *Tools and Techniques for Computer-assisted Text Processing*, en C. S. Butler, (ed.) *Computer and Written Texts*, Basil Blackwell, Oxford, pág. 1-28.

BUTLER, C. S. (ed.) (1992), *Computer and Written Texts*, Basil Blackwell, Oxford.

EGIDO, A. (1990) *La página y el lienzo: Sobre las relaciones entre poesía y pintura*, en *Fronteras de la poesía en el Barroco*, Crítica, Barcelona, pág. 164-197.

FAULHABER, Ch. (1991a), *Textual Criticism in the 21st Century*, "Romance

Philology", XLV, 1, pág. 123-148.

_____ (1991b), *Informática y filología española: Observaciones acerca de la coyuntura actual y una descripción de* BETA (Biblioteca Española de Textos Antiguos), Simposio de la lengua española. Ciencia y tecnologia.

_____ (1994) *La* Text Encoding Iniciative *y su aplicación a la codificación textual y explotación*", en *Actas del Congreso de la Lengua Española*, Instituto Cervantes, Madrid, pág. 331-340.

GOLDFARB, C. F. (1993), *The SGML Handbook*, Oxford University Press, Oxford.

HUYNH-ARMANET, V. y PINEIRA-TRESMONTANT, C. (1988), *La description contextuelle de mots et l'ordinateur*, "Literary and Linguistics Computing", 4, 1, pág. 12-18.

INFANTES DE MIGUEL, V. (1980), *La textura del poema: disposición gráfica y voluntad creadora*, "1616, Anuario de la Sociedad Española de Literatura General y Comparada", III, pág. 82-89.

JOHANSSON, S. y STENSTRÖM A. (1991), *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín.

JÖRDER, O. (1936), *Die Formen des Sonetts bei Lope de Vega*, "Beihefte zur Zeitschrift für Romanische Philologie", 86, Halle / Saale, Max Niemeyer Verlag.

LARA, L. F. (1994), *Teoría y método en el* Diccionario del español de México, en *Actas del Congreso de la Lengua Española*, Instituto Cervantes, Madrid, pág. 660-665.

LARA, L. F. y HAM CHANDE, R. (1974), *Base estadística del* Diccionario del español de México, "Nueva Revista de Filología Hispánica", XXIII, pág. 245-267.

LEECH, G. (1991), *The state of the art in corpus linguistics*, en K. Aijmer y B. Altenberg, (eds) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Longman, Londres, pág. 8-29.

LEECH, G. y FLIGELSTONE, S. (1992), *Computers and corpus analysis*, en Ch. S. BUTLER, (ed.) *Computer and Written Texts*, Blackwell, Oxford-Cambridge, pág. 115-140.

MACKENZIE, D. (1994), *Problemas de transcripción textual electrónica: lenguas, dialectos, máquinas*, en *Actas del Congreso de la Lengua Española*, Instituto Cervantes, Madrid, pág. 341-344.

MARCOS MARIN, F. (1991a). *Computers and Text Editing: A Review of Tools, an Introduction to UNITE and Some Observations Concerning its Application to Old Spanish Texts*, "Romance Philology", XLV, 1, pág. 102-122.

_____ (1991b), *Archivos Digitales*, Simposio de la lengua española. Ciencia y tecnologia.

_____ (1994), *Estándares y estándar: ADMYTE, el archivo digital de manuscritos y textos españoles y sus soluciones para codificar e intercambiar datos textuales*, en *Actas del Congreso de la Lengua Española*, Instituto Cervantes, Madrid, pág. 345-359.

NERC (1994), *NERC-1 Network of European Reference Corpora. Final Report*, ILC-CNR, Pisa.

RENOUF, A. (1987), *Corpus development*, en J. Sinclair, (ed.) *Looking Up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Langage Dictionary*, Collins, Londres, pág. 1-40.

ROJO, G. (1991), *Frecuencia de fonemas en el español actual*, en *Homenaxe ó profesor Constantino García*, I, Universidade de Santiago de Compostela, Santiago de Compostela, pág. 451-467.

_____ (1993), *La base de datos sintácticos del español actual* "Español Actual", 59, pág. 15-20.

_____ (1994), *Problemas lingüísticos e informáticos en los diccionarios de construcción y régimen*, en *Actas del Congreso de la Lengua Española*, Instituto Cervantes, Madrid, pág. 307-315.

ROMERA CASTILLO, J. (1980), *Poesía figurativa medieval: Vigilan monje hispano latino del siglo X, precursor de la poesía concreto-visual*, "1616, Anuario de la Sociedad Española de Literatura General y Comparada", III, pág. 138-155.

SIMPOSIO DE LA LENGUA ESPAÑOLA. CIENCIA Y TECNOLOGIA, (1991).

SINCLAIR, J. (1987), *Looking Up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Langage Dictionary*, Collins, Londres.

_____ (1991a), *Creación de* corpus, en J. Vidal Beneyto (ed.) *Las industrias de la lengua*, Fundación Germán Sánchez Ruipérez, Pirámide, Madrid, pág. 93-107.

_____ (1991b), *Corpus, concordance, collocations*, Oxford University Press, Oxford.

_____ (1992), *The automatic analysis of corpora*, en J. SVARTVIK, (ed.) *Directions in Corpus Linguistics*, pág. 379-397.

SMITH, J. M. (1987), *The Standard Generalized Markup Language (SGML) for Humanities Publishing*, "Literary and Linguistic Computing", 2, 3, pág. 171-175.

SPERBERG-McQUEEN, C. M. (1991), *Text in Electronic Age: Textual Study and Text Encoding, with Exemples from Medieval Texts*, "Literary and Linguistic Computing", 6, 1, pág. 34-46.

SPERBERG-McQUEEN, C. M. y BURNARD, L. (eds.) (1994), *Guidelines for Electronic Text Encoding and Interchange, (TEI P3)*, Text Encoding Iniciative, Association for Computational Linguistics, Association for Computers and the Humanities, Association for Literary and Linguistic Computing, Chicago-Oxford.

STOPELLI, P. y PICCHI, E. (eds.) (1993), *Letteratura Italiana Zanichelli*, Zanichelli, Bolonia.

SVARTVIK, J. (ed.) (1992), *Directions in Corpus Linguistics. Procceedings of Nobel Symposium 82*, Mouton de Gruyter, Berlín, Nueva York.

TORRUELLA, J. (1992), *Arxiu Informatitzat de Textos Catalans Medievals*,en *Actes del IXè Col·loqui Internacional de Llengua i literatura Catalanes*, II, Publics. de l'Abadia de Montserrat, Barcelona, pág. 239-252.

_____ (1993), *Bases de dades per a textos medievals*, en *Actes du XXe Congrès International de Linguistique et Philologie Romanes*, IV, Francke Verlag, Tübingen, pág. 747-760.

ZAMPOLLI, A. (1991), *Corpora de* referencia, en J. Vidal Beneyto (ed.), *Las industrias de la lengua*, Fundación Germán Sánchez Ruipérez, Pirámide, Madrid, pág. 119-124.

Els corpus lingüístics orals

Joaquim Llisterri
(Universitat Autònoma de Barcelona)

1. Els corpus orals

Per tal de definir a què ens referim en parlar de corpus orals, pot ésser útil de situar-los dins el conjunt del que s'anomenen els "recursos lingüístics" i explicar, des d'una perspectiva històrica, quin ha estat el seu desenvolupament.

1.1. *Els recursos lingüístics*

Tant la descripció de la llengua com el desenvolupament de les tecnologies del llenguatge i de les seves aplicacions en el camp que s'ha anomenat "indústries de la llengua" o "enginyeria lingüística" requereixen la creació de recursos lingüístics en suport informàtic que siguin accessibles a la comunitat investigadora i als qui s'encarreguen més directament de l'elaboració de productes. Aquests recursos consisteixen bàsicament en diccionaris -incloent els reculls terminològics-, gramàtiques i corpus, juntament amb les eines necessàries per a utilitzar-los i extreure'n la informació rellevant en cada cas. Els corpus -tant orals com escrits-, formen doncs part dels recursos lingüístics, i

constitueixen mostres molt àmplies de la llengua de les quals es pot obtenir informació tant lingüística com estadística en els diversos nivells d'anàlisi que habitualment s'utilitzen en la descripció del llenguatge.

1.2. *El sorgiment dels corpus orals*

Pot considerar-se que els corpus orals tal com els entenem actualment sorgeixen com a resultat de la confluència de tres tradicions: per una banda la fonètica experimental, per una altra les tecnologies de la parla i, finalment, la lingüística de corpus.

1.2.1. La fonètica experimental

Des del seu naixement a principis d'aquest segle, la fonètica experimental ha fet ús de la noció de corpus, entès com un conjunt controlat de realitzacions fonètiques. Amb això es relaciona amb altres disciplines lingüístiques, amb les quals va estar molt lligada des del principi, com ara la dialectologia. Cal recordar, per exemple, que la tesi de l'Abbé Rousselot, considerat el fundador modern de la fonètica experimental, era un estudi sincrònic i diacrònic de la parla de Cellefrouin, el seu lloc de naixement (Rousselot, 1892). L'ús dels instruments propis de la recerca en fonètica fa que sigui del tot necessari partir de les realitzacions d'un o més parlants, i com més s'ha avançat en els estudis, més s'ha vist la necessitat de controlar l'aparició de diverses variables que poden influir en els elements segmentals o suprasegmentals de la parla. Així, des del punt de vista de la fonètica, un corpus ha d'ésser en primer lloc dissenyat específicament en funció d'allò que es vulgui estudiar; en segon lloc, ha d'estar enregistrat en unes circumstàncies que en permetin l'estudi experimental mitjançant tècniques d'anàlisi acústica, que habitualment són molt sensibles a les pertorbacions introduïdes pels sorolls de l'ambient.

1.2.2. Les tecnologies de la parla

Per una altra banda, a partir dels anys 70 es varen començar a desenvolupar les possibilitats de portar a terme aplicacions pràctiques de les tecnologies de la parla, sobretot en el camp del reconeixement. Un sistema de reconeixement de parla requereix una fase d'entrenament, durant la qual s'adquireixen els models que després es comparen amb el senyal acústic que es vol reconèixer. Com més dades s'introdueixen en aquesta etapa d'entrenament, més garanties hi ha d'arribar a un sistema amb una taxa d'error baixa a l'hora de reconèixer. Cal considerar també que si un sistema ha d'ésser utilitzat per diverses persones, l'entrenament ha de tenir en compte aquest factor i fer-se amb

enunciats produïts per parlants diferents. Per altra banda, un cop es disposa d'una versió entrenada del reconeixedor, cal verificar els seus resultats abans de convertir-lo en un producte comercial. Això requereix també un nombre elevat de realitzacions fonètiques de molts parlants. Per aquests motius, sorgí la necessitat de crear bases de dades orals a gran escala, concebudes com un conjunt de realitzacions fonètiques que permetessin tant l'entrenament com l'avaluació dels sistemes de reconeixement de parla.

1.2.3. La lingüística de corpus

Pot considerar-se que la lingüística de corpus es va desenvolupar a partir dels anys 60 al marge de les tecnologies de la parla i de la fonètica experimental; la idea que orientà els seus principis és que la descripció de la llengua no es pot dur a terme considerant únicament la intuïció d'un parlant natiu, sinó que s'ha de basar en un conjunt de produccions reals. Un corpus s'entén en aquesta tradició com un conjunt ampli de dades reals de la llengua estudiada, i pot consistir tant en textos escrits com en transcripcions ortogràfiques de la llengua parlada.

1.3. *La convergència dels objectius i dels mètodes*

Si la fonètica experimental i les tecnologies de la parla varen confluir ben aviat, com a conseqüència de la possibilitat d'utilitzar les bases de dades pensades per al reconeixement en l'estudi de la variabilitat contextual dels al·lòfons o la variació inter- i intra-locutor, la lingüística de corpus i el que sovint es defineix com "the speech community" no han començat a apropar-se fins fa ben poc.

Mentre que, com s'ha vist, des de la lingüística de corpus s'ha entès com a "corpus oral" la transcripció ortogràfica - més o menys enriquida amb diferents tipus d'anotació, tal com es comentarà més endavant- d'una sèrie d'enregistraments obtinguts en condicions el més natural possibles, en l'àmbit de la fonètica experimental i de les tecnologies de la parla el més essencial d'un corpus ha estat el senyal sonor, sobre el qual es treballa directament per tal de modelar les característiques articulatòries o acústiques de la parla o d'entrenar i avaluar sistemes de reconeixement. Les diferències més importants entre les dues tradicions es resumeixen a la taula 1.

	Lingüística de corpus	Tecnologies de la parla i fonètica experimental
Materials	Parla espontània, no preparada (*unelicited speech*)	Corpus controlat (*elicited speech*)
Àmbit	Discurs, diàleg	Enunciat
Enregistrament	Entorn natural	Entorn controlat
Transcripció	Transcripció ortogràfica enriquida	Transcripció fonètica i ortogràfica alineada amb el senyal sonor
Orientació	Representació simbòlica, categorial	Senyal sonor, representació temporal contínua

Taula 1: Diferències en la concepció dels corpus orals en la lingüística de corpus i en tecnologies de la parla i fonètica experimental (Llisterri, 1994a)

Tot i aquestes diferències històriques, ha sorgit un interès des de la fonètica experimental i les tecnologies de la parla per disposar de dades obtingudes en situacions el més natural possibles, controlant, però, la bona qualitat tècnica de l'enregistrament. Això és degut probablement a la possibilitat de posar a punt aplicacions que han de funcionar en diferents entorns reals, permetent el diàleg entre l'home i la màquina per a realitzar tasques concretes. Tal com assenyala Teubert (1993, pàg. 4) "the speech community has commenced to express their interest in large spoken language corpora. Even general purpose corpora of impromptu, unrehearsed, unscripted, non elicited informal conversations now seem to arouse some interest in speech research as they can be used as test-beds for speech recognition systems".

Les tecnologies de la parla requereixen cada vegada més que els materials utilitzats no es limitin al senyal sonor, sinó que continguin una representació ortogràfica, una representació fonètica i també altres nivells d'anotació. Només cal pensar, per exemple, en la importància de disposar d'una bona anàlisi sintàctica si a partir d'un text escrit s'ha de generar automàticament la seva entonació, tal com és el cas en la conversió de text a parla. En els sistemes de reconeixement s'ha introduït també la noció de "model de llenguatge", equiparable a grans trets a una gramàtica que contribueix al bon funcionament del sistema complementant el tractament purament acústic del senyal. En paraules de Moore (1991, pàg. 3) "for many purposes (especially in speech technology) it

has become clear that speech data can be very useful if it is accompanied by machine-readable annotations consisting. at the very least, of an orthographic transcription with paragraph or phrase level pointers into the acoustic data".

Així com les tecnologies de la parla s'han anat apropant gradualment als interessos de la lingüística de corpus, aquesta també ha trobat elements d'interès en el treball que s'ha portat a terme partint del senyal sonor. La possibilitat de processar digitalment la parla per al seu emmagatzemament i de segmentar (semi)automàticament el senyal i sincronitzar-lo amb la representació ortogràfica ha desvetllat l'interès dels qui fins no fa gaire treballaven únicament amb la transcripció escrita d'enregistraments. Un resultat d'aquest interès són les recomanacions del NERC (*Network of European Reference Corpora*) d'incloure el senyal digitalitzat en qualsevol corpus de llengua oral. (Sinclair, 1993, pàg. 65-70).

Finalment, com assenyalen Church i Mercer (1993), la recerca sobre la llengua en la tradició de la lingüística de corpus ha estat essencialment basada en mètodes heurístics (*knowledge-based*), mentre que l'aproximació dominant en reconeixement de parla ha estat l'estadística des del moment en què es va disposar de grans quantitats de dades. El tractament de problemes clàssics de la lingüística computacional com l'assignació automàtica de parts de l'oració (*tagging*) o l'anàlisi sintàctica automàtica (*parsing*) no s'ha fet amb mètodes estadístics fins fa poc temps. La confluència en una mateixa metodologia de dos camps que tradicionalment havien estat separats ha contribuït, i ben segurament continuarà contribuint, a la integració entre el processament del llenguatge natural i el processament de la parla.

S'observa doncs com les diferències entre els "corpus orals" (*spoken corpora*) tal com havien estat entesos per la lingüística de corpus i les "bases de dades orals" (*speech data bases*) utilitzades en les tecnologies de la parla s'esborren progressivament tant pel que fa a la convergència d'objectius com de metodologies. Tot i així, és important de ressaltar els aspectes específics de cada tipus de material, tal com s'intenta fer en la tipologia que es presenta a continuació.

2. Tipologia de corpus orals[1]

Després d'aquesta breu reflexió històrica, sembla clar que l'aplicació que es vol donar a un corpus i la tradició en la qual s'insereix han condicionat i segueixen condicionant encara el seu disseny. Considerant els materials de què es disposa actualment, es podria intentar d'establir una tipologia de corpus, distingint els inventaris fonètics i fonològics i els corpus per a la descripció fonètica (2.1), els corpus orals per al desenvolupament d'aplicacions tecnològiques (2.2), i els que tradicionalment s'han anomenat corpus de llengua oral (2.3).

2.1. *Inventaris fonètics i fonològics i corpus orals per a la descripció fonètica*

2.1.1. Inventaris fonètics i fonològics

S'inclouen en aquests grups els repertoris d'inventaris fonètics i fonològics de llengües organitzats en forma de bases de dades per tal de fer possible l'estudi dels universals lingüístics i de la tipologia lingüística en l'àmbit de la fonètica i la fonologia. Per a aquesta mena de treballs s'han recollit inventaris de segments de les llengües del món extrets de descripcions publicades. En alguns casos la base de dades inclou només els inventaris, com per exemple el SPA -*Stanford Phonological Archive* (Greenberg et al. (eds.) 1978) o l'UPSID-*UCLA Phonological Segment Inventory Database* (Maddieson, 1991; Maddieson i Precoda, 1990)

2.1.2. Corpus per a la descripció fonètica comparada

Tot i que no són els més nombrosos, alguns corpus dels que es tracten en aquest apartat es plantegen con a objectiu la descripció fonètica comparada; són més exhaustius

[1] Queda fora de l'abast d'aquest treball presentar un recull exhaustiu dels corpus orals desenvolupats fins ara, i per aquest motiu només s'esmenten els projectes més representatius en cadascun dels àmbits. Per a una visió generals dels projectes als Estats Units pot consultar-se Lamel (1992); un resum de les principals inciatives japoneses ha estat publicat a *NESCA - The European Speech Communication Association Newsletter* 13 (1994), pàg. 11-15; Badia et al. (1994) recull corpus textuals i orals per al català i Arrarte i Llisterri (1994) fa el mateix per al castellà. NESCA, el butlletí de la *European Speech Communication Association* publica habitualment informació sobre nous projectes en aquest àmbit: per a informació sobre com obtenir-lo hom pot adreçar-se a ESCA, ICP - Université Stendhal, BP 25 X, 38400 Grenoble Cedex, France. Tel (33.76) 82.43.36. Fx (33.76) 82.43.35. Correu electrònic: esca@icp.grenet.fr; servidor WWW: htpp://ophale.icp.grenet.fr/esca/esca.html. A través d'ESCA poden obtenir-se també les actes dels diversos congressos *Eurospeech* i dels *Workshops* organitzats per l'associació.

que els inventaris als quals s'ha fet referència, i contenen una varietat de materials més àmplia. Probablement el millor exemple sigui el projecte IRIS -*Immigrant Voices in Swedish-Phonetic Models* (Engstrand, 1987); el febrer de 1987 es disposava de materials enregistrats per a unes 100 llengües. En principi, per a cada llengua s'intenta de recollir síl·labes aïllades que representin els contrastos fonètics i fonològics propis de la llengua, parells mínims, frases llegides a diferent velocitat d'elocució, dos textos llegits, una explicació lliure d'un dels textos, un monòleg lliure sobre les activitats quotidianes de l'informant, i una lectura de *La Tramuntana i el Sol,* el text oficial de l'Associació Fonètica Internacional.

Cal esmentar també dos productes comercials recents: la *Kay Phonetic Database* desenvolupada per Kay Elemetrics Corp. el 1991 i l'*Oxford Acoustic Phonetic Database* (Pickering i Rosner, 1993; MacKay, 1994); no cobreixen una gamma tan àmplia de llengües com els inventaris de Stanford o de Califòrnia, però ofereixen en canvi el senyal sonor digitalitzat fent possible una anàlisi acústica directa de contrastos fonètics en diverses llengües.

2.2. *Corpus orals per a aplicacions en tecnologies de la parla*

Es fa referència aquí a corpus que solen contenir una àmplia varietat de realitzacions fonètiques de la llengua, abastant des del segments aïllats fins a la parla espontània, incloent mots aïllats, mots en frases marc, frases i textos llegits. Un element comú a aquests corpus és que l'enregistrament es realitza sempre en condicions acústiques molt controlades, en suport digital i seguint determinats estàndards que es detallen més endavant (3.2). Els enregistraments s'organitzen en fitxers ben identificats en una estructura de base de dades, i sovint s'acompanyen de la transcripció fonètica sincronitzada amb el senyal sonor.

Per altra banda, els continguts dels corpus dissenyats des de la perspectiva de les tecnologies de la parla varien en funció de l'objectiu del corpus. Poden trobar-se des dels corpus que tenen com a funció entrenar i avaluar sistemes de reconeixement o construir productes que facin possible el dictat automàtic, fins als que s'utilitzen per a entrenar i avaluar sistemes de diàleg home-màquina en aplicacions molt concretes com pot ser la reserva de bitllets d'avió. Per aquest motiu, s'ha intentat de separar els corpus útils per a aplicacions generals dels que estan lligats a una aplicació específica.

2.2.1. Corpus per a aplicacions generals

2.2.1.1. Corpus monolingües

Tot i que els projectes com BDSONS, PhonDat, *Albayzín*, TIMIT, BREF o WSJ-CSR que es descriuen més endavant en aquest apartat es varen concebre inicialment per a aplicacions en les tecnologies de la parla, sobretot per a l'entrenament i l'avaluació de sistemes de reconeixement, la riquesa dels materials i l'elevat nombre de locutors els fa indubtablement útils per a la recerca més bàsica en fonètica. Per exemple s'han publicat recentment estudis centrats en la variació fonètica en funció del sexe i la base dialectal dels parlants partint d'un corpus com TIMIT, que ofereix una diversitat de locutors difícilment abastable per a un únic investigador (Keating et al., 1994).

En les principals llengües europees s'han desenvolupat corpus monolingües que, com s'indicava més amunt, permeten tant la descripció fonètica com la utilització en aplicacions tecnològiques, especialment l'entrenament i la verificació de sistemes de reconeixement de parla. Pot destacar-se per al francès BDSONS - *Base de données des sons du français* (Carré et al., 1984; Dolmazon, 1994), actualment disponible en 7 CD-ROMs i PhonDat (Kohler, 1991; Draxler et al., 1993) per a l'alemany.

En ambdós casos es tracta d'iniciatives que han agrupat diversos centres de recerca en el marc d'un projecte nacional, i que han donat com a resultat uns corpus amb materials llegits en els quals predominen les frases aïllades i els textos curts, encara que també s'hi trobin paraules aïllades, especialment les més relacionades amb les necessitats del desenvolupament de productes en reconeixement de la parla, com ara els dígits i els números de telèfon; contenen, a més, un material interessant que són les anomenades "frases fonèticament equilibrades", un conjunt reduït d'enunciats dissenyats de manera que contenen els segments fonètics o fonològics de la llengua amb la freqüència d'aparició pròpia de la llengua oral.

Un corpus amb una filosofia similar s'està desenvolupant per al castellà. Es tracta del projecte *Albayzín*[2] (Casacuberta et al., 1992; Moreno et al., 1993; Díaz et al., 1993), coordinat per la Universitat Politècnica de Catalunya, en el qual participen la Universitat

[2] Subvencionat per la *Comisión Interministerial de Ciencia y Tecnología* (TIC91-1488-C06). Per a més informació sobre el projecte: Dr. Ciment Nadeu, Departament de Teoria del Senyal i Comunicació, Escola Tècnica Superior d'Enginyers de Telecomunicació, Universitat Politècnica de Catalunya, Gran Capità s/n 08034, Barcelona. Fax: (93) 401.64.47. Correu electrònic: nadeu@tsc.upc.es

Autònoma de Barcelona, la Universitat de Granada, la Universitat Politècnica de Madrid i la Universitat Politècnica de València. Els materials es divideixen en tres subcorpus: un de tipus fonètic, que conté 200 frases fonèticament equilibrades i 500 frases fonèticament controlades; un altre que conté 3900 frases que formen part de la consulta oral a una base de dades geogràfica; i un tercer, en el qual s'enregistren elements dels dos subcorpus anteriors mitjançant l'efecte Lombard, consistent en enviar soroll al locutor a través d'uns auriculars, per aconseguir les característiques de la producció de la parla en ambients sorollosos. Participen en els enregistraments un total de 300 locutors. Està previst que la versió final del corpus es distribueixi en 5 CD-ROMs.

Potser, però, l'exemple més clàssic de corpus general en aquest àmbit sigui TIMIT - *DARPA Acoustic Phonetic Continuous Speech Corpus* (Lamel, Kassel i Seneff, 1986; Zue, Seneff i Glass, 1990). El corpus consisteix en 450 frases fonèticament equilibrades, 1890 frases natural i dues frases anomenades "de calibració dialectal", que presenten els fenòmens fonètics més importants per tal de determinar la base dialectal del parlant. El corpus ha estat enregistrat amb 630 locutors, i conté, a més del senyal acústic, la transcripció fonètica i l'ortogràfica alineades amb el senyal. TIMIT es distribueix en CD-ROM des del 1990.

Cal referir-se també a BREF i WSJ-CSR, que tenen en comú el fet d'estar constituïts per textos periodístics. BREF -*A Database of Read Text in French* (Lamel, Gauvain i Eskénazi, 1991) conté 11.000 textos seleccionats de *Le Monde*, i han participat en els enregistraments 120 locutors, cadascun dels quals ha produït entre 5.000 i 10.000 paraules. El WSJ-CSR -*Wall Street Journal Continuous Speech Recognition Corpus* (Paul i Baker, 1992) conté també textos periodístics, extrets aquest cas del *Wall Street Journal*. Per una banda, es recull la lectura de l'article i, per una altra, es demana als locutors que dictin un article de característiques similars al llegit: amb això es pretén d'obtenir un corpus que permeti l'entrenament de sistemes de dictat automàtic. Es treballa amb 160 locutors, i existeixen ja més de 30 CD-ROMs amb els resultats del projecte.

2.2.1.2. Corpus multilingües

Si els projectes que s'han esmentat fins ara en aquest apartat se centren en una llengua, el corpus conegut com EUROM.1 (Sherwood i Fuller, 1992) és de naturalesa

multilingüe. Existeix actualment en alemany, anglès, castellà,[3] danès, francès, italià, holandès, noruec i suec, amb les versions portuguesa i grega en desenvolupament: de moment, estan disponibles en CD-ROM les versions italiana i anglesa. El corpus consta de mots consonant-vocal-consonant o consonant-vocal-consonant-vocal, segons la llengua, que contenen les consonants inicials i finals -i en algun cas medials- en un context format per les vocals /i, a, u/, les vocals en paraules, 100 números, els mots en cinc frases marc diferents, 40 textos curts d'unes cinc frases cadascun amb continguts equivalents en cada llengua i 50 frases seleccionades per a augmentar la cobertura fonètica. Els enregistraments han estat fets per 74 parlants de cada llengua, que llegeixen parts diferents del corpus. D'aquests parlants, se n'han seleccionat 4 per llengua per tal de recollir també el senyal laringogràfic.

2.2.2. Corpus per a aplicacions específiques

Entre el corpus que s'han creat per a aplicacions específiques en el camp de les tecnologies de la parla, destaquen els que s'orienten cap a aquelles aplicacions que requereixen la interacció de l'usuari amb el sistema. Per aquest motiu, no n'hi ha prou amb recollir textos llegits, sinó que calen diàlegs el més reals possibles que reflecteixin com es porten a terme determinades operacions, com ara fer una reserva de bitllet per telèfon.

El paradigma de recollida de dades en aquest cas es coneix com a *Wizard of Oz*; consisteix en imitar un sistema real mitjançant un operador ocult que proporciona a un usuari que creu comunicar-se amb un ordinador totes les respostes que donaria el sistema final. Amb aquest mètode s'obté informació sobre tots els nivells lingüístics, sobre les estratègies de diàleg, i sobre tots aquells fenòmens propis d'una interacció oral per a realitzar una tasca concreta.

El corpus més conegut en aquest àmbit és el recollit en el projecte ATIS -*Air Traffic Information Systems Corpora* (Zue et al., 1991), del qual es desenvolupa també una versió francesa (Bonneau-Maynard et al., 1993). Consisteix, com s'ha explicat, en

[3] La versió castellana d'EUROM.1 es va realitzar el 1993 coordinada per A. Moreno (Dra. Asunción Moreno, Departament de Teoria del Senyal i Comunicació, Escola Tècnica Superior d'Enginyers de Telecomunicació, Universitat Politècnica de Catalunya, Gran Capità s/n 08034, Barcelona. Fax: (93) 401.64.47. Correu electrònic: amoreno@tsc.upc.es) en el marc del projecte ESPRIT 6819 SAM-A, *Speech Technology Assessment in Multilingual Applications*, amb la participació de la Universitat Politècnica de Catalunya i la Universitat Autònoma de Barcelona (Moreno, 1993; Llisterri *et. al.*, 1993). Vegeu també la informació al servidor www: http://www.phon.ucl.ac.uk/resource/eurom.html

l'enregistrament d'una interacció simulada amb un sistema de reserva de vols, i és accessible en CD-ROM. Un corpus similar -encara que en aquest cas només conté la transcripció ortogràfica- orientat al desenvolupament de sistemes de comprensió de la parla natural és VOYAGER (Zue, Seneff i Glass, 1990; Glass et al., 1993); l'àmbit d'aplicació és la informació necessària per a viatjar per una zona determinada, i la versió anglesa del corpus recull diàlegs espontanis de 90 locutors, corresponents a 20 minuts d'interacció amb el sistema simulat.

2.3. *Corpus per a l'estudi de la llengua oral*[4]

Entrem ja finalment en els que, en la tradició de la lingüística de corpus, s'han designat com a corpus de llengua oral (*spoken language*), ben diferenciats, tal com s'intentava de fer veure al principi d'aquest treball, de les bases de dades orals (*speech data bases*). Es fa referència aquí a transcripcions ortogràfiques de la llengua oral- enriquides amb diversos tipus d'anotació- útils per a la descripció lingüística a tots els nivells i també per a l'elaboració de diccionaris. Pel que fa a aquest darrer aspecte, pot esmentar-se, per exemple, que en el corpus que serveix de base al *Collins COBUILD* s'inclouen transcripcions de converses informals, de classes universitàries i de debats i entrevistes radiofòniques (Renouf, 1987).

A l'hora de determinar el contingut d'aquest tipus de corpus se sol considerar en primer lloc la variació estilística i de registre. S'inclouen habitualment transcripcions dels mitjans de comunicació i d'entrevistes enregistrades *in situ* en diversos contextos naturals.

---

[4] Pot trobar-se una informació més exhaustiva sobre altres corpus que contenen transcripcions ortogràfiques de llengua oral a Edwards (1993b) [accessible en format electrònic mitjançant ftp anònim a cogsci.berkeley.edu -fitxer en format comprimit a directori "pub": "CorpusSurvey.Z" o per correu electrònic a a listserv@tamvm1.tamu.edu mitjjançant la comanda "get corpora faq linguist"], Taylor, Leech i Fligelstone (1991) [accessible per correu electrònic a listserv@brownvm.brown.edu mitjançant la comanda "get survey corpora humanist", o per ftp anònim a nora.hd.uib.no (129.177.24.42) buscant el fitxer pub/icame/survey.corpora] i a l'apèndix de Leech (1991). Altenberg (1991) [disponible en versió electrònica al servidor del ICAME: fileserv@nora.hd.uib.no] constitueix una font important d'informació per a l'anglès Edwards (1993b) indica també la manera d'accedir a catàlegs com els de l'*Oxford Text Archive*, del *International Computer Archive of Modern English* (ICAME), del *Center for Electronic Texts in the Humanities* (CETH), o al *Georgetown University Catalog of Projects in Electronics Text* (CPET).

Entre les iniciatives més conegudes cal esmentar el LLC -*London-Lund Corpus of Spoken English* (Greenbaum i Svartvik, 1990) basat en la part oral del *Survey of English Usage Corpus (1953-1987)*; conté 500.000 paraules, i la transcripció ortogràfica va acompanyada d'una transcripció prosòdica. El SEC -*Lancaster/IBM Spoken English Corpus* (Garside, Leech i Sampson (eds.) 1987; Knowles, Taylor i Williams, 1992) conté, a més de la transcripció ortogràfica, fonètica i prosòdica, 52.000 paraules de text etiquetat i analitzat. Knowles i Lawrence (1987) és una bona mostra del tipus de recerca que pot portar-se a terme disposant d'un corpus d'aquestes característiques en el camp de l'assignació automàtica de la prosòdia en relació amb la informació gramatical codificada en el text.

El *Corpus Oral de Referencia del Español Contemporáneo*[5] (Marcos Marín, 1991; Marcos Marín, Ballester i Santamaría, 1993) és un recull de 1.100.000 paraules transcrites ortogràficament seguint les normes de codificació textual de la TEI (*Text Encoding Initiative*). Els textos s'agrupen en diversos àmbits -administratius, científics, jurídics, conversacionals i familiars- i abasten una àmplia varietat de temes.

Entre els projectes en desenvolupament cal esmentar el CSAE -*Santa Barbara Corpus of Spoken American English* (Chafe, DuBois i Thompson, 1991), en el marc del qual es pretenen recollir 200.000 paraules en converses. És important destacar també entre els projectes en curs el Corpus del Català Contemporani de la Universitat de Barcelona, que es descriu amb més detall en aquest mateix volum.

3. La constitució de corpus orals

Aquest breu repàs als principals projectes en el camp dels corpus orals permet, un cop més, d'adonar-se que el procés de constitució d'un corpus ve condicionat per l'objectiu que es vulgui assolir i per les aplicacions posteriors del corpus. Tot i així, és possible d'identificar una sèrie d'etapes que es resumeixen en aquest apartat (Carré, 1991, 1992). Alhora, es descriuen també alguns dels estàndards desenvolupats en el marc del projecte ESPRIT SAM 2589 (*Multilingual Speech Input/Output Assessment, Methodology and*

[5] El corpus és accessible per ftp anònim a lola@lllj.uam.es (150.244.8.2).

Standardisation) que s'han convertit en els habituals en els grups de recerca europeus per a dur a terme l'adquisició de bases de dades orientades a les tecnologies de la parla. Es fa referència també als estàndards que han de sorgir com a resultat del projecte LRE 61-100 EAGLES (*Expert Advisory Group on Language Engineering Standards*). Els objectius globals d'aquests projectes es descriuen en l'apartat 4, dedicat a les principals iniciatives en el camp de l'estandardització.

3.1. *Definició del corpus*

3.1.1. Disseny del contingut

Tal com s'ha tingut oportunitat de veure, el contingut d'un corpus depèn essencialment de l'objectiu i de les aplicacions a què es destini. Per tant, poques recomanacions generals es poden donar sobre aquest punt, donada la gran varietat de corpus que s'ha pogut observar. En el marc d'EAGLES s'està treballant en tipologies de corpus, tant escrits com orals, i també en la manera de definir diverses classes de bases de dades orals. Algunes indicacions sobre els criteris de disseny de diferents corpus es troben a Atkins, Clear i Ostler (1992) i a Oostdijk (1988), referides especialment a la lingüística de corpus entesa tal com s'ha descrit més amunt.

3.1.2. Selecció dels locutors

La selecció de parlants ve també condicionada pels objectius finals del corpus. Mentre que en l'àmbit de la lingüística de corpus es solen tenir en compte criteris predominantment sociolingüístics, en fonètica i en tecnologies de la parla es considera sobretot la representativitat de la població. En el camp específic dels corpus per a aplicacions tecnològiques, SAM utilitza com a criteris de classificació el sexe, l'edat, l'alçada, el pes, la llengua materna, l'accent, el grup ètnic, el nivell d'educació, els hàbits de consum de tabac i les possibles patologies de veu i de la producció de la parla. El grup de treball sobre Llengua Oral de EAGLES ha de proposar també algunes indicacions, que aniran en el sentit de les propostes de SAM.

3.2. *Adquisició de les dades*

El procediment d'adquisició de les dades depèn igualment dels objectius del corpus.

Tradicionalment, en els corpus de llengua oral elaborats des de la lingüística de corpus s'ha procurat d'obtenir enregistraments que fossin útils per a una transcripció auditiva,

sense considerar les possibilitats de treball directe sobre el senyal sonor. En canvi, des de la perspectiva de la fonètica experimental i les tecnologies de la parla, ha predominat l'interès en disposar d'enregistraments en excel·lents condicions acústiques, encara que amb això disminueixi la naturalitat de la situació de parla en què s'obtenen. Actualment estan a disposició dels investigadors sistemes com el DAT (*Digital Audio Tape*) que proporcionen una bona qualitat si l'entorn és adequat i que permeten un enregistrament i una recuperació fàcil del material.

En el marc del projecte SAM s'ha desenvolupat el programa EUROPEC -*European Programme d'Enregistrement de Corpus* (Zeiliger i Serignat, 1991) per a l'adquisició de bases de dades orals, emprat amb èxit en l'enregistrament d'EUROM.1. EUROPEC fa possible la presentació visual dels textos que s'han d'enregistrar, assegura la correcta associació entre la representació ortogràfica i els fitxers que contenen el senyal acústic, i permet de guardar el senyal en un suport digital, que pot ser el disc dur d'un ordinador. El programa està adaptat a l'estació de treball SESAM definida també pels participants en SAM ; la configuració es basa en un IBM-PC-AT-3 amb un processador Intel 80286. 512 kb de RAM, disc dur de 30 Mb i lector de disquets de 1,2 Mb, una targeta gràfica VGA o EGA, un lector de CD-ROM CM-100 o CM 135 de Philips i una placa de processament digital de senyal OROS-AU21 o AU-22 (UCL, 1992); pot veure's que es tracta d'un equip seleccionat amb l'intent d'oferir el màxim de compatibilitat i simplicitat amb els mínims necessaris per al bon funcionament del sistema. El futur treball d'EAGLES en aquest camp es basa en els estàndards de SAM.

3.3. *Preparació de les dades*

3.3.1. Transcripció ortogràfica

La transcripció ortogràfica del corpus sol ésser el primer pas en la preparació de les dades per tal de fer-les accessibles als usuaris. Tot i que pot semblar una operació trivial, no ho és tant en el cas de corpus que recullen llengua oral espontània. En la tradició de l'anàlisi del discurs i de la conversa s'han creat diversos sistemes de notació que enriqueixen la representació ortogràfica amb els elements necessaris per a l'anàlisi (vegeu, per exemple, Edwards, 1992, 1993a; Du Bois et al., 1993; Gumperz i Berenz, 1993 i Ochs, 1979). També des de la lingüística de corpus han sorgit sistemes per incloure en la representació ortogràfica diversos aspectes de la llengua oral. Per exemple, en el *British*

National Corpus (BNC, 1991) es codifiquen els torns de paraula, la superposició d'enunciats, els límits entre enunciats, els diferents tipus de pausa, les formes no estàndards, els elements paralingüístics i no verbals i les incerteses del transcriptor.

En el grup de treball de Corpus Textuals d'EAGLES es desenvolupen recomanacions sobre la codificació textual, basades en les propostes de la TEI (*Text Encoding Initiative*) i també es discuteixen propostes sobre la transcripció de la llengua oral que parteixen de les recomanacions de la TEI i del NERC (*Network of European Reference Corpora*) exposades més endavant (Llisterri, 1994a).

3.3.2. Transcripció fonètica

En referir-se a la transcripció fonètica, cal primer de tot plantejar-se la qüestió dels nivells de transcripció. Es podria, en principi, distingir un nivell en el qual es fa una representació fonèmica -transcripció ampla- corresponent a la forma canònica de les paraules aïllades (*citation form*): en un altre nivell, es podria disposar d'una representació fonètica - transcripció estreta - corresponent a la realització fonètica de l'enunciat. Aquesta és l'orientació adoptada per EAGLES, i té els seus precedents en propostes com la de Barry i Fourcin (1992) o de Tillmann i Pompino-Marschall (1993).

3.3.2.1. Transcripció dels elements segmentals

Pel que fa als sistemes de notació, la comunitat fonètica ha utilitzat tradicionalment l'Alfabet Fonètic Internacional (IPA, 1993). Quan el 1989 es va portar a terme la revisió de l'AFI, es va crear un grup de treball específicament dedicat als problemes de transmissió i intercanvi electrònic de textos transcrits. D'aquí va sorgir la idea d'assignar un número i un nom a cada símbol i diacrític de l'AFI (Esling, 1988,1990). Aquesta codificació s'utilitza, per exemple, en el projecte LRE 61-004 ONOMASTICA -*Multi-Language Pronunciation Dictionary of Proper Names and Place Names*[6] (Schmidt et al., 1993).

Per a resoldre els problemes derivats de l'emmagatzemament i intercanvi en suport informàtic de textos transcrits mitjançant l'AFI, s'ha desenvolupat SAMPA -*SAM Phonetic Alphabet* (UCL, 1992; Wells et al., 1992), un alfabet fonètic en el qual els

[6] El projecte, iniciat el gener de 1993, és coordinat per: Professor Mervyn A. Jack, University of Edinburgh, CSTR, South Bridge, Edinburgh EH1 1HN, United Kingdom. Fax: (44.31) 226.27.30. Servidor www: http://www.ccir.ed.ac.uk Participen, entre d'altres, la Universitat Politècnica de Madrid i *Telefónica.*

símbols de l'AFI reben una codificació en ASCII (*American Standard Code for Information Interchange*). SAMPA és un sistema orientat principalment cap a la representació fonèmica, i s'ha d'entendre que els seus símbols no tenen un valor comú entre llengües ni representen un únic so de la mateixa llengua, sinó que reflecteixen oposicions distintives a l'interior de cada llengua. Per això es fa referència molts cops a la transcripció fonotípica, que constitueix una representació al·lofònica derivada per regles contextuals a partir de la forma canònica de la paraula. Aquesta és, per exemple, la forma que adopten les transcripcions d'EUROM.1. SAMPA ha estat adaptat a les llengües en les que es va treballar en el projecte SAM -alemany, anglès, danès, francès, holandès, italià, noruec i suec- i desenvolupat també per al castellà[7] en el marc del projecte ESPRIT 6819 SAM-A - *Speech Technology Assessment in Multilingual Applications* (Mariño i Llisterri, 1993); tal com assenyala Wells (1989) és fàcilment adaptable a les altres llengües de la Unió Europea.

També com a part dels treballs de SAM s'ha desenvolupat SAMTRA -*Transcription Verification and Phoneme/Diphoneme Analysis Software* (Braun, 1992), una eina que permet verificar que la transcripció fonètica s'ha realitzat amb els símbols propis de SAMPA i alhora permet de calcular la distribució estadística de fonemes i difonemes en un corpus.

Existeixen, però, altres sistemes de transcripció fonètica segmental basats en alfabets compatibles amb les necessitats de la transmissió i l'intercanvi electrònic d'informació.

Val la pena esmentar PHONASCII, emprat en el projecte CHILDES -*Child Language Data Exchange System*, que consisteix en un alfabet per a la transcripció fonèmica anomenat UNIBET i un alfabet per a la transcripció fonètica amb els seus corresponents diacrítics i amb marques per als elements suprasegmentals (Allen, 1988).

3.3.2.2. Transcripció dels elements suprasegmentals

Com s'ha vist en l'apartat 3.3.1., en la tradició de l'anàlisi del discurs i de la conversa existeixen ja convencions per a enriquir la transcripció ortogràfica amb informació prosòdica, i també es donen indicacions en aquest sentit en les propostes de la TEI i del NERC. Alguns corpus de llengua oral que s'han descrit a 2.3. com el LLC i el SEC

[7] Cal esmentar aquí el treball previ d'A. Quilis i E. Enríquez en el projecte ESPRIT 2104 POLYGLOT 1. Per a més informació sobre SAMPA vegeu el servidor www: http://www.phono.ucl.ac.uk/home/sampa/home.htm

inclouen també anotació prosòdica (Knowles, 1991; Wichmann, 1991). L'Alfabet Fonètic Internacional ofereix també símbols per a la representació dels elements suprasegmentals (Bruce, 1988, 1989), igual que SAMPA.

Tot i així, s'han desenvolupat altres sistemes de representació prosòdica especialment orientats al corpus i a les bases de dades orals (per a una revisió vegi's Llisterri, 1994b). Alguns d'aquests sistemes -PROSPA, SAMSINT, SAMPROSA i INTSINT- han estat discutits i avaluats en el grup dedicat a la prosòdia del projecte SAM i es presenten de forma resumida a Wells et al. (1992) i a Gibbon (1989).

PROSPA fou desenvolupat per Selting i Gibbon (Selting, 1987) per a l'anàlisi del discurs i de la conversa, i ofereix una transcripció ampla de la melodia i dels moviments locals deguts als accents. SAMSINT -*SAM System for Intonation Transcription* (Wells et al., 1992) permet transcriure moviments melòdics a l'interior d'unitats entonatives mitjançant símbols associats a codis ASCII. SAMPROSA, proposat per Gibbon i presentat a Wells et al. (1992), es basa en la transcripció prosòdica de SAMPA, en la codificació utilitzada en PROSPA, i pren com a base teòrica la fonologia autosegmental.

Els símbols utilitzats representen moviments tonals, tons nuclears, accents, pauses i límits entre unitats prosòdiques. Finalment, INTSINT -*International Transcription System for Intonation* (Hirst i Di Cristo, en premsa; Hirst, 1991; Hirst, 1994) es fonamenta en la representació del contorn melòdic com una seqüència de punts (*target points*) situats a diferents nivells; la descripció d'aquests nivells pot fer-se de manera relativa en relació als punts anteriors, o de manera absoluta en relació a tota una unitat entonativa. Un dels avantatges del sistema és que és automatitzable, partint d'un programa d'estilització de contorns melòdics que defineix els punts que posteriorment es codifiquen mitjançant els símbols d'INTSINT, separant així la representació fonètica de la representació prosòdica. L'aplicació a diverses llengües ha donat resultats encoratjadors (Hirst, Nicolas i Espesser, 1991; Hirst et al., 1993) i aquesta és la tècnica de codificació prosòdica d'una part d'EUROM.1 que s'utilitza en el projecte MULTEXT al qual es fa referència més endavant (Hirst, Ide i Véronis, 1994).

Un altre sistema creat per tal de respondre a les necessitats de codificar prosòdicament les bases de dades és TOBI -*Tone and Break Indices* (Silverman et al., 1992), inspirat en la fonologia prosòdica i desenvolupat principalment per a l'anglès americà, encara que el sistema es pretengui universal. La transcripció es divideix en quatre nivells (*tiers*): el de la

representació ortogràfica, el nivell en el qual es representen els índexs, el nivell tonal, i un nivell per a comentaris del transcriptor. Els índexs (*break index*), codificats numèricament en una escala del 0 al 4, indiquen el grau de coherència o separació entre paraules adjacents, mentre que els tons marquen els moviments melòdics, tant a l'interior de les unitats melòdiques com en els límits entre unitats i també els accents tonals (*pitch accent*); es codifiquen per mitjà de les lletres H (*high* -alt) i L (*low* -baix), a les quals s'afegeixen símbols per a indicar el seu abast. TOBI pot utilitzar-se conjuntament amb el programa comercial d'anàlisi acústica Waves™.

Cal afegir només que les recomanacions d'EAGLES pel que fa a la transcripció fonètica dels elements segmentals i dels suprasegmentals segueixen molt de prop les propostes de SAM.

3.3.3. Alineació temporal, segmentació i etiquetat

Mitjançant l'alineació temporal (*time alignment*) establim la correspondència entre el senyal acústic i la seva representació simbòlica, tant si és ortogràfica com fonètica. Per a això és necessària una segmentació, entesa com aquella operació que es realitza sobre el senyal sonor per tal d'introduir marques que assenyalin el principi i el final de cadascuna de les unitats fonètiques. Ambdues operacions estan estretament lligades a l'etiquetat (*labelling*), consistent en identificar cadascuna de les unitats amb un símbol de transcripció.

A fi de realitzar aquestes operacions, el primer que ens cal és un conjunt de programes que ens permetin d'accedir al senyal i visualitzar-lo de diverses maneres. En el projecte SAM s'ha desenvolupat una eina per a l'anàlisi acústica del senyal denominada PTS-*Progiciel de Traitement de Signal* (Caerou et al., 1992), adaptada a l'estació de treball SESAM, però existeixen també en el mercat diferents alternatives comercials per a l'anàlisi acústica de la parla amb diferents graus de complexitat i per a diferents entoms informàtics.

La segmentació es pot portar a terme o bé de manera manual o bé de manera semi-automàtica, tot i que aquesta darrera necessita una verificació posterior. El programa PTS esmentat abans permet fer una segmentació manual dels enunciats, però també és possible aquesta operació amb programes comercialitzats. Les possibilitats de la segmentació automàtica obren noves perspectives a la lingüística de corpus, permetent de tractar de

manera eficient i ràpida grans quantitats de dades, encara que, de moment, sigui necessari un cert grau de verificació manual del procés.

Tal com indiquen Barry i Fourcin (1992) l'etiquetat d'un corpus pot realitzar-se a diferents nivells, que comprenen, segons la proposta d'aquests autors el nivell físic -en el qual es defineixen propietats acústiques del senyal com la periodicitat, els canvis espectrals, els sorolls d'alta freqüència-, el nivell acústic fonètic -en el qual s'empren etiquetes corresponents a esdeveniments com oclusions, explosions, aspiracions-, el nivell de transcripció fonètica estreta, el nivell de transcripció fonèmica -definit en termes dels segments que són distintius a la llengua o en termes de les formes canòniques de les paraules- i el nivell prosòdic. El nivell de transcripció que es determini vindrà donat per les aplicacions del corpus.

L'etiquetat és una operació que també pot realitzar-se de manera manual o (semi)automàtica; en ambdós casos es porta a terme una alineació entre el senyal sonor i les etiquetes que indiquen la categoria fonètica de les unitats. Naturalment, els programes d'etiquetat automàtic necessiten un entrenament previ i un "coneixement" de les categories fonètiques de la llengua, però juntament amb les possibilitats de segmentació, faciliten, com es deia abans, el tractament de quantitats important de dades (vegi's, com a exemple de treballs recents en aquest àmbit, Angelini et al., 1993; Blomberg i Carlson, 1993; Eisen, 1993; De Ginestel-Mailland, de Calmès i Perennou, 1993; o Hernáez, Barandiarán i Monte, 1993).

S'han desenvolupat en el marc de SAM programes d'etiquetat semi-automàtic per al danès (DKISALA; Andersen i Dalsgaard, 1992), el noruec (ELABSEG; Svendsen i Kvale, 1992) i el francès (SAPHO; IRIT, 1991), juntament amb un programa (ELSA; CRIN-INRIA, 1992) que compara l'etiquetat manual i l'automàtic per a verificar la validesa d'aquest darrer.

3.4. *Gestió i tractament de les dades*

Per tal que la informació continguda en el corpus sigui fàcilment accessible, cal disposar de programes de gestió de les dades, que permetin organitzar-les i recuperar-les atenent a diversos criteris com pot ser l'enunciat, les característiques del locutor o la presència de determinats elements. El programa RISE del projecte SAM (Castagneri i

Senia, 1990) permet de portar a terme aquests funcions en el corpus EUROM.1 al qual s'ha fet referència en l'apartat 2.2.1.2.

Per altra banda, cal també considerar una qüestió més general com són les eines per a l'explotació del corpus, però el tema ultrapassa els objectius d'aquesta presentació. Esmentarem que, pel que fa als corpus textuals, inclosos els que contenen transcripcions de llengua oral, es va iniciar el gener de 1994 el projecte LRE 62-050 MULTEXT - *Multilingual Text Tools and Corpora,*[8] que té com a objectiu final la difusió pública d'un conjunt multilingüe de corpus acompanyats d'eines per al seu tractament, tant pel que fa a l'anotació -segmentadors, analitzadors morfològics, desambiguadors de parts de l'oració, eines per a l'assignació automàtica de marques prosòdiques- com a l'explotació -indexació, recuperació de la informació i tractament estadístic-. Algunes de les eines desenvolupades en el projecte SAM (SAM, 1992) permeten igualment el tractament dels materials de les bases de dades orals. En els grups de treball d'EAGLES es tracta també el tema des de la perspectiva textual i des de la perspectiva de les bases de dades orals.

3.5. *Disseminació de les dades: documentació i suport*

Una etapa important en la constitució d'un corpus oral és la documentació, que ha de contenir, com a mínim, informació sobre el contingut del corpus, l'enregistrament, els locutors i l'organització i gestió del material. Sembla clar que com més completa sigui la documentació més fàcilment utilitzable serà el corpus per als usuaris finals. EAGLES té entre els seus objectius establir els mínims necessaris per a la documentació d'un corpus.

Cal plantejar-se, en última instància, en quin suport final s'oferirà el corpus als usuaris.

L'alternativa més habitual és el CD-ROM (*Compact Disc-Read Only Memory*) tant pel que fa a les bases de dades orals (Garofolo i Pallet, 1989) com als corpus textuals. Un cop definit el suport, queda encara organitzar una bona infraestructura de distribució i

---

[8] Per a més informació sobre el projecte hom pot adreçar-se al seu coordinador: Professor Jean Véronis, Laboratoire Parole et Langage, URA 261 CNRS, Université de Provence, 29, Avenue Robert Schuman, F-1361 Aix-en-Provence Cedex, France. Fax (33.42) 20.59.05. Correu electrònic: multext@univ-aix.fr. Com a participants associats formen part del consorci del projecte la Fundació Bosch Gimpera (Universitat de Barcelona) i la Universitat Autònoma de Barcelona.

Consulteu també els servidors www: http://etext.virginia.edu/TEI.html; http://www.tei.uic.edu/orgs/tei; http://info.ox.ac.uk/archive/teilite

d'intercanvi de materials. En l'apartat 4.3, dedicat a les iniciatives de disseminació, es fa referència a centres com el LDC (*Linguistic Data Consortium*) o a una experiència en curs com RELATOR.

4. Iniciatives en corpus orals: estàndards i disseminació

Es presenten en aquest apartat algunes de les principals iniciatives relacionades l'estandarditzacio i la difusió de corpus orals. En primer lloc es fa referència als projectes d'estandarditzacio que tenen per objectiu tant els corpus textuals com els orals, i en segon lloc es presenten els que més específicament es dediquen als corpus orals. Finalment, es recull informació sobre centres o projectes relacionats sobretot amb la disseminació de materials.[9]

4.1. *Iniciatives dedicades a corpus textuals i orals*

4.1.1.- TEI, *Text Encoding Initiative*[10]

La TEI és un projecte internacional iniciat el 1988, coordinat per la *Association for the Computers and the Humanities* (ACH), la *Association for Computational Linguistics* (ACL) i la *Association for Literacy and Linguistic Computing* (ALLC) i finançat per la Direcció General XIII de la Comissió Europea, el govern americà i la *Andrew W. Mellon Foundation.*

L'objectiu de la TEI és desenvolupar i difondre un format clarament definit que permeti l'intercanvi de textos en suport informàtic. Per a assolir-lo utilitza un llenguatge estàndard conegut com SGML (*Standard Generalized Mark-up Language*), que permet

---

9 Edwards (1993b) constitueix una font d'informació excel·lent que complementa les dades presentades en aquest apartat. Inclou indicacions sobre com accedir a centres i associacions, a bulletins de discussió i distribució per correu electrònic, i a arxius de textos.

10 La informació sobre la TEI pot obtenir-se adreçant-se a: Mr. Lou Burnard, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, UK. Fax (44.865) 273.275. Correu electrònic: lou@vax.ox.ac.uk. Existeix també un butlletí electrònic de discussió sobre la TEI, l'adreça del qual és: TEI-L@uicvm.uic.edu; per a subscriure's cal enviar el següent missatge: "subscribe tei-l <nom>". La versió P3 de les *TEI Guidelines* pot obtenir-se electrònicament enviant un misssatge a listserv@uicvm.uic.edu; amb "get teip3 package" s'obté el document el versió SGML; amb "get p3ascii package" s'obté una versió ASCII, i amb "get p3all package" les dues versions. La versió publicada pot adquirir-se a: TEI Orders, Oxford University Computing Services, 13 Banbury Road, Oxford OX2 6NN.

diferenciar clarament el text de les marques de codificació que indiquen la seva estructura (Bryan, 1988).

El treball es realitza en quatre comissions dedicades a documentació, representació textual, anàlisi i interpretació, i metallenguatge i sintaxi, i els resultats es presenten en una sèrie de guies per a la codificació i intercanvi de textos en format electrònic editades per C. M. Sperberg-McQueen (Universitat d'Illinois, Chicago) i L. Burnard (Universitat d'Oxford). L'última versió disponible és la versió P3 (Sperberg-McQueen i Burnard (eds.) 1994).

Els estàndards de la TEI descriuen la manera de documentar un text que s'inclou en un corpus, i defineixen també diverses etiquetes per a cadascuna de les parts d'un text (capítols, seccions, notes, etc.); alhora, s'han desenvolupat recomanacions per a la descripció morfològica i sintàctica i, el que ens interessa més aquí, per a la transcripció ortogràfica de la llengua oral. Pel que fa a aquest aspecte, es distingeixen els elements estructurals següents: informació contextual, informació temporal, enunciats, pauses, elements vocalitzats semi-lèxics i no lèxics, esdeveniments cinètics, altres tipus d'esdeveniments comunicatius i elements presentats en forma escrita al parlant. La TEI també proporciona recomanacions per a la transcripció de la superposició de torns de paraula, de la forma de les paraules quan aquesta no és estàndard, dels trets anomenats paralingüístics -velocitat d'elocució, intensitat, rang tonal, tensió, ritme i qualitat de veu- i dels fenòmens propis de la parla espontània. Pel que fa a la representació fonètica, la TEI recomana la utilització de l'Alfabet Fonètic Internacional. Les indicacions sobre la transcripció de la llengua oral es troben en el capítol 11 de la versió P3 (Sperberg-McQueen i Burnard (eds.) 1994).

4.1.2.- NERC, *Network of European Reference Corpora*[11]

NERC és una iniciativa finançada per la Comissió Europea (1991-1993), en la qual varen participar sis centres de recerca europeus -Universitat de Birmingham, Institut de Lexicologia Neerlandesa de Leiden, Universitat de Málaga, Institut de la Llengua Alemanya a Mannheim, Institut Nacional de la Llengua Francesa i Institut de Lingüística Computacional de Pisa- amb l'objectiu de buscar una aproximació científica i

[11] Per a obtenir més informació sobre NERC cal adreçar-se al seu coordinador: Professore Antonio Zampolli, Istituto di Linguistica Computazionale, CNR Università di Pisa, Via della Faggiola 32, 56100 Pisa, Itàlia. Fax (39.50)58.90.55. Correu electrònic: glottolo@icnucevm.cnuce.cnr.it

metodològica comuna al disseny de corpus i determinar les millors estratègies per a la construcció de corpus tant en el nivell nacional com en l'internacional. La coordinació del projecte ha estat a càrrec d'A. Zampolli (*Istituto di Linguistica Computazionale* -CNR, Pisa), i els resultats poden veure's en l'informe final (NERC, 1994). El resultat d'alguns dels treballs portats a terme per al castellà en el marc del NERC es presenten a Alvar i Villena (coord.) (1994).

Entre els treballs del NERC s'inclouen una sèrie de recomanacions sobre el format dels corpus de llengua oral, entesos en la tradició de la lingüística de corpus a la qual s'ha fet referència al principi. Els capítols 3B i 5.2 de l'informe final es dediquen, respectivament, a la representació textual i a l'anotació fonètica i prosòdica de la llengua oral (Sinclair, 1994). Pel que fa a la transcripció de la llengua oral, es recomanen les convencions desenvolupades per JP French en el marc del projecte COBUILD (French 1991, 1992); en aquest sistema es consideren quatre nivells de representació, en cadascun dels quals s'introdueix una codificació més rica per tal de reflectir de manera més acurada la realitat fonètica del text transcrit. Els quatre nivells adoptats pel NERC són els següents:

Nivell I: representació ortogràfica amb signes de puntuació i sense cap informació sobre la interacció entre els parlants.

Nivell II: representació ortogràfica augmentada amb informació sobre la identitat dels parlants, el canvis de torn de paraula i sobre elements no verbals.

Nivell III: inclou els límits entre unitats melòdiques i la codificació de les síl·labes tòniques, juntament amb indicacions sobre el solapament de parlants.

Nivell IV: inclou anotació sobre les característiques tonals de les síl·labes per tal de codificar la informació prosòdica, i l'alineació entre el senyal acústic i la transcripció fonèmica. En aquest nivell, el NERC recomana que s'inclogui una representació digital del senyal acompanyada d'una representació espectrogràfica i de la corba melòdica.

Una acurada anàlisi portada a terme per Payne (1992) permet d'establir que el sistema adoptat pel NERC és compatible amb el de la TEI.

4.1.3. EAGLES, *Expert Advisory Group on Language Engineering Standards*[12]

EAGLES (1992-1995) és un grup de treball promogut per la Direcció General XIII de la Comissió Europea en el marc de les accions horitzontals del programa LRE (*Linguistic Research & Engineering*). La supervisió del programa corre a càrrec d'un consell de direcció en el qual estan representats els principals projectes europeus amb finançament comunitari i les associacions europees relacionades amb les tecnologies de la parla i el processament del llenguatge natural. De la coordinació del grup es responsabilitza A. Zampolli (*Istituto di Linguistica Computazionale* - CNR, Pisa).

L'activitat d'EAGLES es basa en cinc grups de treball formats per experts tant del món de la universitat com de l'empresa: Corpus Textuals (amb seu al *Instituto Cervantes*, Alcalá de Henares), Lèxics Computacionals (amb seu a GSI ERLI, París), Formalismes Lingüístics (amb seu al DFKI -*Deutsches Forschungzentrum für Künstliche Intelligenz*, Saarbrueken), Avaluació (amb seu al CST -*Center for Sprogteknologi*, Copenhaguen) i Llengua Oral (amb seu a Vocalis Ltd, Cambridge). Cada grup de treball té un president i un organisme que actua com a seu.

L'objectiu general d'EAGLES és la definició d'especificacions i d'orientacions per a la descripció i la representació de recursos lingüístics, i el desenvolupament de mètodes per a l'avaluació de productes i serveis lingüístics. Aquesta tasca s'ha de portar a terme creant un consens i implicant en el treball els principals projectes europeus en el camp de l'enginyeria lingüística.

Existeixen dos grups de treball a EAGLES que incideixen directament en el desenvolupament de corpus: el grup de Corpus Textuals i el de Llengua Oral. El primer -presidit per A. Zampolli- té un programa de treball encaminat a la definició d'estàndards

[12] La informació sobre EAGLES pot obtenir-se adreçant-se a: Sr. G Arrarte o J. Llisterri, Área de Investigación, Instituto Cervantes, Libreros 23, 28801 Alcalá de Henares, Madrid. Fax: (91) 883.50.10. Correu electrònic: eagles@cervantes.es o també en el servidor www: http://www.ilc.pi.cnr.it/EAGLES/home.html o també en el servidor www: http://www.idg.pi.cnr.it/EAGLES/home.html Vegeu també les presentacions publicades al *DGXIII Magazine* (Brinkhoff, 1993) i a *ELSNews* (EAGLES, 1993). *ELSNews* és una publicació d'ELSNET - *European Network of Excellence in Language and Speech*, una xarxa temàtica promoguda en el marc del programa ESPRIT que agrupa a més de cinquanta centres acadèmics i també gairebé cinquanta centres industrials; per a informació sobre ELSNET cal adreçar-se a: ELSNET, OTS. Utrecht University, Trans 10, 3512 JK. Utrecht. The Netherlands. Fax (31.30) 536.000. Correu Electrònic: elsnet@let.ruu.nl, Buccleuch Place, Edinburgh EH8 9LW, UK. Fax (44.31) 650.45.87. Correu electrònic: elsnet@ed.ac.uk. Servidor WWW: htpp://www.cogsci.ed.ac.uk/elsnet/home.html

pel que fa a la tipologia de textos i de corpus, representació textual, anotació lingüística, documentació i disseminació, eines per al tractament de corpus i corpus paral·lels. En el grup de Llengua Oral -presidit per R. Moore (RSE-DRA, Malvern, Anglaterra)- es pretenen consolidar els resultats aconseguits en el marc dels projectes ESPRIT SAM: el treball sobre corpus orals es centra en les àrees següents: disseny i representació, caracterització i descripció lingüística, caracterització i descripció física i formats i eines.

Existeix també un subgrup de treball comú als dos grups esmentats, específicament dedicat a la transcripció de corpus orals. Els resultats de la primera fase de treball d'EAGLES seran accessibles a finals de 1994.

4.2. *Iniciatives dedicades a corpus orals*

4.2.1. SAM, *Multilingual Speech Input/Output Assessment, Methodology and Standardisation*[13]

El projecte ESPRIT 2589 SAM *Multilingual Speech Input/Output Assessment, Methodology and Standardisation* es va desenvolupar entre 1989 i 1992, com a continuació del projecte ESPRIT 1541 del mateix nom iniciat el 1986, i va seguir fins el 1993 com a ESPRIT 6819 SAM-A *Speech Technology Assessment in Multilingual Applications,* encara que amb uns objectius més reduïts. La coordinació del projecte, en el qual han intervingut 28 centres de recerca europeus, ha anat a càrrec d'A. Fourcin (*University College London*, Anglaterra).

SAM ha tingut com a objectiu el desenvolupament d'estàndards europeus en tecnologies de la parla i s'ha centrat en tres àrees de treball: avaluació de sistemes de síntesi, avaluació de sistemes de reconeixement, i creació de bases de dades orals (Fourcin i Dolmazon, 1991). Pel que fa a aquest darrer aspecte, en el marc de SAM s'han desenvolupat els corpus coneguts com EUROM.0 i EUROM.1 (*cf.* 2.2.1.2), juntament amb una sèrie d'eines per a l'adquisició, transcripció, anotació i organització d'aquestes bases de dades (SAM 1992a, b i c). En l'apartat 3, dedicat a les etapes en la constitució d'un corpus, s'ha fet una breu descripció de diversos resultats del projecte.

---

[13] La informació sobre el projecte SAM pot obtenir-se adreçant-se a: Professor Adrian Fourcin, Department of Phonetics and Linguistics, University College London, Wolfson House, 4 Stephenson Way, London NW1, 2HE, UK. Fax: (44.71) 383.07.52. Correu electrònic: adrian@phonetics.ucl.ac.uk.

4.2.2. COCOSDA, *Coordinating Committee for Speech Databases and Speech Input/Output Systems Assessment*[14]

COCOSDA és un comité internacional sorgit el 1991 com a resultat d'una reunió de treball a Chiavari (Itàlia) ((Castagneri ed.), 1991) i que ha continuat les seves activitats amb *Workshops* a Banff (Canadà) (Jones i Mariani (eds.) 1992), a Berlin (1993) i a Yokohama (1994). El seu objectiu és coordinar les activitats en el camp de l'avaluació de sistemes de síntesi i reconeixement i en el de la creació i disseminació de bases de dades orals i els problemes derivats de la necessitat d'etiquetar-les. La coordinació del grup és actualment a càrrec de A. Fourcin (*University College London*, Anglaterra).
COCOSDA es divideix en tres grups de treball, dedicats a cadascun dels temes esmentats abans; el de síntesi és coordinat per L. Pols (Universitat d'Amsterdam), el de reconeixement per G. Castagneri (CSELT, Torí) i el de corpus i etiquetat per D. Pallet (NIST, EEUU).

Per tal de donar unitat i coherència a les activitats europees ha sorgit una iniciativa finançada dins el programa LRE (*Linguistic Research & Engineering*) conegut com a *EuroCocosda -European Interface to COCOSDA*. Entre els projectes d'EuroCocosda es compta la constitució de dos corpus orals: TED (*Transnational English Database*), obtingut a partir de les gravacions de les comunicacions presentades al congrés *Eurospeech'93* celebrat a Berlin i la part europea de POLYPHONE, un corpus recollit a través de trucades telefòniques per a aplicacions en aquest àmbit.

4.3. *Iniciatives per a la disseminació de corpus*

4.3.1. LDC, *Linguistic Data Consortium*[15]

El *Linguistic Data Consortium* és una agrupació americana d'universitats, empreses i centres del govern organitzada i finançada per l'*Advanced Research Projects Agency*

[14] Per a més informació sobre COCOSDA i EUROCOCOSDA: Professor Adrian Fourcin, Department of Phonetics and Linguistics, University College London, Wolfson House, 4 Stephenson Way, London NW1, 2HE, UK. Fax: (44.71) 383.07.52. Correu electrònic: cocos@phonetics.ucl.ac.uk. o bé euro@phon.ucl.ac.uk. Existeix un butlletí electrònic de discussió sobre temes generals (cocosda@atr.co.jp) i un d'específicament dedicat a corpus (cocosda_corp@atr.co.jp). Servidor www: http://www.itl.atr.co.jp/cocosda

[15] Per entrar en contacte amb el LDC cal adreçar-se a: The Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305. Fax: (1.215).573.21.75. Correu electrònic: ldc@unagi.cis.upenn.edu. Servidor www: http://www.cis.upenn.edu/~ldc

(ARPA). La funció del consorci és la distribució de recursos lingüístics, coordinant la recollida de materials entre diversos centres i encarregant-se després de fer-los públics entre els socis. La seu del LDC és a la Universitat de Pennsylvania i és dirigit per M. Liberman. Actualment compta amb 70 membres, que contribueixen al manteniment del centre amb quotes anuals diferents segons si es tracta d'una universitat o una empresa.

Cal assenyalar que el LDC no distribueix únicament corpus americans, sinó que també té en el seu catàleg corpus provinents de projectes europeus. La taula següent resumeix els corpus orals distribuïts a través del LDC:[16]

Nom	CDs	Locutors	Contingut	Finançament
TIDIGITS	3	326	Dígits llegits	Texas Instruments
TIMIT	1	630	Frases llegides	ARPA
NTIMIT	2	630	Frases llegides	NYNEX
RM1	4	144	Frases llegides	ARPA
RM2	2	4	Frases llegides	ARPA
ATIS0	6	36	Frases llegides i parla espontània	ARPA
ATIS2	4	351	Parla espontània	ARPA
TI-46	1	46	Paraules aïllades llegides	Texas Instruments
Road Rally	1	136	Frases llegides i parla espontània	DoD
WSJ0	12		Textos llegits	ARPA
ATIS	8	250	Parla espontània	ARPA
MAPTASK	8	64	Parla espontània	HCRC Edimburg
Switchboard	26	550	Diàlegs espontanis per telèfon	ARPA
SB-Credit Card	1	69	Diàlegs espontanis per telèfon	ARPA

16 La taula es basa en la presentació del LDC realitzada per J. Godfrey al Workshop de COCOSDA celebrat a Berlin el 24 de setembre de 1993.

OGI-MLT	2	200	Parla espontània per telèfon	OGI
OGI-SPL	1		Lectura i parla espontània per telèfon	OGI
WSJ-CSR1	32	124	Frases llegides	ARPA
ATIS3	8		Parla espontània	ARPA
YOHO	2		Autentificació de parlants	Govern Americà
KING	2		Identificació de locutor	Govern Americà
SWBSPKRS	2-3		Identificació de locutor	LDC
BU-FM Radio	2		Lectura	NSF/LDC
NYNEX-IW	2		Lectura, recollida per telèfon	LDC
BRAMSHILL	12		Conversa, recollida per telèfon	Govern Britànic
BREF		120	Lectura de textos	LIMSI

Taula 2. Corpus orals distribuïts en CD-ROM pel *Linguistic Data Consortium*.

Entre els actuals projectes del LDC cal esmentar COMLEX (*Common Lexical Database of English*), consistent en un lexicó que, a més d'informació sintàctica i semàntica, ha d'incloure la transcripció fonètica, la lectura de les paraules per part d'un nombre reduït de locutors i la concordança entre les formes aïllades i les formes en parla contínua. El LDC també es proposa coordinar l'adquisició i distribució de POLYPHONE, un corpus multilingüe recollit per via telefònica amb uns 5.000 parlants per llengua que conté entre 20 i 40 enunciats per parlant, consistents en material fonèticament equilibrat, dígits, i una sèrie de paraules per a desenvolupar aplicacions que permetin efectuar verbalment operacions habituals com per exemple tornar a marcar un número de telèfon.

4.3.2. ECI, *European Corpus Initiative*[17]

La ECI és una iniciativa sorgida de la *Association for Computational Linguistics* (ACL) per tal de difondre en CD-ROM corpus de diverses llengües europees codificats segons els estàndards de la TEI. El material recollit fins ara es distribueix en un CD-ROM que, amb el títol de *The European Corpus Initiative Multilingual Corpus I* (ECI/MCI), conté aproximadament 93 milions de paraules i es divideix en 48 subcorpus en 26 llengües diferents. Aquests subcorpus contenen materials extrets de fonts escrites, excepte les transcripcions de llengua oral del *Corpus Oral de Referencia del Español Contemporáneo* (descrit a l'apartat 2.3) i transcripcions de programes de ràdio en holandès.

4.3.3. RELATOR, *European Network of Repositories for Linguistic Resources*[18]

Es tracta d'un projecte iniciat el gener de 1994, finançat en el marc del programa LRE (*Linguistic Research & Engineering*) de la Direcció General XIII de la Comissió Europea. L'objectiu és definir un marc organitzatiu per a la creació, recopilació, verificació, normalització i redistribució de recursos lingüístics, tant escrits com orals. El projecte és coordinat per A. Zampolli (*Istituto di Linguistica Computazionale* -CNR Pisa) i hi participen el LIMSI-CNRS (Orsay, França), el DFKI -*Deutsches Forschungzentrum für Künstliche Intelligenz* a Saarbrueken, la Universitat d'Edimburg, el *Center for Sprogteknologi* de Copenhaguen i l'*Institut de la Communication Parlée* de Grenoble.

5. Conclusions

Al llarg d'aquest repàs, necessàriament limitat, als diversos tipus de corpus orals de què disposem actualment, a les etapes en la constitució d'un corpus oral i a les principals iniciatives, s'ha pogut advertir que els corpus orals constitueixen recursos lingüístics

[17] Per a més informació hom pot adreçar-se a: Professor Henry S. Thompson, Center for Cognitive Science, University of Edinburgh, 2 Buccleuch Place Edinburgh EH8 9LW, Scotland. Fax: (44.31) 650.44.28. Correu electrònic: eucorp@cogsci.edinburgh.ac.uk. També pot obtenir-se informació sobre el corpus i la manera d'adquirir-lo per ftp anònim a scott.cogsci.ed.ac.uk/pub/elsnet/eci o al servidor WWW: http:/www.cogsci.ed.ac.uk/elsnet/eci.html

[18] Més informació sobre el projecte pot obtenir-se del seu coordinador: Professore Antonio Zampolli, Istituto di Linguistica Computazionale, CNR Università di Pisa, Via della Faggiola 32, 56100 Pisa, Itàlia. Fax (39.50)58.90.55. Correu electrònic: glottolo@icnucevm.cnuce.cnr.it. Sevidor www: http:// cristal.icp.grenet.fr/Relator/homepage.html

d'importància cabdal tant en el camp de la lingüística com en el de les indústries de la llengua. Per una banda, els corpus de llengua oral permeten la descripció de la llengua en tots els nivells d'anàlisi i fan també possible l'estudi de la variació geogràfica, social i d'estil, juntament amb l'observació de l'ús de la llengua en diverses situations. Per una altra, els corpus orientats a les tecnologies de la parla fan possible obtenir el coneixement lingüístic necessari per a la síntesi, entrenar i verificar sistemes de reconeixement, i dissenyar i entrenar sistemes de diàleg entre l'home i la màquina amb capacitat per a comprendre la parla natural, encara que sigui en un context restringit; s'ha vist també que, en aquest mateix àmbit, cada cop són més necessaris els corpus que proporcionin models de llenguatge als sistemes de reconeixement de parla contínua. Finalment, els corpus orals són a la base del desenvolupament d'altres recursos lingüístics com diccionaris de pronúncia o diccionaris que tinguin en compte la realitat de la llengua parlada.

Tot i que existeix una clara consciència de la necessitat de corpus orals, el procés de recollida és sovint costós i requereix l'esforç coordinat de diverses institucions o grups, recolzat per una font de finançament important. Però encara és més costosa la preparació del corpus per a la seva utilització posterior. La situació ideal d'un corpus oral és que estigui segmentat, fonèticament etiquetat i prosòdicament anotat; si és necessari també un treball sobre el text, aquest hauria d'estar codificat i amb una anotació lingüística que inclogués la categoria gramatical de cada mot (*tagging*) i l'estructura sintàctica dels enunciats (*parsing*). Encara que totes aquestes operacions es poden realitzar de manera (semi)automàtica, hi ha sempre un procés de verificació manual que consumeix alhora temps i recursos.

Seria desitjable que un corpus, igual que els altres recursos lingüístics, fos reutilitzable, ja que, com s'acaba de veure, constitueix una inversió important. No només haurien de poder-se emprar en nous projectes els materials sense processar -tant si es tracta del senyal sonor com de la transcripció ortogràfica de la llengua parlada-, sinó que també hauria de ser factible recuperar l'etiquetat i l'anotació a tots els nivells en què s'hagin realitzat. Això només es possible si es treballa d'acord amb estàndards comuns, alguns dels quals s'han presentat més amunt. El problema per a l'investigador és que, en determinats àmbits, no existeix un únic estàndard. Mentre que algunes propostes han tingut una difusió molt àmplia i poden considerar-se acceptades per gairebé tota la comunitat científica -per exemple la codificació de la TEI en corpus textuals o els

estàndards desenvolupats pel projecte SAM pel que fa a bases de dades orals-, en molts camps encara no s'ha imposat un sistema adoptat per la majoria.

És important, per això, que els nous projectes que sorgeixin, o els que encara estan en fase de desenvolupament, es portin a terme tenint en compte les iniciatives relacionades amb l'estandardització a les que s'ha fet referència, comptant amb la informació adequada sobre els estàndards emergents en cada camp. Això no només possibilita la reutilització dels corpus, sinó que també fa més fàcil l'intercanvi de materials en un context molt ampli.

Finalment, a més de la coordinació amb projectes i iniciatives internacionals, és també desitjable d'evitar la duplicació d'esforços per a la mateixa llengua, incentivant la col·laboració i la programació conjunta en el treball dels grups d'un mateix àmbit lingüístic. En el cas del català, la jornada de treball sobre "Compatibilitat i accessibilitat dels corpus en llengua catalana", celebrada el 6 de maig de 1994 a la Universitat Pompeu Fabra, ha estat un pas endavant important i s'espera que tingui una continuïtat en altres trobades similars.

Bibliografia

ALTENBERG, B. (1991), *A bibliography of publications relating to English computer corpora*, dins, S. Johansson i A. Stenström, (eds.) *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín, pàg. 355-396.

ALVAR EZQUERRA, M. i VILLENA PONSODA, J. A. (coords.) (1994), *Estudios para un corpus del español*, "Anejo 7 de Analecta Malacitana".

ALLEN, G. D. (1988), *The PHONASCII System*, "Journal of the International Phonetic Association", 18, 1, pàg. 9-25.

ANDERSEN, O. i DALSGAARD, P. (1992), *DKISALA V1.1- Users Guide (SAM-IES-059)*, dins *SAM User Guide to ETR Tools*. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. ref. SAM-UCL-G007.

ANGELINI, B., BRUGNARA, F., FALAVIGNA, D., GIULIANI, D., GRETTER, R. i OMOLOGO, M. (1993), *Automatic segmentation and labelling of English and Italian speech databases*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 653-656.

ARRARTE, G. i LLISTERRI, J. (1994), *Informe sobre recursos lingüísticos para el español (I): Corpus escritos y orales disponibles y en desarrollo en España*, Instituto Cervantes, Madrid.

ATKINS, S., CLEAR, J. i OSTLER, N. (1992), *Corpus design criteria*, "Literary and Linguistic Computing", 7, 1, pàg. 1-16.

BADIA, T., CABRÉ, M. T., LLISTERRI, J. i DE YZAGUIRRE, Ll. (1994), *Recursos en llengua catalana: estat de la qüestió*, "Jornada de Compatibilitat i

accessibilitat dels corpus de dades en llengua catalana", Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

BARRY, W. J. i FOURCIN, A. J. (1992), *Levels of Labelling*, "Computer Speech and Language", 6, pàg. 1-14.

BLOMBERG, M. i CARLSON, R. (1993), *Labelling of speech given its text representation*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 3, pàg. 1775-1778.

British National Corpus - Spoken Corpus Transcription Guide TGCW (1991), 21, 18.

BONNEAU-MAYNARD, H., GAUVAIN, J. L., GOODINE, D., LAMEL, L. F., POLIFRONI, J. i SENEFF, S. (1993), *A French version of the MIT-ATIS system: portability issues*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 3, pàg. 2059-2062.

BRINKHOFF, N. (1993), *Towards standards in language engineering: EAGLES*, "DGXIII Magazine", pàg. 25-27.

BRUCE, G. (1988), *2.3. Suprasegmental categories and 2.4. The symbolization of temporal events*, "Journal of the International Phonetic Association", 18, 2, pàg. 75-76.

_____ (1989), *Report from the IPA working group on suprasegmental categories*, "Lund University Department of Linguistics, General Linguistics, Phonetics, Working Papers", 35, pàg. 25-40.

BRYAN, M. (1988), *SGML: An Author's Guide to the Standard Generalized Markup Language*. Wokingham, Addison-Wesley.

CAEROU, J. C., DOLMAZON, J. M., EL BADMOUSSI, A., JONES, K. i BARRY, B. (1992), *PTS Software V.4.40, user manual*, dins *SAM User Guide to ETR Tools*,

ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

CARRÉ, R. (1991), *Los bancos de sonidos*, dins J. Vidal Beneyto, (dir.) *Las industrias de la lengua*, [versió cast., a cura de M. Alvar et al., Fundación Sánchez Ruipérez-Pirámide, Salamanca-Madrid, pàg. 108-118.]

_____ (1992), *Speech Databases* dins W. A. Ainsworth, (ed.) *Advances in Speech, Hearing and Language Processing. A Research Annual*, 2, Jai Press, Londres, pàg. 199-216.

CASACUBERTA, F., GARCÍA, R., LLISTERRI, J., NADEU, C., PARDO, J. M. i RUBIO, A. (1992), *Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)*, "Procesamiento del Lenguaje Natural, Boletín", 12, pàg. 35-42.

CASTAGNERI, G. (ed.) (1991), *Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods*.

CASTAGNERI, G. i SENIA, F. (1990), *SAM-RISE v 1.1 - User guide. Relational Interface for Speech Evaluation (SAM-CT-105)* dins *SAM User Guide to ETR Tools*, ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

CHAFE, W. L., DU BOIS, J. W. i THOMPSON, S. A. (1991), *Towards a New Corpus of Spoken American English*, dins K. Aijmer, i B. Altenberg (eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Longman, Londres, pàg. 65-82.

CHURCH, K. W. i MERCER, R. L. (1993), *Introduction to the Special Issue on Computational Linguistics Using Large Corpora*, "Computational Linguistics", 19. 1, pàg. 1-24.

CRIN-INRIA (1992), *ELSA-ESPRIT labelling system assessment software. User's guide for V2.4 (SAM-UCL/CRIN-042)*, dins *SAM User Guide to ETR Tools*. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

DE GINESTEL-MAILLAND, A., DE CALMÈS, M. i PÉRENNOU, G. (1993), *Multi-Level Transcription of Speech Corpora from Orthographic Forms* dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 2, pàg. 1441-1444.

DÍAZ, J. E., RUBIO, A. J., PEINADO, A. M., SEGARRA, E., PRIETO, N. i CASACUBERTA, F. (1993), *Development of task-oriented Spanish speech corpora*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology.*

DOLMAZON, J. M. (1994), *BDSONS*, dins L. Jackson-Eve (comp.) *Language Engineering Convention, CNIT*, La Défense, París, pàg. 89-90.

DRAXLER, C., TILLMANN, H. G. i EISEN, B. (1993), *Prolog Tools for Accessing the PhonDat Database of Spoken German* dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 191-194.

DU BOIS, J. W., SCHUETZE-COBURN, S., CUMMING, S. i PAOLINO, D. (1993), *Outline of discourse transcription*, dins J. A. Edwards, i M. D. Lampert, (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 45-89.

EAGLES (1993), *EAGLES Working Groups Report: Text Corpora Working Group / Spoken Language Working Group*, "ELSNews", 2, 2, pàg. 4-5.

EDWARDS, J. A. (1992), *Design principles in the transcription of spoken discourse* dins J. Svartvik, (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, Mouton de Gruyter, Berlín, pàg. 129-147.

_____ (1993a), *Principles and Contrasting Systems of Discourse Transcription*, dins J. A. Edwards, M. D. Lampert, (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 3-31.

_____ (1993b), *Survey of Electronic Corpora and Related Resources for Language Researchers*, dins J. A. Edwards, i M. D. Lampert, (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 263-310.

EISEN, B. (1993), *Reliability of speech segmentation and labelling at different levels of transcription* dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 673-676.

ENGSTRAND, O. (1987), *The IRIS speech data base-a status report*, "RULL, Reports from the Uppsala University Department of Linguistics", 17, pàg. 121-126.

ESLING, J. H. (1988), *Computer coding of IPA symbols and detailed phonetic representations of computer databases*, "Journal of the International Phonetic Association", 18, 2, pàg. 99-106.

_____ (1990), *Computer Coding of the IPA: Supplementary Report*, "Journal of the International Phonetic Association", 20, 1, pàg. 22-26.

FOURCIN, A., DOLMAZON, J. M. (1991), *Speech knowledge, standards and assessment*, dins *Actes du XIIème Congrès International des Sciences Phonétiques*. 5, Aix-en-Provence, Université de Provence, pàg. 430-433.

FRENCH, J. P. (1991), *Updated notes for soundprint transcribers*, "Working paper", october 1991, University of Birmingham, NERC-WP 4-47.

_____ (1992), *Transcription proposals: multilevel system*, "Working paper", october 1992, University of Birmingham, NERC-WP 4-50.

GAROFOLO, J. S., PALLET, D. S. (1989), *Use of CD-ROM for speech database storage and exchange* dins J. P. Tubach i J. J. Mariani (eds.) *Eurospeech 89. European Conference on Speech Communication and Technology*, CEP Consultants Ltd., Edinburg, pàg. 309-312.

GARSIDE, R., LEECH, G. i SAMPSON, G. (eds.) (1987), *The Computational Analysis of English: A corpus-based approach*, Longman, Londres.

GIBBON, D. (1989), *Survey of Prosodic Labelling for EC Languages. SAM-UBI-1.90, dins ESPRIT 2589 (SAM) Interim Report, Year 1,* ref. SAM-UCL G002, University College London.

GLASS, J., GOODINE, D., PHILLIPS, M., SAKAI, S., SENEFF, S. i ZUE, V. (1993), *A bilingual Voyager system*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 3, pàg. 2063-2065.

GREENBAUM, S. i SVARTVIK, J. (1990), *The London-Lund Corpus of Spoken English* dins J. Svartvik, (ed.) *The London-Lund Corpus of Spoken English. Description and Research*, Lund University Press, Lund, pàg. 11-63.

GREENBERG, J. A., FERGUSON, C. A. i MORAVCSIK, E. A. (eds.) *Universals of Human Language*, 2, Phonology, Stanford University Press, Stanford.

GUMPERZ, J. J. i BERENZ, N. (1993), *Transcribing Conversational Exchanges*, dins J. A. Edwards, i M. D. Lampert, (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 91-121.

HERNÁEZ, I., BARANDIARÁN, J. i MONTE, E. (1993), *A segmentation algorithm based on acoustical features using a self organizing neural network*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 661-664.

HIRST, D. (1991), *Intonation models: towards a third generation*, dins *Actes du XIIème Congrès International des Sciences Phonétiques*, Université de Provence, Aix-en-Provence, 1, pàg. 305-310.

HIRST, D. i DI CRISTO, A. (en premsa), *A survey of intonation systems* dins D. Hirst i A. Di Cristo, (eds.) *Intonation Systems. A Survey of 20 Languages*, Cambridge University Press, Cambridge.

HIRST, D., DI CRISTO, A., LE BESNERAIS, M., NAJIM, Z., NICOLAS, P. i ROMÉAS, P. (1993), *Multilingual modelling of intonation patterns*, dins D. House i P. Touati, (eds.) *Proceedings of an ESCA Workshop on Prosody*, "Lund University Department of Linguistics and Phonetics, Working Papers", 41, pàg. 204-207.

HIRST, D. J., IDE, N. i VÉRONIS, J. (1994), *Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project*, dins *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, Mohonk Mountain House, New Paltz, Nova York, pàg. 77-80.

HIRST, D. J., NICOLAS, P. i ESPESSER, R. (1991), *Coding the Fo of a continuous text in French: An experimental approach* dins *Actes du XIIème Congrès International des Sciences Phonétiques*, Université de Provence, Aix-en-Provence, 5, pàg. 234-237.

IPA (1993), *IPA Chart, revised to 1993*, "Journal of the International Phonetic Association", 23, 1.

IRIT (1991), *SAPHO. Installing and running IRIT-SALA (DSP Step, SAPHO Step) Version 2, Installing and running EVAL Software Version 1 (SAM-IRIT-10)* dins *SAM User Guide to ETR Tools*, ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

JONES, K. i MARIANI, J. (eds.) (1992), *Proceedings of the 1992 Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment*. Monday, Banff Springs Hotel, Banff.

KEATING, P. A., BYRD, D., FLEMMING, E. i TODAKA, Y. (1994), *Phonetic analysis of word and segment variation using the TIMIT corpus of American English*, "Speech Communication", 14, 2, pàg. 131-142.

KNOWLES, G. (1991), *Prosodic labelling: the problem of tone group boundaries*. dins S. Johansson i A. Stenström. (eds.) *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín, pàg. 149-163.

KNOWLES, G. i LAWRENCE, L. (1987), *Automatic intonation assignment* dins R. Garside, G. Leech, i G. Sampson, (eds.) *The Computational Analysis of English: A Corpus-based Approach*, Longman, Londres, pàg. 139-148.

KNOWLES, G., TAYLOR, L. i WILLIAMS, B. (1992), *A Corpus of Formal British English Speech*, Longman, Londres.

KOHLER, K. (1991), *Phonetic data bases for German* dins *Actes du XIIème Congrès International des Sciences Phonétiques*, Université de Provence, Aix-en-Provence, 2, pàg. 466-469.

LAMEL, L. F. (1992), *Report on Speech Corpora Development in the U.S.*, "NESCA - The European Speech Communication Association Newsletter", 8, pàg. 7-10.

LAMEL, L. F., GAUVAIN, J. L. i ESKÉNAZI, M. (1991), *BREF, a Large Vocabulary Spoken Corpus for French*. dins *Eurospeech 91. 2nd European Conference on Speech Communication and Technology*, 2, pàg. 505-508.

LAMEL, L. F., KASSEL, R. H. i SENEFF, S. (1986), *Speech database development: Design and analysis of the acoustic-phonetic corpus*, dins *Proceedings of the DARPA Speech Recognition Workshop.*

LEECH, G. (1991), *The State of the Art in Corpus Linguistics*, dins K. Aijmer i B. Altenberg, (eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Longman, Londres, pàg. 8-29.

LLISTERRI, J. (1994a), *EAGLES Spoken Texts, Draft Working Paper*. Draft Technical Report, Madrid, EAG-CSG/IR-T7.1.

_____ (1994b), *Prosody Encoding Survey*, WP 1 Specifications and Standards. T1.5. Markup Specifications. Deliverable 1.5.3. Final version, LRE project 62-050 MULTEXT.

LLISTERRI, J., AGUILAR, L., BLECUA, B., MACHUCA, M. J., DE LA MOTA, C., RÍOS, A., MORENO, A. i SALAVEDRA, J. (1993), *Spanish EUROM 1: Phonetic Contents*, Report D6 Appendix X, SAM-A/UPC/002, ESPRIT PROJECT 6819.

MacKAY, I. (1994), *Review of The Oxford Acoustic Phonetic Database on Compact Disk*, *Linguist List*, 5-256, "NESCA, Newsletter of the European Speech Communication Association", 14, pàg. 12-13.

MADDIESON, I. (1991), *Testing the universality of phonological generalizations with a phonetically specified segment database: results and limitations*, "UCLA Working Papers in Phonetics", 78, pàg. 11-25.

MADDIESON, I. i PRECODA, K. (1990), *Updating UPSID*, "UCLA Working Papers in Phonetics", 74, pàg. 104-111.

MARCOS MARÍN, F. (1991), *Corpus oral de referencia de la lengua española contemporánea* dins F. Marcos Marín, *Archivos Digitales*, Sociedad Estatal del V Centenario, Área de Industrias de la Lengua, pàg. 1-25.

MARCOS MARÍN, F., BALLESTER, A. i SANTAMARÍA, C. (1993), *Transcription Conventions used for the Corpus of Spoken Contemporary Spanish*, "Literary and Linguistic Computing 8", 4, pàg. 283-292.

MARIÑO, J. B. i LLISTERRI, J. (1993), *Spanish adaptation of SAMPA and automatic phonetic transcription*, dins *SAM-A/UPC/001/v1 20th April 1993*, ESPRIT PROJECT 6819.

MOORE, R. K. (1991), *User Needs in Speech Research*, dins *Proceedings of the Workshop on European Textual Corpora*, Pisa.

MORENO, A. (1993), *EUROM-1 Spanish Database*, Report D6, SAM-A/UPC/003.

MORENO, A., POCH, D., BONAFONTE, A., LLEIDA, E., LLISTERRI, J., MARIÑO, J. B. i NADEU, C. (1993), *ALBAYZIN Speech Database: Design of the Phonetic Corpus*, dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 175-178.

NERC (1994), *NERC-1 Network of European Reference Corpora. Final Report*, ILC - CNR, Pisa.

OCHS, E. (1979), *Transcription as Theory* dins E. Ochs, i B. B. Schieffelin, (eds.), *Developmental Pragmatics*, Academic Press, Nova York, pàg. 43-72.

OOSTDIJK, N. (1988), *A corpus linguistic approach to linguistic variation*, "Literary and Linguistic Computing" 3, 1, pàg. 12-25.

PAUL, D. B. i BAKER, J. M. (1992), *The design for the Wall Street Journal - based CSR Corpus*, dins *Proceedings of the Fifth DARPA Workshop on Speech and Natural Language*.

PAYNE, J. (1992), *Report on the compatibility of JP French's spoken corpus transcription conventions with the TEI guidelines for transcription of spoken texts*, "Working paper, COBUILD Birmingham and IDS Mannheim. December 1992", NERC-WP 8/WP 4-122.

PICKERING, J. B. i ROSNER, B. S. (1993), *The Oxford Acoustic Phonetic Database on Compact Disk*, Oxford University Press, Oxford.

RENOUF, A. (1987), *Corpus development*, dins J. Sinclair, (ed.) *Looking Up. An Account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Langage Dictionary*, Collins, Londres, pàg. 1-40.

ROUSSELOT, P. J. (1892), *Les modifications phonétiques du langage étudiées dans le patois d'une famille de Cellefrouin*, H. Welter, París.

SAM (1992), *User Guide to ETR Tools.* ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref., SAM-UCL-G007.

SCHMIDT, M. S., SCOTT, C. i JACK, M. A. (1993), *Phonetic transcription standards for European names (ONOMASTICA)* dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, 1, pàg. 279-282.

SELTING, M. (1987), *Descriptive categories for the auditive analysis of intonation in conversation*, "Journal of Pragmatics", 11, pàg. 777-791.

SHERWOOD, T. i FULLER, H. (1992), *Guide to EUROM.1 Speech Database,* Doc n. SAM-NPL-102.

SILVERMAN, K., BECKMAN, M., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., PIERREHUMBERT, J. i HIRSCHBERG, J. (1992), *TOBI: A standard for labelling English prosody*, dins *Proceedings of the Second International Conference on Spoken Language Processing, ICSLP-92*, Banff, pàg. 867-870.

SINCLAIR, J., (1994), *Spoken Language, Phonetic/Phonemic and Prosodic Annotation* dins *NERC-1 Network of European Reference Corpora*, ILC - CNR, Pisa.

SPERBERG-McQUEEN, C. M. i BURNARD, L. (eds.) (1994), *Guidelines for Electronic Text Encoding and Interchange, (TEI P3) Text Encoding Iniciatives*, Association for Computational Linguistics, Association for Computers and the Humanities i Association for Literary and Linguistic Computing, Chicago - Oxford.

SVENDSEN, T. i KVALE, K. (1992), *ELABSEG V2.5, Users Manual* dins *SAM User Guide to ETR Tools*, ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

TAYLOR, L., LEECH, G. i FLIGELSTONE, S. (1991), *A survey of English machine-readable texts*, dins S. Johansson i A. Stenström (eds.) *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín, pàg. 319-354.

TEUBERT, W. (1993), *Phonetic, Phonemic and Prosodic Annotation*, IDS Mannheim, NERC-WP 8-171.

TILLMANN, H. G. i POMPINO-MARSCHALL, B. (1993), *Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems* dins *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*, Berlín, 3, pàg. 1691-1694.

UCL (1992), *Speech acquisition and Annotation Protocols and Index of Mnemonics (SAM-UCL-018)- Section IV: SAMPA*, dins *SAM User Guide to ETR Tools*. ESPRIT PROJECT 2589, (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

WELLS, J. C. (1989), *Computer-coded phonemic notation of individual languages of the European Community*, "Journal of the International Phonetic Association", 19, 1, pàg. 31-54.

WELLS, J. C., BARRY, W., GRICE, M., FOURCIN, A. i GIBBON, D. (1992), *Standard Computer-Compatible Transcription. SAM Stage Report Sen.3 SAM UCL-037*, dins SAM (1992), *ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation*, University College London, Londres.

WICHMANN, A. (1991), *A study of up-arrows in the Lancaster/IBM Spoken English Corpus*, dins S. Johansson i A. Stenström, (eds.) *English Computer Corpora. Selected Papers and Research Guide*, Mouton de Gruyter, Berlín, pàg. 165-178.

ZEILIGER, J. i SERIGNAT, J. F. (1991), *Europec software V.4.1 User's Guide (SAM-ICP-045)* dins *SAM User Guide to ETR Tools.*, ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation, ref. SAM-UCL-G007.

ZUE, V., GLASS, J., GOODINE, D., HIRSCHMAN, L., LEUNG, H., PHILLIPS, M., POLIFRONI, J. i SENEFF, S. (1991), *The MIT ATIS system: Preliminary development, spontaneous speech data collection and performance evaluation*, dins *Eurospeech 91. 2nd European Conference on Speech Communication and Technology*, 2, pàg. 537-540.

ZUE, V., SENEFF, S. i GLASS, J. (1990), *Speech database development at MIT: TIMIT and beyond*, "Speech Communication" 9, 4, pàg. 351-356.

El *Diccionari del català contemporani* i el *Corpus textual informatitzat de la llengua catalana**

Joaquim Rafel i Fontanals
(Universitat de Barcelona)
(Institut d'Estudis Catalans)

1. Origen i justificació del projecte

El nom de *Diccionari del català contemporani* designa un programa de recerca de l'Institut d'Estudis Catalans que sorgí del desig de la Secció Filològica de poder incorporar a l'elaboració d'un futur diccionari descriptiu de la llengua els avenços científics, metodològics i tecnològics que la lexicografia ha assumit en els darrers anys. Una de les principals preocupacions de la Secció Filològica en aquells moments devien ésser justament els aspectes metodològics i tecnològics, perquè a principis de 1983, quan jo no era encara membre d'aquesta corporació acadèmica, m'encarregaren un informe sobre la utilització de la informàtica en un projecte d'elaboració d'un nou diccionari de la llengua moderna; aquest informe fou acabat el juliol de 1983. A l'hora de redactar-lo vaig

* En altres llocs hem donat informació sobre l'origen i el desenvolupament d'aquest projecte: *Cap a un diccionari del català contemporani*, dins *Segon Congrés Internacional de la Llengua Catalana, IV, Àrea 3: Lingüística Social*, Palma, 1992, pàg. 589-595. *El Corpus textual automatitzat de la llengua catalana* en col·laboració amb J. M. Solanellas dins *Actas de las II Jornadas Españolas de Documentación Automatizada. Ponencias y comunicaciones*, 1986, pàg. 147-161. El *Diccionari del català contemporani*, "Serra d'Or", 1987, pàg. 428-431. *El* Corpus textual informatitzat de la llengua catalana *i el* Diccionari del català contemporani. *Un proyecto del Institut d'Estudis Catalans*, Anthropos, 1988, pàg. V-VII. *El* Diccionari del català contemporani*: treballs realitzats i previsions de futur*, "Llengua & Literatura", 5, 1992-93 [1994], pàg. 733-737. *Le* Diccionari del català contemporani *et le* Corpus textual informatitzat de la llengua catalana. *Brève description du projet et état des travaux*, dins *Actes du XX Congrès International de Linguistique et Philologie Romanes*, IV, Tübingen-Basel, 1993, pàg. 816-821.

voler destacar, entre moltes altres qüestions a què ara no al·ludiré, que la utilització de mitjans informàtics per part del lingüista no es redueix a l'adopció d'un instrument de treball o d'un mitjà auxiliar qualsevol, sinó que obliga a replantejar el mètode i a sotmetre els propis coneixements a noves formulacions. En aquest informe vaig proposar de constituir un corpus textual informatitzat de la llengua catalana com a fase inicial, o, fins i tot, prèvia a allò que podem considerar pròpiament l'elaboració d'un diccionari. Aquesta proposta es basava en arguments com els que segueixen: la utilització dels ordinadors electrònics en la fase de recol·lecció i organització de les dades pot contribuir d'una manera decisiva a encaminar la tasca del lexicògraf cap a l'objectivitat que proporciona el discurs reflexiu a partir d'unes dades no sotmeses *a priori* a cap restricció més o menys arbitrària, i cap a l'exhaustivitat, la claredat i l'explicitud; el gran volum d'informació requerit per a poder aconseguir aquests objectius no pot ésser sotmès a un tractament manual; el caràcter exhaustiu de la documentació de què pot disposar el gabinet lexicogràfic per aquests procediments permet, d'altra banda, pensar a enfocar l'elaboració dels articles d'un diccionari a partir de mètodes diferents dels tradicionals (per exemple, l'establiment dels significats pot deixar d'ésser apriorístic; els sentits de cada mot poden ésser fixats d'acord amb els seus usos diversos, que es reflecteixen en les diferents ocurrències arxivades de cada un; les classificacions tradicionals poden ésser posades en qüestió a favor d'altres que estiguin més d'acord amb la realitat que hom vol descriure i que el lexicògraf pot tenir al davant en forma de relació exhaustiva d'ocurrències d'un mot amb els seus contextos, etc.). El concurs d'aquest tipus de mitjans permet, d'altra banda, d'exemplificar els diferents valors i usos de cada mot i les diferents estructures sintàctiques de què forma part amb frases preses dels textos -literaris, científics, tècnics, periodístics, etc.- que constitueixen el corpus; s'evita així, a més, l'exemple -sovint poc feliç- inventat pel lexicògraf mateix. No cal dir, afegiré encara, que un plantejament com aquest allunya l'activitat lexicogràfica de la lingüística especulativa i del dogmatisme, i l'apropa a la recerca basada en l'anàlisi racional de les dades empíriques. Com tots sabem, més d'un lingüista o lexicògraf ha posat de relleu en més d'una ocasió el valor del context a l'hora d'establir els significats dels mots; només esmentaré com a testimoni d'aquesta actitud una breu citació d'un treball de Robert-Léon Wagner, en què aquest lingüista s'expressa de la manera següent: «*Police* se prête à une analyse phonologique ou morphologique. Sémantiquement ce mot n'existe pas en soi. Sans l'aide de contextes,

impossible d'en rien dire. Pour définir correctement ses valeurs d'emploi successives il faut classer ces contextes chronologiquement et par domaines...» (*Reflexions naïves à propos des dictionnaires*, "Cahiers de Lexicologie", XXVII, 2, 1975, pàg. 82). En definitiva, un plantejament com aquest es declara tributari d'uns conceptes de mot i de significat com els que suggereix d'una manera poètica la frase que Josep Ferraté i Móra escrigué en el pròleg del llibre de Xavier Benguerel que agrupa els dos anomenats pel mateix autor «assaigs novel·lístics» *L'home dins el mirall* i *La màscara*: «La paraula és -ha de ser- com la pell, cenyida i elàstica, d'un cos vivent i sempre mòbil: la significació.» (Barcelona, 1986, pàg. 17).

La tardor de 1983 la Secció Filològica de l'Institut d'Estudis Catalans prengué l'acord d'iniciar les gestions necessàries per a poder tirar endavant un projecte de creació d'un corpus textual de la llengua catalana com el que era definit en el projecte esmentat, dins l'àmbit d'un programa de treball més ampli que prengué el nom de *Diccionari del català contemporani*; a continuació aquest programa fou assumit per l'Institut.

2. Característiques generals del corpus

A l'hora d'establir les característiques d'un corpus general de referència de la llengua, a més d'una sèrie de qüestions més específiques de què no podré ocupar-me al llarg d'aquesta breu exposició, han d'ésser-ne determinats tres aspectes fonamentals: el temporal, el qualitatiu i el quantitatiu.

a) L'aspecte temporal es refereix a les dates extremes dels textos que han de formar part del corpus. En el cas d'aquest, es tractava de determinar quina era la data més adequada per a constituir el punt de partida d'un corpus que havia de contenir textos publicats fins al moment més recent possible. Després de diverses propostes s'acordà de fixar aquest terme *a quo* al voltant de 1833, com a data simbòlica representativa de la represa de l'ús literari del català en l'època moderna, amb la qual cosa l'abast temporal del corpus és de 155 anys, si mantenim el tancament a 1988, que és la data dels darrers textos seleccionats.

b) L'aspecte qualitatiu es refereix al tipus de text o de llengua que s'ha de prendre en consideració per a formar part del corpus. La caracterització general dels textos tinguts en compte és la de textos o obres publicats, sigui quina sigui la seva forma; des d'aquest

punt de vista només s'ha fet excepció en el cas de la correspondència i d'escriptures notarials (contractes i testaments), que han estat tinguts en compte malgrat la seva condició de textos manuscrits o mecanoscrits, no publicats. Al marge d'aquesta qüestió, contràriament al que havia estat habitual en corpus ja establerts per a d'altres llengües, a l'hora de definir el que ara presentem decidírem de prendre en consideració no solament textos de caràcter literari, sinó de qualsevol altre tipus, com ja es desprèn de l'esment que acabo de fer a la correspondència i a les escriptures notarials. Pel que fa a la proporció entre la llengua literària i la no literària, inicialment havíem previst que fos del 60% i el 40%, respectivament, però a l'hora de seleccionar els textos a introduir en el corpus, comprovàrem que la gran riquesa i l'extraordinària varietat de les obres o publicacions de diversa naturalesa que no cabien sota la denominació de llengua literària ens obligava a invertir pràcticament l'ordre d'aquests percentatges, donant així una importància o un pes més gran als textos no literaris, a fi de mantenir una representativitat adequada, en certa manera proporcional al volum de textos publicats i a la varietat temàtica, com precisaré d'aquí a un moment, en parlar del procés de selecció.

c) Pel que fa a l'aspecte quantitatiu, que es refereix al que pròpiament anomenem extensió del corpus -la quantitat o volum de text de què ha de constar-, és evident que no pot ésser establert *a priori* amb exactitud en un corpus que, com el que ara presento, no té com a característica principal el nombre de mots de què consta, sinó el seu caràcter representatiu i equilibrat, tant des de la dimensió temporal, com des del punt de vista de la diversitat que anomenem genèricament tipològica; però és evident que s'ha de prendre una referència orientativa a partir de la qual poder treballar. Aquesta referència orientativa inicial fou la xifra de quaranta milions de mots del text, però, a l'hora de seleccionar les obres que havien de formar part del corpus aplicant els criteris de selecció establerts, ens adonàrem que aquesta previsió es feia curta. L'estimació actual, en un moment en què tenim la major part del text del corpus introduït a l'ordinador, i, per tant, comptabilitzat, és d'entre cinquanta-tres i cinquanta-quatre milions de mots.

A part d'aquestes característiques generals, podríem parlar d'altres aspectes més concrets o particulars, si disposéssim d'un temps més dilatat. De tota manera, potser és ineludible de fer referència al fet que les obres preses en consideració s'incorporen íntegrament al corpus; només en casos molt comptats, i per raons suficientment determinades, s'ha introduït una part d'una obra extensa. Per una altra banda, el text font,

que serveix de referència definitiva, ha estat sempre la primera edició de l'obra, i ha estat integrat al corpus sense fer-hi cap correcció ni modificació.

3. Descripció dels treballs

Des del punt de vista de la realització dels treballs, el *Corpus textual informatitzat de la llengua catalana* ha estat organitzat en tres fases, que corresponen a tres àrees de treball diferenciades:

a) Selecció de les obres que han de formar part del corpus.

b) Introducció dels textos a l'ordinador.

c) Lematització.

a) La selecció de les obres que han de constituir el corpus és una operació complexa; aquesta selecció és, d'altra banda, l'única fase de tot el procés que podria introduir un punt de subjectivitat en un projecte que té com a principi la màxima objectivitat en el tractament de les dades. Amb això no vull dir que, a fi d'evitar el possible caràcter subjectiu, la tria s'hagi fet d'una manera arbitrària, ni molt menys, sinó que s'ha seguit un mètode rigorós que no podré exposar aquí més que en els seus aspectes generals. L'objectiu ha estat, com ja he dit, que el corpus tingui el màxim grau de representativitat possible de totes les publicacions en llengua catalana dins la franja temporal establerta. Per aconseguir aquest objectiu s'ha dividit el total de l'extensió temporal del corpus (1833-1988) en vint-i-tres grups cronològics (vuit grups de deu anys cada un entre 1833 i 1913, i quinze grups de cinc anys cada un entre 1914 i 1988). A l'hora de fer la selecció s'ha procurat que, no solament d'una manera general, sinó també dins cada un d'aquests grups, hi hagi una representació adequada de cada tipus de text, entenent per tipus cada un dels gèneres (narrativa, teatre, poesia i assaig) pel que fa a la llengua literària, i cada una de les deu àrees temàtiques establertes, subdividides, la major part, en diverses subàrees, també fins a deu, pel que fa a la llengua no literària (aquestes àrees temàtiques, que, per qüestions pràctiques, segueixen de prop els criteris de classificació decimal de les biblioteques, són: 0. Correspondència, 1. Filosofia, 2. Religió i teologia, 3. Ciències socials, 4. Premsa, 5. Ciències pures i naturals, 6. Ciències aplicades, 7. Belles arts, divertiments, jocs i esports, 8. Llengua i literatura, 9. Història i geografia). Entenem que la representació és adequada des del punt de vista cronològic i temàtic quan podem

garantir fins a un cert punt una relació de proporcionalitat entre la quantitat de text seleccionat per a cada tipus i cada grup cronològic i el volum de text de cada tipus publicat dins cada grup cronològic.

A aquests criteris generals se n'hi afegeixen uns altres de complementaris, com són la consideració de les diferències dialectals, o bé la temàtica diferent dins els textos corresponents a un mateix gènere literari, per exemple. A fi de poder assegurar al màxim l'efectivitat d'aquests criteris, s'han hagut de desenvolupar activitats específiques de naturalesa diversa, que van des de l'organització d'uns repertoris informatitzats d'autors i obres, que ens han permès de creuar dates de publicació, tipus de text i autors, a fi d'objectivar al màxim les decisions, fins a les visites a les biblioteques de diversos indrets dels Països Catalans, tant a fi d'obtenir dades sobre publicacions, com amb l'objectiu de localitzar els originals dels llibres, diaris o fullets, quan aquests havien entrat ja en la fase de preselecció. No parlaré de més detalls sobre aquesta fase, però he volgut esmentar-ne alguns aspectes, perquè pugueu valorar-ne la complexitat. Només afegiré que el *Repertori d'autors i obres* (RAO) constituït amb aquesta finalitat té el volum següent: 6.273 autors (d'entre els quals n'han estat seleccionats 1.458 per a formar part del corpus) i 26.121 obres (de les quals n'han estat seleccionades 3.302, que es reparteixen en 2.296 corresponents a la llengua no literària i 1.006 a la llengua literària).

b) La incorporació de les dades textuals al suport informàtic és també una operació complexa; em referiré, per tant, només als seus aspectes més rellevants. Després de seleccionar les obres que han de formar part del corpus, cada una ha d'ésser objecte d'un estudi individualitzat a fi de posar de relleu les seves característiques més específiques, les quals poden tenir algun tipus de repercussió en aquesta fase d'introducció de la informació en l'ordinador. El text s'introdueix a través d'una operació de teclat, amb incorporació simultània d'una codificació que en permet el tractament informàtic adequat, segons les previsions del projecte. Dins aquesta fase es realitzen també diversos tipus de revisions del text introduït, a fi de garantir-ne, per una banda, la fidelitat a l'original, que, com he dit ja més amunt, és un objectiu fonamental, i, per una altra banda, la correcta incorporació dels codis que han de permetre una interpretació adequada de les dades per part dels programes informàtics que les han de tractar. En realitat, en aquesta fase es du a terme, doncs, no solament la incorporació a l'ordinador de les dades textuals, degudament codificades, sinó també la validació de la informació introduïda. Aquesta fase

culmina amb l'aplicació dels programes de segmentació del text en mots, els quals donen lloc a uns fitxers de mots, a partir dels quals hom pot obtenir diversos tipus de llistats (classificacions diverses, informació sobre freqüències, concordances). Notem, però, que fins aquí hem utilitzat el terme *mot* amb dos valors diferents: el mot del text, o ocurrència, i el mot considerat com a cadena de caràcters (mot gràfic o grafia): d'acord amb aquest segon criteri, tot allò que correntment consideraríem mots repetits queda reduït a una sola unitat en aquests fitxers de mots; això ens permet de dir que una obra, com, per exemple, *Nosaltres els valencians*, de Joan Fuster, té 69.914 mots (en el sentit de mots del text) i també que té 9.753 mots (en el sentit d'unitats diferents des del punt de vista gràfic).

Deixeu-me dir aquí que tots els programes informàtics, tant els relacionats amb les fases del treball que ja he comentat, com els referents a les que esmentaré a continuació, han estat elaborats específicament dins les tasques del mateix projecte per a aquesta finalitat.

c) Tots aquells que esteu més o menys familiaritzats amb la problemàtica que és objecte d'aquest col·loqui sabeu que molts dels corpus que existeixen presenten les dades tal com resulten al final de l'operació que acabo d'esmentar (segmentació del text): en canvi, en altres projectes d'aquesta naturalesa s'ha escomès una nova operació que representa un pas més en l'anàlisi de les dades: es tracta de la *lematització.*

No puc entrar ara a comentar tota una problemàtica específica que hi ha en relació amb la conveniència o no de lematitzar un corpus; només aquest aspecte podria ja ésser motiu de debat per a un col·loqui més ampli. Deixaré dit, però, que hi ha qui defensa la idea de no lematitzar els corpus, a fi de treballar amb una informació més verge, sense cap altre tipus de manipulació que les que he descrit fins ara. No és aquesta la posició adoptada en el *Corpus textual informatitzat de la llengua catalana*, de què us parlo ara, sinó que des del primer moment s'ha previst la lematització com una de les fases de constitució del corpus.

Per a aquells que no estiguin familiaritzats amb aquest concepte, només diré que la lematització representa una primera anàlisi de caràcter lingüístic de les dades; en la seva especificitat, però, pot ésser de naturalesa diversa, segons els criteris utilitzats; i, per una altra banda, el procés es pot dur a terme a través de més d'un procediment diferent. Faré només una referència breu als objectius del procés de lematització del nostre corpus i al tipus de procediment utilitzat, sense entrar a fer cap comentari sobre altres possibilitats, cosa que ens allunyaria de la finalitat específica d'aquesta dissertació.

A través de l'operació de lematització s'aconsegueixen dos objectius, que en realitat són un conseqüència de l'altre. Cada una de les ocurrències de cada mot gràfic és categoritzada gramaticalment i associada a una forma de referència anomenada *lema*: per exemple, davant cada una de les ocurrències concretes de la grafia *cap*, a la vista de l'entorn contextual, es determina a quina possibilitat de categorització correspon de les quatre que, en principi, té aquesta seqüència de caràcters: *cap* (substantiu masculí), *cabre* (verb), *cap* (preposició), i *cap* (adjectiu); en el primer cas és una forma de *singular*, en el segon de *tercera persona del present d'indicatiu*, i en el tercer i en el quart no té caracterització morfològica, perquè correspon al que correntment anomenem mots invariables. Així, per una banda, es desambigüen gramaticalment les formes homògrafes, i, per una altra banda, s'agrupen sota un mateix lema els diferents components d'una sèrie inflectiva. En un corpus sense lematitzar un usuari podria demanar informació només a través de la grafia *cap*, i la informació que obtindria sobre les diferents ocurrències d'aquesta grafia seria indiscriminada des del punt de vista lingüístic; és a dir, rebria, sense diferenciar, les formes corresponents al substantiu masculí, les corresponents al verb, a la preposició o a l'adjectiu. En un corpus lematitzat, en canvi, d'acord amb aquests principis, l'usuari pot consultar les dades agrupades adequadament partint del substantiu *cap*, del verb *cabre*, de la preposició *cap*, o de l'adjectiu *cap*, perquè la base de dades corresponent conté aquesta informació.

Cal advertir, però, que aquest procés no entra en l'anàlisi semàntica; és a dir, en el cas de l'exemple que acabo de posar a la vostra consideració no són destriats els diferents significats que poden correspondre a les diferents ocurrències del substantiu *cap*. No m'entretindré ara a justificar aquesta opció; però no deixaré de dir que és coherent amb els comentaris que he fet més amunt a propòsit del procediment per a establir els significats dels mots: no és sinó a través d'una anàlisi detinguda dels contextos, que podem determinar els diferents significats associats a un mot.

En aquest punt del procés arribem a tenir quatre unitats de treball diferenciades:

OCURRÈNCIA (unitat del text)

ex.: CAP, amb la indicació de la localització (obra, pàgina, línia, núm. d'ordre dins la línia)

FORMA GRÀFICA (unitat gràfica)

ex.: CAP

FORMA GRAMATICAL (unitat morfològica)
ex.: CAP singular
CAP 3a pers. present d'indicatiu
CAP (forma invariable)
LEMA (unitat del lèxic)
ex.: CAP substantiu masculí
CABRE verb
CAP preposició
CAP adjectiu invariable (gènere) i defectiu (nombre)

Després de realitzar aquesta operació, podem dir, per exemple, que una obra com la que he esmentat abans (*Nosaltres els valencians*, de Joan Fuster), amb 69.914 ocurrències i 9.753 formes gràfiques, conté 10.660 formes gramaticals i 6.476 lemes.

Cal afegir, encara que sigui tan breument, que el mateix procés de lematització implica l'agrupació sota un mateix lema de qualsevol variant gràfica, que podem trobar sobretot en textos no normalitzats (en la il·lustració de la figura 3 podem veure, per exemple, que, associades al lema CAP, apareixen grafies com *cab, cáp, câp, cabs, cáps);* a més, s'ha de tenir en compte que, a l'hora de definir el procediment, vam optar per classificar sota un mateix lema tots els derivats apreciatius i intensius, per als quals, doncs, no es crea un nou lema (vegeu també a la figura 3 les formes *cabet, caparró, caparronet, caparrot, caperrot, capet, cabets, caparrins, caparronets, caparrons, caparrots, caperrots, capets, cabot*).

Pel que fa al procediment utilitzat en la lematització del corpus, diré que és un procediment semiautomatitzat, en el curs del qual l'ordinador fa un tractament determinat de la informació, que després ha d'ésser completat amb la intervenció humana. L'element clau en aquesta operació és el que anomenem *Diccionari bàsic informatitzat* (DBI), que consta, en la seva estructura lògica, d'un inventari de lemes i un inventari de formes; l'inventari de lemes conté totes les entrades que figuren en el *Diccionari general de la llengua catalana* de Pompeu Fabra i les que figuren en el *Diccionari de la llengua catalana* d'Enciclopèdia Catalana, i l'inventari de formes conté totes les formes flexionades que corresponen als dits lemes; cada una d'aquestes formes flexionals està relacionada adequadament amb el lema, o, eventualment, els lemes a què correspon. El procés

automàtic és el següent: per a cada obra introduïda, l'ordinador llegeix una per una les formes gràfiques que troba en el fitxer de mots sense lematitzar; a continuació explora el fitxer de formes del DBI per veure si hi troba alguna forma gramatical que coincideixi en la seva grafia amb la forma gràfica que en aquell moment considera, i, a partir de la forma o les formes que ha trobat, accedeix al lema o als lemes del fitxer de lemes del DBI que hi estan relacionats. A partir d'aquesta informació, en dos casos el mateix programa atribueix directament el lema a totes les ocurrències de la forma que havia pres en consideració: *a)* si només ha trobat una solució de lematització en el DBI, *b)* si, havent trobat més d'una solució de lematització, n'hi ha una que ha estat prèviament prioritzada en el sistema a causa de la seva major freqüència (com, per exemple, *de* com a preposició, que s'ha prioritzat enfront del substantiu que correspon al nom de la lletra d); en els altres casos no atribueix encara cap lema. A continuació l'ordinador munta un llistat, anomenat *proposta de lematització*, en què, per a cada forma gràfica, ofereix tota la informació que ha obtingut en el curs del procés que acabo de descriure, seguida de la relació de contextos en què apareix la forma en qüestió en l'obra que es lematitza (vegeu-ne unes mostres en les figures 1 i 2); a partir d'aquí donem pas a una fase d'intervenció humana, en el curs de la qual, a la vista d'aquest llistat i examinant-ne cada un dels contextos, es determina:

1. Si l'atribució del lema i de la forma, en el cas d'assignació automàtica per part de l'ordinador, ha estat l'adequada.

2. La solució adient per a cada ocurrència en el cas que l'ordinador en proposi més d'una per a una forma gràfica determinada, o bé en el cas que no en proposi cap, perquè no ha trobat en el DBI cap forma que correspongui a aquella grafia.

La informació nova que resulta d'aquesta darrera fase és incorporada a l'ordinador a través del programa corresponent, previst per a aquesta funció, i amb això culmina el procés de lematització pròpiament dit.

No cal dir que en el curs d'aquest procés apareixen un cert nombre de formes gramaticals o gràfiques i també un cert nombre de lemes que no es troben en el DBI originari, perquè no figuren com a entrades en els diccionaris que han servit de punt de partida (si es tracta de lemes), o perquè no es corresponen amb cap de les formes que han estat preses en consideració a l'hora de constituir el DBI seguint les previsions de la gramàtica i de l'ortografia de la llengua (si es tracta de formes). Aquests nous lemes i

aquestes noves formes -degudament codificades per a mantenir la informació sobre el seu origen- passen a formar part del DBI a mesura que apareixen, i a partir d'aquest moment l'ordinador ja els reconeix com a lemes o com a formes a tenir en compte en la fase automàtica del procés de lematització. Així, en la il·lustració de la figura 1 podem observar que, a part dels lemes i de les formes del DBI originari (en aquest cas les que figuren codificades com DFA -*Diccionari Fabra*-), hi ha unes altres formes i uns altres lemes, codificats com a DCC (*Diccionari del Català Contemporani*), que han sorgit durant el procés de constitució del corpus; per exemple, *cap* com a adverbi, la sigla CAP, l'abreviatura de *capítol*, una forma femenina ocasional, que apareix a l'expressió *a la fi i a la cap* en una obra determinada. Després, encara, d'haver lematitzat aquesta obra, hem hagut de crear un nou lema CAP, substantiu femení, per donar raó de l'ús que trobem en la frase següent: "podeu dirigir-vos a *les antigues caps* o companyes" (*Avui*, núm. 1092, any 1979). Des de l'inici dels treballs de lematització del corpus fins avui s'han hagut de donar d'alta 72.885 lemes nous i 251.796 formes noves, que han anat apareixent en els textos tractats, en virtut d'aquest fet, i no figuren a les fonts lexicogràfiques utilitzades per a constituir el DBI. Aquestes xifres, sumades a l'extensió del DBI originari (88.067 lemes i 631.287 formes), fan un total de 160.952 lemes i 883.083 formes, que constitueixen el DBI en aquest moment.

4. La *Base de dades textual de la llengua catalana*

Una vegada acabat el procés de lematització de cada obra i després d'haver-ne validat els resultats, aquests passen a formar part d'una única base de dades constituïda per tots aquells elements que són fonamentals per a l'explotació del corpus: la *Base de dades textual de la llengua catalana* (BDTLC).

Com a resultat del procés de lematització, s'incorporen a la base de dades, per a cada una de les ocurrències, les informacions següents: la forma gràfica, la localització de l'ocurrència (codi de l'obra, pàgina, línia i núm. d'ordre del mot dins la línia), el codi morfològic i el nombre del lema que li ha estat atribuït. D'altra banda, a fi de poder reconstruir el context a partir de la base de dades, s'hi incorporen també una sèrie d'elements procedents de l'*Arxiu de textos per obra* (ATO) -arxiu que conté el text de cada obra emmagatzemat de manera seqüencial, amb tots els seus elements, els originaris

i els que procedeixen de la codificació-; aquestes dades són, bàsicament, per una banda, tots aquells codis que corresponen al que podríem considerar *grosso modo* signes de puntuació (punts, comes, punts i comes, cometes, signes d'interrogació i d'exclamació, etc.) i els codis lògics (com, per exemple, els que ens indiquen final de línia o de paràgraf i altres), i, per una altra banda, les parts del text originari que no han estat tingudes en compte durant el procés de lematització, ja que en la fase d'introducció havien estat codificades com a no analitzables (nombres expressats en xifres, citacions d'altres autors o en altres llengües, etc.), o bé com a noms propis, que no són tinguts en compte en el procés de lematització, però que cal recuperar a l'hora de reconstruir el text.

Per a l'explotació de la base de dades han estat elaborats una sèrie de programes que permeten, per una banda, l'elaboració de llistats diversos, i, per una altra banda, la consulta interactiva.

Els llistats que hom pot obtenir en aquest moment es refereixen a diversos tipus d'ordenacions (alfabètica directa, alfabètica inversa, per freqüències, amb distribució de les freqüències segons criteris cronològics o tipològics, segons el tipus de llengua, els codis gramaticals, etc.).

El sistema de consulta interactiva permet accedir a la base de dades a través del *lema*, a través de la *forma* o a través de la *localització*; aquesta darrera modalitat, però, i, en part la segona, tenen un interès exclusivament intern per a la validació i el manteniment de la base de dades. Per una altra banda, el sistema permet accedir o bé al conjunt de tot el corpus, o bé a una part de les dades, que poden ésser seleccionades per l'usuari, com un subcorpus; aquest subcorpus pot ésser definit a partir de criteris cronològics, a partir de criteris tipològics, o d'ambdós alhora; a més, si la consulta afecta un lema o una forma que té una freqüència molt elevada en el corpus, i, per tant, fóra molt feixuc de consultar-ne tots els contextos, el programa permet de fer-ne una selecció prèvia, de caràcter aleatori, a partir d'un tant per cent del total d'ocurrències, o d'un nombre concret, fixats per l'usuari.

A través del sistema de consulta, podem obtenir, per exemple, si entrem per un lema determinat, la freqüència total del lema, les formes que té associades i la freqüència de cada una (figura 3), i, si accedim a una d'aquestes formes, les diferents ocurrències concretes que presenta en el corpus (figura 4); podem veure a continuació el context que correspon a cada una d'aquestes ocurrències, amb una sèrie de dades relacionades, com

són, l'autor i el títol de l'obra a què correspon cada ocurrència, l'any de publicació, el tipus de text (literari/no literari i les seves subdivisions), la localització específica (pàgina, línia, número d'ordre dins la línia), i les dades morfosintàctiques de la forma i del lema (figura 5).

Pel que fa al context corresponent a cada ocurrència examinada, en primera instància apareix l'extensió corresponent a tres línies físiques de l'edició de referència, del text originari; però l'usuari pot accedir immediatament al paràgraf sencer, si el context inicial és insuficient (figura 6).

Per a poder proporcionar tota aquesta informació, el programa de consulta accedeix inicialment a la BDTLC, que conté només la informació més indispensable, sense repeticions innecessàries, a fi de fer les consultes el més àgils i ràpides possible, i, a través d'aquestes dades, accedeix a les del DBI i del RAO. El programa obté del DBI dades com la grafia i el codi gramatical del lema, i el codi de procedència; del RAO, obté el nom de l'autor, el títol de l'obra, l'any de publicació, i altres dades relacionades amb l'obra (tipus de llengua, etc.).

5. Execució i estat actual del projecte

L'execució del projecte que acabo de descriure sumàriament s'inicià al començ de 1985 amb recursos limitats. A part d'aspectes materials, com el condicionament de locals i l'establiment d'una infraestructura informàtica adequada per al tractament de les dades, i de qüestions com la formació de personal especialitzat, durant els quatre primers anys es treballà fonamentalment en la preparació del *Repertori d'autors i obres*, en la selecció de les obres que han de formar part del corpus, en el disseny i elaboració del sistema d'introducció i verificació de dades i en el disseny i elaboració del sistema de lematització semiautomatitzada.

En l'apartat que he dedicat a la descripció dels treballs del corpus, ja m'he referit al RAO com un instrument necessari per a poder portar a terme, amb les màximes garanties de representativitat i equilibri, la selecció de les obres per a integrar en el corpus; doncs bé, la realització d'aquest fitxer informatitzat s'emprengué des del primer moment i aviat estigué en condicions d'ésser utilitzat en la fase de selecció. La selecció pròpiament dita s'inicià també molt al començament, però es perllongà al llarg d'aquests anys, perquè és

una tasca lenta i feixuga, tal com es desprèn de la breu descripció que n'he fet més amunt; ara és totalment acabada, i, tal com he dit també, el nombre d'obres seleccionades és de 3.302 (2.296 corresponents a textos no literaris i 1.006 a textos literaris).

Durant aquests primers anys es realitzà i es perfeccionà el disseny del sistema d'introducció i verificació de dades i s'elaboraren els programes per a executar-lo: això implicà la presa d'una sèrie de decisions a propòsit del tipus de codificació a utilitzar, íntimament relacionades amb la naturalesa dels programes que haurien de tractar el text. El resultat d'aquests treballs foren, per una banda, els criteris per a la introducció del text a l'ordinador, que es materialitzen en un manual d'introducció de dades, i, per una altra banda, en els programes d'introducció pròpiament dita i d'esmena de dades, i els programes de verificació automàtica, que permeten detectar automàticament la presència de certs possibles errors, de llistat del text introduït, per a una verificació manual, i de separació de mots, que permet localitzar certs errors altrament difícils de trobar, i, a més, en el seu procés de realització dóna lloc als fitxers de mots sense lematitzar que serveixen d'*input* en la fase de lematització.

També durant aquest període inicial s'abordaren els diferents aspectes relatius al procediment de lematització. Aquesta operació, des del punt de vista lingüístic, és aparentment simple, però presenta una sèrie de problemes davant els quals s'han de prendre decisions, tant a l'hora d'establir els lemes de referència, com a l'hora de determinar l'atribució de determinades formes a un lema o a un altre; el conjunt d'aquests problemes arriba a ésser prou important; cada un d'ells ha estat estudiat detingudament i les decisions que s'han pogut prendre com a resultat d'aquest estudi han donat lloc a un manual de lematització, que conté explícitament les normes que s'han de posar en pràctica durant el procés, a fi d'assegurar al màxim la unitat de criteri. A part, però, dels criteris lingüístics, el procés de lematització requeria l'establiment d'uns instruments informàtics que permetessin la realització d'aquesta operació d'una manera semiautomatitzada; això implicava, per una banda, la constitució d'un diccionari de màquina, el DBI, a què ja he fet referència més amunt, i, per una altra banda, l'elaboració dels programes informàtics adequats per a aquesta finalitat; doncs bé, a finals de 1988 s'havien aconseguit tots aquests objectius i s'havia començat a treballar productivament en la fase d'introducció de dades, de tal manera que a final d'aquest any comptàvem amb 6.500.000 mots introduïts i verificats.

A principi de 1989, havent establert clarament aquestes fases prèvies (selecció d'obres, criteris d'introducció i verificació, criteris de lematització) i havent resolt els aspectes tècnics que planteja un projecte d'aquesta naturalesa, un increment important dels recursos destinats a aquesta finalitat (gràcies a un conveni signat entre l'Institut d'Estudis Catalans i la Secretaria d'Estat d'Universitats i Investigació) permeté d'accelerar la realització dels treballs d'introducció i de lematització fins a finals de 1992, en què els recursos dedicats al projecte disminuïren. En les taules que segueixen podem observar l'evolució del volum de text introduït i lematitzat al llarg d'aquests darrers anys:

TEXT INTRODUIT			
Data	*Llengua literària*	*Llengua no literària*	*Total*
31-XII-88	5.941.272	577.051	6.518.330
31-XII-89	7.316.331	5.862.195	13.178.526
31-XII-90	8.692.489	15.857.027	24.550.416
31-XII-91	9.258.576	27.054.538	36.313.114
31-XII-92	17.047.496	29.286.521	46.334.017
31-XII-93	18.110.247	29.234.544	47.344.791

TEXT LEMATITZAT			
Data	*Llengua literària*	*Llengua no literària*	*Total*
31-XII-89	433.036	1.376.952	1.809.988
31-XII-90	2.486.692	7.629.382	10.116.074
31-XII-91	2.635.242	19.980.080	22.615.322
31-XII-92	2.746.863	27.427.599	30.174.426
31-XII-93	4.245.262	29.234.544	33.479.262

Pel que fa a la situació actual, és la següent: la part del corpus corresponent a la llengua no literària està completament acabada; els 29.234.544 mots introduïts i lematitzats constitueixen els seus efectius; queda, però, en canvi, una part important de text literari

per introduir (uns 7.000.000 de mots), i una part encara més important per lematitzar (uns 21.000.000 de mots).

Cal afegir, per a completar aquesta visió del procés d'execució del projecte al llarg del temps, que durant els anys 1990 i 1991 es dugueren a terme els treballs encaminats a la creació de la base de dades que he descrit en l'apartat anterior, que és el que permet pròpiament l'explotació del corpus; a final de 1992 quedà la base de dades completament constituïda, i a partir d'aquest moment s'hi incorporen d'una manera automàtica els resultats del procés de lematització a mesura que es donen per bons.

Figura 1. Fragment de la proposta de lematització de la forma *cap*, corresponent a l'obra *Laia* de Salvador Espriu. Per una banda hi veiem tots els lemes a què pot correspondre aquesta forma, i, per una altra banda, podem observar que l'atribució de lema i forma s'ha fet, en aquest cas, manualment (columna de la dreta).

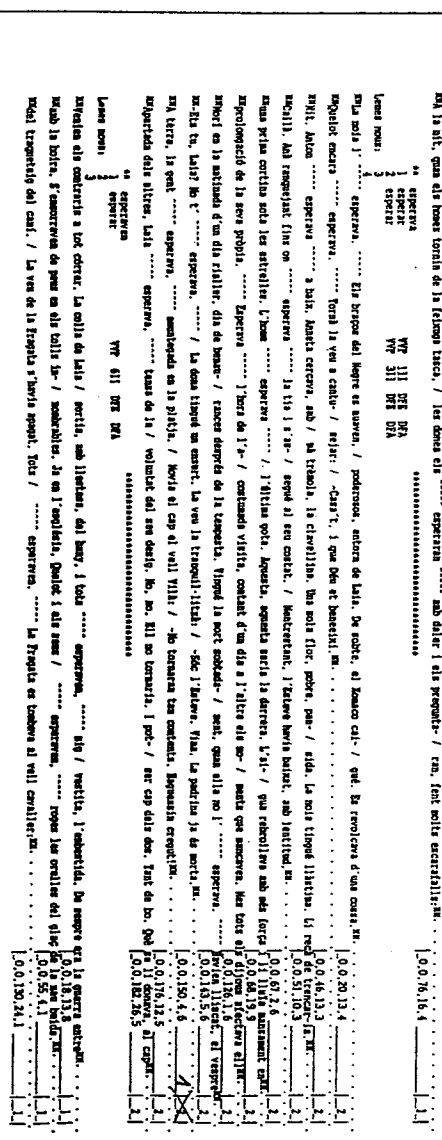


Figura 2. Fragment de la proposta de lematització (diverses formes) de l'obra *Laia* de Salvador Espriu. En la forma *esperava* podem observar l'atribució automàtica a l'opció prioritzada (tercera persona en comptes de la primera), que en un sol cas ha hagut d'ésser esmenada manualment.

BDTLC RECONSTRUCCIÓ SEGONS L'ORIGINAL A PARTIR D'UN LEMA

Lema: cap Cat. gram.: M
Codi: 37.921 Freq. lema: 18.016

Selecció de la forma

Forma	C.M.	Freq. abs.
cap	FS	1
cab	S	5
cap	S	15.842
cáp	S	6
câp	S	2
cabs	P	1
caps	P	2.004
cáps	P	4
cabet	DS	3
caparró	DS	40
caparronet	DS	3
caparrot	DS	12
caperrot	DS	1
capet	DS	17
cabets	DP	1
caparrins	DP	1
caparronets	DP	1
caparrons	DP	7
caparrots	DP	2
caperrots	DP	1
capets	DP	8
cab	MET	1
cabet	MET	1
cabot	MET	1
cap	MET	45
caparró	MET	2
caps	MET	4

Figura 3. Repertori de formes corresponents al lema CAP, tal com ens les ofereix el programa de consulta de la base de dades. Hi podem observar, diferenciades pels codis corresponents, les formes de singular i de plural, els derivats apreciatius (amb el codi D) i les formes que han aparegut en contextos metalingüístics; també hi podem apreciar a simple vista les variants gràfiques. Per a cada una de les formes diferenciades se'ns ofereix la freqüència que té en el corpus, és a dir, el nombre de vegades que hi apareix.

```
BDTLC   RECONSTRUCCIÓ SEGONS L'ORIGINAL A PARTIR D'UN LEMA

                Selecció de les localitzacions de la forma
Forma: cap                      Codi morf.: S          Freq. abs.: 15.842
Lema : cap                      Cat. gram.: M          Codi lema : 37.921

 Obra    Nivl Niv2 Localitzacions          Obra    Nivl Niv2 Localitzacions
    1           18,7,9                        1           59,26,7
                24,19,3                                   63,25,4
                24,27,13                                  73,21,3
                26,26,6                                   81,10,3
                30,25,4                                   86,6,1
                39,27,1                                   92,4,6
                43,25,7                                   92,13,2
                46,5,5                                    92,22,7
                50,2,5                                    138,6,4
                52,3,5                                    146,23,11
                58,5,8                                    150,11,7      +

F13 Finestra d'opcions
F12 Selec. totes les localitzacions              SELECCIÓ PER OBRA :
F1 Menú inicial   F2 P. anterior   F5 Següent forma
```

Figura 4. El programa de consulta ens pot donar la relació completa de totes les localitzacions de qualsevol forma. Aquesta figura ens mostra les 22 primeres localitzacions de les 15.842 que corresponen a la forma *cap* (singular del substantiu masculí). La localització se'ns ofereix amb el codi de l'obra, el nombre de la pàgina i de la línia i el nombre d'ordre del mot dins la línia.

```
  BDTLC   RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor : Joan Fuster
Títol : Nosaltres, els valencians
Obra  : 79      N.1:      N.2:      Localització : 9,6,3

Any d'edició : 1962                 Forma . .: cap
Tipus public.: Assaig               C. morf. : S   Singular
Total d'ocurrències   : 64.914      Lema  . .: cap
Total de lemes usats  : 6.476       C. Gram. : M   Nom masculí
Total de formes usades: 10.660      Núm. lema: 37.921

  C O N T E X T

   ¤¤ públic del País Valencià, hi ha severs retrats a l'oli de Felip V
 penjats cap per avall: la ira vernacla es projecta romànticament sobre el
 primer Borbó espanyol, i la ¤¤


 F22 Per paràgrafs   F24 Consulta segons l'entrada
 F1 Menú inicial   F2 P. anterior   F6 Context de nova localització
```

Figura 5. Ací veiem el context en què es troba una de les ocurrències de *cap* (singular del substantiu masculí), triada per l'usuari del sistema de consulta (concretament la que apareix a l'obra 79 -*Nosaltres els valencians*, de Joan Fuster-, (pàgina 9, línia 6, mot número 3); el programa ens ofereix el text corresponent a tres línies de l'original.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor : Joan Fuster
Títol : Nosaltres, els valencians
Obra : 79 N.1: N.2: Localització : 9,6,3

C O N T E X T

» L'anàlisi d'aquest fenomen ens portaria lluny, massa lluny. La recensió i l'examen de més detalls i de més condicions de la mentalitat del valencià actual podrien ocupar-nos encara pàgines i pàgines. Ho deixarem córrer, però. El desconcert "nacional" dels valencians no és cap secret per a ningú. I seria pueril de voler-lo explicar amb la mitja dotzena de tòpics que són de rigor en tals casos. Sobretot, seria ridícul de transferir totes les responsabilitats -posat que "responsabilitats" sigui la fórmula justa- a l'"altre". Res més senzill ni més confortable, per a un poble com per a un individu, que considerar-se "víctima" i atribuir l'origen de les seves desgràcies o dels seus errors a una dolorosa interferència aliena. A casa d'algun il·lustre patriota enragé, i fins i tot en algun museu públic del País Valencià, hi ha severs retrats a l'oli de Felip V penjats cap per avall: la ira vernacla es projecta romànticament sobre el primer Borbó espanyol, i la inversió dels quadres és una venjança simbòlica ben significativa. La innocència i la falta de sentit històric que aquestes actituds suposen, resulten més còmiques que simpàtiques -i com a simpàtiques, ja ho són! Però la veritable qüestió és tota una altra. «

Figura 6. Context corresponent a un paràgraf sencer de l'original, al qual l'usuari pot accedir si el context de tres línies que ofereix inicialment el programa (vegeu fig. 5) no és suficient per a la interpretació adequada.

Els materials de llengua oral del corpus de català contemporani de la UB (CUB)*

Emili Boix Fuster
(Universitat de Barcelona)

0. Introducció

En primer lloc presentaré les motivacions i la història del projecte global del CUB. En segon lloc descriuré els objectius i l'estat de l'elaboració de les diferents parts orals d'aquest corpus, especialment el corpus oral de registres, conversa i varietats socioculturals. Maria Teresa Cabré i Lluís de Yzaguirre en descriuran els materials escrits en un altre capítol d'aquest volum. Com que el primer Club-1 tingué lloc el desembre del 1993 l'exposició era sobretot programàtica. El setembre de 1995, en cloure la redacció d'aquest text, podem oferir més dades sobre la feina feta.

1. La història i les motivacions del CUB

1.1. *Les motivacions d'un corpus lingüístic als països de llengua catalana*

Entre els qui treballen sobre la llengua un nombre creixent de persones és conscient que la recerca lingüística no pot basar-se exclusivament ni en la intuïció (masculina o femenina) dels investigadors ni en la dels companys de departament universitari. A vegades ens oblidem que la gent, simplement, *parla*.

* El projecte *Variació en el llenguatge: corpus oral i escrit de català contemporani* rep finançament de la CIRIT (CS93-1017) i de la DGICYT (PB 90-0505).

Les teories lingüístiques, els models de descripció i d'explicació lingüística, s'han de contrastar i confrontar constantment amb les dades, unes dades que no siguin amputades o maquillades sinó fidels al màxim als usos lingüístics reals. Uns usos reals en què l'enunciat i l'enunciació constitueixen un tot. Hi hauria d'haver doncs un vaivé entre els models, d'una banda i les bases de dades, de l'altra. Hi hauria d'haver uns moviments d'anada i tornada entre el treball inductiu i el treball hipoteticodeductiu.

Entre els enfeinats en la llengua també molts són conscients que les intervencions prescriptives, freqüents i alhora necessàries a l'hora d'establir criteris normatius, han de tenir en compte unes recerques descriptives de base amb dades prou representatives. Em temo, però, que aquest interès pels usos lingüístics continua despertant suspicàcies o, si voleu, veritables pors, en molts parlants de comunitats amb contactes i conflictes de llengües tan intensos i extensos com els que tenim en la nostra. Tants anys de subordinació lingüística afavoreixen el manteniment d'ideologies prescriptivistes a ultrança. L'heterogeneïtat interna s'evita en mor d'una pretesa unitat simbòlica davant de l'exterior, un exterior mai no especificat del tot.

Els usos orals reals són encara una caixa negra massa ignota, o massa temuda perquè la seva heterogeneïtat se'ns escapa de les mans; una caixa negra que, tanmateix, cal obrir especialment en el cas de la llengua catalana després de les dimensions espectaculars dels canvis demogràfics i socials i, per tant, també lingüístics, que ha experimentat i continua experimentant la nostra societat durant aquest segle.[1]

Finalment em sembla que, entre els qui treballem amb i per la llengua, força gent és conscient també que la tradició de recerca en dialectes geogràfics -tan valuosa com ha

[1] Thomas (1991) mostra les arrels socials del purisme, sobretot en societats europees. En societats en situacions de contacte de llengües el tema de la norma és subjecte a passions. Els diccionaris no són percebuts doncs com a objectes de consum sinó com a elements simbòlics, en la mesura que indiquen quins són els usos considerats més legítims, més autentificadors. Les elits socials quebequeses, per exemple, mostraven actituds negatives envers un diccionari recent (*Dictionnaire Québecquois d'Aujourd'hui*), perquè es limitava a descriure l'ús i, en canvi, no oferia una norma, una proposta de llengua estàndard (Martel 1992). De fet les elits interessades per la llengua, aquelles que exerceixen de "policia lingüística" es mostren més interessades doncs en la subjectivitat de les opinions que no pas en l'objectivitat - si aquesta mai existeix- dels fets. La manca d'uns usos legítims establerts, d'un 'bon usage' socialment hegemònic, expliquen aquesta inseguretat lingüística de base i, com a conseqüència, aquesta demanda social de normes lingüístiques inequívoques i decidides de manera centralitzada. Processos de normativització menys centralistes i menys dirigistes, com els de l'anglès, no deixen tanmateix d'ésser efectius per controlar i fixar amb flexibilitat certs usos lingüístics.

estat-, en centrar-se en parlars rurals, s'ha de modificar i coordinar amb d'altres per descriure com parla la majoria dels catalanoparlants a les acaballes del mil·leni.

No cal veure la terra des de l'espai a la pel·lícula de la NASA *Blue Planet* -la franja de la Mediterrània catalana hi apareix com un passadís ribetejat de llums- per reconèixer que la llengua catalana és la pròpia d'uns països de població sobretot urbana. Amb prou feines una mica més d'un sis per cent de la població viu de l'agricultura i de la ramaderia (Cardús 1991). Tenim encara, però, poques dades de la variació urbana de la llengua catalana. Començant pel mateix Milà i Fontanals (1890) fins arribar a les descripcions i prescripcions fabrianes, és cert que s'ha partit del català central i barceloní. Es ben cert també, alhora, que les dades dialectals (socials, generacionals, geogràfiques) d'aquesta zona són molt i molt insuficients en comparació a la diversitat i quantitat de la població que hi viu. Per exemple, fa pocs anys en una bibliografia exhaustiva de la tradició dialectològica catalana hi trobàvem trenta-quatre referències a l'Alguer però només catorze de la ciutat de Barcelona (Colomina 1990).

D'una banda en aquesta zona pràcticament tots els catalanoparlants són castellanoparlants i, de l'altra, el català esdevé llengua coneguda per bona part de la població que no el té com a llengua apresa a casa. El contacte de llengües és doncs intensíssim i els seus efectes s'entrebarregen amb els del canvi lingüístic no induït. A hores d'ara fer un diccionari d'ús del català, amb fonaments prou seriosos, esdevé quasi impossible perquè no disposem de dades empíriques suficients. Sorprèn doncs la rapidesa amb què especialistes del ram avaluen l'ús de la llengua, com succeeix en l'exemple següent. A l'*Avui* del diumenge 5 de desembre de 1993 entrevistaven Arnau Puig, un dels responsables de la primera Bíblia Catalana interconfessional. El periodista li plantejava la pregunta següent: ¿"Com han resolt la pluralitat dialectal?" La resposta fou la següent: "[...] hem volgut fer una Bíblia que no donés preferència a cap àrea dialectal. Ho puc il·lustrar amb un exemple: "A Barcelona «*el llibre és sobre la taula*», però a moltes àrees catalanoparlants «*el llibre és damunt la taula*»". El responsable de la nova versió bíblica està descrivint el model de llengua a què aspira, però no descriu què es el català a la regió de Barcelona. Per exemple, ¿Qui gosa dir que l'única variabilitat en aquests enunciats es troba en l'ús de la preposició *damunt* o *sobre* ? ¿Quin és l'ús real dels verbs *ser* i *estar*? De la mateixa manera que no sabem quin és l'ús dels verbs *ser* i *estar* a la major part del domini, no coneixem de manera contrastable l'ús de moltes d'altres formes

variables en el català contemporani.

La major part del CUB descriu el català de Barcelona i la seva regió. Aquesta prioritat en l'estudi de la metròpoli barcelonina -raonada i raonable des de molts punts de vista, que no pretenem explicar ara amb detall- no exclou sinó que vol estimular en d'altres zones dels països de llengua catalana treballs similars o comparables al que estem començant a fer aquí.

1.2. *L'especificitat d'un corpus als països de llengua catalana*

Les recerques encara pendents de la variació lingüística del món urbà de llengua catalana han de tenir en compte els dos aspectes següents:

(1) Les experiències de la majoria de recerques internacionals sobre variació urbana no es poden adaptar directament a les necessitats dels nostres països perquè s'han realitzat en comunitats monolingües o amb una llengua establerta força indiscutida, que molt sovint, com és d'esperar, és l'anglès. A l'àmbit de llengua catalana ens plantegem les mateixes preguntes amb matisos específics. En els països de llengua catalana, i sobretot a Catalunya, el procés de bilingüització avança i des de fa pocs anys no sols en un sentit castellanitzador. Ens hem de plantejar qüestions sobre el contacte de llengües que han estat secundàries o absents en les recerques sobre varietats urbanes de la tradició anglosaxona dominant: quin és el lligam entre el canvi induït pel contacte de llengües i el canvi autònom?, quins factors socials expliquen la distribució i valoració de les diferents varietats del repertori?, quina evolució imbricada tenen?

Per aquesta raó ens interessen les recerques sobre el repertori lingüístic urbà de Brussel·les (Witte i Baetens Beardsmore 1988) de Montréal (Thibault 1990) i de ciutats bilingües de Suïssa (Kolde 1981).

(2) L'estudi del repertori lingüístic urbà, com ha assenyalat ja A. Tuson (1988) demana un treball coordinat amb especialistes en les llengües en contacte i amb especialistes en ciències socials. D'una banda cal augmentar el treball conjunt amb els els hispanistes i estudiosos de les varietats del castellà als països de llengua catalana. D'altra banda cal investigar en equip amb els sociòlegs i antropòlegs de les grans zones urbanes dels nostres països. No debades, i no pas per atzar sinó per necessitat, els grans *urban language surveys*, començant pel clàssic de Labov a NovaYork (1966) s'han basat en grans recerques sociològiques anteriors.

1.3. *La proposta del CUB*

Precisament l'aparició d'un estudi sociològic innovador i ambiciós féu el 1990 plantejar que era possible investigar les varietats de l'àrea urbana més important de llengua catalana. Les dues *Enquestes de la Regió Metropolitana de Barcelona* dels anys 1986 i 1990, permeten per primera vegada situar de manera congruent i sistemàtica la variació lingüística en un marc social de base. Aquesta enquesta proporciona dades sobre la distribució del català i castellà com a primera llengua a la regió de Barcelona, (vegeu especialment Subirats et al. 1992). Aquesta oportunitat fou el detonant de la primera proposta d'un corpus del català a la zona barcelonina. Així, més endavant els onze professors de la Secció de Llengua del Departament de Filologia Catalana es decidiren a engegar un projecte de corpus (des d'ara CUB), de més abast geogràfic, subvencionat fins a l'actualitat, per la CICYT del Ministerio de Educación y Ciencia, i per la Direcció General d'Universitats de la Generalitat de Catalunya, amb l'objectiu de recollir materials representatius del repertori de varietats geogràfiques, històriques, socials i funcionals del català contemporani.

El projecte és *complex*, tècnicament i humanament perquè exigeix una inversió pressupostària (sempre difícil en períodes de crisi com l'actual) i un esforç constant, perseverant i regular de coordinació entre especialistes en diferents ciències de llenguatge. Un esforç contracorrent perquè encara estem massa avesats a treballar com a franctiradors, dins la vella tradició humanista de 'tants caps, tants barrets' dominant a les 'repúbliques de coronels' universitàries.

El projecte, però, és *possible*, tècnicament, informàticament, com hem comprovat en les visites *in situ* a grups de treball en corpus lingüístics a Gran Bretanya (sobretot el projecte modèlic del *British National Corpus* de la Universitat de Lancaster, Oxford University Press i Longman) i al Canadà, i perquè la nostra tradició de recerca en la llengua oral ens proporciona amb escreix un pòsit metodològic (Montoya 1992).

Per acabar, el projecte és *necessari* o fins i tot *prioritari* si és que ens hem de prendre seriosament les línies de recerca proclamades pels nostres mateixos poders públics, tant catalans com europeus.

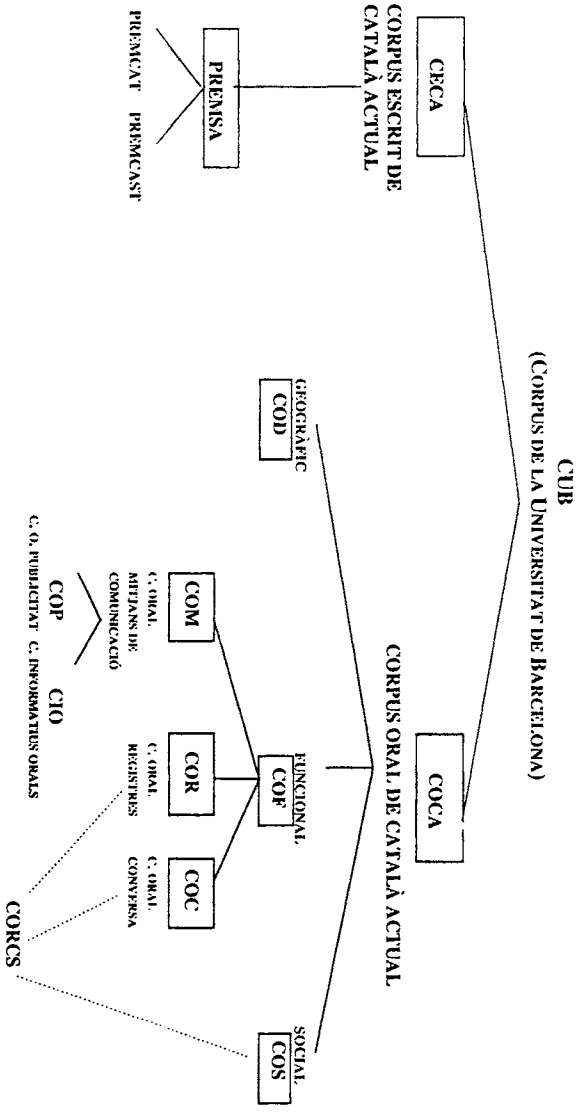
A continuació presentem les característiques principals del CUB. Aquesta presentació és fruit de les aportacions de tot l'equip de professors i becaris de la secció de Lingüística

Catalana que col·laboren en el projecte. Es de justícia que agraïm aquí la col·laboració de molts estudiants de segon i tercer cicle en la recollida de les dades. Estem convençuts que la seva feina serveix per a servir-los també: les dades del corpus són a la disposició per a recerques universitàries, tesis de llicenciatura i de doctorat.

2. Presentació general de les parts del corpus

D'una banda presentaré què és aquest arxiu de corpus, què pretèn ésser i amb qui es pot entrar en contacte per conèixer-ne més detalls. D'altra banda pretenc convidar a la crítica d'aquest projecte, un projecte dinàmic, de llarga durada, si els pressupostos no ens fallen, i un projecte força heterogeni.

FIG 1


CUB
(CORPUS DE LA UNIVERSITAT DE BARCELONA)
CECA
CORPUS ESCRIT DE CATALÀ ACTUAL
PRENSA
PREMCAT PREMCAST
COCA
CORPUS ORAL DE CATALÀ ACTUAL
GEOGRÀFIC
COD
FUNCIONAL
COF
SOCIAL
COS
COM
C. ORAL MITJANS DE COMUNICACIÓ
COR
C. ORAL REGISTRES
COC
C. ORAL CONVERSA
COP
C. O. PUBLICITAT
CIO
C. INFORMATIUS ORALS
CORCS

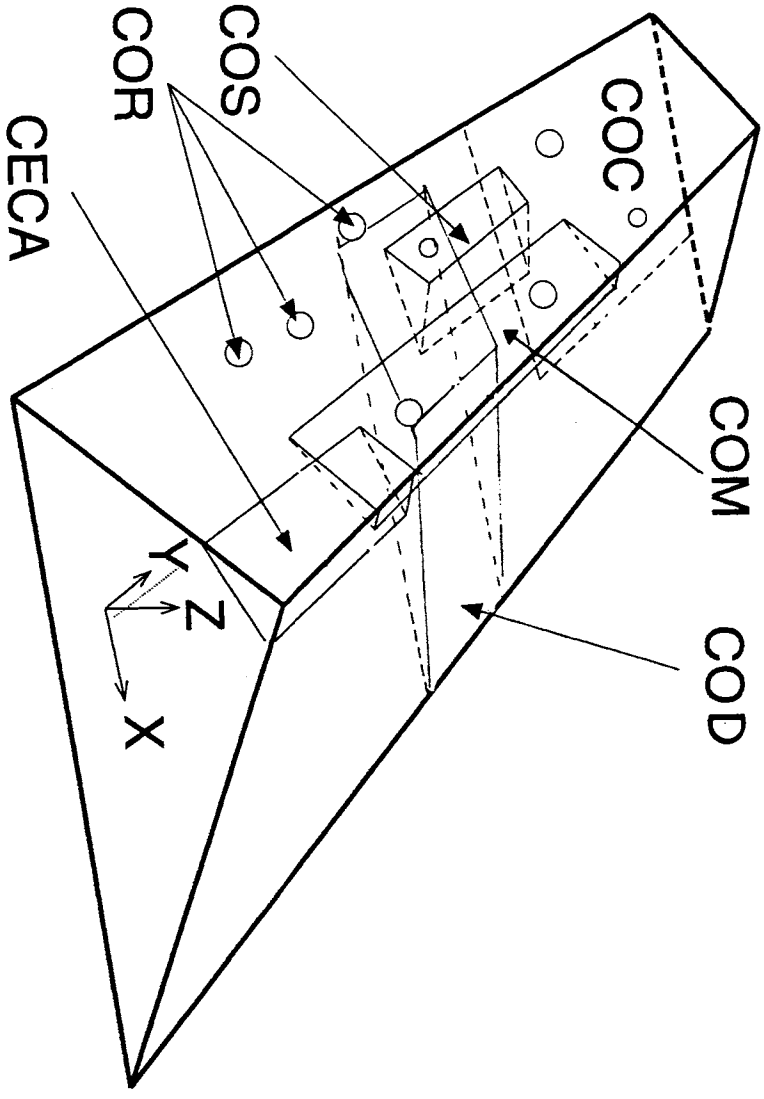
El CUB és un projecte de treball departamental coordinat, a partir de fonts de finançament unificats i de directrius d'informatització comunes. Lluís de Yzaguirre és el responsable màxim dels programes informàtics del projecte.

El CUB és un projecte col·lectiu força obert, integrador, poc dogmàtic i, potser per això mateix, a vegades fins i tot un xic magmàtic. Els diferents mòduls que l'integren i que apareixen en la figura 1 no obeeixen a un disseny travat en tots els seus aspectes, que s'hagi dibuixat des d'un inici. *Ad hoc* s'estableixen mecanismes d'homogeneïtzació, que poden permetre recerques de caire transversal en els diferents subcorpus.

Com us mostra la figura 1, aquest CUB, corpus de català, es divideix temàticament en dues grans branques: d'una banda, un *corpus escrit de català actual* (CECA), i, de l'altra, un *corpus oral de català actual* (COCA). Aquest subcorpus oral o COCA té tres grans components. En primer lloc el CUB consta d'un un corpus oral geodialectal (COD), és a dir un corpus de parla *segons les característiques d'origen geogràfic de l'usuari*. A l'interior d'aquest COT diferenciem el COD, que recull varietats segons el territori i el COS, que en recull segons els trets socioculturals dels parlants. En segon lloc, el CUB comprèn un corpus oral de registres o funcional (COF), establert *segons les funcions de la parla en diferents contextos socials*, és a dir *segons les característiques de l'ús*. Cadascun d'aquests subcorpus, com mostra la figura 1, es subdivideix en d'altres parts més petites. Dins les varietats funcionals, d'esquerra a dreta trobem un corpus oral de mitjans de comunicació (COM), el qual inclou dues parts, un corpus de publicitat (COP) i un Corpus d'Informatius Orals (CIO), un corpus oral de registres (COR) i un corpus de conversa (COC). Finalment, el CUB consta d'un altre corpus segons les característiques de l'usuari, *segons els trets socioculturals dels parlants* (COS). El COR, COC i COS constitueixen un grup de treball: el CORCS.

La figura 2 vol situar d'una manera coherent aquests diferents subcorpus, veure quines relacions poden tenir. Amb aquesta figura irregular, volem representar un espai ideal que ocupen les diferents varietats del repertori d'una llengua. Situem aquestes varietats mitjançant tres eixos, cadascun dels quals en representa una magnitud segons criteris de classificació diferents: criteri d'origen geogràfic (eix X), de formalitat (eix Y) i de control en la producció (eix Z).

FIG 2



La consideració de cadascuna d'aquestes magnituds (certament n'hi ha d'altres que no prenem en consideració) permet establir la situació relativa d'una varietat en el conjunt de la llengua. Es ben conegut que aquesta variació és multidimensional, no pas dicotòmica.[2]

Criteri geogràfic o territorial. A la base de la figura 2, en aquest eix horitzontal *X*, se situa la gradació de varietats geogràfiques a què pertany la mostra de parla recollida. El CUB, en el seu conjunt, està esbiaixat cap al català central entorn de la ciutat de Barcelona, és a dir cap a al costat esquerre de la figura. Els materials escrits, amb més control de la producció - i en menor grau els materials de mitjans de comunicació oral- estan més ancorats cap a models normatius més supradialectals. Només el COD té un abast geogràfic representatiu de tot el domini.

Criteri de grau de formalitat. A l'eix vertical *Z*, classifiquem les varietats segons el grau de formalitat de la situació de parla.

Un fragment de parla segons Irvine (1986, or. 1979) és més formal (1) si hi ha una gran estructuració del codi emprat, (2) si hi ha una consistència elevada en les regles que s'hi apliquen (regles de co-presència), (3) si hi ha un èmfasi central en la distància social entre els participants -més en els rols públics i definits dels participants que no pas en els seus rols privats; i (4) si tendeix a haver-hi un focus central d'atenció.

La formalitat és precisament la dimensió amb més poder classificatori de textos de les sis que proposà Biber per a l'anglès a *Variation across speech and writing* (1988). Aquest autor etiqueta aquesta dimensió "Informational versus Involved Production", que podríem traduir com "Producció informativa en contrast a producció involucrada o participativa". Els pols d'aquesta dimensió són els següents: D'una banda, en un extrem de la gradació hi ha un tipus de discurs amb objectius interactius, afectius (amb a) i participatius, que normalment té lloc amb restricció de temps real i de comprensió; d'altra banda hi ha un tipus de discurs amb un fort component referencial o informatiu, que ha estat elaborat acuradament i que normalment, tendeix a aproximar-se a models normatius orals o

[2] Biber (1988) identifica sis dimensions que són subjacents als paràmetres de variació. Aquestes dimensions són necessàriament abstraccions, generalitzacions a partir de la consideració global de seixanta-set formes lingüístiques aïllades.

escrits, la varietat estàndard. Les varietats estàndard, tant les orals com, encara més, les escrites, se situen en els punts més elevats d'aquesta gradació. Les varietats més informals es troben a la base, en canvi, i són més heterogènies. Evidentment en català hi ha buits ben coneguts en el coneixement i reconeixement d'aquestes normes de referència (Lamuela 1993), que faciliten que manquin models de variació funcional. Els enregistraments del COR se situen en una llarga gradació de possibilitats pel que fa a la formalitat. El COC proporciona materials al màxim informals, mentre que el COM i tot el CECA ofereixen materials més formals -situats doncs en la part superior de la figura fins arribar al vèrtex d'aquesta.

Criteri de grau de control de la producció. En el tercer eix Y situem les varietats segons el grau d'atenció, de control o monitorització de la producció de la parla que hi presten els parlants. No postulem pas que hi ha hagi un pol de llengua més natural, a l'estil de la *varietat vernacular* laboviana, perquè qualsevol fragment de parla varia considerablement, segons el context de situació, i no sols en resposta al control per part del parlant. El que sí postulem és que aquestes varietats amb poc control -com ocorre en la conversa cara a cara entre amics- són les que proporcionen dades més vàlides per a l'estudi del canvi lingüístic en la mesura que és en aquestes dades on trobem de manera més sistemàtica l'heterogeneïtat estructurada. De fet, doncs, aquest eix representa una gradació metodològica. Com més cap al pol 0 ens trobem, més recollim dades naturals, que ens permeten intentar solucionar la paradoxa de l'observador.

Presentem cadascun dels mòduls o parts del CUB a partir d'aquests diferents criteris de classificació. En cada cas n'assenyalarem els objectius, les dades que es volen recollir, i les dades de què es disposa ara mateix. En presentar-los triem com a fil conductor l'eix metodològic, en la mesura que ens permet identificar els diferents grups de treball operatius. Per aquesta raó presentem conjuntament el COC, el COR i el COS tot i que temàticament els haguéssim encasellat separadament. Tots tres formen el CORCS (Corpus Oral de Registres, Conversa i Varietats Socioculturals).

2.1. *El Corpus Oral Dialectal*

El Corpus Oral Dialectal té com a objectiu confegir un corpus de realitzacions orals de tot el domini lingüístic, adreçat a sectors de classe mitjana de totes les capitals de comarca

i d'algunes poblacions de menys de 600 habitants, pertanyents a tres franges d'edat: 15-19, 30-45 i 55-70. Hi ha un mínim de dos informants per localitat. Les dades lingüístiques es recullen mitjançant qüestionaris lingüístics que eliciten informació fonètica i morfològica (pronominal, nominal i verbal) i mitjançant textos lliures de cinc minuts. Es volen estudiar realitzacions fonètiques, sobretot aspectes morfològics (flexió nominal, flexió pronominal i flexió verbal regular) i alguns aspectes sintàctics com les combinacions pronominals. Amb aquests materials s'està constituint una base de dades digitalitzada amb transcripció fonètica i ortogràfica. Aquestes transcripcions seran presentades sincronitzadament.

En aquests moments el material acumulat existent és el següent: 110 cintes audio (qüestionari de pronoms febles i textos lliures) de localitats diverses del domini lingüístic, 35 cintes audio (qüestionari global i textos lliures) de caps de comarca de l'àrea nord-occidental, 5 cintes audio (formes adjectivals) de l'àrea xipella, 86 cintes DAT (qüestionari global, textos lliures) d'informants de 30 a 45 anys de 43 caps de comarca d'una part de tot el domini lingüístic; i 70 cintes DAT (qüestionari global i textos lliures) de 12 caps de comarca i 12 localitats de menys de 600 habitants d'una part de l'àrea nord-occidental (3 franges d'edat).

El material transcrit existent és el següent. S'han introduït en una base de dades les respostes (2.924 entrades) a un qüestionari de pronoms corresponents a 18 localitats. Es disposa de la transcripció ortogràfica i fonètica de textos lliures corresponents als dos informants de set localitats (123 minuts). S'han transcrit els qüestionaris globals corresponents als informants de l'àrea nord-occidental (franja d'edat de 30 a 45 anys). S'han transcrit parcialment els qüestionaris globals corresponents a dotze informants de les tres franges d'edat de l'Alt Urgell.

2.2. *El Corpus Oral de Publicitat*

El Corpus Oral de Publicitat recull i analitza dades d'anuncis publicitaris de ràdio i televisió dels anys 1992 i 1993. El desembre de 1993 s'havien recollit vint minuts de publicitat radiofònica de Catalunya Ràdio, Ràdio 4 i Ràdio Barcelona i més d'una hora i mitja de publicitat de TV3 i TVE a Catalunya. Tot el material s'està transcrivint a tres nivells: (1) una transcripció ortogràfica convencional, (2) una transcripció fonètica i (3)

una transcripció de la corba entonativa en què consten els grups fònics dins de cada enunciat i l'estructura de l'enunciat.

2.3. *El Corpus d'Informatius Orals*

El Corpus d'Informatius Orals té com a objectiu marcar el seguiment que els informatius de la ràdio i televisió públiques catalanes fan de la Proposta d'Estàndard Oral de la Secció Filològica de l'Institut d'Estudis Catalans. El desembre de 1993 aquest corpus estava format per setze hores de ràdio i televisió gravades entre desembre de 1992 i febrer de 1993, amb programes que cobreixen diversos graus d'elaboració, majoritàriament amb locutors metropolitans. Un 25% del material estava digitalitzat i un 5% transcrit. Aproximadament hi havia 4000 ocurrències d'uns 1400 mots, de manera que es preveia que hi havia un total de 50.000 ocurrències. Els fitxers de veu estan digitalitzats a 22 kilocicles delimitats paraula per paraula. Pel que fa al format escrit els fitxers han estat transcrits ortogràficament i es troben preparats per al tractament lexicomètric i per al marcatge gramatical, lematització, transcripció fonètica, etc.

2.4. *El corpus oral de registres, conversa i varietats socioculturals*

Passo ara a descriure els subcorpus amb dades relativament menys controlades (COS,COR i COC), és a dir, els tres subcorpus que s'agrupen sota l'acrònim de CORCS, és a dir el corpus oral de registres, conversa i varietats socioculturals. Exposaré quins són els objectius i dissenys dels diferents subcorpus, quins són els procediments de recollida i de tractament de dades i quin és l'estat actual de la feina.

2.4.1. Objectius i dissenys del CORCS

El CORCS té per objectiu confegir una base de dades lingüístiques orals de català central contemporani de la regió de Barcelona. En el CORCS, cadascun dels subcorpus ha estat dissenyat de forma autònoma, és a dir amb objectius, procediments de recollida de dades i dimensions de les mostres propis i diferents entre ells, però sense perdre de vista la coordinació necessària per tal de facilitar-ne l'explotació conjunta posterior. Tot seguit presentem el disseny i les característiques principals que presentarà aquest corpus una vegada acabat:

A) Subcorpus de varietats funcionals o registres (COR)

El COR està dissenyat perquè contingui trenta mostres de trenta minuts de durada mitjana d'interaccions seleccionades combinant dos criteris:

(1) a partir d'una graella confeccionada sobre la base de quatre variables dicotòmiques de variació contextual, proposades per Lewis i Carroll a *Language and situation. Language varieties and the social contexts* (Londres, Routledge & Kegan Paul, 1978):

camp (no especialitzat/especialitzat)

mode (oral no preparat/preparat)

tenor (informatiu/interactiu)

to (formal/informal)

(2) a partir d'una selecció de mostres d'esdeveniments comunicatius rellevants en la comunitat catalanoparlant.

B) Subcorpus de conversa espontània

Conté trenta mostres de trenta minuts de durada d'interaccions de camp no especialitzat i no preparades, interactives i de to informal, del gènere conversacional. L'edat dels parlants oscil·larà entre els 18-30 anys i els quaranta anys i aquests pertanyeran preferentment a capes socials mitjanes i amb estudis mitjans o superiors.

C) Subcorpus de varietats socioculturals

Està dissenyat per contenir vuitanta mostres de parlants pertanyents a les diferents capes socials de la regió de Barcelona, seleccionats a partir de les categories proposades per l'Enquesta Metropolitana de la Regió Metropolitana de Barcelona. Els parlants seleccionats són de primera llengua catalana o bilingües familiars. Evidentment, tant en aquest subcorpus com en els altres, apareixen un nombre important de fragments de parla de llengua castellana, que oferim als qui estudien el castellà regional. Cadascuna d'aquestes mostres conté quinze minuts d'interaccions. Està previst que aquest subcorpus disposi de dos tipus de materials:[3]

[3] Hom disposa com a complement del COS d'un total de vuitanta-una entrevistes d'aproximadament una hora cadascuna procedents dels enregistraments (retransmesos o no retransmesos), tant en video com transcrites ortogràficament en paper i disquet, del programa *Ciutadans* de TV3, en què es recollien opinions de sectors especialment significatius de la societat catalana. Aquests enregistraments, que van acompanyats de dades dels parlants, seran incorporats progressivament al conjunt del COS.

1) Cinquanta entrevistes semidirigides de final obert, entorn d'un guió d'història de vida. El quadre 1 mostra quins són els grups socials entrevistats al COS.

2) Trenta converses espontànies homogènies des del punt de vista contextual.

Quadre 1
Mostres dels grups socials enregistrats al COS.

GRUP SOCIAL	Percent. Població	18-25 anys	35-45 anys	60-70 anys	Totals
Tècnics alts i professionals liberals	6,9	1	1	1	3
Empleats i tècnics mitjans	22,5	13	2 2	2	10
Moderns	1,5	--	1 --	--	1
Petits empresaris	3,3	1	1	1	3
Comerciants i artesans	9,2	1 1	1 1	1	5
Contramestres i capatassos	3,1	1	1	--	2
Treballadors manuals en actiu	31,6	1 4	3 1	1	10
Treballadors inactius	20,9	--		2 4	6
Totals absoluts	100,0	14	14	12	40

2.4.2. Recollida i tractament de les dades en el CORCS

Les dades del CORCS estan essent recollides en el marc de diferents assignatures de segon i tercer cicle i arriben al Departament en forma de cintes magnetofòniques i transcripcions escrites. L'objectiu del tractament del material recollit és fer les dades més fàcilment consultables en un format uniforme. Per aconseguir-ho se segueixen les passes següents:

1. *Unificació dels materials* rebuts a un sol suport informàtic (WP 2.1 de Macintosh)
2. *Homogeneïtzació i correcció inicial* de divergències i errors en les transcripcions detectables sense l'ajut dels textos orals.

Una vegada acabat aquest segon estadi, els textos del CORCS són ja utilitzables i. de fet ja estan essent utilitzats, per a tot un seguit de recerques que no exigeixen la sincronització estricta de l'escrit i de l'oral. Aleshores les fases posteriors són:

3. *Conversió dels textos corregits* a base de dades de relacions (Fox-Base)
4. *Digitalització dels enregistraments*
5. *Sincronització o paral·lelització* dels textos escrits amb els enregistraments digitalitzats
6. *Verificació final* de les transcripcions en relació amb l'oral.

Cal assenyalar que la correcció del CORCS no contempla la codificació del material per a tota mena d'investigacions. La correcció té per objecte assegurar el màxim de fidelitat entre les versions escrites i orals i d'introduir uns mínims comuns de transcripció que permetin un ús multidisciplinar de la base de dades.

2.4.3. Estat actual del CORCS?

El material del CORCS està essent tractat de forma sectorial segons la seva data d'arribada i el subcorpus al qual s'adscriu.

En conjunt en l'actualitat disposem de 246,5 hores d'enregistrament, desglossades de la manera següent:

24,5 hores del COC (49 textos x 30 minuts/text)
8 hores del COR (16 textos x 30 minuts/text)
67 hores del COS(67 textos de converses semidirigides x 1 hora/text)
80 hores del Programa *Ciutadans* de TV3

Quadre 2
Estat actual del CORCS en relació al nombre de textos de cada subcorpus

CORPUS	UNIFIC.	HOMOG.	CONV.	DIGIT.
COC	45	30	11	11
COR	16			
COS	147	37		

Estat actual del CORCS en relació al nombre d'hores de cada subcorpus

CORPUS	UNIFIC.	HOMOG.	CONV.	DIGIT.
COC	27,5	15	5,5	5,5
COR	8			
COS	150,5	52		

UNIF: unificació del format informàtic (conversió a WordPerfect 2.1 per a MAC)
HOMOG: homogeneïtzació i correcció
CONV: conversió en bases de dades
DIGIT: digitalització

Els parlants enregistrats al CORCS en l'actualitat ja ultrapassen els cinc-cents, bé que la seva participació és extremadament variable: des d'un torn ocasional en una conversa fins a un monòleg de llarga durada en una història de vida. L'estat actual d'aquests textos queda reflectit en quadre 2. *Unif.* significa unificació en WordPerfect 2.1. per a MAC. *Homog.* significa homogeneïtzació i correcció. *Conv.* significa conversió a base de dades i *Digit.* significa digitalització.

En aquests moments s'ha arribat a digitalitzar gairebé sis hores del corpus de conversa, cosa que ha permès disposar ja del primer CD Rom amb veu digitalitzada.

Aquesta veu digitalitzada pot ara començar a ser sincronitzada o paral·lelitzada amb la part escrita.

2.4.4. Explotació i difusió del CORCS

Per a l'ús docent intern el CORCS ja està essent utilitzat des de les seves primeres etapes per a la formació pràctica dels estudiants de segon i tercer cicle mitjançant l'organització del treball de camp de disseny de la recerca i recollida de dades, la transcripció dels textos orals, la digitalització de la veu, la correcció i la supervisió dels textos i d'altres.

Podem esmentar a tall d'exemple algunes de les línies de recerca per al CORCS:

- la descripció de la gramàtica del català actual
- la comparació de la llengua oral amb la llengua escrita en totes les seves vessants
- l'estudi de les diferències entre les diverses varietats funcionals i socials de la llengua.

No tenim elements de judici, per exemple, per descriure quin és el català dels sectors populars de la regió metropolitana de Barcelona.

- la investigació tant quantitativa/i qualitativa del canvi lingüístic. Les conseqüències lingüístiques dels grans canvis demogràfics i socials de la zona de Barcelona encara ens són desconeguts.
- les ideologies lingüístiques. Una part important d'entrevistes semidirigides s'adreçaren a cònjuges de famílies lingüísticament mixtes. El qüestionari seguit permet descriure i descobrir, mitjançant històries de vida, els factors relacionats amb les pautes de transmissió lingüística intergeneracional.
- la recerca en contacte de llengües
- l'estructura de la interacció en català

Hi ha prevista la difusió del CORCS en diferents formats, de manera que el públic interessat pugui adquirir-lo en aquell que s'adigui més amb les seves necessitats i capacitats operatives:

- Quatre CD-Roms que permetran la consulta simulània de les formes gràfiques i orals o un nombre de disquets amb capacitat equivalent.
- Un paquet de cintes magnetofòniques amb les mostres orals seleccionades.
- Un llibre que contindrà tots els textos escrits

3. Conclusions

Els projectes de corpus són complexos, necessaris i possibles. Aquest projecte de corpus (el CUB) que acabem de descriure serà també útil. EL CUB facilita que els investigadors facin tant recerques específiques de cada subcorpus com algunes recerques transversals que posin en relació els elements d'interès de cada subcorpus.

Aquestes recerques transversals entre les parts del CUB són més factibles en els aspectes lingüístics i textuals -sempre hi ha text en el corpus, sigui quina sigui - i menys factibles en els aspectes contextuals i sociolingüístics perquè el projecte no s'ha construït sobre un disseny unificat sinó modular.

L'objectiu últim d'aquesta ponència és donar compte d'un treball de recerca de base per a la descripció de la llengua catalana. Aquest corpus és un projecte endegat amb diners públics per treballar en una riquesa natural com és la llengua. Aquest corpus en formació és, per tant, obert també a l'intercanvi de dades i de metodologia amb altres centres públics de recerca.

Bibliografia

BIBER, D. i FINEGAN, E. (eds.) (1994), *Sociolinguistic Perspectives on Register*, Oxford University Press, Oxford.

CARDÚS, S. (1991), *Els Països Catalans en xifres*, Fundació Jaume Bofill.

COLOMINA, J. (1990), *Bibliografia de dialectologia catalana*, "A Sol Post. Estudis de Llengua i Literatura", 1, (1990), pàg. 75-131.

IRVINE, J. T. (1984), *Formality and Informality in Communicative Events*, dins *Language in Use. Readings in Sociolinguistics*, (a cura de J. Baugh i J. Sherzer), Prentice-Hall, Englewood Cliffs, pàg. 211-228.

KOLDE, G. (1981), *Sprachkontakte in Gemischtsprachigen Städten*, Franz Steiner, Wiesbaden.

LAMUELA, X. (1994), *Establiment i estandardització de llengües*, Ed. 62, Barcelona.

MILÀ FONTANALS, M. (1890), *Catalán contemporaneo. Lenguaje de Barcelona*, dins *Obras Completas*, III, Barcelona, pàg. 511-544.

MARTEL, P. et al. (1992), *Dictionnaire de fréquences des mots du français parlé au Québec*, Peter Lang, Nova York.

MONTOYA, B. (1992), *Per una recerca de la llengua parlada en català*, dins *Miscel.ània Joan Fuster. Estudis de Llengua i Literatura*, V, (a cura d'A. Ferrando i A. G. Hauf), Publics. de l'Abadia de Montserrat, Barcelona, pàg. 391-417.

SUBIRATS, M. et al. (1992), *Enquesta metropolitana 1986. Condicions de vida i hàbits de la població de l'àrea metropolitana de Barcelona, 20.Transmissió i coneixement de la llengua catalana a l'àrea metropolitana de Barcelona*, Institut d'Estudis

Metropolitans, Barcelona.

THIBAULT, P. i VINCENT, D. (1990), *Un corpus de français parlé. Montréal 1984: historique, méthodes et perspectives de recherche*, Université Laval, Quebec.

TUSÓN, A. (1987), *El repertori lingüístic de la ciutat de Barcelona*, dins *Formazione dell'insegnante di lingue in ambiente di lingue in contatto*, Bargatto, Roma, pàg. 63-85.

WITTE, E. i BAETENS BEARDSMORE, H. (1987), *The Interdisciplinary Study of Urban Bilingualism in Brussels*, Multilingual Matters, Clevedon.

El projecte CECA
(Corpus escrit de català)*

M. Teresa Cabré, Lluís de Yzaguirre i Mercè Lorente
(Universitat Pompeu Fabra)

1. Presentació

El CECA, subcorpus escrit del CUB (Corpus de la Universitat de Barcelona) dirigit pels professors M. Teresa Cabré, Mercè Lorente i Lluís de Yzaguirre, es va dissenyar tenint en compte que els materials constituïts per altres organismes, com l'Institut d'Estudis Catalans, no permetien de dur a terme algunes aplicacions científiques que es preveu de realitzar amb aquests materials, específicament estudis sobre neologia catalana i en contrast amb altres llengües, i anàlisi del text periodístic.

Amb aquesta idea, es va preveure de constituir inicialment un doble arxiu de premsa escrita:

1) L'arxiu CEDICA, integrat per premsa catalana, que havia d'incloure premsa de tres subarxius:
   a) premsa editada a Barcelona
   b) premsa comarcal
   c) premsa catalana d'abast general
2) L'arxiu CEDICAST, format per text de premsa en llengua castellana, també amb dos subarxius:
   a) premsa en llengua castellana editada a Barcelona
   b) premsa en llengua castellana editada fora de Barcelona

* El projecte *Variació en el llenguatge: corpus oral i escrit de català contemporani* rep finançament de la CIRIT (CS93-1017) i de la DGICYT (PB 90-0505).

Aquest doble arxiu s'havia de complementar amb el subcorpus CETV (Corpus de Textos de TV3), format per un recull de textos escrits per a ser llegits, actualment en fase de disseny.

El projecte inicial previst s'ha tancat avui dia en un arxiu de premsa en llengua catalana del diari AVUI, CECA. i és previst de complementar-lo en fases successives amb les aportacions dels nous Projectes "Llenguatges especialitzats" i "OBNEB" que duen a terme actualment els membres de l'equip a l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, tant pel que fa a la compleció de dades, com a l'elaboració d'eines d'explotació.

En aquesta fase, l'equip ha comptat amb la cooperació de diversos becaris: Xavier Solé, Roland Pearson, Zulema Borràs i Carme Bach.

2. Estat actual del Projecte CECA

CECA inclou els textos corresponents a 120 dies consecutius del Diari AVUI (del 17 de febrer de 1993 al 25 de juny del mateix any), que corresponen a 7.000.000 de formes (excloses les procedents de seccions del diari repetitives: cartellera i borsa). Les dades, emmagatzemades en un disc òptic, ocupen 50 Megaoctets de memòria.

Les dades estan estructurades en format ASCII i distribuïdes en 120 fitxers separats, cada un dels quals correspon a un dia de diari i ocupa aproximadament 500K de memòria.

Cada fitxer, integrat per una part d'identificació (l'encapçalament) i el contingut (el text) s'identifica per un encapçalament de 6 caràcters numèrics, que corresponen a les dades corresponents d'any/mes/dia del diari.

El text de cada fitxer s'ha etiquetat amb les marques següents:

a) pàgina

b) secció del diari

c) marques tipogràfiques (negreta i cursiva)

d) altres delimitadors

Del fitxer de text sencer s'han generat, per necessitats d'investigació, altres fitxers alternatius, emmagatzemats per fragments de text: frases i mots.

3. Adquisició de les dades

En una primera fase, quan no existia la possibilitat d'obtenir el diari AVUI a través d'Internet, es van explorar dues vies possibles per adquirir les dades, refusada la possibilitat de la via manual:

a) l'escannerització

b) l'obtenció dels materials en suport magnètic

La primera via es va descartar pel temps i l'esforç que requeria a causa de la mala qualitat de les edicions i, gràcies a la generositat dels directius del diari AVUI, es van obtenir els textos en disquet de fotocomposició. Amb un programa de neteja dels codis no rellevants (conservant-ne alguns de pertinents: final de pàgina, marques tipogràfiques, etc.) es va elaborar el fitxer definitiu de CECA, actualment en disc òptic.

4. Projectes previstos o en curs d'elaboració

Del projecte CECA se n'han d'obtenir pròximament els següents productes:

Un llistat lèxic general de freqüències

Llistats selectius de freqüències per seccions

Un diccionari de neologismes de premsa

Un diccionari de manlleus

Tres són les línies de recerca fonamentals que es desenvolupen a partir del projecte CECA:

a) La creativitat lèxica a través de la premsa

b) Morfologia lèxica del català: processos morfològics i projecció sintàctica del lèxic

c) Enginyeria lingüística: analitzadors i generadors

5. Programes d'explotació de les dades

Els programes que fins ara permeten explotar els materials són els següents:

TACT

Word Cruncher

Le Concordeur
Programes realitzats *ad hoc* en Pascal

6. Mostra dels materials de CECA

Inclou un fragment del diari amb marques estructurals, una versió sobreetiquetada (sense desambiguar) i una versió parcialment desambiguada. Per interpretar els codis i per veure el llistat de les regles de desambiguació que s'han usat (documents massa extensos per adjuntar-los), connecteu-vos via Internet a l'URL "http://www.iula.upf.es"

```
<Data=930227>
<Pàgina=2>
<Secció=MÓN>
<Text=930227-01>
<Part=títol>
<frase=1>
Rabat vol finançament de la CE per eradicar els conreus de
droga i frenar l'emigració./
<Part=text>
<frase=2>
"Maria Favà", "corresponsal" RABAT ./
<frase=3>
Marroc confia en l'ajuda econòmica de la Comunitat Europea
(CE)
per eradicar els conreus de droga del Rif i per posar fi a
l'emigració clandestina.
<frase=4>
Així de clar ho va dir ahir al vespre Karim Lamrani, el
primer
ministre marroquí, durant un sopar ofert en honor de Jacques
Delors, president de la Comissió Europea, que està en visita
oficial al Marroc./
```

```
<frase=5>
Des del setembre passat, el govern marroquí ha enviat al
nord
del país 3.000 policies, una esquadrilla d'helicòpters i una
brigada d'inspectors fiscals per lluitar contra la droga,
l'emigració i el contraban.
<frase=6>
I el que ara demana als Dotze són diners per potenciar un
pla
de desenvolupament del Rif, la zona on es conrea la droga i
d'on surten molts dels futurs emigrants il·legals.
<frase=7>
Aquest pla s'ha xifrat en 20.000 milions de dirhams (uns
260.000 milions de pessetes)./
<\Pàgina=2>
```

Text sobreetiquetat

```
Rabat vol      finançament de          la          CE     per
&NB;  &N5MS;  &N5MS;       &P;         &AFS;       &N5FS; &P;
      &VDR3S;              &N5FS;&ETL; &RE3FS;             &N5MS;
                                  &N5FS;&ETN;

eradicar els      conreus de          droga   i           frenar
&VI;      &AMP;   &N5MP;  &P;         &N5FS;  &C;         &VI;
          &RE36P;         &N5FS;&ETL; &V8R6S; &N5FS;&ETL;


l'      emigració./
&AMS;   &N5FS;
&RE3MS;
```

```
«Maria Favà, » «corresponsal» RABAT ./
 &N4;  &NC;      &J6S;          &NB;

Marroc confia  en     l'       ajuda   econòmica de          la
&N5MS; &V8R6S; &P;    &AMS;    &N5FS;  &JQFS;     &P;         &AFS;
               &AMS;  &RE3MS;  &V8R6S;            &N5FS;&ETL; &RE3FS;
               &RV                                            &N5FS;&ETN;

Comunitat Europea (CE)     per     eradicar els     conreus de
&N5FS;    &JFS;    &N5FS;  &P;     &VI;     &AMP;   &N5MP;  &P;
                           &N5MS;           &RE36P;         &N5FS;&ETL;

droga   d   el    Rif  i           per    posar fi     a
&N5FS;  &P; &AMS; &NO; &C;         &P;    &VI;  &N5FS; &P;
&V8R6S;               &N3FS;&ETL; &N5MS;       &JQMS; &N5FS;&ETL;

l'      emigració clandestina.
&AMS;   &N5FS;    &JQFS;
&RE3MS;

Així de          clar  ho      va       dir    ahir a   l
&D4; &P;         &JMS; &RE36S; &V6DR3S; &VI;   &D4; &P; &AMS;
     &N5FS;&ETL;             &JQMS;   &N5MS;

vespre Karim Lamrani, el      primer ministre marroquí, durant
&N5MS; &NM;  &NC;     &AMS;   &D4;   &N5MS;   &JMS;     &P;
                      &RE3MS; &JOMS;
                              &N5MS;

un      sopar  ofert  en    honor  de          Jacques Delors,
&JN3MS; &VI;   &VCMS; &P;   &N5MS; &P;         &NPM;   &NC;
&RN3MS; &N5MS;        &AMS;        &N5FS;&ETL;
```

```
&AMS;                &R
```

```
president de          la          Comissió Europea, que    està
&N5MS;    &P;         &AFS;       &N5FS;   &JFS;    &C;    &VDR3S;
          &N5FS;&ETL; &RE3FS;                       &RR66;
                      &N5FS;&ETN;                   &D4;
```

```
en   visita  oficial a   l     Marroc./
&P;  &N5FS;  &J6S;   &P; &AMS; &N5MS;
&AMS; &V8R6S;
&RV
```

```
Des         d   el    setembre passat, el      govern marroquí
&P;         &P; &AMS; &N5MS;   &VCMS;  &AMS;   &N5MS; &JMS;
&N5FP;&ETL;                    &JMS;   &RE3MS;
```

```
ha       enviat a   l     nord   d   el    país   3000   policies,
&I;      &VCMS; &P; &AMS; &N5MS; &P; &AMS; &VJA6S;&JC6P; &N5FP;
&V6DR3S; &N5MS;                             &N5MS;
```

```
d'  inspectors fiscals per    lluitar contra la          droga,
&P; &JMP;      &J6P;   &P;    &VI;    &P;    &AFS;       &N5FS;
                       &N5MS;                &RE3FS;     &V8R6S;
                                             &N5FS;&ETN;
```

```
l'      emigració i          el      contraban.
&AMS;   &N5FS;    &C;        &AMS;   &N5MS;
&RE3MS;           &N5FS;&ETL; &RE3MS;
```

```
I          el      que    ara    demana  a   ls    Dotze  són
&C;        &AMS;   &C;    &D4;   &V8R6S; &P; &AMP; &JC6P; &V6DR3P;
&N5FS;&ETL; &RE3MS; &RR66; &N5FS;                  &N566;
                   &D4;
```

```
diners per    potenciar un      pla    de          desenvolupament
&N5MP; &P;    &VI;       &JN3MS; &JQMS; &P;         &N5MS;
       &N5MS;            &RN3MS; &N5MS; &N5FS;&ETL;
                         &AMS;   &D4;

d   el    Rif, la          zona   on   es          conrea  la
&P; &AMS; &NO; &AFS;       &N5FS; &D4; &AMS;&EDB; &V8R6S; &AFS;
               &RE3FS;                 &RE366;             &RE3FS;
               &N5FS;&ETN;             &N5FP;&ETL;         &N5FS;&ETN;

droga   i          d'  en   surten  molts  d   els   futurs
&N5FS;  &C;        &P; &D4; &VDR3P; &JFMP; &P; &AMP; &JMP;
&V8R6S; &N5FS;&ETL;                &RFMP;

emigrants il·legals.
&J6P;     &JQ6P;

Aquest pla    s'         ha       xifrat en    20000  milions
&JDMS; &JQMS; &AFS;&EDB; &I;      &VCMS; &P;   &JC6P; &N5MP;
&RDMS; &N5MS; &RE36S;    &76DR3S;        &AMS;
       &D4;                               &RV

de          dirhams (uns     260000 milions de          pessetes).
&P;         &N5MP;   &JN3MP; &JC6P; &N5MP;  &P;         &N5FP;
&N5FS;&ETL;          &RN3MP;                 &N5FS;&ETL;
                     &AMP;

* Mots       = 175 (175)
* Errors     = 0 (0%)
* Lemes      = 271
* Coeficient = 154
```

Text parcialment desambiguat

```
Rabat vol      finançament de  la           CE      per
&NB; &N5MS;    &N5MS;        &P; &AFS;&R0032; &N5FS; &P;&R1000;
      &VDR3S;

eradicar els          conreus de  droga          i  frenar l'
&VI;     &AMP;&R0032; &N5MP;  &P; &N5FS;&R1010; &C; &VI;   &AMS;&R0032;

emigració./
&N5FS;

«Maria Favà, » «corresponsal» RABAT ./
 &N4;  &NC;      &J6S;          &NB;

Marroc confia  en          l'           ajuda          econòmica
&N5MS; &V8R6S; &P;&R1020; &AMS;&R1020; &N5FS;&R1030; &JQFS;

de  la           Comunitat Europea (CE)    per         eradicar
&P; &AFS;&R0032; &N5FS;    &JFS;    &N5FS; &P;&R1000; &VI;

els          conreus de  droga         d   el    Rif  i   per
&AMP;&R0032; &N5MP;  &P; &N5FS;&R1010; &P; &AMS; &NO; &C; &P;&R1000;

posar fi     a   l'           emigració clandestina.
&VI;  &N5FS; &P; &AMS;&R0032; &N5FS;    &JQFS;
      &JQMS;

Així de  clar  ho      va       dir         ahir a   l     vespre
&D4; &P; &JMS; &RE36S; &V6DR3S; &VI;&R0130; &D4; &P; &AMS; &N5MS;

Karim Lamrani, el           primer        ministre marroquí,
&NM;  &NC;      &AMS;&R0082; &JOMS;&R1040; &N5MS;   &JMS;
```

```
durant un          sopar        ofert  en        honor  de
&P;    &AMS;&R1050; &N5MS;&R1050; &VCMS; &P;&R1060; &N5MS; &P;

Jacques Delors, president de  la          Comissió Europea,
&NPM;   &NC;    &N5MS;    &P; &AFS;&R0032; &N5FS;   &JFS;

que          està    en         visita        oficial a   l
&RR66;&R0080; &VDR3S; &P;&R1070; &N5FS;&R1070; &J6S;   &P; &AMS;

Marroc./
&N5MS;

Des d   el          setembre passat, el          govern
&P; &P; &AMS;&R0032; &N5MS;   &VCMS;  &AMS;&R0032; &N5MS;
                                     &JMS;

marroquí ha  enviat a   l     nord   d   el    país
&JMS;    &I; &VCMS; &P; &AMS; &N5MS; &P; &AMS; &N5MS;&R1030;

3000   policies, una          esquadrilla d'  helicòpters i
&JC6P; &N5FP;    &AFS;&R0032; &N5FS;      &P; &N5MP;      &C;

una          brigada d'  inspectors fiscals per        lluitar
&AFS;&R0032; &N5FS;  &P; &JMP;      &J6P;   &P;&R1000; &VI;

contra la                  droga,        l'          emigració
&P;    &AFS;&R1080;&R1081; &N5FS;&R1081; &AMS;&R0032; &N5FS;

i   el          contraban.
&C; &AMS;&R0032; &N5MS;
```

```
I   el          que          ara    demana  a
&C; &AMS;&R0082; &RR66;&R1090; &D4;   &V8R6S; &P;
                                              &N5FS;

ls    Dotze  són      diners per        potenciar un
&AMP; &JC6P; &V6DR3P; &N5MP; &P;&R1000; &VI;      &AMS;&R0032;
      &N566;

pla           de         desenvolupament d   el    Rif, la
&N5MS;&R1100; &P;&R1100; &N5MS;          &P; &AMS; &NO; &AFS;&R0030;

zona   on   es      conrea  la          droga        i   d'
&N5FS; &D4; &RE366; &V8R6S; &AFS;&R1120; &N5FS;&R1300; &C; &P;

on   surten  molts        d   els   futurs emigrants il·legals.
&D4; &VDR3P; &RFMP;&R1110; &P; &AMP; &JMP;  &J6P;     &JQ6P;

Aquest pla    s'      ha  xifrat en    20000  milions de  dirhams
&JDMS; &JQMS; &RE36S; &I; &VCMS; &P;   &JC6P; &N5MP;  &P; &N5MP;
&RDMS; &N5MS;                    &AMS;
       &D4;                      &RV

uns          260000 milions de  pessetes)./
&AMP;&R0032; &JC6P; &N5MP;  &P; &N5FP;

* Mots       = 175 (175)
* Errors     = 0 (0%)
* Lemes      = 185
* Coeficient = 105
```

TAULA RODONA

Recerques i aplicacions a partir de corpus lingüístics*

Lluís Payrató

Bona tarda. Gràcies als ponents, primer de tot, per haver acceptat la invitació a participar en aquesta taula rodona, que té com a finalitat cloure la jornada arrodonint els temes tractats al matí.

El símbol que figura en el tríptic que vam editar per a aquest primer col·loqui lingüístic de la Universitat de Barcelona (CLUB) permet, aparentment, interpretacions diverses: segons com es miri pot semblar una A, o, segons com, s'hi veuen quatre barrots... En realitat és el signe dels alquimistes, representa el concepte d'*amalgama*. I en aquest sentit aquesta taula rodona voldria ser una mena d'amalgama, metafòrica, evidentment, del que s'ha fet aquest matí al llarg de cadascuna de les quatre ponències.

La comissió organitzadora ha previst un desenvolupament determinat de la taula rodona, a fi que no es converteixi en una sessió massa desestructurada. Explicaré molt breument aquest desenvolupament i donaré de seguida la paraula als ponents.

Hi haurà un torn inicial de cinc minuts aproximadament per a cadascun dels participants, en el mateix ordre que les ponències del matí. Els continguts d'aquesta primera intervenció tindran a veure amb el títol genèric de la taula: "Recerques i aplicacions a partir de corpus lingüístics." A continuació hi haurà un torn tancat de preguntes als participants que tindrà tres eixos, que no tenen com a finalitat encadenar la

* La transcripció d'aquesta taula rodona ha estat obra de Lídia Torres i Carme Bach, i l'edició de Lluís Payrató, que va fer-hi de moderador.

realització de la taula, sinó simplement indicar uns camins per al seu desenvolupament. Aquests tres eixos, formulats en forma de pregunta, són els següents:

1) ¿És veritat que en lingüística, i en les ciències del llenguatge en general, s'ha passat d'una posició predominantment contrària als corpus a una de predominantment favorable? ¿Estem assistint al naixement d'una lingüística de corpus?

2) ¿Fins a quin punt són determinants per al disseny d'un corpus la recerca i les aplicacions que s'han projectat dur a terme?, o dit amb altres paraules: ¿Fins a quin punt és possible un corpus més aviat neutre o obert, que no prejutgi una recerca i unes aplicacions exclusives?

3) ¿De quina manera la presentació material del corpus i les vies per accedir-hi limiten o faciliten les recerques i aplicacions que se'n poden derivar?

Repeteixo que no es tracta, per descomptat, de cenyir-nos només exclusivament a aquests temes. En sortiran d'altres, però aquesta estructuració vol donar simplement uns camins perquè el desenvolupament de la taula sigui més o menys endreçat.

A continuació hi haurà un torn reservat a la taula per si algú vol introduir un tema addicional, i finalment hi haurà un torn obert a tot el públic assistent.

Acabo la meva intervenció. A part d'agrair als ponents la seva presència, voldria demanar ara, simplement, que mirin de cenyir-se al temps acordat i de no excedir-se en aquest primer torn inicial que hem situat al voltant dels cinc minuts per a cadascú. De manera que, en l'ordre que s'han fet les ponències aquest matí, té la paraula el professor José Manuel Blecua.

José Manuel Blecua

Procuraré ajustarme a los cuatro minutos y medio. Como he prometido esta mañana me gustaría presentar rapidísimamente los problemas que el proyecto EAGLES tiene en estos momentos con referencia a los corpus textuales. EAGLES es una de las tantas siglas de la Unión Europea que corresponde a un grupo de expertos asesor en materia de estándares para la ingeniería lingüística. En EAGLES se trata de elaborar unos estándares que permitan que los grandes recursos lingüísticos, diccionarios electrónicos, corpus, gramáticas computacionales, puedan ser transportados y reutilizados.

EAGLES está dividido en cinco grandes áreas: corpus textuales, léxicos computacionales, formalismos gramaticales, evaluación de productos de terminología

lingüística y lengua hablada. No voy a hablar más que de corpus textuales porque el profesor Llisterri esta mañana ya ha hecho suficiente alusión a los problemas de la lengua oral.

Cada grupo de trabajo, que está instalado en el Instituto Cervantes, tiene unos subgrupos que se dedican a tareas específicas. La enumeración de estas tareas específicas es simplemente el resumen de todo lo que hemos hablado esta mañana y de los grandes problemas que hoy tienen los corpus textuales. Primero el problema de las tipologías de los corpus y de los textos, problema que yo creo que no va a haber manera de resolver nunca. Pero esto tendrá la ventaja de que de un corpus se podrá extraer automáticamente un subcorpus especial. En segundo lugar el problema de la representación textual. Las estructuras y fenómenos tipográficos en un texto tienen que tener una codificación uniforme, como veíamos largamente esta mañana. Un tercer problema que ha surgido a propósito del corpus de Madison es la notación de palabras aisladas. Toda palabra de un corpus tiene que tener una notación sintáctica, semántica, morfológica, y a ser posible esta notación tiene que ser automática y revisable. Sin este proceso de *tagging*, como se llama técnicamente, es imposible realizar un corpus.

A todos ustedes les son muy familiares los fenómenos de ambigüedad que aparecen en cualquier texto. El otro problema, que explicaba Joan Torruella esta mañana. la pregunta de la Dra. Cabré, son las especificaciones sobre las características de la documentación, es decir, la definición del documento. Porque sin esta definición no es posible trasladar un documento que no se sabe en qué consiste.

Hablamos esta mañana de un fenómeno que es muy interesante, y es que un corpus no se concibe aisladamente sin unas herramientas que permitan su explotación. La construcción de estas herramientas es tan importante como la construcción del corpus mismo. El problema de los corpus de textos paralelos es fundamental. Un texto traducido a varios idiomas aparece en correlación, alineado, y hay que representar estos emparejamientos. Y por último está el problema de que les hablaba el profesor Llisterri: cómo se representan, transcriben y codifican los textos hablados.

Estas cuestiones que trata el proyecto EAGLES tendrán que estar acabadas en 1995. Y en 1995 estos estándares se introducirán como recomendaciones en los trabajos de ingeniería lingüística de la comunidad. Seguirá luego un proceso complejo, intentar convencer a la comunidad científica de que estos son los estándares que hay que utilizar,

y que Dios nos de más suerte que a los de Madison, como veíamos esta mañana, con sus famosas normas. Por lo menos que nos sigan un poco más.

Joaquim Llisterri

Bé, jo intentaré cenyir-me als temes que havíem de tractar en aquesta taula, potser llançar unes primeres aproximacions i després ja veurem cap a on va el debat.

Com que el tema eren recerques i aplicacions de corpus lingüístics, m'ha semblat que podia començar fent una llista de coses per a les quals em sembla que poden servir els corpus orals, que era el tema que jo havia de tractar. Intentant organitzar aquesta llista, jo crec que hi ha tres grans àmbits d'aplicació dels corpus orals: el primer seria la descripció lingüística, el segon seria el de les tecnologies de la parla i el tercer el desenvolupament d'altres recursos lingüístics a partir d'un corpus.

El que es pot fer en descripció lingüística quan es té un corpus, crec que ha quedat molt ben exemplificat aquest matí quan ens han presentat el corpus de la Universitat de Barcelona.

El que es pot fer en tecnologia de la parla, bàsicament ho he intentat documentar també en la meva intervenció del matí: entrenar i avaluar sistemes de reconeixement; dissenyar, entrenar i avaluar sistemes de diàleg home-màquina, i potser un aspecte que he deixat una mica de banda, però que enllaça amb la qüestió dels corpus paral·lels als quals s'ha referit el Dr. Blecua, la qüestió de la traducció automàtica de converses telefòniques: per portar-la a terme fa falta un bon corpus de converses telefòniques reals.

Pel que fa al desenvolupament dels recursos lingüístics, en el camp dels corpus orals específicament, per exemple el Linguistic Data Consortium que presentava aquest matí, té un projecte entre les mans, que es diu Comlex, que és un lexicó, un diccionari, però que a més a més portarà incorporada la transcripció fonètica de cada paraula. I això em direu que no és cap novetat. Però a més a més portarà també incorporada la pronúncia real de tres o quatre parlants de totes aquestes paraules i portarà també una altra cosa que em sembla interessant, un mecanisme d'associació entre la forma aïllada de la paraula tal com la pronunciaria qualsevol d'aquests parlants que llegís una a una les paraules del diccionari i la forma que adquireix aquesta paraula en parla contínua.

A part d'això, dins l'àrea de desenvolupament de recursos dels corpus orals, també hi ha tot el camp de l'ensenyament de la llengua assistit per ordinador, del qual ja hem parlat.

Joaquim Rafel

Si ens hem de remuntar a les qüestions que aquest matí hem deixat necessàriament de banda per no incidir en els temes de la taula rodona d'aquesta tarda, una de les qüestions de què no he parlat explícitament, i que en certa manera també està recollida en les preguntes que vénen a continuació, fa referència a la diversitat d'aplicacions d'un corpus lingüístic. Quan es dissenya un corpus lingüístic, s'ha de pensar més aviat en l'aplicació concreta que es vol aconseguir o bé cal contemplar la possibilitat d'aplicacions diverses? No entraré massa en aquesta qüestió, però, lligant potser amb el que deia al matí, i remuntant-me a algunes qüestions sobre les quals vaig haver de reflexionar quan fèiem el primer projecte del que va ser l'origen del corpus que he descrit i que he mostrat en la seva realització i en els resultats d'aquest moment, hi ha alguns aspectes, després d'estudiar a fons la situació que en aquells moments de l'any 83 hi havia en el camp dels corpus, que lliguen d'alguna manera amb algunes de les qüestions a què aquest matí ens hem referit diversos ponents.

Jo vaig arribar aleshores a tres conclusions importants, que, bàsicament i resumint-les, diria que són: per una banda, la necessitat que hi hagi una política científica coherent per a poder abordar la realització del corpus; i per una altra banda, la necessitat d'una estabilitat i una continuïtat en la recerca, sense les qual és impossible la realització de corpus mínimament coherents i mínimament importants. Dic això perquè la paraula corpus, com tots sabem, s'aplica a realitzacions de molt diversa naturalesa i magnitud, de molt diverses intencionalitats i de molts diversos objectius; sense aquesta continuïtat en la recerca i sense aquesta estabilitat és impossible treballar profitosament en aquest terreny. I encara l'últim aspecte, que es relaciona més amb el que he començat dient, és la necessitat d'elaborar corpus que siguin realment aprofitables en diverses direccions i per a diverses finalitats. Dic això, potser d'una manera una mica absoluta, sabent, com acabo de dir, que hi ha corpus de molt diversa naturalesa i magnitud. Quan un corpus és extens, i per tant, vol dir que requereix uns recursos relativament importants per a la seva realització (tant recursos humans, com econòmics, com anys de dedicació, etc.), gairebé és una necessitat

intrínseca que el treball pugui ser aprofitable per a diverses finalitats. És tenint en compte aquestes consideracions que en el moment de fer la proposta concreta que va donar lloc al projecte que he exposat, vaig dir que un projecte d'aquesta naturalesa, concebut amb criteris moderns, no podia limitar-se a una acció puntual generada exclusivament per aquest motiu i encaminada a satisfer només aquesta necessitat. Per això he fet la precisió aquest matí que el corpus que he presentat és un corpus encaminat no exclusivament a la redacció d'un diccionari, perquè una acció com aquesta només es justifica dins una empresa d'una gran amplitud de mires, posada al servei de la recerca lingüística catalana en general. Seria doncs absurd que la complexa infrastructura que calia posar a punt per a la producció dels materials necessaris per al diccionari que es projecta acabés la seva funció amb la consecució d'aquest objectiu, i no es preveiés, d'una banda, la possibilitat de servir diversos fins i, d'altra banda, la continuïtat d'aquest treball, com podria fer d'alguna manera una institució que es posés al servei de la recerca lingüística en general.

Alguna d'aquestes afirmacions que he fet estan documentades en opinions de lingüistes prestigiosos que s'han dedicat a aquestes qüestions i recordo fins i tot alguna anècdota. Em sabria greu equivocar-me en alguna dada, però recordo que el professor J, de Kock va haver de realitzar uns treballs i volia partir del corpus informatitzat que havien utilitzat A. Juilland i J. Chang Rodríguez per a elaborar el diccionari de freqüències del castellà a què aquest matí ha al·ludit el Dr. Blecua, si no ho recordo malament. J. de Kock va tenir accés als materials informatitzats i va comprovar que, pel fet d'haver introduït aquell corpus en l'ordinador mirant només la finalitat de realitzar el diccionari de freqüències del castellà, no li servien per a la finalitat que ell volia aconseguir, que jo en aquests moments no recordo quina era. Va haver de tornar a introduir les dades a l'ordinador d'una manera adequada per a la realització del seu treball.

És probable que al llarg del desenvolupament d'aquesta taula rodona surtin opinions en contra i a favor dels corpus dissenyats específicament per a una finalitat concreta. Hi ha, però, exemples clars que demostren que un corpus que s'ha introduït en suport informàtic pensant només en un objectiu concret, per a una finalitat concreta, menyspreant dades, amb un disseny que impossibilita o dificulta que s'apliqui a altres finalitats, implica una duplicació de l'esforç per a tornar a introduir aquell corpus o un altre de semblant, quan no hauria estat necessari. No vull dir que amb això per a determinats corpus de finalitats molt específiques, la manera d'introduir les dades a l'ordinador, la

selecció de les dades, etc., no pugui estar encaminada a aquesta finalitat. En resum, la conclusió és que no es poden fer afirmacions generals sense dir a quina classe de corpus ens referim, sinó que cal precisar més. Però això no sé si ho podrem fer en el temps que queda aquesta tarda. Potser algun dia haurem d'organitzar un col·loqui molt més ampli per a tractar aquestes qüestions.

M. Teresa Cabré

Respecte al subcorpus escrit del corpus de català de la Universitat de Barcelona, i seguint l'esquema que ha introduït el professor Llisterri ara fa un moment, quan deia que d'aplicacions dels corpus se'n podien fer diverses, es podrien establir els tres grans grups esmentats de descripció lingüística, tecnologies de la parla i desenvolupament d'altres productes lingüístics.

Nosaltres hem de dir que pel que fa a l'escrit en principi només tenim previst fer aplicacions en el terreny de la descripció lingüística i del desenvolupament d'altres productes lingüístics. Suposo que el col·lega Emili Boix parlarà d'altres terrenys: el terreny de la fonètica i de la veu. Pel que fa a aquests dos apartats, ja està previst que la manera com tenim recollit i estructurat el material en suport magnètic ens permeti fer edicions del material.

Nosaltres partim de la base que, com que fet i fet el corpus del Departament s'ha constituït bàsicament per a finalitats de recerca, una de les primeres coses que hem de proporcionar són materials perquè els investigadors catalans, els col·legues i els especialistes, els estudiants que vulguin engegar una tesi doctoral puguin ja disposar d'aquests materials, precisament per fer aquesta tesi o treball de recerca, i per tant tenim previst d'entrada explotar aquests materials en el sentit de treure inventaris selectius, fonamentalment, primer, centrats en el lèxic, perquè en el lèxic de moment seria més fàcil de fer diccionaris de neologismes, diccionaris de freqüències, un inventari de manlleus, etc. Això no exclou tampoc la possibilitat que aquests índexs selectius els fem acompanyats de determinades informacions gramaticals, contextuals, etc., o que els fem simplement descontextualitzats, com després ja diré.

En segon lloc, pel que fa a la descripció lingüística és obvi que un corpus d'aquestes característiques permet de fer molts estudis tant centrats en el lèxic com en la morfologia, com en la sintaxi, com estudis de semàntica.

Com deia abans, el blanc és l'element que ens permet diferenciar d'entrada sense cap mena d'etiquetatge el que són les unitats lèxiques brutes, per dir-ho d'alguna manera, perquè tots som molt conscients que moltes vegades un segment que va de blanc a blanc inclou més d'una paraula i a l'inrevés, diferents segments que introdueixen blancs entremig constitueixen una sola unitat. De moment, encara no tenim previst un etiquetatge que ens permeti recuperar la informació en el sentit aquest de fragments llargs. que corresponen a una unitat, o de fragments enganxats, que corresponen a més d'una unitat. Però tot i això, està previst que ho abordem. I evidentment el nostre corpus tampoc no està etiquetat de manera que puguem extreure'n directament segments morfològics, perquè no el tenim etiquetat, cosa que tampoc no vol dir que no ho fem en un moment determinat, però d'aquella manera que us deia al matí. Més aviat ho farem d'una manera molt pràctica, és a dir, que comptant ja amb el material que tenim a la màquina, fent-ne sortir llistats freqüencials, llistats inversos, llistats directes, llistats de context, etc., anirem en tot cas estudiant quina és la forma més ràpida o més rendible d'introduir determinades etiquetes, perquè després puguem recuperar la informació des de molts punts de vista i seleccionar-la també des d'aquests punts de vista.

A part d'estudis teòrics, que es podrien engegar tant pel que fa a les eines de tractament informàtic de les dades com a la seva explotació, ja he dit al matí que estem constituint alguns materials que ens serveixen de materials de referència. Per exemple el primer i el més bàsic és un diccionari desplegat morfològicament, que ja el tenim en suport magnètic. Està previst que l'anem actualitzant d'una manera més o menys sistemàtica. Altres projectes que tenim en marxa, també en el terreny informàtic, són un lematitzador i un programa d'etiquetatge de locucions i frases fetes.

Una de les explotacions que ens agradaria fer del corpus conjuntament amb altres universitats que ja hi estan treballant és un programa que pugui establir el que són les unitats més llargues que el mot i que en canvi constitueixen un lexema, com serien les locucions, o allò que Benveniste anomena *sinapsis*. Des d'aquest punt de vista també està previst que en un període relativament mitjà, no pas relativament curt, treballem sobre un detector de sintagmes de tipus terminològic. La majoria dels que heu treballat en terminologia sabeu que un dels problemes fonamentals de la terminologia per al reconeixement de les unitats terminològiques dintre d'un domini determinat és el fet que es complica molt la qüestió perquè hi ha molt poques paraules simples i en canvi molts

grups de paraules complexes, d'associacions o combinacions de paraules que constitueixen una sola unitat terminològica. Dintre d'aquesta perspectiva, tenim previst realitzar un projecte que pugui arribar a establir, dintre d'un domini temàtic determinat, quines podrien ser les unitats terminològiques d'aquest domini. I també encara en el terreny de les realitzacions tenim molt avançat ja un programa, que realitzem conjuntament amb el Dr. de Yzaguirre, de detecció semiautomàtica de neologismes. Vist que teníem els textos dels diaris en suport magnètic, hem creat un programa que llegeix aquests textos en suport magnètic a base d'aquest diccionari desplegat morfològicament, que diem que és de referència. Va comparant totes aquelles unitats del diari que no reconeix amb el seu diccionari de referència. Diem que és un programa semiautomàtic, només, de detecció de neologismes, perquè ens ofereix tota una sèrie de possibilitats que no troba al diccionari, algunes de les quals són rellevants, mentre que les altres no tenen cap tipus d'importància. Les que no són importants són simplement unitats que són errates del diari, per això no es troben en el diccionari, o bé unitats com per exemple noms propis, que tot i així tenim previst d'estudiar de la forma més automàtica possible. Comptem per exemple que les universitats ens proporcionin llistes, evidentment del tot anònimes i en combinacions diverses (perquè cap nom propi correspongui a cap persona), simplement llistes de noms propis amb les quals puguem anar nodrint precisament aquest detector de neologismes, per tal que automàticament puguem bandejar els noms propis que no han de figurar en un diccionari, sense la necessitat de considerar-los, i així ens ofereixi una llista una mica més neta de possibles neologismes que es troben a la premsa, alguns dels quals són simplement mancances del diccionari, com una forma flexiva que no figurava en aquest diccionari. Però d'altres, en canvi, sí que suposen realment paraules noves. Estem treballant de moment pel que fa a l'explotació. Estem treballant en aquesta línia.

Emili Boix

Finalment parlaré de les recerques i aplicacions de la part més oral del corpus del Departament. És comú assenyalar en qualsevol explicació sobre recerca científica, que el punt de vista explica l'objecte. Els que estudiem l'oral, que en bona part és espontani i en bona part és oral provinent dels mitjans de comunicació, que és més controlat, treballem amb un material en brut i extremadament heterogeni. No és un oral recollit en situacions controlables sinó experimentals.

A la majoria dels que treballem amb l'oral ens interessa tant com s'utilitza la llengua com qui la utilitza. Utilitzant el símbol de la pasta de dents que posava Joaquim Llisterri quan parlava de la síntesi i reconeixement, no ens interessa mirar tant com fem entrar la pasta dins del tub i com la traiem, com qui utilitza aquesta pasta de dents, perquè ens interessa tot aquest context. El nostre punt de partida és que l'enunciació i l'enunciat van lligats. Aquesta és la postura típica dels que podem anomenar lingüistes impurs, perquè ens trobem a les fronteres amb altres matèries. La majoria dels que treballem l'oral estem en aquesta situació, encara que no tots.

Parlaré de les recerques que estem fent els que treballem ara i aquí en el projecte. Per no fer volar coloms, parlaré de la tasca d'anàlisi de les dades recollides dels mitjans de comunicació. Aquesta tasca la porta sobretot Lluís de Yzaguirre, que com s'ha vist porta moltes més tasques informàtiques. Un dels objectius fonamentals és analitzar fins a quin punt la població segueix les propostes ortològiques. Es tracta de mirar fins a quin punt l'ús espontani reprodueix propostes prescriptives. És un punt de vista, per cert, bastant oposat al que fem altra gent del corpus oral, però és complementari en la mesura que les dades poden servir per a una visió o per a una altra. Aquest és l'interès principal que en bona part tenen els materials que recull Lluís de Yzaguirre. Seria un exemple d'anàlisi de la part del corpus que jo anomenava territorial o geogràfic. Tant se val l'adjectiu. Estudia variació geogràfica.

Ja he dit que en aquests moments estem treballant al voltant del sistema pronominal. Era un estudi concret que ja provenia d'anys anteriors i que es continua. Són recerques que no tenen una gran ambició explicativa. Són més aviat descriptives. No són recerques d'un gran nivell teòric però segueixen temes que considero rellevants perquè el canvi del català, tal com assenyalava aquest matí, jo crec que és molt fort.

El fet mateix que hi hagi molts nous catalanoparlants, en el sentit que hi ha molta gent que parla català que no el té com a primera llengua, fa que hi hagi uns canvis accelerats. Aquesta és una hipòtesi meva, evidentment, vull dir que no tinc dades empíriques per assenyalar fins a quin punt és accelerat o no.

Tota la part més espontània del corpus: el corpus de conversa, el corpus de registres i el corpus de varietats socials, proporciona dades extremadament riques. En aquests tres subcorpus s'està treballant l'àrea del contacte de llengües, l'estudi d'alternances de codi, l'estudi de manlleus i interferències, un camp en el qual hauríem de treballar conjuntament

amb certs projectes que s'estan elaborant, i no sé fins a quin punt ja s'han posat en marxa, sobre el castellà de la mateixa zona. Negar que se'n parla seria negar una realitat. És tan sols un suggeriment.

També hi ha tota l'àrea de pragmàtica i de discurs: actes de parla, dixi, marcadors discursius, estructura de conversa, és a dir, tot el camp interactiu, els processos comunicatius...

Després hi hauria tot el camp de la variació social, que és el camp específic del corpus de varietat sòcio-cultural, descripció de varietats generacionals... Aquest seria el tema del canvi lingüístic. És un punt en el qual jo hi estic particularment interessat. Cal veure i mesurar fins a quin punt el català canvia. No podem seguir dient coses que són només pures repeticions del que creiem que ha de ser el català. En posaré un exemple i quedarà clarissim. Tinc aquí la cita d'una entrevista que el diari *Avui* publicava el diumenge 5 de desembre, en què entrevistaven un dels responsables d'un llibre que ha estat un best seller al llarg de la història. Es tracta d'una Bíblia catalana interconfessional, de protestants i catòlics. I el periodista plantejava a un dels responsables com es resolia la pluralitat dialectal. La Bíblia és un text típic pel que fa al plantejament dels problemes de l'estàndard, de tota la vida, és ben conegut. Algú recordarà potser aquesta entrevista. I un dels responsables responia: "Hem volgut fer una Bíblia que no donés preferència a cap àrea dialectal." Ho puc il·lustrar amb un exemple: A Barcelona el llibre es posa sobre la taula, però a moltes àrees catalanoparlants el llibre és damunt la taula. El que respon està dient que la Bíblia s'ha fet en català estàndard.

Tornant al projecte, finalment hi ha el camp de les variacions contextuals de normativització i estàndard, que també és un indici del canvi que sofreix la llengua catalana, de fins a quin punt la gent té inseguretat lingüística, de quines ideologies lingüístiques té... En aquest tema el Departament segueix la línia de les jornades de llengua normativa que són precedents, si voleu, d'aquest col·loqui.

Lluís Payrató:

Gràcies als ponents per cenyir-se al temps acordat. Tanquem aquesta fase de la taula rodona i obrim la següent fase que havia anunciat abans. És un torn tancat de preguntes als participants. A diferència de la fase que hem tancat ara no hi ha cap ordre fix

d'intervencions i participació, de manera que cadascú pot prendre la paraula quan ho cregui oportú.

He anunciat abans tres preguntes que volen conduir el desenvolupament de la taula. La primera és la següent, tal com està formulada: ¿És veritat que en lingüística i en les ciències del llenguatge en general s'ha passat d'una posició predominantment contrària als corpus a una de predominantment favorable? ¿Estem assistint al naixement d'una lingüística de corpus?

El tractament que els ponents faran de cada tema determinarà la importància que té. Hi ha un temps màxim previst per a cadascun dels temes, però això ja el calcularem a mesura que es vagi desenvolupant la taula. De manera que s'obre el torn per als ponents.

Joaquim Llisterri

Per començar el debat, puc dir que jo no crec que estiguem assistint al naixement d'una lingüística de corpus, perquè a mi em sembla que la lingüística de corpus, en la tradició de Nimega, Lancaster, Birmingham, etc., és una línia absolutament consolidada. Jo crec que potser estem assistint al naixement d'una *altra* lingüística de corpus.

Una mica en la línia del que deia aquest matí, el que sempre s'ha anomenat lingüística de corpus era una cosa molt de lingüistes, basada en mètodes heurístics i en l'extracció del coneixement de corpus, partint de regles, etc.

Jo crec que ara estem assistint al sorgiment d'una altra lingüística de corpus en la qual es contempla el seu processament estadístic per treure els models de llenguatge que es necessiten per a aplicacions tecnològiques. Jo crec que de lingüística de corpus clàssica n'hi ha, però que n'està sorgint una altra...

M. Teresa Cabré

No, no t'ho discutiré pas. Però em sembla que quan Lluís Payrató formulava aquesta pregunta tancada als membres de la mesa considerava que tot el que s'ha fet en corpus durant aquesta segona meitat del segle vint formava part d'aquesta nova lingüística de corpus.

Però és cert, i jo penso que tu has introduït un element molt interessant, que dintre precisament de la lingüística de corpus ja hi comença a haver unes referències que són clàssiques. Per tant no estem assistint en aquest moment al naixement d'una lingüística de

corpus, sinó que ja hi ha tot un pòsit que es pot considerar estàndard, de referència i, per tant, molt clàssic.

A mi m'agradaria destacar de totes maneres que, si s'ha afavorit la constitució de grans corpus al segle vint, això no ha estat pas gratuït, sinó perquè han canviat els recursos de què podíem disposar per fer-ho. I ha canviat una mica la visió de les coses. A mi em sembla que els dialectòlegs tradicionals ja en feien, grosso modo i entre cometes, de lingüística de corpus, si ho entenem en el nivell dels conceptes més bàsics d'allò que podria ser la construcció d'un model a partir simplement de la intuïció, sense necessitat de dades de referència i en canvi amb la necessitat de recollir elements que provenen de l'expressió real. D'això se n'havia fet tota la vida. El que passa és que no n'havíem dit lingüística de corpus, i a més a més ho fèiem d'una manera absolutament parcial, molt localitzada en unes determinades ambientacions i a les explotacions que se'n feien posteriorment. És el cas de la dialectologia, que es treballava més des de la perspectiva diacrònica o històrica que no pas des d'una descripció sistemàtica de les dades.

Per tant, deixaria de banda que hi ha una prehistòria i em sembla que convindria parlar de les noves tecnologies tant de la informació com de la documentació, i del nostre canvi de tarannà respecte a aquestes coses. Per exemple, em sembla fonamental el fet que valorem d'una manera important en aquest moment no només la informació sinó la manera com la tenim estructurada. El fet del naixement dels ordinadors i les futures generacions d'aquests en el fons ha condicionat la possibilitat de constituir un tipus de corpus o uns altres, i també l'aparició d'aquestes noves tecnologies d'explotació dels materials d'un corpus. És un element que hem de tenir en compte per explicar-nos perquè ha explotat d'una forma tan fulminant i espectacular el que avui anomenem la lingüística de corpus.

Lluís Payrató

Hi ha alguna altra intervenció pel que fa a aquesta primera pregunta?

Joaquim Rafel:

De fet volia intervenir en una línia molt semblant a la de M. Teresa Cabré.

És clar que aquesta pregunta es pot entendre de maneres molt diverses. És evident que s'ha afavorit la creació de corpus, com acaba de dir la meva col·lega fa un minut, però és

precisament per les possibilitats que permet la tecnologia, per una banda, i per una determinada orientació no tant de la lingüística, sinó dels polítics, que han afavorit una determinada línia de treball. Sense uns recursos importants que han de provenir de l'administració, no es podria progressar. És obvi en tot cas el canvi de mentalitat, que el veig més en la mentalitat de l'administració, que, a partir d'un moment determinat, ha afavorit la creació de corpus.

De corpus, en aquest sentit, se n'han creat sempre, el que passa és que han estat escassos perquè eren el resultat de l'esforç ingent, si voleu, i de la dedicació de personal i de recursos molt importants sense l'ajut de la tecnologia de la moderna informàtica.

Ara em venia al cap el Thesaurus Lingua Latinae. No conec prou bé aquest projecte, però vaig poder-lo conèixer una mica quan fa un parell d'anys vàrem tenir aquí una de les seves responsables. És un corpus que inclou tota la literatura clàssica llatina i bona part de la literatura llatina medieval. I està fet tot manualment. Es va projectar a finals del segle passat i aleshores ja es va programar que l'elaboració d'aquest Thesaurus duraria cent anys, i encara ara se segueix escrupolosament el pla de treball. Em vaig quedar parat quan m'ho van explicar. Això és un exemple d'una planificació que ha resistit el pas dels anys. La planificació és un dels aspectes que jo he retret abans i no sé fins a quin punt entra dintre d'aquestes preguntes. Però aquest exemple, que no conec des de dintre i del qual no puc donar detalls, quan el vaig sentir exposar em va colpir vivament.

José Manuel Blecua

Yo quería añadir un aspecto. Normalmente, los profesores de lengua o filología sólo vemos la lengua desde nuestra perspectiva. En estos momentos, en la Comunidad Europea, trabaja más gente con un ordenador que en la agricultura. Nosotros tenemos que ver que este tipo de trabajos que realizamos sitúan la investigación de aplicaciones ofimáticas por ejemplo a un nivel precompetitivo y permiten que en el futuro se desarrollen una serie de aplicaciones que hacen el trabajo más racional, que permiten una documentación más eficiente, que permiten una competitividad, es decir, que no es una investigación solamente inmanente, sino que es absolutamente trascendente para facilitar la vida del ciudadano.

Todos estos trabajos de tipo probabilístico que se realizan con grandes corpus en estos momentos tienen unas aplicaciones inmediatas: los correctores estilísticos, muchas

aplicaciones de reconocimiento de habla, que citaba Joaquín esta mañana... Todo esto va a pasar muy pronto al uso diario de la oficina, de la casa, e incluso de los críos en la escuela, y lo que yo quisiera que vieran ustedes hoy es que debajo de esto por primera vez los profesores de lengua, en nuestras investigaciones, no en nuestras clases. nos podemos sentir un poco útiles con la sociedad que nos paga.

Emili Boix

Responent a la pregunta plantejada per Lluís Payrató sobre fins a quin punt hi ha o no una lingüística de corpus, des de la meva perspectiva, com a persona que treballa en aquesta lingüística impura que he anomenat abans, tinc un cert escepticisme. Si lingüística de corpus vol dir una lingüística amb una concepció teòrica, jo no crec que hi hagi una concepció teòrica fonamental. En realitat hi ha una nova eina que ha generat tots uns canvis que no cal ser sociòleg per veure. Hi ha una maquinària molt gran, amb unes possibilitats que obliga a treballar amb grans dades, dades que permeten confrontar les hipòtesis d'una manera molt més fàcil, amb dades molt més vàlides i representatives. Però des d'un punt de vista teòric, no crec que hi sigui.

M'agradaria que algú dels que hi treballa més em pogués dir fins a quin punt hi ha unes convergències teòriques. Jo crec que n'hi podria haver com a resultat del treball en equip al qual obliga el treball amb grans infrastructures. És un canvi que pot facilitar que entre gent que treballa amb models molt diferents hi hagi certes convergències. Potser no faig una interpretació teòrica a gran nivell, però no sé veure en aquests moments que el que s'anomena lingüística de corpus existeixi més enllà d'aquesta cohesió.

M. Teresa Cabré

És com tu dius. Quan constituïm un corpus el que fem, simplement, és conjuntar, arreglar i posar en comú tota una sèrie de dades tractades d'una manera determinada, en suport magnètic, perquè després les puguem explotar d'una manera multidimensional, com ja ha assenyalat Joaquim Rafel. La multifuncionalitat dels corpus és un tema que hauríem de reprendre després. Una cosa és la multifuncionalitat dels corpus i l'altra són els projectes, més teòrics de la lingüística computacional. Per tant hauríem de fer d'entrada aquesta distinció.

De totes maneres, m'agradaria deixar sobre la taula la meva opinió sobre el fet que quan nosaltres parlem de lingüística de corpus ho fem dintre d'una lingüística aplicada, per tant en una perspectiva filosòfica de recull d'unes unitats que no són mai les unitats reals. A mi m'ha agradat molt la distinció que ha fet el Dr. Blecua aquest matí. Quan nosaltres construïm un corpus no podem pensar que el que estem fent és agafar la realitat i posar-la en una màquina, no és veritat. Estem creant un altre objecte que és un objecte abstracte i per tant un objecte que té a veure amb la realitat. Però és una representació determinada de la realitat, amb codis o sense codis. Per tant estem construint un objecte de recerca que no es correspon exactament a la realitat. És, en el fons, com la relació entre el mapa i el territori, que ja es va fer en la lingüística clàssica. Però sobretot el que voldria dir és que aquesta perspectiva de la constitució de corpus en la línia de la lingüística aplicada em sembla que no està pas en contradicció amb cap mena de perspectiva de la lingüística teòrica. Jo penso que hem d'anar aprenent a no fer aquestes distincions tan taxatives entre el que és teòric i el que és aplicat, perquè, fet i fet, si una perspectiva no es nodreix de l'altra difícilment podrem progressar en matèria de llenguatge. Penso que és molt vàlid fer modalització del coneixement lingüístic i partir purament d'intuïcions, això s'ha de continuar conservant, però en canvi el fet de poder disposar de les dades dintre d'un corpus ens permet, d'una banda, afinar molt més les nostres hipòtesis d'entrada, i, de l'altra, comprovar determinades afirmacions i fins i tot trobar aquells contraexemples que només amb la intuïció de vegades ens costa molt més de trobar. Per tant, no veig que hi hagi una contradicció profunda entre lingüística de corpus aplicada i lingüística teòrica.

D'una altra banda, en el terreny de la multidimensionalitat, en aquest fons clàssic de què ha parlat Joaquim Llisterri, cal que no oblidem que projectes com el de Birmingham o el Cobuild ja han estat summament explotats, amb productes de tipus comercial com són diccionaris, gramàtiques, gramàtiques d'aprenentatge i fins i tot eines lingüístiques de descripció, etc. És a dir, que ja comencem a tenir models fins i tot comercialitzats que provenen d'aquesta creació de recursos clàssics.

Lluís Payrató

Molt bé, ja hem exhaurit el temps que estava previst per a aquest tema. A més a més, la intervenció de M. Teresa Cabré enllaça, em sembla que d'una manera clara, amb la pròxima pregunta, que és, tal com està formulada, la següent: ¿Fins a quin punt és

determinant per al disseny d'un corpus la recerca i les aplicacions que s'han projectat de dur a terme? ¿Fins a quin punt és possible un corpus més aviat neutre o obert que no prejutgi unes recerques i unes aplicacions exclusives?

Torna a haver-hi un torn obert.

Joaquim Rafel

Crec que se'ns ha fet una pregunta amb dues part diferenciades, l'una no implica l'altra. La primera part, sobre fins a quin punt són determinants per al disseny d'un corpus la recerca i les aplicacions que s'han projectat de dur a terme, és evident que a l'hora de dissenyar un corpus les aplicacions s'han de tenir en compte. S'ha de fer el disseny d'una manera molt afinada perquè aquestes aplicacions, que en principi són l'objectiu principal d'aquest corpus, es puguin realitzar. Si el disseny no es fa d'una manera adequada, després et pots trobar que no pots treure els resultats que et pensaves que trauries. Això no implica que un disseny molt adequat per a unes aplicacions concretes converteixi aquestes aplicacions en exclusives.

Com sembla que en la segona manera de formular la pregunta s'indicava, ¿com pot ser adequat un corpus per a un tipus d'aplicacions i a la vegada tenir aquelles condicions que fa un moment posava com a gairebé necessàries, sobretot per a un corpus que costi un esforç important per constituir-lo, sigui esforç humà, econòmic... ? Parlo de corpus en el sentit d'un corpus relativament important. Ho dic perquè en aquests moments s'està utilitzant el nom de corpus per allò que abans se'n deia recollida de materials, per a un treball que volia dir, suposem, recollir unes tres-centes frases. D'això ara també se'n diu corpus. No em refereixo a aquests corpus, que, si només serveixen per extreure un tipus d'aplicació no passa res. Em refereixo a un corpus que ha costat anys de treball, equips de persones que han hagut de treballar molt per trobar solucions a tots aquells problemes del corpus. Si un corpus realitzat aplicant tot el coneixement i tot el saber d'aquests equips d'experts, i després d'utilitzar uns recursos econòmics importants, resulta que només serveix per a l'aplicació concreta que s'ha pensat, crec que això no és admissible. Encara més perquè es pot fer de manera que sigui obert, neutre, però que serveixi per a l'aplicació que s'havia previst i per a moltes altres. Ja s'han posat aquí exemples de casos en què això es produeix. L'aplicació que s'ha projectat per a un corpus no hauria de restringir les possibilitats d'aplicació del corpus. Aquesta és la meva opinió.

Joaquim Llisterri

M'agradaria començar reintroduint una mica la noció d'aplicacions comercials que surten dels corpus, amb les quals ha acabat la seva intervenció la Dra. Cabré. Jo intentava repassar per exemple la llista de corpus orals que actualment distribueix el Linguistic Data Consortium, que és el que tenim més a mà, i amb el que qualsevol persona pot començar a treballar. El primer que distribueix són dos mil cinc-cents dígits obtinguts per Texas Instruments Són sis mil tres-centes frases aïllades, fonèticament equilibrades. etc. Distribueix corpus de quinze mil frases extretes de converses gravades en una oficina de distribució de recursos navals de l'exèrcit americà. Distribueix l'ATIS, per exemple. que són dotze mil frases extretes de diàlegs imitant la interacció amb un operador mecànic que fa reserves de bitllets, que és el que hauria de substituir l'humà quan aquest sistema estigués en marxa. El Wall Street Journal Corpus, que és un corpus de lectura d'articles de diari, fa un total de trenta-vuit mil frases.

Si tota aquesta mena de corpus, que és el que ara hi ha en CD-ROM, com a mínim oral, i és el que es pot trobar, els donem a un lingüista i, comptant que són orals. a un fonetista, segur que en traurà coses, però no sé si en traurà gaires. Amb això vull dir que jo trauria moltes més coses d'un corpus de per exemple dues-centes frases fonèticamnet equilibrades del català, si hagués de fer descripció del català, que no pas d'un corpus de tres-cents diàlegs d'un senyor interactuant en la taquilla d'Ibèria al passeig de Gràcia. Tinc la impressió que dels corpus que s'estan venent, a part d'aquests grans corpus clàssics, el Brown, el Cobuild, etc., la majoria estan aplicats a fer una cosa determinada. Hi ha molts corpus de dígits, per exemple, pel que explicava aquest matí, perquè les companyies telefòniques estan interessades a tenir dígits perquè els interessen telèfons que es puguin marcar sense tenir les mans lliures, per als cotxes, per exemple. Hi ha molts corpus que ja són la cosa més blasfema que un es pot imaginar quan és fonetista. Són corpus gravats amb ambient de soroll d'oficina, amb un senyor teclejant al costat, etc. El primer que ensenyen a un fonetista és que no s'ha de gravar amb soroll al voltant. Però si s'ha de dissenyar l'aplicació d'un reconeixedor de veu, per exemple un sistema per dictar cartes, on hi hagi algú que estigui treballant, per exemple en una oficina real, no es pot entrenar cap reconeixedor amb un corpus netíssim, enregistrat en una sala insonoritzada. Se l'ha d'entrenar amb soroll ambiental. En aquests moments s'està treballant per trobar estratègies per separar el soroll del senyal.

D'altra banda, pel que fa al que he sentit aquest matí, em sembla que gran part del que tenim fins ara es podria fer relativament accessible d'una manera fàcil i ràpida. i no m'estranyaria que hi haguessin sortides comercials. Per exemple, he sentit dir que Word Perfect està desenvolupant un corrector gramatical i un corrector d'estil per al català. Estic segur que als de Word Perfect els interessaria molt tenir els cent dies consecutius que hi ha aquí en un diari per avaluar el seu corrector i suposo que fins i tot estarien disposats a pagar. I de coses d'aquest tipus n'hi ha més. Em sembla també que estem arribant a un moment que per veure si podem o no fer coses amb el que tenim val la pena començar a realitzar algunes proves pilot de distribució del que hi ha a equips diferents. És veritat que quan es fa un corpus es fa en funció del que interessa al grup que el crea, però potser es podria provar de començar a distribuir-lo a altres grups per veure què més en poden treure.

M. Teresa Cabré

A mi em sembla que el Dr. Llisterri ha introduït un element nou en la nostra discussió, que és l'explotació general de les dades de què ja disposem.

A mi m'agradaria dir que si el Departament de Filologia Catalana de la Universitat de Barcelona es va plantejar la constitució d'un corpus sobre el català contemporani va ser precisament perquè ens trobem en un moment en què fem gramàtiques catalanes i intentem fer diccionaris i no disposem dels materials que ens permetrien fer-los d'una altra manera.

És cert que a l'Institut d'Estudis Catalans existeix ja en curs d'elaboració el corpus textual de la llengua catalana. Però és un corpus encara no disponible. I ens trobem que nosaltres continuem els nostres articles, les recerques en tesis doctorals, i no tenim les dades que ens permetin dir que un determinat tipus de català és el català que ara es parla, perquè això s'ha convertit en una mena d'estereotip. Sempre anem traient els mateixos exemples, fins i tot s'ha convertit en un eslògan, i enlloc està codificat el català que ara es parla i s'escriu. Perquè és clar que una cosa és l'estàndard i l'altra el català real que escriuen diferents escrivents, en diferents situacions.... Tot això no està recollit i aquest és el mòbil que va fer demanar al Departament recursos per poder constituir un corpus del català contemporani. L'objectiu va ser precisament disposar d'unes dades al més

ràpidament possible perquè poguéssim fer descripcions de la llengua catalana actual i real, perquè no en disposàvem.

Jo penso que per a cada llengua hi ha sempre un gran corpus de referència i després hi ha diversos corpus que permeten obtenir dades, ja que són corpus fiables, sistemàtics, útils, valuosos, etc., però no són el corpus de referència d'una llengua determinada.

És cert que la majoria de corpus s'han de constituir amb fons públics perquè, sobretot la infrastructura, costa molt. Requereix una quantitat important de recursos humans que normalment s'han de pagar perquè, encara que la Universitat de Barcelona disposa de molts estudiants que s'engresquen fàcilment amb un projecte d'aquest tipus, és clar que en el moment que es posa un becari dedicat unes hores determinades exclusivament a un projecte se li ha de concedir un ajut. Tot això ve dels fons públics, no ens enganyem. Així doncs, potser comença a ser hora que ens plantegem el problema que és a la clau de la constitució d'aquests materials. Independentment dels detalls i els obstacles tècnics, realment podem fer rendibles les dades que estem constituint i recollint? Què ens caldria per fer-ho? Què ens caldria per fer-les accessibles, compatibles, etc.? Ens cal enfocar les coses des d'aquest punt de vista, i no pas començant a estudiar els problemes de codificació. Primer hem de fer-nos la pregunta clau, que és si estem disposats a repartir i compartir recursos. Si no hi estem disposats, aleshores ja no cal que ens plantegem res més, perquè estem perdent el temps, això és una discussió de cafè.

Joaquim Rafel

En el tema sobre fins a quin punt és determinant el disseny, des de la meva perspectiva és urgent saber quin nombre de dades necessitem per a la configuració d'un corpus. Joaquim Llisterri ja ho ha citat aquest matí. Un corpus més gran no vol dir que sigui millor, sinó que més gran sovint vol dir pitjor, perquè no és realitzable. No es pot fer per raons pragmàtiques, en el sentit pressupostari no pots anar més enllà. És una pregunta que ens fem els que estudiem aquesta part de corpus espontani. El problema de la representitivitat no el tenim ben clar. Potser falta un treball interdisciplinari amb sociòlegs, que crec que mostren molt interès per aquesta vessant. Potser hi ha problemes de relació entre disciplines. Aquest matí he citat el British National Corpus, que és un gran projecte esponsoritzat que té el suport governamental de la corona. Fins ara nosaltres no ho tenim això en els nostres projectes. Aquest mateix projecte per recollir dades sobre

l'anglès que ara es parla recull cent informants, però sobre un estudi sociològic seriós. I aquestes dades, que anomenen el corpus demogràfic, són set hores per a cada un d'aquests cent informants, set hores de la seva vida, que vol dir que recull dades de mil informants. És un estudi que serveix de base en els materials lexicogràfics. No ens enganyem. Com sempre anem veient que al final l'estímul és crematístic, com moltes coses. Al darrere d'aquest corpus hi ha l'Oxford University Press i la Longman University Press. A mi em seria igual que aquí hi hagués al darrere l'Enciclopèdia Catalana, que deu tenir menys diners que l'Oxford, per descomptat.

No sabem per al tema de la representitivitat quant material necessitem perquè sigui suficient. A Gran Bretanya amb cent informants en tenen prou. Nosaltres no ho sabem, és un tema que s'ha d'anar estudiant. Estem aquí i s'ha de reconèixer.

José Manuel Blecua

Yo querría hacer una meditación en voz alta. Parece que los datos son inocentes. Creo que como tantas veces los datos en lingüística no son inocentes. Creer en la inocencia de los datos es sumamente peligroso, primer punto. Segundo, creo que pasa con esto de los corpus como pasa en el realismo, en la historia de la literatura: uno toma una parte de la realidad y la refleja, cómo en *Fortunata y Jacinta*, el pueblo bajo y la burguesía. Pero sólo refleja una parte de la realidad que necesitamos. Vuelvo a insistir en lo que citaba antes María Teresa Cabré, nosotros sólo construimos modelos, nunca llevamos la realidad al ordenador, gracias a Dios. Y puesto que construimos modelos, lo importante, como decía el Dr. Rafel, es que ese modelo no sólo sirva para uno, sino que sirva para varios. Es aquí donde se plantea uno de los grandes problemas, porque finalidad y diseño de corpus están muy unidos, y es muy difícil a veces separarlo.

Otra cosa es la posibilidad de volver a utilizar un corpus diseñado para una cosa para otras cuestiones, para otras investigaciones. Estas tres meditaciones son las que quería hacer. Pero sobretodo quiero insistir en la primera. No piensen, sobre todo los alumnos, que los datos son inocentes, que no lo son en modo alguno.

Lluís Payrató

Tanquem aquesta segona pregunta i saltem a la tercera i última de les preguntes previstes: ¿De quina manera la presentació material dels corpus i les vies per accedir-hi limiten o faciliten les recerques i aplicacions que se'n poden derivar?

M. Teresa Cabré

Abans s'ha fet referència a les línies prioritàries de finançament de la recerca per als pròxims anys de la Direcció General Tretze de la Comunitat Europea, que jo trobo que ha plantejat una perspectiva que em sembla bastant esperançadora. Així com abans es creia fonamentalment en grans programes que feien tractament de llenguatges, ara d'una forma molt més realista ea va plantejar que les línies prioritàries serien la creació de recursos en suport informàtic que fossin sobretot en un format compatible, el màxim d'estàndards possibles. Aquesta seria una de les línies prioritàries de recerca.

Un dels temes que es planteja a nivell europeu per a l'estructura o emmagatzematge dels corpus és el fet que estiguin etiquetats, és això que anomenem *tagging*, com aconseguir que els ordinadors d'alguna manera llegeixin els textos amb els materials implícits, amb els mateixos coneixements que nosaltres tenim sobre les dades del llenguatge, les dades que nosaltres com a lectors d'un text, pel fet de tenir tants coneixements implícits, ja reconeixem, i ho fem d'una manera absolutament natural. A ningú se li acudirà quan vegi una expressió com *ull de poll* de pensar que són tres coses diferents, perquè no se li acudiran tres idees diferents, sinó que directament quan llegeixi el text només representarà una sola idea.

Tot aquest programa europeu de l'etiquetatge, que suposo que el Dr. Llisterri i el Dr. Blecua coneixen bastant més que nosaltres, em sembla que és un element fonamental perquè després les dades dels corpus puguin ser explotades des de molts punts de vista diferents, i no només des d'una lectura seqüencial.

Lluís Payrató

Alguna altra intervenció?

Joaquim Llisterri

Per completar aquesta informació, de fet dins del grup de corpus textuals d'EAGLES, hi ha un subgrup treballant en anotació lingüística, i aquest grup està intentant partir dels resultats de NERC sobre el que és l'etiquetatge. Sembla que a NERC van trobar una llista raonable de parts de l'oració que funcionava per etiquetar corpus en les nou llengües europees del projecte. I aquesta és la que més o menys EAGLES té intenció de recomanar. Sembla que pensen donar algunes guies sobre *parsings*, sobre anàlisi sintàctica. No va més enllà perquè naturalment ningú avui en dia no gosa donar cap mena de guia semàntica o sobre notació pragmàtica. Sembla que l'etiquetatge anirà per aquí.

Però el que sí que em sembla que és cert, i cap aquí volia anar jo, és que en el proper o el quart programa marc de la comunitat, dins del programa d'enginyeria lingüística, hi haurà una gran part dedicada al desenvolupament de recursos lingüístics. Em sembla que és un programa que val la pena mirar-se i seguir. Jo crec, pel que he vist avui i per altres projectes que hi ha pel país, que estem en condicions de començar a pensar d'anar junts a presentar propostes a aquesta mena de programes.

José Manuel Blecua

Continuando con la línea del Dr. Llisterri y como respuesta de lo que preguntaba la Dra. Cabré, el otro día por casualidad me tocó repasar cómo está la situación de las etiquetas morfológicas en EAGLES. Como decía muy bien Joaquín Llisterri, el proyecto tiene dos fases. En una primera fase se establece un estándar general de grandes categorías, de todos los elementos que son comunes, y después se hace también un estándar para los elementos que no son comunes, que pueden pertenecer a una sola lengua, pero de manera que esto tenga una etiqueta estandarizada. Es un producto complicado. Por ejemplo, es el problema del neutro. Hay lenguas que tienen neutro y hay lenguas que no tienen neutro marcado morfológicamente, lenguas que tienen dual y lenguas que no tienen dual. Es un sistema que es capaz de cubrir las grandes divisiones morfológicas de las nueve lenguas sin entrar en grandes profundidades de manera que las haga intercambiables, facilite la traducción automática, la explotación, la importación, el emparejamiento, etc. Porque después se piensa que en cada lengua se trabaje sobre eso más profundamente, y cada investigador, después, más profundamente todavía.

Lluís Payrató

Alguna altra intervenció?

Tanquem aleshores tota aquesta part i, simplement abans de passar al torn obert per a tothom, em sembla que cal donar la possibilitat a qualsevol dels ponents que introdueixi algun tema no tractat fins ara. Si ningú no creu pertinent fer-ho passarem directament al torn obert per als assistents.

M. Teresa Cabré

Voldria dir una cosa, només aprofitant que em dónes la oportunitat de fer-ho, perquè me n'he descuidat abans i segurament després no sortirà perquè hi haurà un debat.

Aquest problema que hem plantejat aquí sobre la compatibilitat dels corpus que estàvem constituint es va plantejar a la trobada de departaments de filologia catalana, que es va fer a València en ocasió dels premis Octubre. Tots els departaments van valorar aquest problema d'una forma bastant important i es va suggerir de fer una trobada que seria una jornada de treball de les persones que són responsables o treballen en corpus ja constituïts, en corpus en via de constitució i en corpus que s'estan dissenyant i que es pensen desenvolupar. Per si a algú li interessa, perquè realment té pensat d'establir un corpus sobre llengua catalana (aquest matí un professor de la Universitat de Tarragona ha dit que estaven començant a constituir un corpus que no era a cap directori), farem una jornada que es titularà exactament *Compatibilitat i accessibilitat dels corpus*, que estarà patrocinada per la CIRIT. L'objectiu fonamental serà plantejar-nos totes aquestes qüestions tècniques, veure si les dades poden ser realment transferibles d'un sistema a l'altre, compatibles, compartides, etc. Simplement ho dic perquè si algú està interessat perquè se'l convoqui, que ens ho faci saber i el posarem al directori. Gràcies.

Lluís Payrató

Alguna altra intervenció?

Joaquim Rafel

En la mateixa línia de propostes de futur i de coordinació que fa la M. Teresa Cabré, la meva intervenció és per fer una sol·licitud d'informació sobre un tema que aquest matí hem plantejat quan un periodista ha vingut a interrogar-nos, a M. Teresa Cabré i a mi per

a un diari barceloní, per tal d'informar al gran públic sobre el que estem fent en aquesta sessió. Bé, la pregunta és per saber fins a quin punt tota aquesta promoció de les indústries de la llengua per part d'aquesta Comissió Tretze té en compte més llengües que les nou. Aquest és un tema polític. Fins a quin punt a les indústries de la llengua s'hi té en compte el català? Això ho pregunto als qui ho coneixeu més; com per exemple a Joaquim Llisterri, que hi té més relació. Perquè en el fons sabem que si no hi ha diners seguirem fent cadascú la seva recerca de barraqueta. Si algú de la sala em pogués respondre aquesta pregunta... Em temo que la resposta possiblement sigui negativa.

Joaquim Llisterri

Com que ha sortit el meu nom, contesto, però ni de lluny sé de què parlo. Tinc la impressió que la situació oficial a la majoria de projectes és que es treballa amb les nou llengües, entre les quals no hi ha el català. Però a mi em consta que a moltes universitats europees hi ha un interès molt gran per al corpus de l'Institut, que és un corpus que a molta gent li agradaria veure difós, encara que fos parcialment, o els agradaria veure'l vinculat a projectes europeus de constitució de corpus o de lexicografia computacional. Em consta que existeix aquest interès, i segurament al Dr. Rafel li consta molt més que a mi. Amb això vull dir que encara que el català no sigui una de les nou llengües oficials de treball, és clar que hi ha interès a tenir el català en certs projectes. El finançament no sempre és obvi, però, en tot cas, per la feina que es fa sobre el català, hi ha molt d'interès a molts altres llocs fora d'aquí.

José Manuel Blecua

Yo recuerdo haber asistido a alguna reunión en Luxemburgo con Juan Acordagoicoechea, precisamente, y con algún representante de la Generalitat en la que se planteo en qué número de lenguas se podía investigar, que eran tres o cuatro. La solución consistió en no poner un número de lenguas concreto sino en decir en las lenguas europeas. Yo creo que eso quedo muy claro en la reunión de Luxemburgo.

Joaquim Llisterri

La gràcia de tota aquesta mena de projectes sempre és el requisit del multilingüisme. Jo no crec que fos cap handicap un projecte en el qual s'incorporès el català juntament

amb d'altres llengües, amb uns treballs sòlids i importants com qualsevol dels dos que ens han presentat aquest matí, però parlo a títol personal, d'intuïció i pel que em sembla haver escoltat.

Joaquim Rafel

L'únic que conec són els textos que ens van fer arribar a través de Lluís de Yzaguirre, els companys de feina que hi van assistir. I realment recordo una pàgina d'un comissari dels responsables, que és una pàgina antològica des del punt de vista dels catalanoparlants ja que defensa la diversitat lingüística enfront d'una llengua que no esmenta, però que és l'anglès. Evidentment és el que porta la capdavantera en molts temes i, evidentment, en la indústria de la llengua, també. En aquesta pàgina defensa la diversitat i després quan ho concreta només parla de les nou llengües. Aquest és el problema. Jo solament recordo aquesta pàgina. El tema de fons és evidentment un tema polític. No estic demanant a Joaquim Llisterri que representi les autoritats europees. Simplement ho volia plantejar públicament.

Lluís Payrató

Em sembla que és el moment d'obrir el torn obert definitiu. Abans, però, voldria fer uns quants aclariments.

Primer de tot, suposo que l'ideal per al desenvolupament d'una taula rodona i en concret pel que fa a les preguntes dels assistents, és que els temes flueixin de manera coherent. Però això no es pot preveure si no es donen les preguntes per escrit amb anterioritat a la taula. Hem desestimat aquest sistema perquè hem preferit l'espontaneïtat de fer les preguntes directament.

La segona qüestió és que abans de fer cada pregunta el parlant ha d'identificar-se. En tercer lloc, demano que les preguntes o intervencions siguin breus. Per últim, recordo que les intervencions s'enregistraran i es publicaran. Això és un advertiment i no pas una amenaça, no és perquè ningú s'ho pensi tres vegades abans d'intervenir, és simplement perquè sàpiga que després sortirà publicat.

Doncs bé, obrim aquest torn. Es pot fer una pregunta adreçada a un membre en concret de la mesa o al conjunt.

Josep Moran (Universitat de Barcelona)

Em referiré a això que abans havia apuntat l'amic i company Emili Boix sobre la relació entre la teoria i els corpus lingüístics. Suposo que els corpus lingüístics són un material de base, però jo crec que a llarg termini les possibilitats que la tècnica ofereix influiran sobre la teoria. La teoria no neix sense relació amb les possibilitats de treball que hi ha. Avui dia jo crec que, en primer lloc, el corpus permet una informació de base que supera alguns aspectes intuïtius amb què fins ara alguns lingüistes havien treballat. Ho dic no solament pensant en els corpus sincrònics sinó també en els diacrònics. A partir d'ara per saber què es deia en la llengua antiga s'han de buidar textos, i no valdrà dir només "la llengua antiga deia..., o "la llengua antiga deixa de dir..." Aleshores, s'haurà de muntar la teoria partint d'aquesta realitat àmplia i prou representativa. Jo crec que a llarg termini, d'una manera o altra, la gran possibililtat de disposar d'informació codificada ha d'influir evidentment sobre la teoria. Avui dia alguns plantejaments excessivament intuïtius o racionalistes, que en el fons crec que vénen a ser el mateix, s'hauran de replantejar davant de les possibilitats actuals de disposar de corpus grans que ens reflecteixen millor la realitat.

Emili Boix

De fet aquest és un debat etern. Quan jo he fet aquesta invocació de la necessitat de teoria era un crit d'alerta perquè dins la tradició catalana domina una tradició positivista en bona part. I el positivisme té aquest problema. El professor Blecua també ho plantejava. No hi pot haver una resposta definitiva. És eclèctica. Hi ha un vaivé continu entre les induccions i les deduccions, que ara és molt possible gràcies a aquestes dades disponibles i accessibles. Si hi ha programes que facin més ràpid l'accés a l'usuari és molt més fàcil aquest continu moviment de llançadora entre les dades i la teoria. Jo no sé anar molt més enllà d'això. Tens raó. L'únic de què estic convençut és que solament amb la descripció de les dades no s'anirà cap a teories excel·lents. Hi ha algun autor que defensa que quan hi ha moltes dades això fa que aparegui un teoria. No hi tendeixo, sóc una mica reticent al positivisme. Això és tot el que respondria. No és un refús, evidentment, perquè no puc refusar aquests plantejaments.

Joaquim Viaplana (Universitat de Barcelona)

El Dr. Rafel ha fet esment a una necessitat que em sembla, en la mesura del que sigui possible, que és digne de tenir en compte, la necessitat que els corpus siguin capaços de donar informació de més aspectes d'aquells per als quals han estat decididament dissenyats. Això té unes limitacions que el Dr. Blecua precisament ens manifestava. Però en la mesura que un corpus pugui facilitar informació a aquell que està interessat a aconseguir-ne, sembla que aquesta informació l'hauria de poder facilitar des del principi de la seva elaboració. Si aquests corpus estan, com és el cas de la majoria, subvencionats amb fons públics, sembla que ha de ser des del primer moment que han d'estar a disposició d'aquells que els puguin explotar des de la perspectiva que els pugui interessar. En la mesura que els corpus en qüestió siguin capaços de subministrar-los aquesta informació, no em semblaria bé, amb franquesa, i tenint en compte com resulta de llarga l'elaboració d'un corpus, que s'hagués de posposar aquesta informació, a qui hi pugui estar interessat, a la finalització del corpus. Les dades utilitzables, en teoria, almenys des del meu punt de vista, haurien d'estar des d'un primer moment a disposició d'aquells a qui poguessin interessar. Quan dic això vull dir també que els corpus que estan en fase incipient s'haurien de presentar de tal manera que aquesta informació no fos difícil d'obtenir. I en el cas d'aquells corpus que ja estan avançats, em sembla que seria convenient que en la mesura del que fos possible abans d'acabar poguessin facilitar informació als interessats, fins i tot en el cas que això suposés forçar una mica la màquina. Caldria buscar mecanismes, que potser ara no estan previstos, que facilitessin aquesta informació.

Voldria fer un altre aclariment en relació a la pregunta de si estem en una etapa de la lingüística de corpus. Si es tracta d'entendre la lingüística de corpus com una aportació més a l'àmbit de la lingüística, a mi em semblaria convenient de respondre que sí que estem avançant en la línia d'un tipus de corpus superior, millor, òbviament perquè els recursos ens ho permeten, als corpus amb què fins fa uns anys s'havia treballat. Hi ha una cosa que no hem de perdre de vista. En el treball estrictament de lingüística, els corpus, per molt complets que siguin sempre ens deixaran dades fora i aquestes dades s'hauran de buscar per altres sistemes que els que els corpus ens facilitin. I em sembla que això, avui i demà, serà inevitable. Sempre hi ha unes dades de tipus sistemàtic que fan falta, uns forats que sabem que hi són i que no apareixen en els corpus. Si no és per

la via de l'elicitació, pel procediment que creiem més convenient, seria inútil de pretendre que a través d'uns corpus les poguéssim aconseguir.

Joaquim Llisterri

Jo crec, i no és per contestar sinó per recolzar el que ha dit el Dr. Viaplana, perquè em sembla que és crucial, que el fet que un corpus no s'utilitzi el converteix en un fòssil. Jo entenc que en aquests grans projectes dels corpus de referència el disseny és molt sofisticat i que el corpus en si s'ha d'oferir sencer per a certes aplicacions, però em sembla que algunes parts d'alguns d'aquests grans corpus es poden aprofitar per a moltes coses. Per exemple, si no ho recordo malament, crec que el corpus de Londres de l'anglès de fet forma part d'un gran corpus molt més gran de l'anglès. Pel que he sentit aquest matí, crec que amb la meitat del que hi ha al corpus oral del català de la Universitat de Barcelona, em veuria en cor de començar a fer descripció fonètica. Potser sóc un optimista incorregible però els estudis que he vist de descripció fonètica del català treballen amb una vintena part del material que teniu vosaltres. Crec que amb una quarta part del que hi ha en el text, en tindria prou per exemple per verificar un analitzador sintàctic o un analitzador morfològic. Moltes vegades els que fan els dissenys dels grans corpus es pensen que són unes coses que no seran perfectes i magnífiques per als usuaris fins que no estiguin del tot acabats, però molts usuaris que tenim problemes bastant concrets per resoldre podríem fer servir parts d'aquells corpus.

I a més, em sembla, com deia aquest matí, que amb això de la distribució de certes parts dels corpus en les primeres etapes, es poden començar a veure els problemes reals i autèntics de compatibilitat de formats. Em sembla que tothom que hagi tocat un ordinador sap que una persona es pot passar dues hores mirant manuals i assegurant-se que té formats compatibles i en realitat quan poses el teu disquet a l'ordinador del veí, allò no va: o desapareixen els accents, o passa qualsevol cosa i en teoria sobre el paper són dos formats compatibles i magnífics. Si ens hem d'esperar a veure els vint-i-cinc compacts de no sé quin corpus... Jo m'estimaria més que em deixessin el primer durant una temporada i després els altres ja veuríem com funcionen.

Joaquim Rafel

Potser podria fer alguna aportació en referència a la intervenció del Dr. Viaplana, per una banda, perquè m'ha al·ludit directament i per una altra, per altres aspectes que ha tocat.

Potser diré alguna cosa sobre el plantejament de l'autorització de la utilització dels materials d'un corpus d'unes certes dimensions abans de la seva finalització. Aquest és un tema complex en molts aspectes, sobretot en el cas d'un corpus d'aquesta naturalesa, en principi almenys tal com s'ha pensat el corpus que jo he presentat de l'Institut d'Estudis Catalans, quan estigui elaborat, constituït i acabat. Amb el corpus com a tal i la base de dades que he descrit, està programat que s'emprengui un projecte, al marge de l'elaboració d'un diccionari descriptiu o d'altres aplicacions dintre del mateix Institut, que posi el corpus a disposició del públic en general a partir d'un sistema de consulta interactiu. Això implica que qualsevol persona que es pugui connectar a la base de dades a través d'un centre oficial de recerca, o des de casa seva, si és que té la possibilitat de fer-ho, pugui entrar-hi a través d'uns programes que s'elaboraran per a l'explotació de la base de dades, si els recursos ho permeten. Així es podrà tant fer el curiós, com obtenir dades a partir d'un pla de treball. Deixo de banda la qüestió dels costos que això pugui representar i qui els ha de pagar, perquè parlant d'això ens podríem allargar molta estona. Evidentment això que he explicat no es pot fer fins que el corpus estigui acabat de constituir i s'hagin obtingut ja uns recursos per a poder-lo a disposició del públic, perquè una base de dades no està directament a la disposició d'un usuari, sense uns recursos per a la seva distribució.

Una altra cosa és que durant la constitució d'un corpus es puguin utilitzar els materials que hi ha en aquell moment en la base de dades. Això, evidentment, és possible, i, en el cas del corpus de l'Institut, concretament, no solament s'ha permès, sinó que s'ha dut a terme en més d'un cas concret. El que passa és que això es fa quan hi ha una proposta concreta, i la persona que ho necessita ha de saber-ho explicitar. Així, les obres que s'han de publicar de Salvador Espriu s'han beneficiat dels llistats d'índexs diversos i concordances extretes del *Corpus textual informatitzat de la llengua catalana* de l'Institut. També s'ha facilitat material a algunes empreses que necessitaven informació sobre els mots més freqüents en la llengua escrita, i s'han realitzat o estan en curs de realització algunes tesis a partir dels materials del corpus de l'Institut. Hi ha una empresa nord-

americana que duu a terme un programa que té com a finalitat la lectura i la síntesi de la veu per poder-la posar a la disposició dels invidents, a fi i efecte que puguin posar textos escrits en un escànner i sentir el contingut del text a través d'un altaveu; un dels components d'aquest producte que, ajuda a la correcta interpretació dels textos, és una llista de les paraules més freqüents en la llengua que es treballa, llista que l'Institut també els va facilitar. Poso aquests exemples per fer veure la possible utilització dels materials d'aquesta naturalesa. Es pot fer una col·laboració institucional però s'han de prendre unes precaucions, s'ha de fer algun document que especifiqui la col·laboració, etc., perquè entre altres coses hi ha la qüestió dels drets que ha estat ja al·ludida aquest matí, de la qual ara no ens hem d'ocupar. Jo he de confessar que des d'aquest punt de vista estem despullats perquè no hem pres les precaucions legals per garantir d'una manera absoluta que ningú ens pugui buscar alguna responsabilitat per haver introduït aquestes obres en suport informàtic.

Un altre aspecte que no afecta directament el que jo havia exposat és la referència al fet que els corpus no ho solucionen tot. D'alguna manera hi ha dades que se'ns escapen, és evident, i si jo he defensat els corpus textuals, ho he fet perquè era la meva obligació.

Jo crec, tot amb tot, que hi ha una diferència fonamental entre els corpus textuals exhaustius com els que hem estat comentant avui i el buidatge selectiu de textos que, en el millor dels casos, es feia tradicionalment a l'hora de fer diccionaris, com en el diccionari Alcover-Moll, que depenia exclusivament de la competència i de vegades de l'humor o del cap clar que tenia la persona que subratllava una paraula en un llibre o en un text.

Marta Juncadella (Universitat de Barcelona)

Voldria preguntar al Dr. Rafel si d'alguna manera està previst d'anar actualitzant el corpus textual cronològicament. Ja s'ha fet un tipus de selecció qualitativa, però penso que un corpus d'aquesta mena, tan important, valdria la pena que tingués prevista una actualització cronològica, del 88 ençà, en la mesura que això sigui possible.

Joaquim Rafel

No cal que digui que comparteixo plenament aquesta idea i aquesta inquietud. En aquest moment, si previst vol dir planificat, no ho està, pel fet obvi que encara falta molt per acabar el que hi ha planificat. En aquest moment, molts dels que són aquí saben que

estem en una situació d'*impasse* respecte als recursos per continuar. És una qüestió que els polítics tenen entre mans i mentre tot això no estigui resolt i el corpus no estigui acabat no podem planificar la continuació després de 1988, de 1988 a 1993, i després de 1993 per continuar amb aquesta periodització de cinc anys que hem anat fent fins ara. Per això jo he fet una al·lusió a la continuïtat posterior i a l'assistència des d'una institució com l'Institut d'Estudis Catalans dels investigadors que vulguin anar treballant no només amb corpus fins a 1988 sinó amb un corpus que arribi més enllà. Evidentment això forma part de les previsions inicials, però en aquests moments la previsió concreta no pot anar més enllà del que fins ara hi ha planificat, sobretot tenint en compte que no podem assegurar quan s'acabarà perquè ara per ara no sabem els recursos amb què comptem.

Lluís de Yzaguirre (Universitat de Barcelona)

Voldria ampliar algunes informacions, en primer lloc sobre l'efecte integrador i catalitzador que ha tingut el projecte en el Departament: integrador de feines que s'estaven fent i catalitzador de noves feines. Per exemple, hi ha unes rutines de detecció d'errors que el Servei de Llengua Catalana va dissenyar per a uns programes per a les aules d'autoaprenentatge, que s'utilitzaran en el programa de detecció de neologismes. Hi ha una tesi doctoral en curs de confecció d'un diccionari de formants que serà utilitzat també en el procés de detecció de neologismes. Probablement aquest mateix curs es posarà en marxa una tesi per dissenyar un validador sintàctic i, tard o d'hora, en un marge curt de temps, es posarà en marxa una extracció de terminologia i una indexació automàtica de continguts.

En segon lloc s'ha parlat de la reutilització dels materials. Jo voldria fer veure que, tal com els hem presentat, els materials ja en l'estat actual, per exemple tota la part de premsa, tenen un interès evident per a qualsevol estudiós del llenguatge periodístic. Teníem aquest matí entre els assistents alguns col·legues de ciències de la informació, i espero que molt aviat els tinguem com a usuaris dels materials. També pensem que gent interessada en l'estudi dels subllenguatges poden extreure i separar del conjunt dels materials coses com les crítiques televisives o les cartes al director, l'horòscop, el temps, la temàtica d'una àrea especialitzada...

Com que el temps és escàs, voldria aprofitar aquesta mateixa intervenció per dir que el Departament està especialment interessat en qualsevol tesi o tesina que impliqui la

utilització d'aquests materials, especialment si això implica que l'interessat aportarà hores de feina al buidatge de textos; fins i tot, encara que no s'hagi de fer una tesi o una tesina, a qualsevol dels presents o dels absents que tingui hores i no sàpiga com invertir-les, ni que siguin un parell d'hores a la setmana, li trobarem la manera d'ocupar-les.

Finalment també vull adreçar-me a aquest segment més jove dels assistents. Voldria fer saber que la Universitat de Barcelona té un servei de lexicometria obert a qualsevol estudiós, no només de la nostra universitat sinó d'arreu, per ajudar-lo a dissenyar qualsevol tipus d'explotació, textual o lèxica, adreçada a fer una tesina, una tesi, una comunicació a un congrés, el que sigui, i aquest servei és gratuït.

Lluís Payrató

Si algú vol afegir alguna informació...

M. Teresa Cabré

Com que el Dr. de Yzaguirre ha fet aquesta mena de crida als investigadors joves, jo voldria recordar que el primer corpus que va servir per a l'elaboració del diccionari Cobuild tenia x milions d'ocurrències. Però d'aleshores ençà s'ha anat completant aquest corpus, i una de les fonts a través de la qual s'ha anat completant ha estat precisament que tots els estudiants que han volgut fer una tesi o treball de recerca sobre les dades d'aquest corpus han aportat una quantitat semblant d'informació, que és la que ha constituït pròpiament el seu treball de recerca, que s'ha incorporat en el corpus. La cooperació entre els estudiants i els professors que treballen en un determinat corpus, una mica com un servei interactiu, és un fet normal.

Joan Acordagoicoechea (Bibliograf, diccionaris Vox)

En aquest ambient una mica optimista de l'elaboració dels corpus, als quals ens hem afegit tant els catalans com els castellans, no hem d'oblidar algunes coses. Primera, que el corpus neix en l'idioma anglès, que no té una institució normativa, i per tant no hi ha més norma que l'ús, la qual cosa diferencia probablement els criteris de selecció i els possibles conflictes entre la normativa més o menys estricta i l'ús. Segona, i potser faig d'advocat del diable, però a l'últim número de *International Journal Lexicografic*, Della Summers, que és la directora de diccionaris Longman i a més manager director, fa veure com dos

corpus grans com per exemple el Cobuild, de vint milions d'ocurrències, i el Longman i Lancaster, que n'han fet també un altre de vint, no coincideixen en una dada concreta, a saber: quina és la construcció prioritària del verb *decidir*. Mentre que el Cobuild dóna *decidir* + infinitiu, *decide to*, el Longman, que ja arriba a trenta milions, dóna com a prioritari *decidir* + interrogatiu indirecte, *decideixo qui vindrà, quan hi aniré*, etc., traduït, i fa la seva reflexió. Diu que això no vol dir que el seu corpus sigui millor que l'altre. Malgrat tot hi ha un últim o primer element que és la intuïció i la decisió del lexicògraf. Precisament perquè el corpus no pot abraçar-ho tot, s'ha de prendre una decisió sobre com s'intervé, què s'hi posa i com s'hi posa, i per tant, encara que és una eina molt potent que pot separar l'arbitrarietat de la no arbitrarietat, hem de tenir present que sempre hi ha alguna dada o informació que la màquina no donarà, fins i tot en corpus molt grossos. Per tant, no hem de confondre el fet de tenir eines més potents amb el fet d'elaborar productes millors. Evidentment, el fet de tenir una eina més potent possibilita fer millor les coses, però no necessàriament. Es pot tenir una eina molt potent al servei d'un cap no tan potent, i per tant, utilitzar-la malament.

Finalment, per possibles al·lusions, sobre el probable finançament de la Comunitat Europea, jo crec que el català no està vedat, com deia molt bé el professor Blecua. A la reunió es va parlar de les nou llengües oficials, però també d'altres llengües reconegudes. De fet hi ha publicacions de la Comunitat Europea en català. El que passa és que la massa social catalanoparlant no és tant potent com la castellanoparlant, i evidentment els interessos comercials, als quals es referia el professor Llisterri, fan que es doni prioritat a una o a una altra. Però jo crec que si es té una proposta potent, més potent que una possible proposta anglesa o alemanya, els projectes podrien anar endavant. Aquesta és la meva opinió.

Joaquim LListerri

Només una qüestió que he recordat. Una de les llengües en què es va treballar en el projecte SAM d'estàndards en avaluació de la sintesi, reconeixement i bases de dades era el suec, que no és una de les nou llengües. I acabo de recordar també que el mateix Institut d'Estudis Catalans té el projecte ESPRIT ACCOR, que és un projecte sobre coarticulació en les llengües europees. No tinc gaire informació sobre el projecte però

com que hi ha l'Institut no m'estranyaria que el català també fos una de les llengües incloses en el projecte. Per tant, això ens demostra que tècnicament és factible.

Lluís Payrató

Última pregunta, última intervenció.

Joan Pujolar (Universitat Autònoma de Barcelona)

A partir d'un comentari que ha fet Emili Boix sobre la qüestió del corpus de conversa espontània, i això lligaria una mica amb la idea del Dr. Blecua que les dades no són innocents, m'he preguntat com s'ha estructurat i pensat aquest corpus.

Normalment ens trobem que la gent, en qüestió de conversa espontània, són una mica mandrosos. Normalment quan graves una conversa i intentes analitzar-la t'adones que els parlants no fan gaire esforços per adir-se amb les expectatives dels lingüistes. A més, ho dic perquè s'ha comentat que faria falta que es fes un estudi sobre el castellà que ara es parla, i em pregunto fins a quin punt es poden mantenir les dues llengües separades en la conversa espontània, perquè la gent efectivament no les separa. Després tenim el tema dels registres. A l'hora de definir-los s'intentarà que siguin uns registres estrictament catalans. També hi ha la qüestió de l'ús estilístic que es pot fer de les diferents varietats lingüístiques. Em pregunto que es fa quan algú diu alguna cosa com ara un barbarisme, o es passa al castellà: ¿pareu i torneu a començar, com s'aborda aquest tema?

Emili Boix

Pel que fa al tema del català-castellà, la meva resposta deu ser òbvia. Evidentment, les dades que tenim només de conversa espontània proporcionen una gran quantitat de dades del castellà, perquè apareixen continuadament en la vida quotidiana d'aquesta zona. Podríem oferir, com fa en premsa el grup de M. Teresa Cabré, Mercè Lorente i Lluís de Yzaguirre, que té una part de premsa castellana, un material de castellà. Per això, quan deia de coordinar-ho amb el castellà gairebé és perquè tenim alguna cosa per oferir. No separem cap tipus de conversa per la presència del castellà. Pel que fa als registres, no utilitzem uns registres catalans pròpiament sinó que només utilitzem els quatre criteris de Gregory i Carrol que comentava aquest matí. Busquem diferents situacions on es combinin aquests quatre criteris. El que hi ha a vegades és un problema de delimitació de

registres en català, i això sí que és específic de la història del país, que és contingent i específica. M'ha costat entendre la teva pregunta sobre els registres.

Lluís Payrató

Havia dit última intervenció perquè em sembla que, encara que la inscripció sigui gratuïta, més de dues hores de taula rodona és una mica excessiu per a tots. Si hi ha alguna puntualització es pot fer abans d'arribar justament a les dues hores.

Joan Pujolar (Universitat Autònoma de Barcelona)

Jo m'imagino que et pots trobar una situació en què un registre es defineixi precisament per la situació de contacte lingüístic, pels diversos significats que tu pots aportar mitjançant l'ús de diverses llengües...

Emili Boix

Em refereixo als efectes retòrics aconseguits a causa de la combinació, d'alguna manera, de dues o més llengües. Sí que hi ha alternances, fenòmens d'aquest tipus, no pas canvis de codi. També hi hauria tot un material aprofitable, i a més hi ha tota una llarga tradició en el país que ha tractat el tema. Seria el camp d'aplicació de recerca.

Lluís Payrató

Bé, si hi ha alguna altra puntualització, em sembla que ja es pot fer directament i personalment als ponents. En nom de la comissió organitzadora i de la Secció de Lingüística Catalana del Departament, voldria acabar donant les gràcies a tothom, i primer als ponents, per haver-se cenyit al temps previst. Agraeixo també, en nom de la comissió, als col·legues del Departament, als becaris, i a la Secció de Protocol de la Universitat, i molt especialment, a tots els assistents, que hagin fet possible el col·loqui.

En nom de tots, també, bon Nadal, i us esperem al CLUB 2, que tindrà lloc si tot va bé al desembre de 1994.

SEGONA PART

La selecció de dades lingüístiques: una perspectiva sociològica

Cristina Sánchez Miret
(Universitat de Girona)

Com es diu a la presentació, aquest col·loqui està dedicat a la metodologia per a l'anàlisi de corpus lingüístics, i vol ser un fòrum de discussió dels problemes pràctics que sorgeixen en una investigació d'aquest tipus. La selecció de les dades, és a dir, la mostra i la seva representativitat, és un dels primers aspectes que s'ha d'afrontar en la metodologia de qualsevol investigació, no només en les de l'anàlisi de corpus lingüístics. Les decisions preses en aquest sentit, com moltes d'altres preses en el procés investigador, són cabdals per maximitzar la validació de les conclusions.

Com s'han de seleccionar les dades i els informants per què el corpus sigui representatiu?

En aquest article dibuixarem, des de la perspectiva sociològica, algunes de les idees fonamentals que haurien de conformar el disseny de la mostra en els estudis de corpus a la nostra societat.

El primer paràmetre que ha de guiar la selecció de les dades és que aquestes siguin representatives de la població que volem estudiar. Aquesta afirmació, que sembla d'una gran simplesa i obvietat, té implicacions prou complexes per a l'anàlisi de la realitat social.

El segon, en una anàlisi d'aquest tipus, la mostra ha de ser representativa dels "actes comunicatius" que es volen estudiar. En la vida quotidiana els individus parlem contínuament, en moltes situacions, en molts contextos, amb molts individus diferents i no ho fem de la mateixa manera.

En conclusió, per dir-ho de manera simple, la mostra ha de ser representativa de persones i d'esdeveniments. Les decisions preses en el disseny de la investigació en aquestes qüestions són determinants per a la seva validesa ecològica. És a dir, per a la seva capacitat de representació de la realitat i per tant d'explicació.

Cal que ens aturem en aquest concepte de validesa ecològica per entendre les implicacions que tenen en l'explicació de la realitat les decisions preses en el disseny, mostral en aquest cas, de la investigació.

1. La validesa ecològica: complementarietat entre la validesa interna i externa

El problema de la validesa ecològica és un tema clau en tota investigació, perquè en definitiva dóna compte de la seva capacitat d'explicació de la realitat. En la tradició sociològica s'han conreat branques metodològiques diferents les quals han tingut tradicionalment diferents tipus de validesa ecològica.

D'una banda, tenim els procediments que podem englobar sota el nom genèric de quantitatius en els quals s'ha aconseguit amb els perfeccionaments de diferents tècniques principalment estadístiques un gran control de les variables i condicions dels mitjans a estudiar i això ha suposat per a aquest tipus d'investigacions un alt grau de validesa ecològica interna.

El problema ha estat que, per exemple en les grans enquestes, la validesa externa dels resultats s'enfonsa perquè no es pot tenir seguretat de què en la vida quotidiana la gent faci en realitat el que havien dit que feien -no es té constància dels actes al fer mostres d'individus i no de comportaments; també dins d'aquest corrent es poden fer mostres sobre comportaments però no és té la certesa que aquests siguin els mateixos fora del disseny de l'experiment o de la presència de l'investigador.

Per altra banda la metodologia qualitativa lligada al treball de camp tradicional té la problemàtica oposada. En aquest tipus de procediments iniciats pels antropòlegs la validesa externa de les investigacions ha acostumat a ésser molt alta, donat que, per exemple, s'estudien els individus en el seu propi hàbitat i per períodes llargs de temps. Ara bé, la debilitat científica en aquest tipus d'investigació resideix en l'escassa validesa interna donada la problemàtica que presenta en la quantificació, representativitat i

possibilitats de comparació dels resultats.

Que hàgim dibuixat fins aquí aquestes dues línies metodològiques diferenciades no vol dir que siguin oposades, encara que hagin estat llargament presentades així.

La validesa interna és un component essencial de la validesa externa, fins a l'extrem que no es podran eliminar explicacions alternatives dels resultats en la nostra mostra i no la podem extreure de cap interpretació o generalització que se'n derivin. Per altra banda, algunes conclusions depenen més fortament de la generalitat (validació externa) dels resultats que altres.

En les investigacions del corpus lingüístic es pot dur a terme un disseny metodològic que permeti gaudir a les investigacions tant d'un alt grau de validesa ecològica interna com externa. Els elements bàsics de la proposta són dos: les tècniques de mostreig d'Altmann per a la tria dels actes "comunicacionals", i el concepte d'estructura social per a la caracterització de la mostra de persones.

És cert que hi ha un tercer element, la problemàtica de la codificació de les dades. que és un altre dels aspectes claus per assegurar la validesa ecològica d'una investigació, tema que no tractarem en aquesta ponència.

Parlarem en primer lloc de les tècniques de mostreig d'Altmann que ens permeten dibuixar una estratègia adequada per a l'anàlisi dels diferents actes comunicatius de la vida quotidiana d'una societat donada.

En segon lloc, ens centrarem en un dels aspectes més rellevants que, des del meu punt de vista, la sociologia pot aportar a l'estudi dels corpus lingüístics: el coneixement de l'estructura social com a base per a la selecció d'una mostra representativa de la nostra societat.

2. Les tècniques de mostreig d'Altmann

El problema de la validesa ecològica està estretament lligat amb com s'extreuen les mostres de tot tipus d'usos i amb el caràcter d'aquestes mostres. En el cas de la investigació lingüística, haurà de ser l'objectiu de la investigació i les seves característiques les que decideixin la seva conveniència.

La incorporació de les tècniques de mostreig de Jeanne Altmann (1974) ha suposat un gran pas cap endavant en l'assoliment d'una alta validesa ecològica interna dels resultats

en les investigacions qualitatives.

No es pot observar un grup social contínuament, per tant, s'han d'elegir tècniques de mostreig que sempre s'han d'explicitar. Les mostres sistemàtiques permeten crear dades quantitatives i alhora permeten la generalització a d'altres situacions o poblacions. A partir d'aquestes mostres es poden fer freqüències i percentatges de comportaments i d'esdeveniments.

Els tipus de mostres que es poden realitzar són els següents:

Mostres ad libitum: són mostres d'observació general per familiaritzar-se amb l'estructura de la comunitat i amb la seva quotidianitat. Aquestes mostres són molt útils com a pretest, i s'acostumen a aplicar en investigacions on es desconeix la comunitat a analitzar o algun aspecte de la mateixa.

Mostres d'esdeveniments: són també una mostra aleatòria que normalment es recull en mostres ad libitum i que consisteix a observar esdeveniments en un moment donat.

Mostres focals: poden ser de dos tipus: *d'espai* o de *sub-grups o individus*. En aquestes mostres el que es fa és observar en un període pre-fixat de temps un espai concret en el primer cas i en el segon un individu o individus.

Per aplicar aquetes mostres és necessari partir d'aquesta investigació prèvia sobre la comunitat a què abans fèiem referència, a partir de la qual i junt amb el paradigma teòric pertinent, crearem les metacategories i estratègies que han de guiar la nostra investigació segons el marc d'hipòtesi per nosaltres establert.

En certa manera el mostreig proposat per Altmann no és res nou, però la seva importància resideix en el seu ús sistemàtic que permet un disseny d'investigació en el qual sense perdre un alt grau de validesa ecològica externa s'aconsegueix un alt grau de validesa ecològica interna.

3. La contextualització: Estructura social i classes socials

Diferents autors han parlat de la importància del context, concepte del qual existeixen diverses definicions, per interpretar els esdeveniments o interaccions en una societat donada. (Vegeu en la bibliografia autors com : Sapir, Goodwin i Duranti, Duranti, Brown i Yule, Briggs, Cicourel, etc.). Per tant aquest ha estat utilitzat en diversos sentits. Per exemple, el coneixement *a priori* de la comunitat és bàsic perquè si no, l'investigador no

pot controlar en quin sentit poden els entrevistats canviar la seva organització diària per la seva presència.

Briggs (1986) resumeix molt bé una part del problema quan diu que "s'ha d'aprendre a preguntar", s'hauria d'afegir que també s'ha d'aprendre a observar i a interpretar.

Per fer una anàlisi adequada calen dades demogràfiques, etnogràfiques, de l'estructura social i dels mecanismes de poder de la comunitat. Igual que és imprescindible per a l'investigador conèixer les pautes comunicatives d'aquesta comunitat.

Però s'ha d'anar més enllà, donada la naturalesa dels estudis sobre corpus lingüístics cal saber a "qui observem". És a dir, en quina mesura els individus observats ens donen compte del conjunt de la realitat social estudiada. Això només ho podem saber coneixent l'estructura social i el lloc que aquests individus hi ocupen.

Posant el problema al revés, cal que la mostra dissenyada per nosaltres sigui representativa de l'estructura social de la societat estudiada.

4. Estructura social i classes socials

Les societats complexes estan caracteritzades, en diversos graus, per la distribució desigual de diversos béns materials i simbòlics. Aquesta estructura de desigualtats socials és persistent i constitueix una característica de les nostres societats. L'estudi de les desigualtats socials és un dels temes centrals de la sociologia des dels seus inicis. Amb conceptes com el d'estructura social, en el seu sentit genèric, es vol donar compte d'aquestes estructures de desigualtat.

El terme estructura prové del món espacial. El seu significat literal fa referència a una disposició d'elements, en la qual es donen relacions espacials correlacionades amb algun sistema de relació entre aquests elements o entre parts determinades i el conjunt.

En un sentit metafòric, l'estructura és un sistema de distàncies i relacions interpretades figurativament d'un o altre tipus.

Si interpretem l'estructura social en el sentit literal ens trobem amb l'ecologia social. Però cal que el terme social del concepte d'estructura limiti les seves connotacions espaials. Per això, parlar d'estructura social vol dir tractar amb un conjunt de grups o de categories d'individus -però no individus particulars- com a elements d'un sistema.

La tradició teòrica que arrenca de Marx i Weber ha anomenat aquests grups classes socials i, a partir d'aquests autors, la definició del concepte de classe social ha esdevingut fonamental per explicar les desigualtats socials dels individus en una societat donada.

El concepte d'estructura social és més ample que el d'estructura de classes, donat que es poden identificar d'altres grups a partir de criteris com l'edat, l'origen geogràfic, el sexe, etc. Sense negar l'existència d'aquesta diversitat, la identificació de l'estructura de classes d'una societat concreta és l'aspecte més fonamental per entendre com s'estructuren les diferències i desigualtats existents.

En l'experiència de cada dia la classe és una categoria present per ordenar el món social: no és estrany trobar que la gent fa servir termes com ara "classe mitjana" i "classe baixa" per a les descripcions de les seves experiències. I això no és només característic de la nostra època.

Per contrast amb la quotidianitat del terme, on se l'associa molt marcadament amb el concepte prestigi o rang social, en el discurs sociològic el mot pren diferents usos i significats.

Hi ha, però, tres principis que són comuns a totes les concepcions de societats de classe social, que cal remarcar:

Les classes constitueixen un sistema dels grups més amplis en l'estructura social. És a dir, un nombre de grups reduït dividits a conseqüència de la divisió de la societat segons criteris que són importants en la vida social.

La divisió de classes suposa estatus socials connectats amb un sistema de privilegis i discriminacions no determinats per criteris biològics.

La pertinença dels individus a una classe social determinada és relativament permanent.

5. Un exemple: les classes socials a la Regió Metropolitana de Barcelona

La realitat de les classes socials és comprovable d'ençà que constitueixen fenòmens reconeixibles al nivell de l'experiència, és a dir, per Marx i Weber quan es constituïen en grups d'acció.

L'existència d'estils de vida, valors i actituds diferents segons la classe social dels individus és el que ens permet parlar de la seva existència i importància en la nostra

societat.

Els comportaments i les actituds formen part dels estils de vida dels individus i aquests al seu conjunt es corresponen amb l'estructura de posicions objectives que dibuixen el mapa de classes que descriu l'estructura social d'una societat determinada. Per això, no hem partit d'una atribució prèvia del caràcter de classe, sinó que hem utilitzat diversos grups de variables per arribar a construir les classes socials, seguint el criteri de Bourdieu.

El mètode fet servir en la investigació es basa en l'aplicació de les tècniques estadístiques d'anàlisi de correspondències múltiples i de classificació automàtica jeràrquica, que permeten analitzar l'estructura de classes socials des d'una òptica multidimensional. De manera que, ens és possible incorporar a la definició de classe altres variables definidores de desigualtats com el sexe, l'edat o l'origen geogràfic; i, alhora, mantenir la definició de la unitat d'anàlisi com a individual i familiar.

Així mateix permet, a partir de la informació sobre hàbits i comportaments de la població que ofereix la base de dades de l'Enquesta Metropolitana, construir diversos indicadors d'hàbits de classe que també s'incorporen a la definició dels grups socials, i acaben configurant quina és l'estructura social de la Regió Metropolitana de Barcelona.

Els resultats ens mostren una estructura social definida per grups socials que es poden caracteritzar com a classes socials i que s'ajusten a l'esquema hipotètic d'estructura de classes establert en la investigació.

Les diferents anàlisis realitzades ens han permès obtenir dotze grups, definits per l'ocupació de l'entrevistat. Una vegada analitzades les característiques de cadascun d'aquests grups, si bé les categories ocupacionals (posicions en les relacions socials de producció), no es mostren totalment homogènies com a element de classificació dels grups, presenten característiques d'agrupació prou consistents per ésser interpretades en termes de classe; classe que es defineix, alhora, per la posició en les relacions de producció i per la similitud dels seus recursos i dels seus hàbits. Cal assenyalar però, que tres dels grups resultants han quedat al marge d'aquesta classificació.

D'altra banda, malauradament, també es confirma la incapacitat de bases de dades com la nostra de donar informació sobre els col·lectius extrems de l'estructura social.

Existeixen dues grans classes socials, no monolítiques, a la Regió Metropolitana de Barcelona: la classe treballadora i la classe mitjana. Aquesta darrera dividida segons la

propietat o no dels mitjans de producció en dues grans fraccions: la vella classe mitjana i la nova classe mitjana.

Els resultats mostren una classe treballadora dividida, però no a partir del grau de qualificació al treball, sinó a partir de l'edat. Cal dir que, atès les característiques generacionals del col·lectiu més vell dels treballadors, es pot afirmar que la qualificació/no qualificació dels treballadors segueix essent important per explicar desigualtats socials.

Empresaris-autònoms amb assalariats	3,24%
Empresaris-autònoms sense assalariats	4,49%
Comerciants	4,54%
VELLA CLASSE MITJANA	12,27%
Tècnics alts	4,23%
Professionals liberals i directors/gerents	2,73%
Tècnics mitjans i empleats	19,30%
NOVA CLASSE MITJANA	26,26%
Contramestres i capatassos	2,92%
Treballadors en actiu	30,31%
Treballadors vells	16,85%
CLASSE OBRERA	50,08%
Propietaris/ treballadors agraris	2,49%
Sense categoria	8,10%
FF.AA.	0,79%

La classe mitjana i la classe treballadora presenten entre si desigualtats molt notables, que apareixen, sobretot, quan es posen en relació l'una amb l'altra. S'han constatat desigualtats importants entre ambdues classes i fraccions de classe a l'accés als recursos econòmics, a l'accés a la propietat, en l'amplitud de les vivendes, a l'equipament

domèstic, en la propietat de béns de consum; en tots aquests aspectes els recursos de la classe mitjana són molt més abundants que els de la classe treballadora. També s'han observat desigualtats en altres aspectes importants: a l'accés a l'educació, a l'accés a la informació i a l'hàbit de lectura, a l'accés a la cultura ciutadana, en les formes de lleure, als tipus de vacances.

Tots aquests aspectes que recollíem en les hipòtesis queden confirmats per les dades obtingudes. La caracterització de les classes segons aquest conjunt d'aspectes ens dóna, alhora, informació sobre altres aspectes menys evidents.

El primer tret diferenciador, que cal remarcar, és l'efecte de l'envelliment damunt les condicions de vida d'ambdues classes socials: mentre que en la classe mitjana la situació de jubilació no sembla conduir a un canvi dràstic de condicions de vida, en la classe treballadora l'envelliment condueix a un empobriment substancial, de manera que els treballadors manuals ocupats i els jubilats queden inclosos en capes socials diferents.

Des del punt de vista de la composició entre catalans i immigrants hi ha també una diferència notable entre la classe mitjana i la classe treballadora. La classe mitjana es compon, en una proporció molt més elevada, de persones que han nascut a Catalunya i pertanyen a llars on tots els membres són nascuts a Catalunya, mentre que en la classe treballadora la majoria de les persones han nascut fora de Catalunya o són fills d'immigrants. Això, no significa que totes les persones de classe mitjana siguin d'origen català: en ambdós grups hi ha natius i immigrants o fills d'immigrants. Però les proporcions són molt diverses a conseqüència de les característiques socials de la immigració que es va produir fa uns trenta anys.

Amb tot, la tendència actual és la del mestissatge: la capa inferior de la classe mitjana comprèn més persones d'origen immigrant que la capa superior d'aquesta mateixa classe, de manera que tot fa pensar que s'estan produint moviments que tendeixen a desdibuixar els avantatges deguts a les posicions heretades i a igualar les possibilitats de natius i immigrants, sense, però, que això signifiqui que estan igualades.

Per dir-ho d'una altra manera: no és l'origen immigrant el que condiciona la posició en l'estructura social, sinó la pertinença de classe; però, donada la diferent composició de les classes en termes de migrants i natius, és previsible que es mantingui encara en les generacions properes una diferència observable sobretot en l'adscripció lingüística. Aquest fet, no suposa que les persones d'origen immigrant no coneguin el català: entre les

generacions joves s'estén ràpidament el bilingüisme.

Un altre tret diferenciador és el dels hàbits a l'ús de l'espai: mentre que la classe mitjana no reconeix límits a l'ús i consum de l'espai, i sembla tenir a la seva disposició el conjunt de la Regió –encara que després no utilitzi tot el territori amb la mateixa intensitat–, la classe treballadora viu més confinada a un entorn espacial proper al seu domicili, de manera que no explora el territori amb la mateixa llibertat, o, dit d'una altra manera, no sembla reconèixer-lo com a propi en la mateixa mesura amb què ho fa la classe mitjana.

Aquest fet queda corroborat encara per un altre fenomen: la major tendència a viatjar i a sortir a l'estranger per part de la classe mitjana, que mostra que, més enllà de la Regió Metropolitana, l'ús i el coneixement del territori segueixen presentant unes pautes diferenciades per a cadascuna de les dues grans classes.

La seva distribució dins l'espai metropolità no és tampoc homogènia: Barcelona concentra una major proporció de població pertanyent a la classe mitjana que la resta de la Regió, mentre la resta de l'antiga àrea metropolitana concentra prioritàriament els treballadors manuals. La segona corona metropolitana, és a dir, la resta de la Regió una vegada treta l'antiga àrea metropolitana, té una composició més equilibrada. Però les desigualtats no s'estableixen únicament en relació a les dues grans classes, sinó a l'hora de la distribució de les diverses capes i fraccions de la classe mitjana. La segona corona té una concentració més gran, respecte a la resta del territori, de la vella classe mitjana.

Crida també l'atenció la diferència que s'observa en relació al temps lliure, que d'antuvi sembla anar en direcció contrària a la relació assenyalada respecte a l'espai, però que cal observar amb cura. La classe mitjana es considera més ocupada i els grups dels comerciants i dels empresaris sense assalariats mostren tenir molt poc temps lliure. La llibertat respecte a la disposició del temps, que tradicionalment era un dels signes dels grups benestants, apareix ara com a pròpia de la classe treballadora vella. Però, ben mirat, darrera la idea de temps lliure s'insinua la idea de temps buit, un dels fenòmens que rebutja amb més força la nostra cultura.

Hi ha, també, un conjunt de trets que assenyalen diferències de mentalitat entre les dues classes, com per exemple la rapidesa d'adopció de les innovacions: mentre la classe mitjana adopta més ràpidament nous instruments, com l'ordinador, o nous serveis, com les targetes de crèdit, la classe treballadora tarda més temps a adoptar aquestes novetats,

probablement perquè tarda més temps a considerar que estan al seu abast. El mateix succeeix en termes d'assegurances: la classe treballadora segueix mantenint determinats tipus d'assegurances, com la d'enterrament, mentre la classe mitjana fa sobretot assegurances de vida, plans de jubilació i de pensions, etc.

S'observa una diferència notable en relació a les tendències de vot: no hi ha una correspondència estricta entre pertànyer a una classe i votar un partit, però hi ha tendències més acusades a votar uns partits o uns altres segons la classe social. A les eleccions generals de 1989, la classe treballadora va votar sobretot el PSC/PSOE i la classe mitjana Convergència i Unió; però en la classe mitjana hi ha una major tendència a matisar el vot segons la pertinença a una o altra fracció o capa, mentre que en totes les capes de la classe treballadora es manté d'una manera estable la proporció de persones que van votar PSC/PSOE (al voltant del 38%).

Hem assenyalat fins aquí un conjunt de desigualtats globals entre la classe treballadora i la classe mitjana. Cal ara assenyalar uns aspectes en els quals no es manifesten diferències notables: la grandària de les llars i el tipus de família i d'organització familiar. En aquest sentit estan desapareixent les especificitats de classe, s'ha generalitzat la família nuclear, s'ha produït a tots els grups una reducció de la natalitat que situa el nombre de fills entre un i dos.

En canvi, el que sí presenta diferències significatives entre els grups socials és el nombre d'esposes del cap de família que treballen a l'àmbit productiu. És a la classe mitjana, tant a la fracció vella com a la nova on és habitual que més de la tercera part i fins a la meitat de les llars les esposes estiguin ocupades al mercat laboral. En canvi, entre la classe treballadora, a més de la meitat de les llars les esposes del cap de família es declaren només com a mestresses de casa.

6. Les classes socials i els estudis de corpus lingüístics

Al començament de l'article hem intentat deixar clar un primer tipus de relació entre classe social i l'estudi de corpus lingüístics: la necessitat d'incloure la variable classe social per fer una mostra representativa del conjunt poblacional. D'una banda perquè aquesta és una variable clau en l'estructuració de la població, tant com ho poden ser l'edat, el sexe i el nivell educatiu per parlar de les variables que s'utilitzen habitualment.

D'altra banda no podem oblidar que hi ha una llarga tradició, iniciada a l'antiga Unió Soviètica a partir dels estudis de Marr que parlen de la relació entre llengua i classe social. Tradició que podríem dir que culmina en els estudis de Bernstein els anys setanta que expliciten les diferències lingüístiques entre els nens segons la classe social a la qual pertanyen.

Però n'hi ha més. Volem apuntar ara un segon tipus de relació entre classe social i corpus lingüístic, evidentment relacionada amb l'anterior, que la complementa: la necessitat de vincular els usos lingüístics amb la resta de comportaments socials, usos, costums i també condicions materials dels individus. Això només és possible fer-ho d'una manera integrada, multidimensional -fugint dels estudis particulars o segmentats entre les diferents realitats- a partir del denominador comú de la classe social. Per tant, un estudi com el que aquí hem presentat sobre els grups socials de la Regió Metropolitana de Barcelona, ens permetrà contextualitzar, amb una gran riquesa de dades socials, els actes comunicatius fets pels individus, donat que, sabent la seva pertinença de classe, coneixerem bona part de les seves condicions de vida, hàbits i usos, com ha quedat reflectit breuement en l'apartat anterior.

El coneixement de les classes socials d'una societat determinada ens parla de quina és l'estructuració dels individus en grups socials, estructuració que ens permet parlar de quines són les relacions dels individus, els comportaments, les expectatives, etc. Tot això és fonamental per conèixer el veritable abast, la realitat social, de la llengua. Realitzant anàlisis dels corpus lingüístics que tinguin en compte la representació de les classes socials d'una societat determinada sabem quina incidència tenen en la societat els diferents comportaments lingüístics que existeixen; també podem preveure quina serà la seva evolució. Ja hem explicat que a la Regió Metropolitana existeixen dos grans grups diferenciats de classe treballadora, i que un d'ells està fortament caracteritzat per l'envelliment dels seus membres: és de preveure que els comportaments que els caracteritzen, entre ells el del coneixement i ús de la llengua, aniran desapareixent. Així mateix hem vist també com determinats grups socials, els grups de la classe mitjana assalariada, són baluards de nous casos socials que s'expandeixen a la resta de grups de la nostra societat, com resultat d'un efecte d'emmirallament: les implicacions d'aquest fet en la llengua també s'han de tenir en compte. I, així, podríem enumerar tot un seguit de característiques dels grups socials i de fenòmens que es donen en l'estructura social que

poden aportar una gran riquesa a l'estudi del corpus lingüístic d'una societat concreta. Així mateix, l'estudi de la llengua dels individus en relació a la seva pertinença de classe redunda en un millor coneixement dels propis grups socials i de la seva caracterització a la nostra societat.

L'objectiu d'aquesta ponència ha estat posar de manifest la importància en la centralitat del concepte de classe social per entendre i alhora donar compte de la multidimentsionalitat social. Aquesta multidimensionalitat de la diferència social ha de ser la matriu de l'estudi del corpus lingüístic, perquè el resultat d'aquests sense aquesta contextualització sociològica aportaria una mínima part de la seva riquesa per al coneixement de la realitat tant lingüística com social.

Bibliografia

ALTMANN, J. (1974), *Observational Study of Behaviour: Sampling Methods*, "Behaviour", 49, pàg. 227-267.

BRIGGS, CH. L. (1986), *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*, Cambridge University Press, pàg. 93-111 i 115-119.

BROWN, G. i YULE, G. (1983), *Coherence and the Interpretation of Discourse*, dins *Discourse Analysis*, Cambridge University Press, Cambridge, cap. 7.

CICOUREL, A. V. (1988), *Elicitation as a Problem of Discourse*, dins *Sociolinguistics: An International Handbook of the Science of Language and Society*, Walter De Gruyter and Co., pàg. 903-910.

_____ (en premsa), *The Interpenetration of Comunicative Contexts: Examples of medical encounters*, dins *Rethinking Context: Language as an Interactive Phenomenon*, dins A. Duranti i Ch. Goodwin (eds.) Cambridge University Press.

DURANTI, A. (1988), *Etnography of Speaking: Toward a Linguistics of the Praxis*, dins *Linguistics: The Cambridge Survey*, dins J. Frederick (ed.) Newmeyer-Cambridge University Press.

GOODWIN, Ch. (1981), *Notes on the Organization of Engagement*, dins *Conversational Organization*, Academic Press, Inc., pàg. 95-125.

GRICE, H. P. (1975), *Logic and Conversation* dins P. Cole y J. Morgan, (eds.) *Syntax and Semantics*, 3, Speech Acts, Academic Press, pàg. 45-58.

GUMPERZ, J. (1982), *Socio-cultural Knowledge in Conversational Inference*, dins J. Gumperz, *Discourse Strategies*, Cambridge University Press, pàg. 153-171.

OCHS, E. (1979), *Transcription as Theory*, dins E. Ochs i B. B. Schiefflelin (eds.) *Developmental Pragmatics*. Academic Press, Nova York, pàg. 43-72.

SÁNCHEZ, C. (1994), *La definició dels grups socials a la Regió Metropolitana de Barcelona. Un problema teòric i metodològic*. (Tesi Doctoral).

SUBIRATS, M., SÁNCHEZ, C. i DOMÍNGUEZ, M. (1992), *Grups i classes socials a la Regió Metropolitana de Barcelona*, Institut d'Estudis Metropolitans, Barcelona.

VILA, I. (1990), *Llengua, nació i educació* dins J. M. Rotger, et al., *Sociologia de l'educació*, Eumo, Barcelona.

Transcripció del discurs col·loquial*

Lluís Payrató

(Universitat de Barcelona)

1. Introducció: els *perquès* de la transcripció del discurs oral

L'objectiu d'aquest estudi no és altre, tal i com el seu títol suggereix amb obvietat, que el de reflexionar sobre un problema comú a totes les investigacions sobre el discurs oral i, en particular, col·loquial: la seva transcripció, entès aquest terme com un procediment de trasllat o transposició a una forma gràfica (escrita) d'una producció (lingüística, discursiva) originalment oral. L'associació de la paraula *problema* amb el procés de la transcripció no és casual; s'ha utilitzat deliberadament per donar a entendre que en qualsevol investigació sobre el discurs oral cal plantejar, en un moment o altre, la solució a una qüestió concreta: com s'han de manejar unes dades que, per la seva naturalesa, necessiten un tractament més o menys formalitzat per poder ser estudiades.

En canvi, no és un objectiu d'aquest treball, i l'aclariment ja no resulta ara tan obvi, oferir receptes fàcils i ràpides per solucionar els múltiples trencacolls, teòrics i pràctics,

* Aquest treball s'ha beneficiat d'un ajut a la investigació de la DGICYT del Ministerio de Educación y Ciencia (projecte PB90-0505) i de la CIRIT de la Generalitat de Catalunya (CS93-1017). Agraeixo a totes les persones que han intervingut en el procés de correcció de la proposta de transcripció inclosa en aquest article les contribucions i esforços que han fet per millorar-la. Una versió en castellà d'aquest treball va ser presentada com a ponència al *I Simposio sobre el Español Coloquial* (Universidad de Almería. 23-25.XI.1994), mentre que diferents versions anteriors de la proposta de transcripció han estat distribuïdes de forma mecanografiada des de 1992.

que es donen en qualsevol tasca de transcripció. No existeix cap receptari ni existeixen fórmules magistrals o màgiques que resolguin d'una manera automàtica els conflictes derivats de la necessitat de transcriure; no existeix, en definitiva, una transcripció ideal a la qual acollir-se de manera irreflexiva. Per tant, serà sempre responsabilitat última de cada estudi determinar quines són les solucions òptimes dels problemes plantejats, a partir de l'experiència (segons els objectius perseguits i els resultats aconseguits) i aprofitant els sistemes de què disposem actualment.

Els perquès de la transcripció, de fet, semblen més que justificats i evidents, tant en el seu vessant teòric com en el pràctic. Tot i que el progrés tecnològic ens permet comptar avui dia amb tècniques (audiovisuals) que posen en dubte el famós adagi de *Verba volant, scripta manent*, l'anàlisi de molts aspectes del discurs -en especial els "més" verbals o lingüístics- comporta la necessitat d'un suport gràfic (escrit) permanent. Sense aquesta eina, no solament no és possible l'anàlisi per part dels responsables de la investigació, sinó que -malgrat que sovint s'oblidi- es fa impossible el trasllat de les dades, el seu ús per part d'altres investigadors o la discussió de les interpretacions, i prou sabut és que compartir dades i contrastar anàlisis és una fase inevitable en qualsevol disciplina científica. El progrés dels estudis no depèn exclusivament de la seva formalització, però sense ella és inviable. A més a més, cal tenir molt en compte que les dades -els elements *empírics*- en què es fonamenten les investigacions sobre el discurs oral no acaben sent -per dificultats pràctiques de maneig i manipulació- els mateixos discursos, tal com es produeixen originalment, sinó fases o etapes posteriors: *versions*, *transcrites*, d'aquests productes primers. Aquesta consideració estableix una separació gens menyspreable entre l'anàlisi del discurs escrit i l'oral, i atorga al procés de la transcripció un relleu substancial, donat que, per la seva naturalesa, el producte original és irreproduïble amb absoluta fidelitat, sigui quin sigui el mitjà de reproducció: exclusivament escrit o auditiu (els únics disponibles fins fa relativament poc temps) o audiovisual. Fins i tot les reproduccions audiovisuals, en aparença les més fidels, donen en realitat un sol punt de vista (la perspectiva única de la càmera que enregistra), cosa que acaba amb el mite de la seva perfecció; per una altra banda, no eximeixen d'una feina -pel que es refereix al canal verbal- que la versió gràfica escrita fa possible o facilita: possibilitat d'un tractament automatitzat de les dades transcrites, i consegüent trasllat, ús amb d'altres finalitats, o bé aplicació de programes informàtics d'anàlisi de textos.

2. Singularitat de la transcripció del discurs col·loquial

Si bé qualsevol tipus de transcripció de qualsevol tipus de discurs oral sol resultar ja *per se* problemàtica, les dificultats es multipliquen quan es tracta de transcriure seqüències discursives orals col·loquials. L'explicació es troba en la naturalesa de la varietat col·loquial, entesa com una varietat *funcional* o *registre* amb unes característiques que fan que tant enregistrar-la com transcriure-la esdevingui especialment complex. La llengua col·loquial, si ens valem de la classificació dels registres pròpia de la tradició d'estudis britànica segons els factors de camp (tema), mode (canal), tenor (propòsit) i to (formalitat), pot ser qualificada com quotidiana (o genèrica, no específica, no técnica), oral espontània (no escrita ni planificada), interactiva (predominantement, més que informativa) i informal (típica de la interacció entre coneguts, amics, o familiars).[1]

Aquesta delimitació de la *col·loquialitat* representa una abstracció concretable en successius trets que, al seu torn, recullen d'una manera més o meys definida i adequada les formes lingüístiques en què es presenta aquesta modalitat funcional. Per l'estreta relació que tenen amb els problemes de la seva transcripció, entre aquests trets destaquen, en primer lloc, els dos següents:

(1) la multiplicitat de canals comunicatius, típica de la llengua col·loquial (i oral en general), amb els seus aspectes verbals i no verbals (tant vocals com gestuals, amb les unitats respectives).

(2) l'escàs control o consciència del locutor sobre el producte (lingüístic i discursiu) que va elaborant (en una situació espontània i informal), i que es tradueix, per exemple, en una elocució ràpida, relaxada, en uns trets vocals molt variats i en una gestualitat molt més rica que la que sol donar-se en situacions formals.

A continuació, entre els trets atribuïbles al canal verbal o lingüístic, la llengua col·loquial ens planteja un segon repte si volem transcriure'n l'entonació (quan els estudis de què disposem són molt escassos), i un tercer si volem respectar-ne la presentació discursiva típica, és a dir, la pròpia d'un discurs *no planificat* (o poc planificat, o

[1] Un tractament més detallat d'aquestes qüestions es troba a Payrató (1988, 1992), pel que fa al concepte i delimitació de varietat col·loquial, i a Payrató (1996) pel que fa als conceptes de variació funcional i de registre.

planificat sobre la marxa, com es prefereixi), amb vacil·lacions i reorganitzacions, *polifònic* (amb interlocutors diferents i amb les diferents veus de cada locutor), *emocional* (amb encavalcaments entre els interlocutors i una expressivitat sempre destacada) i *exofòric*, és a dir, amb múltiples referències que uneixen íntimament el text amb el seu context situacional de producció/recepció, convertint-lo en un producte molt dependent del context (dependència contextual alta).

3. Respostes als reptes de la transcripció

Cada un dels trets anteriors provoca problemes de diversa mena en el procés de la transcripció, ja sigui per la dificultat de categoritzar certs aspectes (l'entonació o els elements gestuals) o per les dificultats de calibrar la quantitat d'informació necessària sobre d'altres (els vocals o el context situacional). L'escassa tradició d'estudis sobre el tema no permet per ara oferir respostes gaire precises als reptes apuntats, tant més si es té en compte la diversitat de corrents teòrics en què es basen les propostes plantejades.

No és senzill, de fet, classificar els sistemes de transcripció en funció de corrents i escoles, donat que hi ha hagut interseccions múltiples i aprofitaments mutus de les propostes. Des d'un punt de vista històric, cal reconèixer a l'etnometodologia el mèrit d'establir els primers sistemes de transcripció del discurs oral. S'arriba a aquests sistemes al llarg de la dècada del seixanta i sobretot del setanta del nostre segle, i gràcies a l'interès d'aquest corrent sociològic d'arrels nord-americanes per la descripció i l'anàlisi dels mecanismes conversacionals, entesa la conversa com la creació, social, d'una empresa participativa, basada en els coneixements (associats a l'acció) dels parlants. En aquests sistemes de trancripció[2] es manifesta clarament l'interès de l'etnometodologia per múltiples aspectes de la interacció (com per exemple els torns de parla, la sincronització de conductes entre els interlocutors, etc.); en contrapartida, els aspectes purament lingüístics queden sovint més descurats o fins i tot es tracten amb poc rigor.

---

2 Vegeu com a mostres els apartats "Transcript notation" i "Transcript symbols", inclosos respectivament a Atkinson i Heritage (eds.) (1984, pàg. ix-xvi) i Button i Lee (eds.) (1987, pàg. 9-17), recopilacions d'estudis dins d'aquesta tendència. La majoria de les notacions prové de treballs anteriors de Gail Jefferson, a qui s'hauria de considerar, amb tota justícia, com a creadora o impulsora d'aquests sistemes de transcripció.

La sociolingüística interaccional i l'etnografia de la comunicació, corrents paral·lels a l'anterior, presten en canvi una atenció més gran a aquests aspectes, i els seus sistemes són probablement els més perfeccionats en l'actualitat.[3] D'altres tradicions ofereixen propostes específiques per a la transcripció de la sincronització d'elements verbals i no verbals en la conversació (vid. Goodwin 1981), i connecten amb els sistemes de transcripció de la gestualitat i la proxèmica propis de la semiòtica.[4] També la tradició d'estudis psicolingüística i centrada en el llenguatge infantil presenta propostes pròpies, alguna molt difosa (el programa CHILDES, vid. MacWhynney 1991), amb un interès marcat per la interacció entre adults i infants i per les tasques associades amb l'adquisició del llenguatge (vid. Bloom 1993). La tradició més pròpiament lingüística prové del camp de la fonètica i de la dialectologia. En el primer cas ens trobem amb sistemes de transcripció fonètica àmpliament difosos i utilitzats, en especial l'IPA o AFI (de la *International Phonetics Association* o *Association de Phonetique International*. vid. Pullum i Ladusaw 1986) i els propis de tradicions més locals (vid. Quilis 1984). En el segon cas, el de la dialectologia, la transcripció sol ser ortogràfica. Tant en un cas com en l'altre, tanmateix, els sistemes neixen de tradicions interessades pels elements verbals o gramaticals (la fonologia, la morfologia, la sintaxi i el lèxic), i dediquen molt poca atenció als aspectes no verbals i, en general, als discursius, de manera que se situen en el pol oposat als utilitzats per l'etnometodologia.[5] Més recentment, l'interès per la ja denominada sovint *lingüística de corpus* ha provocat l'aparició de múltiples sistemes de transcripció de la llengua oral,[6] i fins i tot de normes comunes i estandarditzades per

[3] Vegeu en especial Ochs (1979), Tannen (1984, pàg. xix), Du Bois (1991), Du Bois et al. (1993) i Gumperz i Berenz (1993). Pel que fa a la comparació entre sistemes, vid. O'Connell i Kowal (1994).

[4] Vegeu en aquesta línia Scherer i Ekman (eds.) (1982) i la recent publicació de Poyatos (1994), on es trobaran també referències dels seus nombrosos treballs anteriors sobre proxèmia i paralingüística vocal i gestual. Pel que fa a qüestions referents a l'enregistrament audiovisual i al tractament de les dades, vid. Goodwin (1993).

[5] En el cas de la tradició espanyola pot consultar-se, especialment, Criado (1980). Cf. també com a panoràmica els estudis inclosos a Quilis (1984) i Cortés (1994).

[6] Vegeu per exemple, en espanyol, el seguit per Marcos Marín (dir.) (1993) i en especial l'utilitzat en el corpus d'espanyol col·loquial de València (Briz i Gómez 1992), adaptat ja directament per a la modalitat col·loquial. Quant a recursos i bases de dades poden consultar-se, com a referències recents, Edwards i Lampert (eds.) (1993), Marcos Marín (1994) i Arrarte i Llisterri (coords.) (1994), aquest últim amb una àmplia col·lecció de corpus orals de l'espanyol. Altres exemples concrets consultables pel seu interès i utilitat són, pel que fa a l'anglès, el sistema utilitzat en el British Corpus (Crowdy 1991) i, pel que fa al francès, el del corpus de francès parlat de Montréal (Thibault i Vincent 1990). La proposta presentada en

afavorir el tractament automatitzat de textos i la seva difusió i anàlisi (per exemple les normes T.E.I., sigles de *Text Encoding Iniciative*, vid. Sperberg-McQueen i Burnard 1992).

Els dos exemples següents il·lustren, encara que evidentment de manera molt fragmentària, i fins i tot anecdòtica, els sistemes apuntats. El primer, en doble versió, ampla (1a) i estreta (1b), correspon a Chafe, Du Bois i Thompson (1991). En la versió ampla destaquen, a primera vista, aspectes com la transcripció d'un grup tonal (o unitat entonativa) per línia, amb la marca de transició (, . ?) i de truncament (--); el truncament de paraula (-), la transcripció de les pauses (amb medició exacta numèrica o aproximada, amb punts suspensius) i del riure (el signe de l'arrova, @). En la versió més estreta. de (1b), es detallen aspectes com l'accent (principal, ^, i secundari, `), l'entonació (barres finals), els allargaments vocàlics (=), les inhalacions (H), les oclusives glotals (%) o els encavalcaments ([ ]). El fragment de (2) mostra una transcripció també de base ortogràfica, com les precedents, però amb signes de puntuació convencionals. En lloc d'una conversació col·loquial, el text correspon ara a una entrevista semidirigida, i no hi apareixen símbols addicionals com els anteriors. Prevista per a un estudi dels anomenats *back channels* (les mostres d'assentiment i seguiment del discurs de l'emissor per part del receptor), inclou aquestes respostes mínimes en l'interior dels textos de l'emissor. Aquest segon exemple il·lustra, així mateix, la fitxa inicial mínima necessària com a capçalera de qualsevol transcripció. En tots dos exemples s'ha omès la convencional numeració de línies al marge esquerre.

aquest treball s'inscriu també en aquesta tradició recent de constitució de corpus lingüístics, amb especial atenció a la llengua oral i col·loquial.

(1a)

G: For most people it's celebration,
for me,
it's it's a time,
to --
to get in bed,
to- to put the mustard plaster on,
a time to take ... fifteen grams of vitamin C a day,
... (2.2) and,
of course,
a lot of herb tea,
when I'd rather be drinking whisky.

K: ... (1.3) You don't drink whisky.
... (1.5) @

G: I would if I wasn't sick.

K: @
... No you wouldn't.

G: Yes I would.
I used to drink bourbon every Christmas.

(1b)

G: For ^most people it's ^celebration, \
for ^me=, _
it's .. it's a% ^ti=me, _
to= --
(H) to `get in ^be=d, V
.. (H) to%- .. to `put the `mustard ^plaster `o=n, V
.. (H) a `time to `take ... (.8) (H) ^fifteen `grams of vitamin `C a ^da=y, V

... (2.2) `a=nd. _
of course, _
.. a `lot of herb ^tea, \
when I'd `rather be drinking ^whisky. \
K: ... (1.3) @^You don't drink ^whisky. \/
... (1.5) @N [(H)]
G: [I] `would if I `wasn't ^si=ck. \
K: @ (H)
... @`No you ^wouldn't. \/
G: .. ^Yes I ^would. \/
I `used to `drink ^bourbon every `Christmas. \

Chafe, Du Bois i Thompson (1991:76-77)

(2)

(No 72'84 1: M.D., 2: Thérèse R., 3: un enfant, 4: sa fille, 5: son mari)
(Pronociation très particulière -- ex. "feuillais" pour "faisais")
(Fréquents manques d'accord entre le sujet et le verbe à la troisième personne du pluriel)
3. Ça ça fait pas-mal de bruit c'est un peu long ça.
2. Oui.
1. Bon: Puis là vous me disiez qu'ils étaient censés: détruire le: le: l'usine en face ( )
2. Oui je le sais pas qu'est-ce-qu'ils vont faire ' l'autre bord ils: Moi j'aurais pensé qu'ils jetaient ça à terre. <1. mais ils:> Parce que: ça a pas été: c'est pas bien bien fameux ' l'autre bord de la rue.
1ç Ça doit vous cacher du soleil ça hein?
2. Non du tout. <1. non?> C'est: ça aide pas-mal à six heures là i: ça cache le: le soleil. <1. ah. OK> Mais seulement que moi je sais pas i: c'est ennuyant tu-sais tu: disons qu'il: ' aurait des magasins on verrait (voirait) mieux tu-sais <1. oui> les affaires.
1. Puis ça ferait plus' de vie.

Thibault i Vincent (1990: 91)

Si s'incrementessin els exemples paral·lels als anteriors de manera que es perdés en caràcter anecdòtic i es guanyés en representativitat, el resultat seria un panorama més aviat heterogeni i fins desconcertant. Tal és una de les conclusions, precisament, de l'estat de la qüestió més recent de què disposem sobre sistemes de transcripció (O'Connell i Kowal 1994), en el qual es detalla, com a mostra, que un signe determinat (la *h* o *H*) arriba a utilitzar-se de vint-i-sis maneres diferents per indicar aspectes diferents (d'índole verbal, prosòdica, paralingüística i extralingüística). Aquesta heterogeneïtat, provocada per la desconnexió entre les múltiples disciplines i subdisciplines interessades per la llengua oral, es reflecteix no solament en la simbologia dels sistemes de transcripció (el seu aspecte més superficial, al capdavall), sinó en els criteris en què es basen, fet que porta a plantejar quins han de ser els seus fonaments i objetius.

4. Concepcions idealitzades de la transcripció

Una altra raó que ajuda a explicar l'heterogeneïtat de les solucions adoptades pels sistemes de transcripció és la pretensió que qualsevol sistema sigui capaç de respondre a totes les demandes (o possibilitats de transcripció) que desperta el discurs oral. En efecte, la concepció idealitzada més comuna -i també més ingènua o *naïf*- d'un sistema de transcripció s'erigeix sobre la creença en l'existència d'una transcripció "perfecta", a la qual s'exigeix la següent sèrie d'atributs:

- *Neutralitat* o *fidelitat*, és a dir, una transcripció no interpretativa, imparcial i objectiva, que no tergiversi els fenòmens propis del discurs oral.
- *Globalitat* o *complexitat*, o, en altres paraules, una transcripció completa, no parcial, que no simplifiqui cap fenomen.
- *Omnifuncionalitat*, que permeti usos (aplicacions) múltiples i diversos.
- *Claretat*, és a dir, un sistema d'aprenentatge i utilització "fàcil" (senzill, còmode), i per descomptat sense ambigüitats i tan econòmic com sigui possible.

Per últim, i com a previsible conseqüència de la creença en un mecanisme perfecte, també se sol demanar a un sistema de transcripció que la seva fonamentació i ús acabi sent universal (o, almenys, que no sigui idiosincràtic) i que, en un vessant més pràctic, no provoqui conflictes informàtics ni en la seva aplicació primera ni en el transport dels arxius.

El problema que es planteja davant una situació com la descrita, la recerca d'una espècie de "pedra filosofal" de la transcripció, és similar al d'una coneguda (i aparent) paradoxa lògica: ¿Què acaba succeint quan es llança un obús capaç de destruir-ho tot contra una guarnició indestructible?

5. Una concepció realista de la transcripció

La resposta al fals dilema anterior resulta bastant evident, donat que les condicions o premisses del problema són lògicament contradictòries: resulta impossible, almenys en el camp de la lògica, afirmar dues proposicions que no poden ser certes al mateix temps. De la mateixa manera, en el terreny que ens ocupa, cal adoptar com a primera premissa d'una transcripció que no es pot conjugar la complexitat amb la simplicitat en termes absoluts. Per consegüent, l'escapatòria de les aparents contradiccions que sorgeixen davant qualsevol transcripció només pot trobar-se amb solucions de compromís, amb equilibris entre opcions dicotòmiques, i tenint en compte en primer lloc els objectius de la transcripció, concebuda com qualsevol altra investigació científica.

Una concepció realista de la transcripció comença, per la raó apuntada, per admetre que es transcriu (i/o es descriu, com a complement) en funció del que es pretén analitzar. La superació de les contradiccions és, per tant, sempre, conjuntural, *ad hoc*, indissociable del *perquè* (i el *per a què*) del conjunt de la investigació. En aquest sentit, és lògic que els sistemes difereixin, com a mínim en abast i simplicitat, si s'utilitzen mètodes qualitatius i mostres reduïdes o limitades (com pot ser el cas, per exemple, d'una tesi doctoral) o si s'utilitzen mètodes més aviat quantitatius i mostres de grans dimensions (com pot ser el cas d'un corpus representatiu d'una comunitat lingüística).

A part la mostra i el mètode, els continguts o objectes d'estudi provoquen també eleccions diverses pel que es refereix al sistema de transcripció. Els múltiples camps abordables per l'anàlisi del discurs actual fan impossible escollir un sistema *a priori*, al marge dels interessos de l'estudi, els quals seran els encarregats de concretar-lo: aspectes estrictament verbals (i encara dins d'aquests, distingint, per exemple, els fonològics, lèxics i sintàctics), aspectes ideològics o argumentatius, aspectes interactius (per exemple els canvis de torn, els anomenats *back-channels* o "respostes mínimes"), aspectes psicolingüístics relacionats amb la producció lingüística (per exemple la interrelació verbal

- no verbal) o, per acabar i intentar evitar una llista encara més prolixa, aspectes psicosociològics (com la construcció del propi individu a través del discurs que va elaborant).

De fet, ni tan sols caldria recordar que els aspectes anteriors poden multiplicar-se, de manera que qualsevol proposta raonable de transcripció haurà d'incorporar, com un dels seus atributs imprescindibles, un mínim grau de flexibilitat. El procés de definició d'un sistema de transcripció (sigui desenvolupant-lo des de zero o, com sembla més lògic, adaptant-ne un de ja existent) és similar, en aquest sentit, als de delimitar una mostra de població o qualsevol corpus de dades, tots dos dependents també dels objetius de la investigació i impossibles de solucionar amb simples "receptes" o fórmules d'aplicació automàtica. L'absurd de la concepció idealitzada de la transcripció es desvela completament quan el procés es deixa de concebre com una simple rutina o pràctica (que s'aplica cegament) i es replanteja com un problema amb un vessant teòric i metodològic i un altre vessant més tècnic i mecànic (que sovint es presenta com l'únic important).

En el vessant més teòric, una transcripció raonable hauria de presentar uns trets justament oposats als de la concepció idealitzada anterior. En contra d'una suposada neutralitat, qualsevol transcripció ha d'acceptar-se com a inevitablement *interpretativa* (o, si es prefereix qualificar així, "teòrica"); en lloc de global, una transcripció adequada és, també necessàriament, *selectiva*, i en lloc d'omnifuncional, *pertinent*, associable amb l'objetiu de la investigació. Aquests trets es tradueixen en les *categories* dels fenòmens abordats per la transcripció i en els diferents *nivells* que s'hi poden distingir, i en aquest apartat són esperables solucions alhora *coherents* (amb la resta de la base teòrica), *fidels* (amb les dades de la interacció) i *flexibles* (que no comprometin en excés la pròpia investigació -en totes les fases- o altres investigacions que puguin aprofitar les dades aportades).

En el vessant més tècnic i mecànic, la transcripció es concreta sobretot en una *simbologia* o *notació* (i en consegüents opcions tipogràfiques). És en aquest vessant on les eleccions han de ser clares i econòmiques, senzilles i no ambigües, i tan respectuoses amb la tradició com sigui possible: no idiosincràtiques, no (innecessàriament) originals i, per tant, comunes i compartibles; tot això plegat implica també que s'adeqüin als tractaments informatitzats (en emmagatzematge, conversió i transport) i als sistemes internacionals estandarditzats de codificació.

6. Criteris de base de la transcripció

Els principis anteriors, exposats de manera sintètica, han estat presents sempre, de forma més o menys explícita i diversa, en les anàlisis dels sistemes de transcripció, almenys des del ja (quasi) clàssic article d'Ochs (1979), segurament el primer estudi en què s'emprèn una reflexió i discussió teòrica rigorosa sobre la transcripció, amb un títol certament inequívoc ("Transcription as theory"). En els escassos treballs posteriors sobre la qüestió, els principis varien, però mantenen un fons comú. Du Bois (1991: 77-97) presenta els criteris en forma de màximes i submàximes (a la manera habitual de molts estudis pragmàtics), i el plantejament es concreta en cinc principis "majors" i vint-i-tres de "menors":

(I) *Definició de categories* : Defineix bones categories ("Define good categories").

(1) Definir categories de transcripció que estableixin les distincions necessàries entre els fenòmens discursius.

(2) Definir categories suficientment explícites.

(3) Definir categories suficientment generals.

(4) Distingir els tipus de dades ("Contrast data types").

(II) *Accesibilitat* : Fes accessible el sistema ("Make the system accessible").

(5) Usar notacions familiars.

(6) Usar notacions motivades.

(7) Usar notacions fàcils d'aprendre.

(8) Segregar les notacions poc familiars ("Segregate unfamiliar notations").[7]

(9) Usar notacions que facilitin (i rendibilitzin al màxim) l'accés a les dades ("Use notations which maximize data access").

(10) Mantenir una aparença consistent (estable) al llarg dels diversos modes d'accés ("Maintain consistent appearance across modes of access").

[7] El criteri es refereix a la utilitat de mantenir separades o apartades del text bàsic de la transcripció les notacions no convencionals, especialitzades (per exemple dels aspectes prosòdics), de manera que no es faci més difícil la lectura i la comprensió del text transcrit pel que fa als aspectes estrictament verbals.

(III) *Robustesa* ("Robustness"): Elabora representacions robustes.
(11) Usar caràcters àmpliament disponibles.
12) Evitar contrastos invisibles.
(13) Evitar contrastos fràgils.

(IV) *Economia* : Elabora representacions econòmiques.
(14) Evitar notacions prolixes ("Avoid verbose notations").
(15) Usar notacions breus per a fenòmens d'alta freqüència.
(16) Usar notacions discriminables per a fenòmens interns de la paraula.
(17) Minimitzar notacions en (posició) interior de paraula.
(18) Usar l'espai de manera significativa.

(V) *Adaptabilitat*: Fes adaptable el sistema.
(19) Permetre transicions sense fissures (suaus, sense solució de continuïtat, "sense costures") entre els (diversos) graus de precisió ("Allow for seamless transition between degrees of delicacy").
(20) Permetre la integració sense fissures de categories de transcripció definides per l'usuari.
(21) Permetre la integració sense fissures de trets de presentació.[8]
(22) Permetre la integració sense fissures d'informació de classificació ("Allow for seamless integration of indexing information").[9]

(23) Permetre la integració sense fissures d'informació de codificació definida per l'usuari.

Ehlich (1993), creador del sistema de transcripció denominat HIAT (*Halbinterpretative Arbeitstranskriptionen*), ofereix un plantejament considerablement

[8] El criteri fa referència a la possibilitat d'emprar elements tipogràfics (en especial tipus de lletra: negreta, cursiva, etc.) en la presentació dels materials transcrits. Cal recordar que no tenen per què coincidir les transcripcions pròpies de l'especialista (el material d'anàlisi, contingut per exemple en una base de dades) amb les utilitzades, per exemple, per presentar una investigació de forma divulgativa o en una publicació.
[9] En especial, la necessària numeració dels materials, en concret de les línies del text transcrit (i, si procedeix, dels torns dels interlocutors).

més sintètic, reduïble als tres principis (o grups de principis) següents:

(a) Simplicitat i validesa ("Simplicity and validity").
(b) Bona "llegibilitat" i "esmenabilitat" ("Good readability and correctability").
(c) Mínim ensinistrament (entrenament) del transcriptor i de l'usuari ("Minimum of transcriber and user training").

Edwards (1993), per la seva part, repassa diversos sistemes anteriors i acaba agrupant els principis en tres grans classes:

(I) Principis de l'elaboració de categories:
(1) Les categories han de ser *sistemàticament discriminables.*
(2) Les categories han de ser *exhaustives.*
(3) Les categories han de ser *sistemàticament contrastives.*

(II) Principis de llegibilitat:
(4) Proximitat de fets relacionats.
(5) Separació visual de fets distints.
(6) Iconicitat temporal - espacial.
(7) Prioritat lògica.
(8) Marcatge mnemotècnic.
(9) Eficiència i caràcter compacte.

(III) Principis per al tractament computacional:
(10) Sistematicitat.
(11) Predictabilitat de la codificació.

O'Connell i Kowal (1994), per últim, revisen i critiquen els sistemes anteriors en funció també dels principis de base que han servit per a la seva elaboració. Les seves conclusions, en forma de suggeriments per a una notació adequada, constitueixen -no podia ser d'una altra manera, evidentement- un nou conjunt de criteris:

(1) Funció unitària de la notació.
(2) Integritat de les paraules com a unitats.
(3) Descripció de fenòmens no fonològics (en lloc de la seva "transcripció").
(4) Mesura exacta (eliminant notacions numèriques aproximades).
(5) Parsimònia ("Parsimony"). en el sentit de circumspecció o parquedat: transcriure només el que contribueixi sistemàticament a l'anàlisi de les dades i presentar al lector només el que faci intel·ligible l'anàlisi de les dades.

No hi ha cap dubte que aquest últim principi, apuntat en l'apartat precedent (6) sota la rúbrica de *pertinència*, conjuntament amb la distinció entre analista i lector, permet una transcripció bastant més econòmica (per a l'analista) i intel·ligible (per al lector) que moltes de les utilitzades actualment. Permet, així mateix, solventar l'aparent paradoxa entre complexitat i simplicitat que també ja ha estat descrita més amunt i que es fa present contínuament en tots els aspectes de la transcripció. Segons aquests autors, l'objetiu principal d'una transcripció no és la "llegibilitat" (o intel·ligibilitat), sinó la facilitat o capacitat d'ús ("usability") de les dades transcrites per als propòsits de l'anàlisi científica. En termes semblants pot tractar-se la distinció entre *transcriure* i *descriure* , útil i aplicable sobretot a fenòmens no lingüístics o prosòdics, i sempre en relació amb els objetius de la investigació. Tanmateix, i aquest és el punt més dèbil o perillós del principi, en cap cas no pot oblidar-se que les dades d'una investigació no haurien de ser tractades mai de forma tan parcial o idiosincràtica que se n'impedís o dificultés l'ús per part d'altres investigadors. La transcripció del discurs oral és una activitat suficientment costosa i conflictiva per prendre precaucions en aquest sentit, i cap comunitat científica pot permetre's el luxe de particularismes (en aspectes tan elementals) que obstrueixin el debat o deixin inaprofitable una font d'informació. Una transcripció adequada, per aquesta raó, acaba situant-se sempre en un terreny de compromís entre els objetius particulars i els *presumibles* (encara que sigui dins d'un camp d'investigació tan ampli com el discurs oral). Una transcripció fonètica on només es simbolitzessin (en alfabet fonètic) els sons vocàlics i es transcrivissin ortogràficament els consonàntics podria servir com a imatge per il·lustrar un cas aberrant que no hauria de tenir paral·lels en una transcripció discursiva.

7. Conclusions: la transcripció com a teoria aplicada

Evitar una nova idealització dels (múltiples) principis precedents, per no recaure en concepcions ja criticades, implica mantenir-los dins dels límits d'allò que és *raonable*, i que en un estudi científic amb base empírica significa el que pot ser demostrat de manera solvent. La discussió de fons -el debat realment *interessant* - sobre els sistemes de transcripció acaba sent sempre una discussió (*teòrica*) sobre els criteris en què han de fonamentar-se, no sobre la simbologia o notació particular que utilitzin. El sistema inclòs com a apèndix d'aquest treball representa una proposta concreta d'un sistema, basat en d'altres d'existents, que intenta conjugar els principis exposats dins dels límits de la raonabilitat, i que ha de ser avaluat -i per consegüent desestimat o millorat- en funció de la seva utilitat i productivitat de cara a la investigació sobre el discurs oral.

La transcripció és una eina *aplicada*, no una pura pràctica o rutina que s'aprèn i s'utilitza de forma automàtica. En aquest sentit, representa un exemple clar d'una concepció determinada de la lingüística aplicada, entesa com a teoria dirigida ("aplicada") a la solució de problemes (amb un vessant pràctic) plantejats per la praxi o ús lingüístic. Un sistema de transcripció del discurs oral cal que es plantegi per això mateix en totes les dimensions, amb els aspectes ja referits de categories, nivells i simbologia. És utòpic pensar en sistemes perfectes o fins i tot exhaustius, i qualsevol investigació ha d'adequar la transcripció de les dades als objetius proposats. L'avaluació del sistema -i la consegüent correcció- vindrà precisament de la seva productivitat, de la seva contribució a l'assoliment dels fins perseguits. Precisament aquest procés d'avaluació és gairebé per començar, donada l'escassa, encara, tradició en l'elaboració i ús de sistemes de transcripció adequats al discurs oral. Basta recordar que la primera versió de l'alfabet fonètic internacional (AFI o IPA) data de 1888, mentre que les primeres versions dels discursius se situen al voltant de 1970 (i les més precises i difoses en la dècada següent). Gairebé un segle de diferència és un lapse apreciable, que invita a la indulgència en la crítica dels sistemes actuals alhora que anima a contribuir a perfeccionar-los.

Bibliografia

ARRARTE, G. i LLISTERRI. J. (1994), *Informe sobre recursos lingüísticos para el español, (I): Corpus escritos y orales disponibles y en desarrollo en España*, Instituto Cervantes, Madrid.

ATKINSON, J. M. i HERITAGE, J. (eds.) (1984), *Structures of Social Action: Studies in Conversation Analysis*, Cambridge University Press, Cambridge.

BLOOM, L. (1993), *Transcription and Coding for Child Language Research: The Parts are More than the Whole*, dins J. A. Edwards i M. D. Lampert (eds.) (1993), *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 149-166.

BRIZ, A. i GÓMEZ, J. R. (1992), *Scheme of Study of Colloquial Spanish: Some Methodological Considerations*, dins F. Moreno-Fernández (ed.) *Sociolinguistics and Stylistic Variation*, "LynX", 3, pàg. 111-124.

BUTTON, G. i LEE, J. R. E. (1987), *Talk and Social Organization*, Multilingual Matters Clevedon.

CHAFE, W. L., DU BOIS, J. W. i THOMPSON, S. A. (1991), *Towards a New Corpus of Spoken American English*, dins K. Aijmer i B. Altenberg (eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Longman, Londres, pàg. 65-82.

CORTÉS, L. (1994), *Tendencias actuales en el estudio del español hablado*, Universidad de Almería, Almería.

CRIADO DE VAL, M. (1980), *Estructura general del coloquio*, SGEL, Madrid.

CROWDY, S. (1980), *Spoken Corpus Design and Transcription*. Document mecanografiat, 1991.

DU BOIS, J. W. (1991), *Transcription Design Principles for Spoken Discourse Research*, "Pragmatics", 1. pàg. 71-106.

DU BOIS, J. W., SCHUETZE-COBURN, S., CUMMING, S. i PAOLINO, D.(1993), *Outline of Discourse Transcription*, dins J. A. Edwards i M. D. Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 45-89.

EDWARDS, J. A. (1993), *Principles and Contrasting Systems of Discourse Transcription*, dins J. A. Edwards i M. D. Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 3-31.

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

EHLICH, K. (1993), *HIAT: A Transcription System for Discourse Data*, dins J. A. Edwards i M. D. Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 123-148.

GOODWIN, Ch. (1981), *Conversational Organization. Interaction between Speakers and Hearers*, "Pragmatics", 3, pàg. 181-209.

GUMPERZ, J. J. i BERENZ, N. (1993), *Transcribing Conversational Exchanges*, dins J. A. Edwards i M. D. Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Lawrence Erlbaum Associates, Hillsdale, pàg. 91-121.

MAcWHINNEY, B. (1991), *The CHILDES Project: Tools for Analyzing Talk*, Lawrence Erlbaum Associates, Hillsdale.

MARCOS MARÍN, F. (dir.) (1993), *Transcription Conventions Used for the Corpus of Spoken Contemporary Spanish*, "Literary and Linguistic Computing", 8, pàg. 283-292.

_____ (1994), Informática y humanidades, Gredos, Madrid.

O'CONNELL, D. C. i KOWAL, S. (1994), *Some Current Transcription Systems for Spoken Discourse. A Critical Analysis*, "Pragmatics", 4, pàg. 81-107.

OCHS, E. (1979), *Transcription as Theory*, dins E. Ochs i B. B. Schieffelin (eds.) *Developmental Pragmatics*, Academic Press, Nova York, pàg. 43-72.

PAYRATÓ, Ll. [1988 (2a. ed., 1990)], *Català col·loquial. Aspectes de l'ús corrent de la llengua catalana*, Universitat de València, València.

_____ (1992), *Pragmática y lenguaje cotidiano. Apuntes sobre el catalán coloquial* "Revista de Filología Románica", 9, pàg. 143-153.

_____ [1996, (en premsa)], *La variació funcional: els registres*, dins Ll. Payrató (ed.) Oral-ment. Estudis de variació funcional.

POYATOS, F. (1994), *La comunicación no verbal, 1. Cultura, lenguaje y conversación*. Istmo, Madrid.

PULLUM, G. K. i W.A. LADUSAW (1986), *Phonetic Symbol Guide*, The University of Chicago Press, Chicago.

QUILIS, A. (1984), *Bibliografía de fonética y fonología españolas*, CSIC / Instituto Miguel de Cervantes, Madrid.

SCHERER, K. R. i EKMAN, P. (1982), *Handbook of Methods in Nonverbal Behavior Research*, Cambridge University Press / Éditions de de la Maison des Sciences de l'Homme, Cambridge / Paris.

SPERBERG-McQUEEN, C. M. i BURNARD, L. (eds.) (1992), *Guidelines for Electronic Text Encoding the Encoding and Interchange*. Document TEI P2, Chicago / Oxford, capítol 3.4.

TANNEN, D. (1984), *Conversational Style: Analyzing Talking about Friends*, Ablex, Norwood.

THIBAULT, P. i VINCENT, D. (1990), *Un corpus de français parlé. Montréal 1984: historiques, méthodes et perspectives de recherche*, Université Laval, Quebec.

Apèndix: Presentació i simbologia d'un sistema de transcripció per a l'anàlisi del discurs oral

0. Presentació

Per bé que al capdavall poden acabar sent eternes, les propostes solen anar envoltades d'un aire de provisionalitat encantador i, és clar, una mica incòmode. Aquest és el cas almenys de la que es presenta a continuació, de manera que quan es publiqui segurament ja es deurà haver modificat en algun aspecte. Tot i així, si val la pena de fer-la pública és, precisament, al marge de la possibilitat d'oferir un sistema raonable, útil i bastant experimentat, pel benefici d'innovacions i millores futures.

Aquesta proposta de transcripció del discurs oral (monologal o conversacional) prové de la necessitat de disposar d'un sistema unificat i rigorós que, per una part, permeti "emmagatzemar" i manejar per escrit el que originalment era oral i, per una altra, faciliti estudis posteriors de diversa mena sobre el material transcrit (gramaticals, sociolingüístics, pragmàtics, discursius). L'origen de la proposta va ser un seminari sobre llengua col·loquial que va sofrir un reduït i selecte grup d'estudiants de la Universitat de Barcelona al llarg del curs acadèmic 1991-1992. Des d'aleshores, ha estat refinada i corregida en nombroses ocasions, gràcies a la col·laboració de molts estudiants i col·legues que l'han posada en pràctica o analitzat, als quals agraeixo públicament els seus suggeriments i comentaris.

0.1. *Criteris i consideracions inicials*

La proposta combina criteris de simplicitat i claredat elementals (en principi, un símbol diferent per a cada categoria o fenomen, i a cada categoria un símbol) amb criteris de respecte a la tradició i de facilitat de maneig automatitzat. Per aquestes raons (sobretot per l'última), en alguns casos s'ofereixen simbologies complementàries o alternatives, encara que se sol suggerir-ne només una. Les fonts fonamentals en què es basa la proposta són les aportacions de la tradició etnometodològica i els treballs de Gumperz i Berenz i de Du Bois (poden consultar-se, en especial, els recents articles de Gumperz i Berenz (1993) i Du Bois et al. (1993), i en general la recopilació sencera d'Edwards i Lampert (eds.) (1993). Tenint en compte les finalitats de les convencions de transcripció, és comprensible i justificable que la majoria de les presentades no siguin originals, sobretot pel que fa a la simbolització.

La proposta es basa en el que s'ha exposat en el text anterior en relació amb els tres aspectes essencials que cal distingir: categories, nivells i simbologia (o notació). Les categories de fenòmens susceptibles de ser transcrits o descrits figuren en la columna esquerra, amb epígrafs numerats, i la simbolització que els correspon a la dreta. A la columna esquerra es distingeixen també, abans de cada entrada i amb xifres romanes, tres nivells o graus de transcripció: (I) ample, (II) intermedi i (III) estret. Si bé el segon sembla el més neutre i en general el més aconsellable com a opció "no marcada" o "per defecte", l'elecció definitiva entre els tres no pot fer-se sense recordar l'explicació anterior sobre la necessitat d'adequar cada transcripció als objetius de la investigació en què s'enquadra; això pot portar a combinar, a la pràctica, els tres nivells, que hauran de ser entesos per consegüent com a tres guies més que com a tres estrats inamovibles. Se sobreentén que tota precisió pròpia d'un nivell superior ho és automàticament dels precedents, i les entrades marcades amb un interrogant (?) s'han de considerar simplement optatives, és a dir, susceptibles de ser afegides a qualsevol nivell (per exemple la transcripció fonètica o la de fenòmens paralingüístics). En l'apartat (10) s'exemplifiquen, de manera breu, algunes convencions, i en l'apartat (11) es recull un conjunt de consideracions finals i recomanacions sobre alguns dels problemes més freqüents que sorgeixen en el procés de la transcripció.

1. Parlants, escenari i escena

1.1. Parlants, escenari i escena:

- (I): descripció inicial esquemàtica
- (II): descripció etnogràfica (model S-P-E-A-K-I-N-G)

1.2. Identificació dels parlants: abreviatura amb majúscula inicial

1.3. Torn de paraula: :

1.4. Continuació de torn: &

2. Seqüències comunicatives (verbals / no verbals)

2.1. Seqüències verbals:

- (I): transcripció ortogràfica convencional en línies numerades al marge esquerre
- (II): sense majúscules (excepte en noms propis) i sense signes de puntuació

2.2. (?) Mateixes seqüències verbals amb un grau de precisió superior o en transcripció fonètica: columnes paral·leles

2.3. (?) Complements no verbals simultanis i no esporàdics de

les seqüències verbals (p. ex. la mirada o la postura):	simbologia addicional en línies consecutives
2.4. (?) Traducció a una altra llengua de les seqüències originals:	línies consecutives i tipus de lletra diferents
2.5. (?) Seqüències comunicatives en més d'una llengua, amb llengua base o sense:	sense marques, o bé (**L2**) text (**L2**) / (**L3**) text (**L3**) o bé tipus de lletra diferents
2.6. (?) Discurs reproduït (citat) i referències metalingüístiques:	" "

3. Unitats

3.1. Unitats entonatives o grups tonals:

- (II): signes de seqüències tonals terminals (/ \ _), i salt de línia (tants grups per línia com sigui possible sense que cap quedi escapçat, tret que sigui més llarg del que permet una línia sencera, iniciada sempre al marge esquerre)

- (III): salt de línia (un grup per línia)

3.2. (III) Grup tonal truncat: --

3.3. Separació de paraules: espai (blanc)

3.4. Paraula truncada: -

4. Aspectes prosòdics

4.1. Seqüències tonals terminals:

- (I) - Descens: . o bé ,
  - Ascens: ¿ ?
  - Manteniment: , o bé ...

- (II)	- Descens:	\	o bé	↓
	- Ascens:	/	o bé	↑
	- Manteniment:	_	o bé	-->

4.2. - (I) Èmfasi: majúscules o bé tipus de lletra

- (II) Ascens mantingut: {(A) text afectat}

- (II) Descens mantingut: {(B) text afectat}

4.3. - (?) Accent principal: *

- (?) Accent secundari: '

4.4. - (II) Allargament (simple): :

- (III) Allargaments (diferents graus): : / :: / :::

5. Pauses i encavalcaments

5.1. (II) Enllaç entre torns:	**(0)**
5.2. - (I) Pausa breu (p < 5"):	, o bé .
- (I) Pausa llarga (p ≥ 5"):	. **(pausa)**
- (II) Pausa breu (p < 1"):	(..)
- (II) Pausa mitjana (1"≤ p < 3"):	(...)
- (II) Pausa llarga (p ≥ 3"):	(... duració en segons)
- (III) Micropauses (0,1"≤ p < 0,3"):	(.)
- (III) Pausa breu (0,3"≤ p < 1"):	(..)
- (III) Pausa mitjana (1"≤ p < 3"):	(... duració en segons)
- (III) Pausa llarga (p ≥ 3"):	(... duració en segons)
5.3. - (II) Encavalcaments:	**[ ]**
- (II) Encavalcaments múltiples:	**[1 1]** o bé **[[ ]]**

6. Aspectes vocals

6.1. - (II) Tempo ràpid ('accelerat', 'allegro'):	{**(AC)** text afectat}
- (III) Tempo molt ràpid ('presto'):	{**(ACC)** text afectat}
- (II) Tempo lent ('desaccelerat', 'largo'):	{**(DC)** text afectat}
- (III) Tempo molt lent ('grave'):	{**(DCC)** text afectat}
6.2. - (II) Intensitat forta ('forte'):	{**(F)** text afectat}
- (III) Intensitat molt forta ('fortissimo'):	{**(FF)** text afectat}

- (II) Intensitat suau ('piano'): {**(P)** text afectat}
- (III) Intensitat molt suau ('pianissimo'): {**(PP)** text afectat}

6.3. - (I) Riure: **(riure)**
- (II) Riure simultani amb elements verbals: {**(@)** text afectat }
- (III) Riure no simultani: **@ / @ @ / @@@** (un signe per síl·laba)

6.4. - (III) Inhalació: **(INH)**
- (III) Exhalació: **(EXH)**

6.5. Sons paralingüístics
- (I): formes escrites (literàries) convencionals

- (II) - Assentiment: **mhm / mhm mhm / hi / hi hi / ha / ha ha**
  - Dubte: **m:: / a:: / e::**
  - Apel·lació: **txist / psist**
  - Demanda de silenci: **ss::**
  - Desacord: **ntx / boah / ps:**
  - Valoració: **pse / psepse**
  - Èmfasi: **fu / buf**
  - Altres: formes escrites (literàries) convencionals

6.6. Altres (no simultanis): (descripció del fenomen i duració)

6.7. Altres (simultanis): {**(fenomen)** text afectat}

7. Aspectes gestuals

7.1. Fenòmens no simultanis: (descripció del fenomen)

7.2. (II) Fenòmens simultanis: {**(fenomen)** text afectat}

7.3. (III) Fenòmens esporàdics, no simultanis i repetits: {(**X**= fenomen) text **X** text **X**}

7.4. Altres: (descripció i duració)

8. Regularitzacions, comentaris i fragments conflictius

8.1. Regularitzacions ortogràfiques: ("text regularitzat")

8.2. (II) Supressió d'elements: (elements suprimits)

8.3. (II) Transcripció fonètica (de fragments): text (/text transcrit/)

8.4. Comentaris del transcriptor: (( ))

8.5. Fragments inintel·ligibles:

- (I): **(incomprensible)**
- (II) - Paraules: **x / xx / xxx** (una **x** per síl·laba)
- Fragments superiors: **xxX** (duració en segons) **Xxx**

8.6. Suposicions, fragments incerts: {(??) fragment incert}

8.7. Altres: ((comentari del transcriptor)) o bé signes complementaris

9. Signes complementaris per a codificacions diverses

9.1. (?) Èmfasi (gràfica) de fenòmens:	tipus de lletra, p. ex. cursiva, negreta o subratllat
9.2. (?) Codificacions fonològiques, morfològiques, sintàctiques o pragmàtiques:	**<CF>** text **<CF>**, **<CM>** text **<CM>**, **<CS>** text **<CS>**, **<CP>** text **<CP>**
9.3. (?) Codificacions d'altres menes:	= = $ $ + + ¡ !
9.4. (?) Alternatives de signes aconsellables (p. ex. en el cas de problemes informàtics):	= = en lloc de [ ] < > en lloc de { }

10. Exemples d'algunes convencions de transcripció

10.1. Seqüències terminals de l'entonació:

Equivalent ortogràfic aproximat	*Transcripció*
(1) Si vol que vingui.	(1) A: si vol que vingui \
(2) Si vol que vingui?	(2) A: si vol que vingui /
(3) Si vol que vingui...	(3) A: si vol que vingui _

10.2. Grups tonals (o unitats entonatives) i seqüències terminals de l'entonació sense pauses (elocució ràpida):

Equivalent ortogràfic aproximat	*Transcripció*
(1) No, la farà bé.	(1) A: no \ la farà bé \
(2) No? La farà bé?	(2) A: no / la farà bé /
(3) No... La farà bé...	(3) A: no_ la farà bé _

És possible qualsevol combinació entre les anteriors, p. ex.:

(4) No... La farà bé?	(4) A: no _ la farà bé /
(5) No? La farà bé.	(5) A: no / la farà bé \
(6) No. La farà bé...	(6) A: no \ la farà bé _

10.3. Grups tonals, seqüències terminals de l'entonació i pauses o allargaments (elocució habitual):

Equivalent ortogràfic aproximat	*Transcripció*
(1) No, la farà bé.	(1) A: no \ (..) la farà bé \
(2) No. La farà bé.	(2) A: no \ (...) la farà bé \
(3) No la farà bé.	(3) A: no la farà bé \
(4) Nooo... la farà bé.	(4) A: no:: \ la farà bé \

10.4. Truncament de grup tonal i de paraula:

Equivalent ortogràfic aproximat	*Transcripció*
(1) Si pen- penses que no vindrà...	(1) A: Si pen- penses que no vindrà_
(2) Si penses, si penses que no vindrà...	(2) A: si penses-- si penses que no vindrà_
(3) Si pen- si penses que no vindrà...	(3) A: si pen- -- si penses que no vindrà _

10.5. Èmfasi (entonatiu, d'intensitat o de ritme), riure simultani i comentaris:

Equivalent ortogràfic aproximat	*Transcripció*
(1) Vés a fer punyetes! (amb un to alt, agut)	(1) A: {(A) vés a fer punyetes \}
(2) Vés a fer punyetes! (amb un to baix, greu)	(2) A: {(B) vés a fer punyetes \}
(3) Vés a fer punyetes! (*forte*, cridant)	(3) A: {(F) vés a fer punyetes \}
(4) Vés a fer punyetes! (*piano*, fluix, suau)	(4) A: {(P) vés a fer punyetes \}
(5) Vés a fer punyetes! (accelerat, ràpid)	(5) A: {(AC) vés a fer punyetes \}

(6) Vés-a-fer-pu-nye-tes! (desaccelerat, lent)	(6) A: {(DC) vés a fer punyetes \}
(7) Vés a fer punyetes! (rient)	(7) A: {(@) vés a fer punyetes \}
(8) Vés a fer punyetes! (fent un gest)	(8) A: {(botifarra) vés a fer punyetes \}
(9) Vés a fer punyetes! (després d'un gest)	(9) A: (somriure) vés a fer punyetes \
(10) Vés a fer punyetes! (a algú en particular)	(10) A: ((a F)) vés a fer punyetes \

11. Consideracions finals

11.1. *Informacions prèvies*

És indispensable, en primer lloc, que qualsevol transcripció vagi encapçalada almenys per una descripció sumària de l'escena i de la situació (física) de l'acte comunicatiu, dels participants que intervenen en l'acte i dels temes tractats. En descripcions més detallades, pot resultar útil l'aplicació d'un model com el de l'etnografia de la comunicació (la xarxa *ètica* de components dels fets comunicatius que sol presentar-se amb l'acròstic *S-P-E-A-K-I-N-G*).

11.2. *Signes de puntuació*

Excepte en nivells elementals o en presentacions divulgatives, no és aconsellable utilitzar en les transcripcions els signes de puntuació tradicionals, donat que indueixen a interpretacions errònies basades en els hàbits de lectura; tampoc majúscules, tret del cas de noms propis (abreujats, parafrasejats o xifrats, si per la naturalesa de la conversació resulta preferible mantenir l'anonimat de persones i llocs). Malgrat tot, si a partir del nivell (II) es decideix presentar la transcripció amb símbols que coincideixin amb signes de puntuació ortogràfics, almenys cal recordar als usuaris que el seu valor no té per què coincidir amb el tradicional del seu ús en textos escrits; així, per exemple, els signes "?", "." i "," no reproduiran el seu valor ortogràfic habitual (pregunta o pausa), sinó que indicaran, respectivament, seqüències terminals de l'entonació ascendents (en preguntes de resposta *sí/no* i algunes altres), descendents (en d'altres tipus de pregunta i en les asseveracions) o de manteniment (quan s'indica que falta informació o es deixa un enunciat en suspens). De fet, les possibles confusions (interpretacions "ortogràfiques" de

la transcripció) s'eviten utilitzant símbols específics per a les pauses (amb diversos graus de precisió) i per a les entonacions ("/", "\" i "_", els més transportables en tractaments automatitzats, o bé "↑", "↓" i "--->", més interpretables o "llegibles").

11.3. *Informacions complementàries i comentaris del transcriptor*

Totes les informacions complementàries que calgui afegir al text transcrit poden presentar-se com a comentaris del transcriptor (que en aquesta faceta es mostra clarament com l'*analista* que en realitat sempre és); així, per exemple: *li vaig dir a J ((nom de dona, hipocorístic)) que se n'anés a S ((poble de la costa de Tarragona))*. Lògicament, no tot el que apareix en actes comunicatius orals pot ser previst, regulat ni transcrit, i per consegüent en determinats casos s'ha d'actuar amb sentit comú i adoptar les solucions que es considerin més convenients en relació amb les finalitats de la transcripció. La típica pregunta de què cal transcriure o comentar només pot contestar-se dient que tot allò que el transcriptor, com a analista, consideri pertinent per al desenvolupament de l'acte comunicatiu, per a la intel·ligibilitat del text transcrit i per a les anàlisis posteriors.

11.4. *Transcripció fonètica*

Acostuma a resultar útil disposar d'una versió del discurs oral en transcripció fonètica, seguint les normes de l'Alfabet Fonètic Internacional o d'altres de també tradicionals. Si es tracta de fragments breus, poden incorporar-se a la transcripció ortogràfica. És aconsellable, en la transcripció fonètica, mantenir la separació de paraules i totes les notacions pròpies de la transcripció ortogràfica rellevants per a la investigació (aspectes prosòdics, pauses, encavalcaments, etc.). Si es disposa de les dues versions, ortogràfica i fonètica, les possibles comparacions es veuen facilitades mantenint el mateix text per línia i la mateixa numeració.

11.5. *Aspectes prosòdics, unitats entonatives i pauses*

El capítol dels aspectes prosòdics, les unitats entonatives o grups tonals i les pauses és un dels més complexos en la transcripció, tant per les dificultats de percepció com per la falta d'estudis sobre aquests fenòmens. Pel que fa a l'entonació, en el sistema de transcripció presentat s'ha simplificat, i només es preveuen marques de seqüències terminals (les tècniques poden consultar-se a la bibliografia apuntada). L'entonació

s'assenyala abans de les pauses i al final de cada unitat entonativa, i es poden afegir marques en els casos en què l'analista ho consideri convenient a fi d'evitar lectures o interpretacions ambigües o errònies. La naturalesa dels grups tonals (o *informatius*, o *tonals-informatius*, etc.) no és, encara avui, clara, cosa que no significa tanmateix que els parlants no elaborin el producte verbal a partir d'unes unitats (denominades de formes diverses) que es van juxtaposant, i el reflex de les quals en el text transcrit facilita una reproducció més fidel de la cadena original. Les pauses, per últim, es poden expressar en dècimes de segon si es disposa dels instruments adequats; si no, poden reproduir-se de manera impressionista, determinant primer la més breu i, a continuació, com a unitats perceptibles subsegüents, la mitjana i les més extenses. Les pauses s'assenyalen després dels símbols d'entonació, i s'atribueixen a un parlant (pausa dins del torn o intervenció) o al conjunt d'interlocutors (pausa entre torns o interrupció, i no atribuïble a cap parlant en concret, que figura en línia a part).

11.6. *Aspectes paralingüístics*

Els aspectes no verbals i no vocals (proxèmics i gestuals, incloent-ne la dimensió temporal o "cronèmica") han de ser transcrits amb sistemes especials si es desitja analitzar-los en detall, sigui de forma autònoma o en la seva interacció amb els aspectes vocals i verbals (vid. les referències bibliogràfiques esmentades anteriorment). En altres tipus de transcripció, com els presentats en la proposta, es poden descriure (més que no transcriure), de la manera apuntada més amunt, els comportaments necessaris per entendre el desenvolupament de la interacció comunicativa, en particular la disposició física dels parlants (en la fitxa inicial) i el seus gestos, postures o desplaçaments rellevants (en l'interior del text, directament o com a comentaris del transcriptor, i distingint si es donen simultàniament o no amb els elements verbals: vid. exemples (8)-(10) de 10.5). Els aspectes no verbals vocals són més senzills de descriure (els extralingüístics, com la tos o un esternut) o de transcriure (els pròpiament paralingüístics, com la intensitat o el ritme), tal com ja s'ha apuntat també en la proposta, i distingint igualment si són simultanis o no amb la verbalitat (amb l'ús de les claus). Si procedeeix, es pot afegir entre parèntesis la duració dels fenòmens, i el sistema és ampliable a aspectes vocals com la qualitat de la veu, per exemple en les citacions o en les seqüències d'atenuació o molt expressives. Respecte als múltiples "sorolls" o "gestos" vocals, etapes prèvies de les

onomatopeies, la transcripció només pot ser aproximada, a l'espera d'estudis que en verifiquin el caràcter sistemàtic i en proposin un sistema coherent de transcripció. El sentit comú convida a proposar formes transcrites diferents per a fenòmens vocals funcionalment diferents (com alguns dels ja apuntats), i a tenir en compte les convencions, més o menys tradicionals, utilitzades en els escrits (novel·les, guions, còmics, etc.) per representar-los.

11.7. *Aspectes de gramàtica normativa*

Una transcripció ortogràfica, en lloc de fonètica, impedeix reflectir moltes pronunciacions ("no ortogràfiques") i implica, naturalment, una pèrdua d'informació. Si no es recorre directament a la transcripció fonètica, que constitueix com és obvi la millor solució però amb un cost notable, els parèntesis poden utilitzar-se per a casos concrets, o bé incloent la transcripció fonètica d'un fragment immediatament després de la seva aparició (per exemple per indicar la pronunciació d'estrangerismes, sigles, etc.), o bé per indicar casos d'elisió (per exemple *emb(a)rassada*, *(ha)gués*, etc.). Aquest procediment ha d'usar-se, tanmateix, amb precaució, per no dificultar en excés la lectura, i no és recomanable per als casos predictibles i sistemàtics. Es poden adoptar solucions semblants, *ad hoc*, per a pronunciacions dialectals, que no siguin pròpies de la llengua estàndar o que no siguin previsibles a partir de la versió ortogràfica, tot plegat en funció dels interessos de l'estudi.

11.8. *Recomanacions pràctiques*

A més a més dels aspectes teòrics i aplicats ja tractats, els següents consells poden servir per facilitar el procés de la transcripció:

(1) Abans de la transcripció:

- Comprovar que *tot* l'enregistrament (i no només una part, per exemple l'inici) té suficient qualitat (sonora) per poder ser transcrit d'acord amb els instruments disponibles (caset, vídeo, dictàfon o ordinador i programes d'anàlisi del so).
- Fer una còpia de seguretat de l'enregistrament original, que serà l'utilitzat per a la transcripció (llevat de la necessitat de comprovar seqüències conflictives).
- Sol·licitar permís als participants en l'acte comunicatiu per poder utilitzar (amb finalitats científiques) el text transcrit, i decidir el grau d'anonimat que se seguirà.

(2) Durant la transcripció:
- Fer còpies de seguretat dels arxius informàtics.
- Seguir rutines per als aspectes mecànics, prestant atenció en cada fase a un sol fenomen (p. ex. els aspectes verbals, l'entonació, les pauses, els encavalcaments, etc.).
- Numerar les línies del text (no les intervencions dels interlocutors) amb un procediment mecànic, no de forma manual.
- No simplificar de manera injustificada cap aspecte del discurs enregistrat, però no hipotecar el procés per fenòmens conflictius irrellevants.
(3) Després de la transcripció:
- Repassar de forma mecànica les diverses categories transcrites, atenent en cada fase a una de sola: aspectes verbals (precisions, regularitzacions), unitats entonatives (delimitació), marques d'entonació, pauses (duració precisa), encavalcaments, etc.
- Repassar, si és possible primer de forma independent i després en equips o grups de transcripció, els aspectes conflictius (amb la versió original de l'enregistrament).

El processament computacional de corpus. Tècniques automàtiques d'anàlisi morfològica i sintàctica

Toni Badia

(Universitat Pompeu Fabra)

1. Introducció[1]

En el moment en què una de les activitats més habituals en el camp de la recerca dels departaments de lingüística i filologia és la compilació de corpus per a l'estudi de la llengua que hi és reflectida, convé observar quines eines la lingüística computacional ha anat elaborant en els seus cinquanta anys mal comptats d'existència. D'entre les moltes coses que s'hi han produït, en podem trobar diverses (moltes, m'atreviria a dir) que ens poden ser enormement útils en l'anàlisi automàtica de corpus.

En aquest article, doncs, pretenem fer un repàs d'algunes de les possibilitats que ofereix la lingüística computacional respecte a l'anàlisi automàtica de textos reals, amb la finalitat de mostrar-les a través d'alguns exemples característics i significatius.[2] Queda

---

[1] Voldria assenyalar el meu agraïment als organitzadors del *2n Col·loqui Lingüístic de la Universitat de Barcelona* (celebrat el 19 de desembre de 1994), Emili Boix, Rosa M. Lloret, Mercè Lorente i Lluís Payrató, per la possibilitat que m'oferiren de presentar llavors i publicar ara aquesta ponència. Aquesta ha estat una oportunitat excepcional per a reflexionar sobre la meva activitat en lingüística computacional en funció de les necessitats explícites dels qui es dediquen al coneixement i descripció de la llengua (i, especialment, de la llengua catalana). A més, voldria agrair a Mercè Lorente i Toni Tuells la lectura detallada que han fet d'una primera versió d'aquest article; amb els seus comentaris han contribuït al fet que la versió que en teniu a les mans contingui molts menys errors i inadequacions dels que originalment hi havia; naturalment, dels que hi romanen només en sóc responsable jo mateix.

[2] Els programes i projectes mencionats en aquest article han de ser vistos com exemples de les estratègies i plantejaments que es discuteixen. En cap cas, s'ha de pensar que un programa és bo o millor que altres pel simple fet de ser mencionat aquí.

molt lluny de les nostres possibilitats (d'espai, però també de coneixement) el poder-ne fer un llistat exhaustiu. Nogensmenys una observació acurada d'allò que s'està fent ens permetrà de valorar la conveniència d'adoptar alguna d'aquestes tècniques en els nostres treballs amb corpus.

L'article s'estructura en quatre parts ben desiguals. En primer lloc, es plantegen unes consideracions generals sobre els corpus i la lingüística computacional. A continuació, en les dues seccions centrals de l'article, s'emprèn l'anàlisi dels programes d'etiquetat morfosintàctic, primer, i els de tractament sintàctic, després. Finalment, l'article acaba amb unes breus conclusions.

2. Consideracions generals

En aquest moment ningú posaria en dubte que per als estudis lingüístics es necessiten corpus de materials recollits d'acord amb uns criteris científics adequats i que, a més, cal saber processar-los de manera que se'n pugui extreure fàcilment la informació que ens n'interessa. Després d'un seguit d'anys en què ha semblat que la intuïció del lingüista era suficient per a l'observació dels fenòmens de la llengua, ara apareix clarament la necessitat de complementar aquesta font d'informació amb l'observació directa de textos produïts realment. Sense ella resulta molt difícil evitar la tendència cap a una descripció lingüística de laboratori, en la qual es dóna prioritat a les qüestions que comporten un repte per als plantejaments teòrics adoptats i, en canvi, es deixen de banda una sèrie d'aspectes que són essencials en l'ús diari de la llengua. La compilació i estudi dels corpus permet acostar-se a la llengua real, tal com és usada efectivament pels parlants.

Aquesta necessitat s'ha vist reforçada per la relativa facilitat amb què des de fa uns pocs anys es poden guardar els textos orals. La simple observació superficial d'aquests textos ha fet adonar de la gran distància que hi ha entre la llengua real i la descripció lingüística tal com queda establerta en les obres gramàtiques descriptives.[3] Però fins i tot des de la perspectiva de la llengua escrita tenim un desconeixement bastant gran de com és usada la llengua realment. Hi ha una sèrie d'elements de la llengua que són francament

[3] Noteu que aquest no és un fenomen restringit a llengües que es troben en una situació general d'inestabilitat com la catalana; vegeu, per exemple, Blanche-Benveniste i Temple (1986) sobre el francès.

poc coneguts. En general tots els aspectes que fan referència a l'ús de les expressions i unitats lingüístiques, tant sintagmàtiques com lèxiques, han estat poc estudiats; però també trobem estructures que són poc conegudes, tot i ser enormement presents en els textos. Un bon exemple el constitueixen els patrons de modificació, especialment en relació a les estructures verbals: tot i que pràcticament no hi ha una oració sense un modificador d'un o altre tipus, aquest és un aspecte (sintàctic i semàntic) que ha estat totalment abandonat pels estudis lingüístics generals. Així ens trobem amb la paradoxa següent: es coneixen amb molt de detall els mecanismes que permeten l'extracció de complements (en contextos interrogatius o d'oracions de relatiu) d'un nivell de profunditat indeterminat, quan la seva presència en textos és realment molt minsa; i en canvi es desconeixen molts dels factors que intervenen en les estructures de modificació, que són presents sempre, en qualsevol tipus de text.[4]

En certa manera, tot i que des de fa força anys ja els lingüistes afirmem (i ens ho creiem, és clar) que la lingüística ha de ser descriptiva i no prescriptiva, de fet fins ara hem estat mancats d'eines que ens permetessin de conèixer realment com és la llengua que s'usa. En certa manera, la voluntat descriptivista del lingüista s'ha vist dificultada (si no impossibilitada del tot) per la impossibilitat de fer una descripció acurada. Això ha fet que en les nostres descripcions haguem hagut de recórrer sobretot a les nostres intuïcions.

Això que ha passat en lingüística general, també ha tingut lloc en el camp de la lingüística computacional, on el problema és potser encara més greu, atès que un dels objectius de l'àrea és produir programes que permetin de tractar automàticament el llenguatge natural. La majoria de sistemes de processament del llenguatge natural s'han basat del tot en els plantejaments de la lingüística teòrica, de manera que els biaixos que observem en els estudis d'aquesta s'han traslladat gairebé fil per randa a aquella. A més, la lingüística computacional, en les seves dimensions més teòriques, s'ha dedicat a tractar qüestions especialment difícils en el camp del processament del llenguatge natural.

Així, s'ha estudiat la complexitat de processament necessària per a tractar els diversos fenòmens lingüístics que apareixen en les llengües naturals; d'aquesta manera, per

[4] En un estudi recent, realitzat com a pas previ a la construcció d'una gramàtica computacional en el marc del projecte europeu LSGRAM (LRE 61029—1994-5), es va posar en evidència que, en un petit corpus periodístic en anglès, no apareixia cap oració de relatiu de profunditat superior a 1, mentre que hi havia pràcticament un adjunt a totes les oracions.

exemple, s'ha estudiat força el tractament de les dependències de llarga distància (pròpies de les oracions de relatiu, interrogatives i amb altres extraccions), ja que comporten una indeterminació en els seus lligams amb la resta dels constituents de l'oració a la qual pertanyen.

També, s'ha intentat determinar quina és la classe de gramàtica necessària per a tractar el llenguatge natural; en concret, els estudis s'han centrat en analitzar fins a quin punt les gramàtiques lliures de context (és a dir, relativament simples des del punt de vista formal) són suficients per a tractar formalment el llenguatge natural. D'aquesta manera s'ha vist que hi ha llengües algunes construccions de les quals no poden ser tractades amb gramàtiques lliures de context; l'holandès i algun dialecte del suís alemany tenen el que se solen anomenar dependències creuades, que impliquen seqüències de complements, primer, i de verbs, després, que no poden ser tractades linealment: el primer nom no depèn de l'últim verb, sinó del primer, el segon ho fa del segon, etc.; d'aquesta manera, si intentem construir l'arbre d'anàlisi d'aquestes construccions ens sortiran relacions creuades.

Finalment, s'ha dedicat un gran esforç, especialment els últims anys, a estudiar l'organització i estructuració de la informació lingüística. Els sistemes habituals de codificació i tractament de la informació lingüística en les gramàtiques formals s'han mostrat inadequats. En aquest camp, s'han estudiat, i s'hi han proposat nous plantejaments, la distribució de tasques entre el lèxic i la gramàtica, la conveniència de convertir les regles gramaticals en principis de caràcter més general, l'adopció de formats nous per a la codificació de la informació lingüística, etc.

Totes aquestes qüestions són enormement importants i han contribuït en gran mesura als avenços que s'han produït en la lingüística computacional dels anys noranta. De totes maneres, no s'ha prestat prou atenció a qüestions que se suposa que tenen menys envergadura teòrica, però que són fenòmens altament freqüents en els textos reals i que, per tant, convé que tot sistema realista de processament tingui resoltes. En aquesta línia, d'entre els que no han estat tractats adequadament, sobresurten les associacions lèxiques, les ambigüitats de tot tipus i la desviació de les regles. Les associacions lèxiques són un fenomen altament estès en els textos reals; de fet el paper importantíssim que tenen en la selecció lèxica inherent a la producció lingüística no ha estat observat fins que no s'han aplicat les tècniques d'observació i anàlisi de corpus. Notem que el concepte d'associació

lèxica engloba tant les relacions lèxiques anomenades col·locacions o idiomatismes, com les preferències de coaparició que tenen unes peces lèxiques determinades (en funció, en part, del tipus de subllenguatge o del tema de què es parla).

El problema de l'ambigüitat és un dels més grans que hi ha en els sistemes de processament del llenguatge natural. En general, es distingeix entre l'ambigüitat lèxica i l'estructural. La primera apareix naturalment quan hi ha mots que tenen més d'un sentit, però el procés creixent de lexicalització de la descripció gramatical crea ambigüitats lèxiques d'origen sintàctic. La segona ocorre quan hi ha estructures sintàctiques que poden ser analitzades de dues maneres; a vegades les dues anàlisis corresponen a significats diferents (és a dir, a una ambigüitat real), però a vegades simplement són resultat de la manera com s'han formulat les regles.

Tot tipus d'ambigüitat és altament problemàtic. Evidentment, un dels problemes és el de buscar una única solució (una única anàlisi) al problema de la interpretació dels fragments de text; en aquest sentit cal saber recórrer al tipus d'informació adequat per a cada classe de problema. Però, a més, l'ambigüitat en el context del processament del llenguatge natural comporta una disminució molt gran de l'eficiència del sistema, perquè implica un augment considerable de l'espai de cerca, de memòria i de processament en general. La simple presència de més d'una entrada lèxica per a alguns mots de la cadena que estem analitzant ja té conseqüències: molt sovint no és fins bastant endavant en el processament que es pot resoldre l'ambigüitat, de manera que mentrestant tenim dues anàlisis que es van arrossegant; i, naturalment, quan es creuen dues situacions com aquesta les anàlisis possibles ja no són dos, sinó quatre (dos a la segona), quan són tres, en resulten vuit (dos a la tercera), i així successivament. És fàcil d'adonar-se que això comporta un augment considerable dels recursos del programa. Però, si a més de les ambigüitats d'origen lèxic hi afegim les de caràcter estructural, el resultat pot ser realment explosiu.[5]

[5] Pot resultar il·lustratiu de recordar una de les meves primeres experiències en aquest camp. En els primers estadis de creació d'una gramàtica d'estructura sintagmàtica per a la sintaxi superficial castellana, l'anàlisi de l'oració *En el proceso de reconstrucción de posguerra Europa se ha ido quedando atrás respecto a otras potencias industriales* donà, davant els nostres incrèduls ulls, més de cent resultats; i la nostra gramàtica no contenia cap error, si de cas només una bona dosi d'ingenuïtat! Això no és simplement una anècdota: reflecteix d'una manera força exacta el tipus de problema amb què han d'afrontar els redactors de gramàtiques.

L'adaptació de les gramàtiques per a tractar textos reals (i no simplement frases de laboratori) implica necessàriament haver de donar resposta al problema de l'anàlisi d'oracions (parcialment) agramaticals o incorrectes. De fet, són moltes les situacions en què apareixen verbs que no tenen tots els seus complements obligatoris, o predicats els complements dels quals no compleixen les restriccions de selecció que aquells els imposen. En tots aquests casos, ens trobem davant desviacions del que és considerat gramatical. Ara bé, la solució no pot consistir simplement en relaxar les regles per a admetre aquest tipus de desviacions, ja que llavors ens trobaríem amb què la gramàtica admetria com a bones oracions que no voldríem que acceptés i, al capdavall, resultaria més ambigua encara.

No obstant, el processament de textos reals és necessari. No podem pensar en una millora del tractament informàtic del llenguatge, sense avançar en aquest sentit. En general, podem establir dues grans finalitats a l'hora de processar els textos: o bé per tal de manipular la informació lingüística que contenen, o bé per tal d'extreure'n part de la informació que contenen. En la primera línia, es troben els programes que impliquen una anàlisi sintàctica i semàntica del text: els programes de traducció automàtica, els de recuperació de la informació, els d'accés a sistemes experts en llenguatge natural, etc. En la segona línia, en canvi, trobem els programes d'anotació de corpus. Tot i que tradicionalment s'han considerat dues línies de treball (fins i tot, dues tradicions) diferents, si no oposades, no és convenient de mirar-les així, sinó de veure-hi els punts de contacte i les interrelacions, ja que parts dels treballs desenvolupats en una tradició són útils per a l'altra. Per altra banda, en adoptar una perspectiva global davant el processament del llenguatge natural, es veuen molt més clarament els límits i els mèrits o possibilitats de cada aplicació concreta.

L'anotació de corpus és certament el primer pas per a poder-ne extreure informació; altrament, seria impossible de fer-hi observacions de caràcter general. Dit d'una altra manera, la possibilitat de processar un corpus a un nivell o altre està estretament lligada al fet que estigui anotat. Considerem què passaria si ens enfrontéssim a un corpus no lematitzat, del qual volguéssim extreure simplement exemples del verb *atreure*: hauríem de fer la cerca per totes les formes d'aquest verb, cosa que, com a mínim, ens complicaria enormement el treball. Una investigació general de caràcter lèxic sobre el verb *atreure* pot ser feta si tenim les formes lèxiques del corpus anotades amb el seu lema. Similarment,

una investigació general sobre estructures de caràcter sintàctic implicarà tenir el corpus anotat (o poder-lo anotar) amb característiques sintàctiques. En general, es pot afirmar que les anotacions del nivell lingüístic que siguin ens obren la porta a la investigació general de corpus en el nivell corresponent.

Naturalment, els corpus poden ser anotats a molts diversos nivells. En principi, no hi ha restriccions al tipus d'anotacions que podem incorporar als corpus. Des de les anotacions més simples de lema i categoria gramatical fins a les anotacions de caràcter semàntic o discursiu, hi ha tota una gradació possible. De totes maneres, en l'actualitat gairebé sempre que es parla d'anotació de corpus es pensa en la lematització i categorització del corpus. En alguns casos, s'han desenvolupat tècniques d'anotació sintàctica dels textos (especialment, de l'estructura de constituents); però encara som lluny del moment en què podrem començar a pensar en etiquetats semàntics sistemàtics.

Davant de qualsevol tasca d'anotació de corpus, evidentment hi ha la possibilitat d'emprendre-la manualment. Ara bé, la lentitud del procés, que implica un cost molt alt, converteix en interessant qualsevol intent d'automatització, sobretot si tenim en compte que les dades emmagatzemades en els corpus actuals poden arribar a xifres altíssimes (de desenes, i fins i tot centenes, de milions de mots). Actualment les tècniques de processament automàtic de corpus estan prou avançades com per poder oferir etiquetadors automàtics força fiables, especialment en relació a la lematització i categorització, i a algunes parcel·les de l'etiquetat sintàctic.[6] Aquests aspectes, ja plenament dins del que anomenem processament del llenguatge natural i, per tant, de la disciplina de la lingüística computacional, són els que tractarem en els següents apartats.

3. Programes automàtics de lematització i desambiguació morfosintàctica

El primer nivell en el processament de corpus és l'anotació morfològica, que comporta l'assignació d'un lema i una categoria sintàctica a cada paraula del text processat. La finalitat essencial d'aquest tipus de programes consisteix en efectuar aquest

[6] Es pot obtenir una bona perspectiva de les possibilitats d'aquest tipus de programes en els dos volums especials de la revista *Computational Linguistics* dedicats a l'explotació de grans corpus (els dos primer números del volum 19, de l'any 1993). De fet, una part important de la informació continguda en aquest article està relacionada amb treballs de recerca ressenyats allí.

etiquetat de forma automàtica. Els programes d'aquestes característiques tenen sempre els següents components:

. un preprocessador,

. un analitzador morfològic i etiquetador, i

. un desambiguador.

En els subapartats que segueixen descriurem els aspectes essencials de cada un d'aquests components.

3.1. *El preprocessador*

El preprocessador consisteix en un programa que prepara el text per a poder ser processat de manera eficient pels components essencials del l'anotador morfològic. Les tasques principals d'un preprocessador són la normalització del text i la detecció d'elements que no han de ser processats posteriorment.

La normalització dels textos és necessària, perquè la forma física amb què apareixen pot ser molt diversa, mentre que l'analitzador demana uns formats unificats per a poder-se aplicar sistemàticament. El procés de normalització suposa tractar els elements de format de text (la seva estructura, títols, notes, majúscules...) de manera que l'analitzador hi pugui passar uniformement. A la vegada, pot tractar també aspectes com les xifres, les dates... que pròpiament no han de ser tractats com a unitats lingüístiques. Aquesta etapa de normalització comporta normalment el que es coneix amb el nom de *marcatge estructural*, és a dir, la codificació amb marques predeterminades (normalment estandaritzades) de les diverses peculiaritats formals que apareixen en el text.[7]

Per altra banda, hi ha una sèrie d'elements en els textos que no han de ser processats, perquè són invariables. En primer lloc, tenim les paraules aïllades invariables (com, per exemple, les preposicions i les conjuncions). Si les tenim llistades a part, com a lemes invariables i les marquem en aquest estadi inicial, evitem que l'analitzador s'entretingui a provar anàlisis inadequades. Però també, tenim expressions fixades que no cal que siguin analitzades (*de mancomú, a mansalva, de sobte...*); en aquest estadi inicial poden ser també detectades i se'ls pot assignar l'etiqueta corresponent.

---

[7] L'estàndard que s'ha imposat actualment per al marcatge estructural és el *SGML* (de *Standard Generalised Markup Language*). Sobre aquest llenguatge, podeu consultar, per exemple, van Herwijnen (1994).

3.2. *L'analitzador morfològic i etiquetador*

Sens dubte l'analitzador morfològic és la peça central d'un anotador morfològic. La seva funció és oferir totes les anàlisis morfològiques possibles de les paraules; així de *traves* n'hauria de donar dues anàlisis: *trava*(N)+*s*(pl) i *trav*-(V)+*es*(pres-ind,2a) i de *salina*, n'hauria de donar tres: *salina*(N), *salí*(A)+*a*(f) i *salin*-(V)+*a*(pres-ind,3a). Els components essencials d'aquest tipus de programes són un paquet de regles i el diccionari de morfemes (arrels i afixos).

Quant a les regles, els analitzadors morfològics poden ser subdividits en dos grans grups segons si usen regles de baix o alt nivell. Les primeres conformen autòmats d'estats finits, mentre que les segones constitueixen gramàtiques lliures de context. Una característica dels autòmats d'estats finits és que les regles han de tenir totes un símbol terminal a la dreta, cosa que fa que no puguin tractar seqüències de constituents que van de dos en dos. Això les fa inadequades per a tractar molts llenguatges (inclosa la sintaxi dels llenguatges naturals), però també més simples formalment i, per tant, més fàcils de processar; a més, resulta possible en general de plantejar l'analitzador de forma determinista, de manera que encara s'augmenta l'eficiència del sistema. En canvi, les regles de les gramàtiques lliures de context poden tenir seqüències de 0 o més símbols terminals o no-terminals a la seva dreta, mentre que a la seva esquerra només hi poden tenir un símbol no-terminal (cosa que les diferencia de les gramàtiques sensibles al context); aquest tipus de regles és el que estem més habituats a veure, ja que és el format de les regles de reescriptura que trobem normalment en sintaxi.

Així com els investigadors coincideixen a afirmar que la sintaxi dels llenguatges naturals no pot ser formalitzada amb una gramàtica d'estats finits, les estructures morfològiques sí que hi poden ser representades. De fet, s'han creat diversos models de tractament morfològic amb gramàtiques de baix nivell (Gazdar i Mellish, 1989, pàg. 62); de totes maneres, en l'actualitat, el que més s'ha imposat és el conegut amb el nom de *morfologia de dos nivells*, desenvolupat per Koskenniemi (1983) a la Universitat de Hèlsinki, i que ha estat usat actualment per a construir analitzadors morfològics per a un nombre força elevat de llengües. La idea central d'aquest tipus d'analitzador és que hi ha dos reconeixedors de caràcters que actuen en paral·lel, l'un reconeixent els caràcters de la cadena del text i l'altre resseguint els caràcters de les entrades lèxiques. En aquest doble recorregut les regles es poden trobar amb variacions entre el que apareix a la cadena

d'entrada i la forma del morfema en el diccionari; per exemple, l'arrel de *lleó* conté una vocal accentuada, que podem no trobar en algunes realitzacions seves en els textos (*lleons*). Les regles de dos nivells, doncs, relacionen els caràcters de la cadena d'entrada (del text) i els de l'entrada del diccionari, atenent únicament als contextos gràfics (o fonològics) de la dreta i de l'esquerra (a més, és clar, de la informació que tenim consignada en el diccionari). Així, una regla de dos nivells podria establir que una *ó* en posició final de morfema ha de ser convertida en *o* davant del morfema de plural *ns*; per a la correcta aplicació d'aquesta regla hauríem de tenir codificades en el diccionari les paraules que formen el plural afegint el morfema *ns*.[8]

Alternativament, s'ha plantejat analitzadors morfològics escrits en gramàtiques lliures de context. En aquest cas, es tracta de formular regles d'estructura de paraula de manera que els mots es construeixin a partir dels seus elements constitutius bàsics (l'arrel i els afixos). Per a assegurar que hi hagi congruència entre el tipus d'arrel i el tipus d'afix, cal establir un sistema de codificació de les arrels i dels afixos que garanteixi que s'aparellin adequadament (que, per exemple, a una arrel verbal de la primera conjugació no se li ajuntin sufixos flexius de la tercera). Aquest plantejament, en definitiva, reprodueix d'una manera bastant minuciosa el sistema de paradigmes flexius. El problema és que es multipliquen enormement el nombre de paradigmes, de manera que es perden moltes de les regularitats que trobem en les classificacions morfològiques clàssiques. Com que un sistema així no permet modificacions ni en la forma de l'arrel ni en la dels sufixos, cal adoptar un sistema de diverses arrels per al mateix lema (encara que les variacions siguin regulars); així, els verbs de la primera conjugació acabats amb *g* hauran de tenir sigui dues arrels (l'una acabada amb *g* i l'altra amb *gu*) o hauran de combinar amb sufixos diferents en alguns casos (per exemple, el sufix *es* es converteix en *ues*). Per a reprendre l'exemple anterior del plural de *lleó*, en aquest tractament necessitaríem dues arrels (una amb accent i l'altra sense) i un sistema de codificació que indiqués que l'arrel sense accent no s'ajunta a cap sufix, mentre que l'altra és la que s'usa per a les formes flexionades; a més, ens caldria indicar que el plural de *lleó* es forma ajuntant el sufix *ns* a l'arrel sense accent (i no, per exemple, amb el sufixos *s* o *os*). Això, com es pot veure, comporta un

[8] Aquest no és necessàriament el millor tractament per a aquest tipus de plural. Només l'adduïm aquí a tall d'exemple.

sistema força complex de codificació lèxica i un nombre elevat de regles, cosa que, naturalment, afecta l'eficiència del sistema.

En tots els casos, els analitzadors morfològics necessiten un diccionari on estigui codificada la informació rellevant de cada peça lèxica. El diccionari conté informació de qualsevol morfema, sigui arrel o afix. Aquesta informació és, sempre també, de dos tipus: la que és necessària per al bon funcionament de les regles de l'analitzador i la que proporciona la informació morfosintàctica amb què anotem el corpus. La primera informació inclou forçosament la classe de mot (o categoria) i la subclasse (segons el tipus de morfemes amb què s'ajunta). Naturalment, la forma concreta que adoptarà aquesta classificació dependrà de l'estratègia usada en la construcció de l'analitzador, ja que hi ha d'haver una adequació important entre el que fan les regles i la informació que conté el diccionari. El diccionari per a un analitzador de dos nivells, per exemple, contindrà informació sobre els caràcters (o fonemes) que sofreixen variació en un morfema, mentre que el d'un analitzador més clàssic contindrà informació sobre el paradigma d'aquell morfema i sobre les restriccions d'ús a què pugui estar subjecte (l'arrel *amagu* combina amb uns sufixos concrets, mentre que *amag* ho fa amb la resta).

A més, però, el diccionari sol contenir informació que pot no ser útil per a l'analitzador, però que és interessant que aparegui en el text etiquetat. Per exemple, les paraules invariables (és a dir, que no són construïdes per cap regla complexa, perquè la seva forma en el text coincideix sempre amb la que tenen al diccionari) han de tenir informació de la seva categoria per a poder ser etiquetades adequadament en el text. Similarment, hi pot haver informació rellevant que vulguem usar en l'etiquetat i que no sigui útil per a l'analitzador (en la majoria de plantejaments, per exemple, el gènere dels noms no és necessari per a l'anàlisi morfològica).

Així, per exemple, el diccionari de *Morfeo*, analitzador morfològic de l'espanyol, que usa una estratègia derivada de la que hem descrit com a gramàtiques lliures de context (Pérez et al., 1994), conté la següent informació:

. arrel primària
. accentuació
. paraula estàndard
. arrel secundària
. categoria morfològica

. model flexiu i/o derivatiu
. codis morfosintàctics
. divisibilitat/indivisibilitat
. camp semàntic
. camps lliures

És clar en aquest exemple que la informació de la majoria dels camps anomenats resulta necessària per al bon funcionament de l'analitzador; de totes maneres, la dels codis morfosintàctics i del camp semàntic és útil només per a l'etiquetat final que volem que adorni els mots del corpus després del procés d'anàlisi.

3.3. *Aplicació de l'estratègia probabilística a l'etiquetat de corpus*

Els programes que es fonamenten en l'*estratègia probabilística* són generalment només etiquetadors. Trobem dues situacions en què aquesta estratègia és realment usada: en llengües que, com l'anglès, tenen una estructura morfològica gairebé inexistent i per a les quals, per tant, l'etiquetat és l'única acció problemàtica; i en llengües morfològicament riques en què procedim a l'etiquetat com un segon pas del procés, després de l'anàlisi morfològica pròpiament dita.

Essencialment, un etiquetador morfològic probabilístic es basa en la noció de canal de la teoria de la informació de Shannon (1948). Suposem que tenim una seqüència de categories C, que introduïm a un canal, que és distorsionador; el resultat que surt del canal és una seqüència de paraules P (que és interpretada com una distorsió de la cadena C inicial). El problema llavors és determinar C a partir de P. Esquemàticament tenim:

$C \rightarrow canal \rightarrow P$

El punt de partida podria ser el càlcul de la probabilitat d'aparició d'una oració, a partir de la probabilitat de cada paraula en la posició que ocupa en l'oració, és a dir, després de les paraules que l'han precedida. Aquest model, extremadament senzill (simplista, fins i tot), rep el nom de *n-gram* (on *n* indica el nombre de mots que es tenen en compte). Un càlcul així de la probabilitat d'una oració no és viable, ja que el nombre de combinacions creix exponencialment amb la longitud de la cadena considerada. Així, amb un vocabulari de 100 mots, cada distribució tindrà 100 valors; llavors, per a cadenes de longitud 1, tindrem exactament 100 distribucions; per a cadenes de longitud 2, les distribucions seran 100^2 (o sia, 10.000); amb cadenes de longitud 3, tindrem 100^3

distribucions (és a dir, 1.000.000); etc. Com que això és totalment inviable, es tendeix a reduir el nombre de paraules que es tenen en compte; és a dir, no es miren totes des de l'inici de l'oració, sinó només les dues, tres... anteriors a la paraula de la qual estem intentant calcular la probabilitat. En aquests casos, parlem de *2-grams*, *3-grams*... Una reducció així, però, encara té problemes importants, especialment pel que fa a la cobertura del model i, per tant, a la seva capacitat d'aprenentatge. La qüestió bàsica és que el model es troba en moltes situacions noves i que, per tant, no han estat apreses anteriorment. Com més gran sigui el nombre de paraules que tinguem en compte més baixa és la probabilitat de què una seqüència concreta no hagi estat trobada mai (és a dir, no formi part de la cobertura del model); així hi ha una relació inversa entre la necessitat de precisió i la de cobertura: com més alt és el valor de *n* més alta és la precisió, però més baixa és la cobertura.

De totes maneres, com que l'objectiu amb què ens enfrontem és l'obtenció d'etiquetes gramaticals (categoria o part de l'oració, i les subclasses que s'hi puguin establir), el problema no és el de la probabilitat d'aparició d'un mot després d'una seqüència determinada de paraules, sinó el de la probabilitat d'aparició d'una categoria determinada després d'una seqüència de categories. És obvi que la probabilitat d'aparició d'un mot d'una categoria normalment no depèn de la presencia en la seqüència que el precedeix d'una sèrie de mots, sinó d'una sèrie de classes de mots (és a dir, de categories). Per tant, el model es converteix en un model de la probabilitat d'aparició de les categories després d'un nombre determinat de categories. En aquest context parlem de *n-pos* (en lloc d'*n-grams*) a partir de l'expressió anglesa *part of speech* (que s'usa habitualment per a referir-se a les categories). Semblantment, utilitzem les expressions de *bi-pos*, *tri-pos*... segons el nombre de categories que observem per a establir les probabilitats.

Naturalment, aquí només podem exposar les idees més simples d'aquest tipus de programes. Aquest és un camp d'investigació en lingüística computacional enormement actiu, en el qual s'estan fent grans avenços tant en els aspectes més teòrics, com en el camp de les aplicacions. Així, dos dels etiquetadors de textos més usats en anglès parteixen d'aquestes tècniques. L'etiquetador CLAWS (Garside, Leech i Sampson, 1987) té com a objectiu assignar una de les etiquetes previstes al sistema (prop de 200) a cada paraula del text que s'etiqueta. Inicialment el sistema assigna a cada mot (considerat aïlladament) una sèrie d'etiquetes candidates amb indicació del seu grau de probabilitat.

Després escull, amb criteris probabilístics, una de les etiquetes possibles en funció del context en què apareix la paraula. Segons els resultats presentats pels mateixos creadors de l'etiquetador, el seu grau d'èxit oscil·la entre el 96% i el 97%. Un plantejament molt similar és el seguit per Church (1988) en el seu etiquetador PoS, el nivell d'èxit documentat del qual oscil·la entre el 95% i el 99%.

3.4. *Desambiguador*

Quan s'apliquen els plantejaments lingüístics clàssics, que hem vist a la secció 3.2, a l'anàlisi morfològica de llengües altament flexives, apareix molta ambigüitat resultant, és a dir, hi ha molts mots que poden ser analitzats de maneres diferents. Conseqüentment, el procés d'etiquetat encara no està acabat, ja que s'ha de procedir a la desambiguació.

L'objectiu del desambiguador és, doncs, doble: eliminar l'ambigüitat que resulta de l'anàlisi morfològica i atorgar a cada mot un únic lema i una única caracterització lingüística. Per a això, es tendeixen a utilitzar dos tipus complementaris d'informació: l'estructura textual i les propietats distribucionals dels mots. Es té en compte l'estructura textual quan, per exemple, es considera que en títols i subtítols el comportament de les estructures lingüístiques és diferent (per exemple, els verbs no són necessàriament els nuclis de les expressions i els noms s'hi comporten de manera diferent).

Per altra banda, a partir de les propietats distribucionals de les categories o classes de paraules es pot determinar a quina categoria pertany cada mot inicialment ambigu. Així, per exemple, sabem que després d'una preposició el mot *la* pot ser un article, però no un pronom feble, o que després de l'auxiliar *haver* hi segueix un participi, i no un nom o un adjectiu (així, en l'expressió *ha rentat*, *rentat* és necessàriament un participi; i a *ha interessat*, *interessat* ha de ser un participi i no un adjectiu). Conseqüentment, a partir de les propietats distribucionals de les categories es pot plantejar la determinació de la categoria sintàctica dels mots que han resultat ambigus en l'anàlisi morfològica.

Aquest procés de desambiguació pot ser plantejat a partir de dues estratègies diferents: la probabilística i la lingüística. En molts casos, s'adopten tècniques probabilístiques, seguint els plantejaments que hem vist anteriorment. Un bon exemple de l'ús d'aquesta estratègia el trobem en el desambiguador de l'analitzador morfològic *Morfeo* (Pérez et al., 1994). El desambiguador consisteix en un filtre de Markov d'ordre 2, que recull la freqüència de transició de cada seqüència de tres categories. L'operació principal del

desambiguador consisteix en la consulta de la taula de freqüències elaborada a partir de l'anàlisi de mostres i en el càlcul subsegüent de la probabilitat de cada un dels camins alternatius entre punts sense ambigüitat. El rendiment documentat d'aquest analitzador comporta una velocitat d'aproximadament 1,63 pàgines per minut; un cop es tenen les taules de freqüència elaborades, el nivell d'èxit és superior al 99%.

La informació que l'analitzador *Morfeo* afegeix a cada mot del corpus al final del procés consisteix en els atributs següents: la forma superficial, el lema, el temps, el mode, la persona, el nombre, el gènere i la categoria, als quals s'adjunta un camp per a incloure-hi informació lèxica específica. A la taula següent incloem el resultat de l'anàlisi de l'oració *Las redes en bus son sencillas de instalar*.[9]

lema	temps	mod	pers	nomb	gènere	categ	inf.lèx.	superf
el_a	0	0	3-pl	pl	fem	art		las
red	0	0	3-pl	pl	fem	n_com		redes
en	0	0	0	0	0	prep		en
bus	0	0	3-sg	sg	mas	n-com		bus
son	pres	ind	3	pl	0	v-ple		son
sencillo	0	0	3pl	pl	fem	adj		sencillas
de	0	0	0	0	0	prep		de
instalar	inf	0	quals	comp	0	v-ple		instalar

Des del punt de vista del tipus de processament, l'alternativa a aquest tipus de desambiguador la consisteixen els desambiguadors basats en regles lingüístiques. Un dels exemples més característics d'aquests processadors el trobem en l'anomenat *English Constraint Grammar* (*ENGCG*) (Karlsson, 1990). De fet, aquest analitzador-desambiguador es basa en la formulació de regles sintàctiques explícites per a dur a terme el procés de desambiguació; d'aquesta manera és un producte que pot ésser considerat tant un analitzador sintàctic com morfològic (vegeu-ne més avall -§4.4.d- una breu descripció). Els rendiments documentats per a aquest analitzador són els següents: l'anàlisi morfològica va a 600-1000 mots per minut, i l'anàlisi sintàctica (que li serveix de desambiguador) a 10-15 mots per segon. El nivell d'encert (calculat a partir de la

[9] Informació extreta del document *Salida estándar para español moderno*, que va ser distribuït a la demostració que es va fer d'aquest etiquetador al congrés de la SEPLN de 1994 (Còrdova, 20-22 de juliol).

desambiguació total i correcta) oscil·la entre el 94 i el 97 %. A la taula següent s'hi pot observar el resultat de l'anàlisi per a l'oració *Bill saw the little dog in the park*:[10]

Bill	Bill	"<Proper> N NOM SG"	@SUBJ	
saw	see	"<SVO> V PAST"	@+FMAINV	
the	the	"DET"	@DN>	
little	little	"A ABS"	@AN>	
dog	dog	"N NOM SG"	@OBJ	
in	in	"PREP"	@<NOM	@ADVL
the	the	"DET"	@DN>	
park	park	"N NOM SG"	@<P	

Així, la informació que aquest anotador incorpora al text consisteix en el lema, la informació morfosintàctica i una anotació de caràcter sintàctic (que indica la funció sintàctica superficial juntament amb les dependències bàsiques).

4. Programes automàtics d'anàlisi sintàctica

4.1. *Introducció*

En general, s'ha tendit a indentificar el processament del llenguatge natural amb l'anàlisi sintàctica. Això és conseqüència d'una sèrie de factors que hi han contribuït, cada un a la seva manera. En primer lloc, la complexitat de l'anàlisi sintàctica és força gran, i clarament superior a la que cal per al processament morfològic. Segonament, la lingüística formal ha donat una gran importància a la sintaxi, que és una de les àrees més estudiades de la lingüística. En tercer lloc, s'ha tendit a pensar que les relacions sintàctiques són les que, de fet, estructuren els textos; i, així, s'ha considerat que l'anàlisi sintàctica és el primer pas per al tractament automàtic del llenguatge natural. Finalment, els investigadors es van adonar fa relativament força temps que les estratègies de processament del llenguatge natural fonamentades únicament en la semàntica no eren adequades en general, sinó que només podien servir per al processament de textos amb subllenguatges molt

[10] Resultat extret de Karlsson (1990, pàg. 172).

restringits. Una bona manera d'aproximar-se a les possibilitats reals del processament sintàctic de textos és recorrent els grans moments de la seva curta història.

4.2. *Breu recorregut per la història del processament sintàctic*

Tradicionalment, s'han adoptat mètodes totalment racionalistes, que normalment han estat adaptats de plantejaments diversos de la lingüística teòrica. Així s'han desenvolupat sistemes en què l'anàlisi sintàctica estava basada en gramàtiques d'estructura sintagmàtica, en gramàtiques de dependència, en sistemes multi-estratals (amb processos de transformació intermitjos), o, més darrerament, en gramàtiques d'unificació. En general, el desenvolupament de les estratègies sintàctiques de la lingüística computacional ha anat variant paral·lelament a l'evolució de la lingüística teòrica. A grans trets, podem distingir quatre grans etapes:

4.2.1. Primera etapa (al voltant dels anys 50 i 60)

El processament està poc guiat per la lingüística teòrica, que encara no ha adoptat els mètodes formals a què ens ha acostumat actualment. L'objectiu immediat de les tasques és la traducció automàtica, que es basa totalment en l'estratègia de la traducció paraula per paraula. D'alguna manera, es pot considerar que en aquests estadis inicials de la lingüística computacional els plantejaments que s'adopten estan relacionats amb els principis del distribucionalisme; de fet, l'única manera de superar (encara que sigui parcialment) les limitacions pròpies del tractament de paraula per paraula és a base de tenir en compte (algunes de les) propietats distribucionals dels mots en els textos.

4.2.2. Segona etapa (de meitat dels anys 60 a meitat dels 70)

En aquest moment s'adopten dues perspectives diferents. Per una banda, hi ha investigadors que adopten tècniques d'anàlisi que no estan fonamentades en la lingüística, sinó en estructures conceptuals i en representacions del coneixement pròpies de la intel·ligència artificial; pròpiament, per tant, no es tracta de programes que exemplifiquin el processament sintàctic. L'autor més característic d'aquesta tendència és Schank (vegeu, per exemple, Schank, 1980). A partir de la recerca dels mots semànticament plens del text s'estableixen les relacions significatives que hi apareixen i es representen mitjançant alguna de les tècniques disponibles (esquemes, marcs...). A la vegada, per a tractar textos de subllenguatges molt específics, s'utilitzen tècniques simplificades de detecció dels mots semànticament plens i de representació de les relacions que aquests mantenen entre

ells; un exemple d'aquesta estratègia el constitueix el sistema de traducció automàtica *Météo*, per a traduir els butlletins de la informació meteorològica entre francès i anglès (Chandioux, 1976).

Per altra banda, es desenvolupen tècniques d'anàlisi basades directament en els conceptes de la programació. En aquesta àrea se segueixen principis pròxims a la sintaxi; de totes maneres, no hi ha una diferenciació clara entre els coneixements lingüístics i els de processament (o *parsing*). L'estadi culminant d'aquestes tècniques el constitueixen les *ATN* (de l'anglès *augmented transition network*, o xarxes de transició augmentades). Les ATN són un formalisme molt potent per a representar estructures lingüístiques i, per tant, per a procedir a anàlisis lingüístiques. Essencialment, una xarxa de transició pretén de reconèixer els mots en una cadena lingüística mitjançant una màquina d'estats finits complementada per registres i operacions amb registres que permeten de superar alguna de les limitacions dels autòmats. Així per exemple, per a tractar la concordança dins el sintagma nominal amb simples màquines d'estats finits, necessitaríem almenys quatre camins diferents (o quatre autòmats paral·lels) que correspondrien a les estructures masculí-singular, femení-singular, masculí-plural i femení-plural. Això, però, ho podem tractar amb una única màquina si hi afegim un parell de registres (un per al gènere i l'altre per al nombre) i obliguem a la comprovació que el valor del registre i el de la paraula que estem considerant són iguals.[11]

Ben aviat apareixen, però, les limitacions de les dues perspectives. En primer lloc, les aproximacions estrictament semàntiques no són capaces de donar compte de propietats essencials de les expressions lingüístiques, especialment respecte a la seva estructura; així, aplicacions que impliquen una anàlisi dels aspectes més formals (com la traducció automàtica) no s'hi poden desenrotllar plenament. Per altra banda, les ATN apareixen com un mecanisme excessivament procedural que barreja totalment el coneixement lingüístic i el de processament.

4.2.3. Tercera etapa (de meitat dels anys 70 a meitat dels 80)

Aquesta és l'etapa d'eclosió de la lingüística computacional. Per primer cop. els investigadors adopten mètodes d'anàlisi propis, elaborats especialment per a l'anàlisi lingüística. Tenim, per tant, enfocaments utilitzables directament en la producció

[11] Trobareu una descripció adequada i assequible de les ATN a Gazdar i Mellish (1989, pàg. 91 i ss).

d'analitzadors sintàctics. Un dels passos essencials que es fan en aquest període és el de la separació clara entre les tècniques de processament (*parsing*) i el coneixement lingüístic. Aquesta separació permet una millora substancial en els plantejaments que es fan, tant des del punt de vista teòric com des del pràctic. De fet, aquesta separació propicia l'adopció de dos principis metodològics que han resultat bàsics en el progrés de la lingüística computacional: la modularitat i la declarativitat.

Entenem que un programa és modular quan està dividit en seccions independents. que es coneixen amb el nom de *mòduls*. En l'elaboració de grans programes, una de les dificultats principals es troba en el fet que tot està relacionat amb tot, de manera que al final el programador ha de tenir en compte més paràmetres o factors dels que realment pot captar; això té, naturalment, conseqüències en el moment de l'elaboració inicial del programa, però també de cara a la seva modificació i manteniment posteriors. En la mesura que el treball és parcel·lable, pot ser dividit en parcel·les autònomes, aquestes dificultats disminueixen; a més, es facilita enormement la reutilització del programa (o d'alguns mòduls del programa). Això, per altra banda, no és possible si no es disposa d'unes estratègies clares d'enllaç entre uns mòduls i els altres.

Si pensem en un programa d'anàlisi sintàctica, hi ha certament molts factors a tenir en compte (regles o principis que interactuen els uns amb els altres, els elements lèxics, les propietats pròpiament lingüístiques i les estratègies de processament...). Si podem pensar en cada un d'aquests aspectes per separat, la tasca de programació se simplifica i resulta molt més adequada i pròxima a les maneres naturals de pensar. Això, però, pressuposa que hi ha mecanismes clars i absolutament definits d'enllaç entre uns mòduls i els altres: per exemple, entre el lèxic i les regles, entre els diversos paquets de regles (si és que, com és habitual, subdividim la gramàtica en grups de regles relativament independents), entre la gramàtica i el procés d'anàlisi...

És característic dels enfocaments declaratius que el programa estableix les condicions perquè ocorrin uns fets determinats (o sia, *declara* els fets). Aquest plantejament és oposat a l'anomenat procedural, pel qual el programa assenyala els passos que s'han de seguir per a resoldre una qüestió determinada. Des de l'òptica declarativa, la regla de construcció de sintagmes nominals $SN \rightarrow Det\ N$ seria formulada tal com s'indica a *a*, mentre que la mateixa regla tindria la formulació de *b* si seguim el plantejament procedural.

a. *Un SN és aquella estructura que està formada per un determinant i un nom, per aquest ordre*

b. *Si en recórrer la cadena d'entrada trobes primer un determinant i després un nom, hauràs trobat un sintagma nominal*

En la mesura que s'han anat independitzant els plantejaments lingüístics dels de càlcul, la descripció lingüística computacional ha anat adoptant cada cop més els plantejaments declaratius. D'aquesta manera s'ha pogut arribar a formulacions gramaticals (o lingüístiques en general) que són totalment independents de l'estratègia seguida per a calcular-les. Aquesta és una de les raons que han contribuït a l'augment en l'ús del llenguatge de programació *Prolog* (de *programació lògica*), que afavoreix totalment la descripció declarativa.

Una altra de les característiques importants d'aquesta etapa és que la lingüística computacional s'ha vist enormement influïda per l'evolució de la teoria lingüística. Això es pot veure clarament si, per exemple, tenim en compte que en el moment en què la lingüística teòrica distingia entre diversos nivells de descripció, la lingüística computacional va adoptar l'estratègia de dividir el procés d'anàlisi sintàctica en diversos nivells, les estructures produïdes pels quals estaven relacionades entre elles mitjançant transformacions (tot reproduint l'estratègia usada en els plantejaments de la gramàtica generativa transformacional).

També és característica d'aquesta etapa la tendència a identificar la gramàtica (és a dir, la codificació del coneixement lingüístic) amb la competència, i l'anàlisi (o càlcul) amb l'actuació. D'aquesta manera es reproduïa en la lingüística computacional la distinció de la lingüística teòrica i també s'hi replicava la preferència dels lingüistes per la competència (en lloc de l'actuació).

Finalment, i també com a reflex dels plantejaments en lingüística teòrica, la descripció sintàctica estava fonamentada sobretot en la gramàtica (representada pels diversos paquets de regles) i, paral·lelament, es donava molt poca importància al lèxic (que no era sinó el lloc on es consignaven els aspectes més idiosincràtics de les paraules).

4.2.4. Quarta etapa (des de finals dels anys 80)

Aquest és, és clar, l'estadi actual. S'hi han produït avenços espectaculars en dues direccions diferents: les tècniques de tractament lingüístic han millorat considerablement i s'han desenvolupat tècniques per a la manipulació de grans corpus.

En el marc del *tractament lingüístic*, s'ha produït una confluència extraordinària de plantejaments adoptats en diverses disciplines. El caràcter multidisciplinar de la lingüística computacional ha quedat més de relleu en aquests darrers anys, en adoptar mètodes i estratègies nascudes al si, com a mínim, de les tres disciplines de la lingüística teòrica, la ciència de la computació i la lògica.[12] Els plantejaments actuals són bàsicament els següents:

S'han desenvolupat tècniques de representació del coneixement lingüístic en estructures tipificades de trets.[13] La informació lingüística no està disseminada en diversos nivells o aspectes descriptius, sinó que aquests conformen un tot integrat. Les informacions fonològica, morfològica, sintàctica i semàntica coexisteixen simultàniament i poden ser tingudes en compte simultàniament. No hi ha dubte que des del punt de vista lingüístic això comporta avantatges indubtables (Badia, 1994). Per només citar-ne un exemple, considerem la interacció entre diversos aspectes en la descripció dels aspectes relacionals dels elements predicatius (p. ex., els verbs). Sempre hi ha una relació entre els valors de la dependència sintàctica i els de la semàntica, de manera que resulta convenient de referir-s'hi simultàniament (en general, a un verb transitiu li corresponen dos papers temàtics, mentre que a un de bitransitiu n'hi corresponen tres). A més, però, hi ha aspectes morfològics amb què això pot interactuar, ja que els participis passats no flexionats poden mantenir les estructures sintàctiques i semàntiques de complementació, mentre que els participis flexionats (en les formes femenines o plurals) no les mantenen, sinó que són subjectes a una reducció sintàctica de complements. Així, tenim una relació entre la forma morfològica d'un mot (el participi passat) i la seva estructura de complementació sintàctica, la qual al seu torn, manté una relació amb l'estructura temàtica.

Des d'un punt de vista tècnic aquestes representacions són fetes amb les estructures matemàtiques conegudes amb el nom de *grafs acíclics orientats*. La potencialitat expressiva d'aquest tipus d'estructures, així com el seu comportament lògic, no han estat plantejats inicialment des de la lingüística, sinó que han sorgit a partir d'estudis en intel·ligència artificial (sobre tècniques de representació del coneixement) i en la branca de

[12] Aquesta interdisciplinarietat queda plenament manifesta al primer capítol de Gazdar i Mellish (1989).
[13] El tractat més complet, en aquest moment, sobre el processament amb estructures de trets tipificades és Carpenter (1992). Podeu veure, a més, les referències que s'hi citen.

la lògica més preocupada per aquests aspectes. Així, tot i aquest és encara un camp en el qual no hi ha resultats definitius, es tendeix a distingir entre tres nivells bàsics: el dels objectes lingüístics, el del llenguatge descriptiu que fem servir per a parlar d'ells, i el de la lògica subjacent a les relacions formulables en aquest darrer. El lingüista treballa normalment en el segon nivell amb l'objectiu de donar compte dels fets propis del primer; ha de suposar l'existència del tercer nivell, perquè és el que li garanteix la coherència dels seus plantejaments (i el que li permet d'afrontar amb èxit plantejaments computacionals en què hi té un paper la inferència).

Una altra de les grans innovacions d'aquest període ha estat la disminució del caràcter procedural de la gramàtica i, paral·lelament, la independència creixent de la declaració del coneixement lingüístic respecte del formalisme en què s'expressa el processament lingüístic. Cada vegada més, es veu la gramàtica com un recull de declaracions de caràcter lingüístic que, juntes, estableixen els fets rellevants lingüísticament per a la descripció d'una llengua. Així hi ha un predomini de l'afirmació en positiu de les característiques dels elements lingüístics que intervenen en la formació d'oracions o textos. Aquest plantejament afavoreix una descripció lingüística deslligada de les necessitats processals, que, per tant, pot ser emprada sense modificacions majors en conjunció amb diversos procediments d'anàlisi.

Durant aquests darrers anys, la lingüística computacional ha estat subjecte al gran procés de canvi que ha sofert la lingüística teòrica que està provocant l'adopció d'una nova perspectiva; ens referim al procés de lexicalització de la descripció lingüística. Després d'uns anys en què la descripció lingüística estava centrada en la formulació de regles sintàctiques (i, en alguns casos, semàntiques), amb un desinterès total per la informació específica de les peces lèxiques, els nous plantejaments han permès de centrar bona part de la descripció lingüística a l'entorn de la paraula (de les peces lèxiques en general). Cal tenir present, però, dues coses. Per una banda, aquest procés no ha comportat un abandó de les observacions de caràcter sintàctic i semàntic que han centrat la lingüística dels darrers 25 o 30 anys; al contrari, els plantejaments lexicalistes han permès d'incorporar tot aquest bagatge científic a la descripció de la informació lingüística associada a les peces lèxiques. I per l'altra, aquest procés ha anat acompanyat per un progrés en les tècniques de codificació i organització lèxiques, sense les quals la lexicalització hauria comportat també dispersió de la informació.

El procés de lexicalització de la informació lingüística ha portat de manera directa un augment considerable de la complexitat del diccionari. Les peces lèxiques ja no han de ser vistes (de fet, ja no poden ser vistes) com independents les unes de les altres, sinó com estretament interrelacionades. Conseqüentment, s'imposen tècniques d'organització d'aquestes relacions, especialment, les que permeten de formular de manera elegant i no reiterativa que diverses peces lèxiques comparteixen un determinat tipus d'informació. Per exemple, cal que el sistema d'organització lèxica permeti de formular que tots els verbs transitius comparteixen una informació (que tenen dos complements d'un tipus determinat), o que totes les formes verbals d'un verb comparteixen una part de la informació que hi va associada. A la vegada, cal que es puguin establir relacions sistemàtiques entre peces lèxiques. Així, cal que hi hagi sistemes generals que relacionin les formes transitives plenes dels verbs amb les pronominals que indiquen impersonalitat. Naturalment, aquest procés de lexicalització no ha estat emprès fins al moment en què les tècniques de representació de la informació han permès de plantejar satisfactòriament aquestes qüestions. Bona part dels plantejaments que ho han fet possible s'han derivat de l'enfocament de les estructures de trets tipificades com a mitjà per a codificar les realitats lingüístiques.

Tots aquests plantejaments que estan conformant la lingüística computacional (i bona part de la lingüística teòrica) actualment, han confluït en la creació de gramàtiques o, millor, formalismes gramaticals en què aquests principis generals es realitzen plenament. En aquest entorn han nascut les que es coneixen amb el nom de *gramàtiques d'unificació*, i que potser més pròpiament hauríem d'anomenar *gramàtiques de restriccions*.[14] Totes elles adopten uns plantejaments bàsicament declaratius de la descripció lingüística, amb un creixent procés de lexicalització. També hi resulta essencial el paper que hi juga la unificació com a eina per a combinar la informació lingüística, tot i que, cada vegada més, s'està imposant la visió que la unificació no és sinó una eina més per a formular restriccions a les estructures lingüístiques ben formades. L'essència de la descripció lingüística, doncs, consistiria en declarar propietats o fets sobre els elements lingüístics, a

[14] Com a formalismes gramaticals d'unificació podem citar *Functional Unification Grammar* (FUG) (Kay, 1984) i PATR-II (Shieber, 1986); d'entre les teories gramaticals destaquen *Lexical-Functional Grammar* (LFG) (Bresnan, 1982), *Generalised Phrase Structure Grammar* (GPSG) (Gazdar *et al.*, 1985) i, sobretot, *Head-Driven Phrase Structure Grammar* (HPSG) (Pollard i Sag, 1987, 1994).

base d'imposar condicions (que normalment anomenem *restriccions*) per a considerar-los ben formats.

L'altre gran progrés dels últims anys en lingüística computacional fa referència a les *tècniques per a manipular grans corpus*. Els progressos en capacitat d'emmagatzematge i de memòria en els ordinadors han fet possible que es puguin tractar grans masses de dades, com les recollides en els corpus amb què ara es compta. Són aquestes tècniques les que han permès els processos d'etiquetat dels corpus que hem comentat més amunt. Per altra banda, la majoria d'aplicacions efectives que s'han construït són dedicades a l'extracció d'informació de corpus, per a la qual l'etiquetat és un primer pas essencial. Els corpus permeten de fer observacions de fets lingüístics a gran escala; això, però, no és possible si no es compta amb eines automàtiques d'extracció d'aquesta informació, ja que el processament de corpus de milions de mots no pot ser fet manualment de cap de les maneres. Així, per exemple, entre les dades que els investigadors extreuen dels corpus hi ha la referent a les associacions entre elements lèxics (com les que apareixen en les col·locacions i altres agrupacions). La immensa majoria de programes dedicats a aquestes tasques d'extracció d'informació usen tècniques probabilístiques, que han adquirit un protagonisme creixent en el camp de la lingüística computacional aplicada.

4.3. *Proposta de classificació dels sistemes de processament sintàctic*

Tot i que hi ha qui ha volgut establir una equació que equiparava els mètodes probabilístics amb l'anàlisi real de corpus i els mètodes racionalistes amb anàlisis lingüístiques de laboratori (Church i Mercer, 1993), aquesta equiparació no és real. De fet, s'estan desenvolupant tècniques d'anàlisi de corpus fonamentades en perspectives explícitament lingüístiques, que estan tenint un nivell d'èxit equiparable al de les aproximacions probabilístiques. Per tant, s'imposa una classificació de les tècniques emprades que reculli totes aquestes possibilitats.

Proposem, doncs, una classificació dels sistemes d'anàlisi que parteixi de la finalitat essencial de l'anàlisi. Segons aquest criteri, distingirem entre tècniques adaptades a textos reals i tècniques de nivell superior. Les primeres estan pensades per a tractar de manera realista amb textos reals, mentre que les segones estan orientades més teòricament i pretenen dur a terme anàlisis lingüístiques de nivell relativament elevat. Cada un d'aquests grups pot ésser encara subdividit, tal com mostrem a la taula següent:

. tècniques adaptades a textos reals o de baix nivell
  . amb formulacions lingüístiques explícites
  . amb formulacions de caràcter probabilístic
. tècniques de nivell superior
  . d'abast purament sintàctic
  . incorporant de manera fonamental l'anàlisi semàntica

Aquesta classificació, que té l'avantatge de ser molt clara, no deixa de ser inadequada, a causa de la simplificació que suposa. De fet, els investigadors tendeixen a pensar que, atès l'estat actual de coneixements, la construcció d'aplicacions concretes ha de partir d'estratègies que plantegin l'ús de tècniques mixtes. Les tècniques de baix nivell són clarament insuficients per a aplicacions que demanin una anàlisi pròpiament semàntica; en canvi, les tècniques de nivell superior no resulten adequades per a emprendre l'anàlisi directa de textos reals. Conseqüentment, s'imposa la interacció entre les dues perspectives, de manera que es distingeixi entre aplicacions i finalitats i es plantegin sistemes complexos amb diversos estadis de processament. D'aquesta manera, una estratègia general per al processament semàntic de textos implicaria un primer estadi d'anàlisi, efectuat amb tècniques de baix nivell, que aniria seguit per un segon estadi, en el qual s'usarien gramàtiques d'alt nivell que incorporin clarament el tractament semàntic.[15]

4.4. *Exemples d'anàlisi gramatical de corpus*

Actualment, podem veure d'una manera realista l'anàlisi gramatical de corpus com determinada per dues variables diferents: el seu enfocament metodològic bàsic i la seva finalitat:

. 2 enfocaments:
  . probabilístic
  . lingüístic

[15] La plataforma de processament impulsada per la Comissió de la Unió Europea, anomenada *Alep*, constitueix un exemple molt clar d'això que diem: en ella, es poden integrar diferents nivells de processament, de manera que s'hi poden efectuar anàlisis de baix nivell que vagin seguides després per anàlisis de nivells superiors (per exemple, amb gramàtiques HPSG).

. 2 finalitats:

. extreure informació dels corpus

. processar fiablement i amb eficiència els corpus

D'aquesta manera, es poden caracteritzar els diferents programes segons els dos paràmetres. De fet, en les ressenyes d'investigació recents s'hi poden trobar exemples de programes que corresponen a cada combinació; d'entre aquestes n'hem escollit una de cada per a presentar-la breument aquí:

. probabilístic per a extreure informació:

. detecció del *pp-attachment* (Hindle i Rooth, 1993)

. probabilístic per a processar:

. traducció anglès-francès (Brown et al., 1993)

. lingüístic per a extreure informació:

. adquisició d'informació de semàntica lèxica (Pustejovsky et al. 1993)

. lingüístic per a processar

. etiquetador sintàctic de corpus (Karlsson, 1990)

A continuació passem a descriure'ls breument, tot intentant de fer-ne sobresortir els aspectes metodològics essencials.[16]

a) detecció del pp-attachment en anglès

Un dels problemes de l'anàlisi sintàctica bàsica en anglès (com també, en gran mesura, en català) és la determinació de la dependència dels sintagmes preposicionals que ocorren després d'una seqüència superficial de VERB (+ DETS) + NOM. Un exemple paradigmàtic d'aquest problema el constitueix l'oració *I saw the man with the telescope*, que permet dues anàlisis diferents segons es resolgui la dependència del sintagma preposicional; en un primer sentit, *with the telescope* depèn de *man* i, per tant, forma part de l'objecte directe (que podria ser parafrasejat per *l'home que portava el telescopi*), mentre que en el segon sentit el sintagma preposicional depèn del verb *saw* i, per tant, ha de ser interpretat com un complement modal del verb (de manera que l'oració podria ser parafrasejada per *vaig veure l'home mitjançant el telescopi*). Des d'un punt de vista del

[16] Òbviament, per a un coneixement detallat del seu funcionament, convé recórrer a les fonts originals que citem.

processament sintàctic, aquesta ambigüitat estructural representa un problema perquè comporta una duplicitat d'anàlisi en la majoria d'oracions que s'analitzen, fins i tot en els casos en què no hi ha una tal ambigüitat per a un observador humà. Els diversos plantejaments que s'han fet per a intentar resoldre el problema no han arribat a proporcionar una solució adequada i eficient.

Hindle i Rooth (1993) proposen d'aplicar a aquest problema la tècnica de la determinació d'associacions lèxiques. De fet, es tracta d'analitzar aquestes relacions entre un verb o un nom i una preposició com casos de col·locacions. L'estratègia bàsica de la seva proposta consisteix a detectar en el corpus les associacions que hi ha entre els verbs i les preposicions i entre els noms i les preposicions. Això s'aconsegueix a partir de diverses passades de detecció pel corpus i de la creació d'una taula auxiliar d'associacions.

Naturalment, el primer pas consisteix en l'etiquetat morfològic i l'anàlisi sintàctica bàsica posterior (que no arriba a determinar les dependències entre els diversos constituents, per a no prejutjar, és clar, el resultat del processament en curs). D'aquesta manera, s'obté una seqüència d'arbres que representen les estructures parcials de les oracions del corpus.

A partir d'aquest resultat, es construeix una taula que conté tots els noms que són el nucli de sintagmes nominals. Per a cada nom, es consigna la preposició que el segueix (si n'hi ha cap) i el verb que el precedeix (si el sintagma nominal n'és l'objecte directe). Una columna addicional de la taula serveix per a marcar els casos en què l'estructura sintàctica determina la solució del problema de l'associació de la preposició (per exemple, quan no hi ha cap verb que s'hi pugui associar, com en el cas dels sintagmes nominals en posició de subjecte).

Un cop obtinguda la taula, s'analitza aquesta en diverses passades amb la finalitat de determinar les associacions lèxiques. És en aquest estadi quan s'utilitzen les tècniques de caràcter probabilístic. Les successives passades per la taula consisteixen en: marcar els noms i verbs que ocorren sense preposició, detectar les preposicions que van sens dubte amb el verb (quan el nom amb el qual podrien competir és un pronom o quan el verb és en veu passiva -excepte per a la preposició *by*), detectar les preposicions que van sens dubte amb el nom (quan la posició del sintagma nominal no permet que hi hagi un verb amb el qual la preposició podria estar associada -per exemple, quan és el subjecte o

qualsevol altre complement preverbal), associar per un índex de probabilitat les preposicions no associades encara (en dues etapes, segons l'índex de probabilitat) i, finalment, associar al nom les preposicions que encara no han estat associades en cap passada anterior.

En la documentació presentada pels autors s'explica el sistema emprat per a avaluar els resultats obtinguts, que consisteix bàsicament en la comparació amb els resultats obtinguts per avaluadors humans en relació a una mostra del corpus. En aquest context, els autors parlen d'una associació correcta per part del programa en un 80% dels casos aproximadament.

b) traducció anglès-francès

Brown et al. (1993) proposen una sèrie de models estadístics del procés de traducció entre textos (és a dir, entre conjunts de parells d'oracions en dues llengües que són la traducció les unes de l'altres).

El seu punt de partida són les tècniques d'alineament de corpus, que consisteixen en relacionar els elements lingüístics d'un text amb els de l'altre. D'aquesta manera, si el procés té èxit s'aconsegueix no només que les oracions dels dos textos estiguin aparellades, sinó que ho siguin també els elements lingüístics que les formen (és a dir, el que els tècnics en traducció anomenen les unitats bàsiques de traducció).

En aquest treball els autors presenten alguns models probabilístics de la relació de traducció, els més adequats dels quals poden donar compte dels aparellaments d'un a molts (és a dir, quan un mot d'una llengua correspon a diversos mots en l'altra) i dels creuaments (o sia, quan l'ordre dels elements no es correspon en les dues llengües, com en el cas de la posició de l'adjectiu en anglès i francès). Es tracta d'investigacions en curs que difícilment poden ser considerades encara com totalment adequades; de totes maneres mostren amb claredat les possibilitats (i els límits) actuals dels plantejaments probabilístics en relació al processament de corpus.

c) adquisició d'informació semàntica

La informació de semàntica lèxica és una de les que més ha resistit els intents de formalització i, també, de tractament automàtic. La manca, per una banda, de teories formalitzades de semàntica lèxica i, per l'altra, de dades massives sobre l'ús de les

paraules ha fet que aquest sigui un dels camps lingüístics en què el processament automàtic es troba en uns estadis més endarrerits. Pustejovsky et al. (1993) presenten una sèrie de tècniques relacionades entre elles per a adquirir informació de semàntica lèxica a partir de corpus textuals.

El seu punt de partida és la teoria de semàntica lèxica coneguda sota el nom de *generative lexicon* (Pustejovsky. 1991), que presenta una sèrie d'eines generatives per a derivar recursivament nous sentits per als elements lèxics d'una llengua. El sistema suposa quatre nivells diferents de descripció semàntica: l'estructura d'arguments, l'estructura de quàlies, l'estructura d'herència lèxica i l'estructura eventiva. Un d'aquests components, l'estructura de quàlies, caracteritza el sentit d'un mot a partir de diverses subestructures: la constitutiva (que estableix la relació entre un objecte i les seves parts constitutives), la formal (que permet d'inscriure l'objecte en un camp més ampli), la tèlica (que n'indica la finalitat i la funció), i l'agentiva (que especifica els factors que intervenen en el seu origen).

La finalitat del projecte exposat en l'article consisteix a detectar les estructures de semàntica lèxica a partir de l'anàlisi de les definicions de diccionaris en suport magnètic. Naturalment, la teoria semàntica del lèxic generatiu és la que permet d'estructurar la informació extreta i, fins i tot, de guiar el procés mateix de l'extracció.

Les etapes bàsiques de l'anàlisi són les següents: en primer lloc, es detecten les paraules desconegudes (que altrament farien fracassar els primer nivells del processament); a continuació, s'etiqueta (morfològicament) el corpus; després, se'l processa sintàcticament, de manera que se n'obté una anàlisi sintàctica parcial (representada per una estructura plana, és a dir, sense determinar totes les dependències dels constituents majors); en quart lloc, es procedeix al reconeixement dels compostos nominals i a determinar-ne l'estructura; després, es generen les relacions taxonòmiques a partir de la informació sobre col·locacions; i, finalment, s'extreu la informació relacionada amb els quàlies dels noms.

Els aspectes més nous d'aquest programa de recerca deriven de la interacció entre un plantejament teòric molt elaborat i les tècniques de detecció de la informació continguda en els corpus. S'hi fa evident que fins i tot en un camp tant poc procliu a la formalització com és la semàntica lèxica es poden utilitzar tècniques adequades per al tractament de corpus.

d) etiquetador sintàctic de corpus

Karlsson (1990) ens presenta l'analitzador de textos anglesos *English Constraint Grammar* (ENGCG). La característica essencial d'aquest analitzador és que utilitza eines de baix nivell (tant des del punt de vista del processament, com de la codificació lèxica). Això permet que estigui intrínsecament lligat amb un analitzador morfològic (que utilitza la tècnica de dos nivells) i que, com conseqüència, junts ofereixin una eina força potent des d'una perspectiva relativament nova en el camp.

L'estructura general del sistema suposa d'una manera molt estàndard els següents passos:

. preprocessament
. anàlisi morfològica
. desambiguació morfològica local
. projecció morfosintàctica
. anàlisi sintàctica

Després de l'etapa preparatòria de preprocessament (encarregada de normalitzar el text), el primer gran pas és el de l'anàlisi morfològica, que s'efectua en base a un analitzador de dos nivells (vegeu més amunt, §3.2). Un cop generades totes les estructures morfològiques possibles d'un mot, es planteja la possibilitat de desambiguar des d'una perspectiva estrictament local; bàsicament es comparen les estructures obtingudes i s'avalua la possibilitat que alguna d'elles sigui exclosa (en funció de principis com el que estableix que en igualtat de circumstàncies és preferible l'anàlisi que implica el mínim nombre de límits de composició, o sia la paraula més simple des del punt de vista de la composició). La quarta etapa consisteix en l'assignació de funcions sintàctiques als elements lèxics que queden; d'aquesta manera es prepara el terreny per a l'aplicació de les restriccions pròpies de l'anàlisi sintàctica.

La darrera etapa consisteix en l'anàlisi sintàctica pròpiament dita, la qual consisteix en tres processos independents: desambiguació morfològica contextual, determinació dels límits oracionals i desambiguació de les funcions sintàctiques superficials. En el primer procés, es produeix la desambiguació contextual típica a partir de la seqüència de classes de mots en la qual està inserit el que estem analitzant. A la vegada, la determinació dels límits oracionals és essencial per a poder aplicar les diverses restriccions, tant les de caràcter estrictament morfològic com les sintàctiques. Finalment, el tercer procés elimina

interpretacions a partir de l'aplicació de restriccions sintàctiques formulades en termes de seqüències permissibles de funcions sintàctiques superficials. En un apartat anterior, §3.4, hem presentat un exemple del que seria el resultat de l'anàlisi d'una petita oració.

L'aspecte més interessant i nou d'aquest plantejament és que veu com un tot el processament morfològic i el sintàctic (almenys, el sintàctic de baix nivell). Això naturalment implica utilitzar tècniques similars i compatibles, però a la vegada permet de maximitzar els resultats, ja que comporta un estalvi de processament. Tot analitzador morfològic ha d'anar acompanyat per un desambiguador, el qual ha d'usar informació lingüística general i, sobretot, sintàctica. En la majoria de plantejaments aquesta informació sintàctica utilitzada no queda fixada en el text que s'està processant, simplement es perd. Llavors si s'ha de procedir a l'anàlisi del text, en els estadis posteriors s'ha de reproduir o recalcular. La desambiguació morfològica de ENGCG, en canvi, és feta a partir del coneixement lingüístic sintàctic explícit, cosa que permet d'etiquetar amb característiques de sintaxi superficial els mots del text.

4.5. *Estratègies generals per al processament de corpus*

Com dèiem més amunt (§4.3), per a processar sintàcticament corpus de manera fiable i eficient calen tècniques mixtes. Aquest és un fet àmpliament reconegut avui dia, de manera que ja han passat gairebé a la història les discussions quasi gremials entre les aproximacions estadístiques i les lingüístiques.

Les tècniques mixtes han de ser incorporades tant en el tractament a baix nivell com en el d'alt nivell. Quant al primer, sembla convenient de donar preferència (quan sigui possible) a les estratègies explícitament lingüístiques, complementades per tècniques probabilístiques sempre que calgui. Respecte al segon, cal tenir clar que no és adequat per a totes les tasques, especialment per a les més properes a l'anàlisi morfològica i a la sintàctica de primer nivell; les tècniques d'alt nivell són útils, fins i tot imprescindibles, en alguns casos (quan cal arribar a la forma lògica o representació semàntica, quan convé d'establir les relacions sintàctiques profundes...), però no ho són sempre.

Aquesta integració de tècniques que proposem ha de funcionar en dues direccions diferents, però complementàries: en el disseny general de sistemes i en el millorament de les estratègies d'anàlisi (incloent l'anotació dels diccionaris i regles).

Quant al *disseny general de sistemes*, convé tenir present un principi general fonamental: hem de tenir una visió global del processament, des del punt de partida (és a dir, el text real), fins allà on volem arribar. Les eines de què ens dotem han d'estar plantejades de manera que puguin ser inserides en aquest procés general. Això correspon a un dels principis essencials en programació, el de la modularitat; naturalment, convé fragmentar els processos, de manera que cada tasca pugui ser programada independentment, però a la vegada convé que els enllaços entre processos siguin explicitats (altrament podem estar duplicant feines innecessàriament).

Per altra banda, és convenient de recórrer a tècniques que han anat mostrant la seva eficiència al llarg de la història del processament sintàctic. En aquest sentit, podem apuntar a la incorporació d'un bon preprocessament, que normalitzi el text abans de processar-lo, que detecti els mots que no tenim en el nostre diccionari, que detecti les expressions semifixades, col·locacions... També és convenient de prestar atenció a l'organització general del procés: és important d'anticipar tanta informació com sigui possible (és a dir, resoldre els problemes tan aviat en el procés com es pugui), cal tractar l'ambigüitat de manera econòmica –no creant estructures arbòries independents per a cada solució o no utilitzant categories massa indeterminades–, s'ha de ser escrupolosament modular i tant declaratiu com sigui possible, és bàsic de dedicar esforços a la racionalització del diccionari... Finalment, també és important de considerar l'ús de tècniques relativament perifèriques al sistema, com permetre la relaxació de regles (que fa possible l'obtenció d'anàlisis parcials, quan una anàlisi global fracassa), o introduir (limitadament, és clar) la interactivitat amb l'usuari del programa.

En relació al *millorament de les tècniques d'anàlisi*, cal assenyalar que actualment és possible de millorar amb informació probabilística el comportament dels analitzadors sintàctics. Aquest enfocament ha pres dues direccions diferents: per una banda, s'anoten probabilísticament els tipus de dades lingüístiques que manipula l'analitzador (regles i entrades lèxiques), i, per l'altra, es combina la informació probabilística amb la que manipula directament l'analitzador.

En el primer sentit, es poden graduar les entrades lèxiques, segons els seus factors d'aparició, o les regles, segons el seu índex d'aplicació; d'aquesta manera, es redueix el nombre d'intents fracassats en l'anàlisi; alternativament, es poden graduar les estructures resultants del procés d'anàlisi (segons les entrades lèxiques o les regles usades,

l'estructura obtinguda té un o altre índex de probabilitat). Sens dubte, la primera aproximació permet una reducció de les dades manipulades durant el procés, mentre que la segona només fa possible escollir entre els resultats obtinguts.[17]

Per altra banda, en la segona direcció, trobem analitzadors com el proposat per Briscoe i Carroll (1993), que millora el poder dels analitzadors amb taula en introduir a la taula informació probabilística adquirida automàticament.

Finalment, l'adopció de les tendències lexicalistes actuals (tant en lingüística teòrica com computacional) facilita la racionalització del procés de construcció de la gramàtica i la interacció entre els diversos mòduls. Un primer factor essencial és que les gramàtiques fortament lexicalitzades poden ser construïdes de manera que les entrades lèxiques siguin l'elements d'enllaç entre els diversos mòduls. Això implica, naturalment, que el lèxic global de tot el sistema ha de ser concebut com un tot, independentment del fet que hi hagi parts de la informació lèxica que només siguin accedides en un determinat mòdul del procés; d'aquesta manera, ens acostem a un plantejament, que es va imposant recentment, que consisteix a distingir entre el diccionari general del sistema (que és extern a tots els mòduls i que conté la informació, convenientment estructurada, necessària per a tots ells) i els diccionaris particulars als quals ha d'accedir cada mòdul en concret: el disseny del sistema, llavors, ha de permetre de buidar la part del lèxic rellevant per a un procés determinat abans que aquest comenci a aplicar-se.

Els diversos estudis sobre avaluació de sistemes computacionals que s'han dut a terme darrerament fan pensar que és molt més fàcil de treballar al voltant del lèxic que del sistema de regles de la gramàtica. Per una banda, la informació codificada lèxicament resulta molt més inspeccionable i, per tant, intel·ligible. Per altra banda, és molt més fàcil de transportar d'un sistema a un altre la informació lèxica que la que queda distribuïda en

[17] En el formalisme usat al programa de traducció automàtica Eurotra, format, en els seus aspectes més essencials, per un esquelet lliure de context i molt poc eficient (a causa de l'espai que ocupaven les dades manipulades en el procés d'anàlisi), s'hi van introduir en els últims temps unes eines per a avaluar probabilísticament les dades que manipulava l'analitzador. Sens dubte, les que implicaven una reducció de les dades manipulades en el temps de processament són les que van produir un augment més gran en l'eficiència global del procés.

un nombre indeterminat de regles o principis.[18] D'aquesta manera, la lexicalització dels sistemes computacionals ha contribuït a la facilitat de reutilització dels recursos creats.[19]

5. Conclusions

Després d'aquesta ràpida panoràmica que hem anat resseguint podem avançar unes primeres conclusions, que fem constar de manera esquemàtica:

- el processament gramatical dels corpus de manera automàtica està encara en estadis força incipients, tot i que els resultats són força satisfactoris pel que fa al tractament morfològic i sintàctic superficial,
- les diverses tècniques que s'usen o s'experimenten no han de ser vistes com a antagòniques, sinó com a complementàries; això és cert també pel que fa a la disjuntiva entre tècniques probabilístiques i tècniques explícitament lingüístiques.
- les úniques aplicacions reals i eficients que avui ens podem imaginar impliquen l'ús de tècniques mixtes, tant si es tracta d'extreure informació de corpus com de processar-los de manera eficient,
- el pes principal del sistema de processament l'han de dur les tècniques de baix nivell, a les quals s'incorporen mòduls sofisticats quan cal (especialment quan cal recórrer a l'anàlisi semàntica),
- és preferible que les tècniques de baix nivell utilitzin estratègies lingüístiques per a garantir l'enllaç amb els mòduls clarament lingüístics d'anàlisi sintàctica i semàntica i per a aprofitar directament el coneixement en què es basen,
- les gramàtiques d'unificació (fortament lexicalistes) són avui per avui les més adequades per a la formulació dels mòduls d'alt nivell, especialment aquelles que han sorgit de les preocupacions directament computacionals, i
- l'estructura del sistema ha de ser modular, però altament organitzada (preferentment,

[18] Òbviament, aquest tipus de consideracions han propiciat el procés de lexicalització de la gramàtica (Badia, 1994).

[19] Com ja hem mencionat més amunt, aquest procés no ha estat exclusiu de la lingüística computacional, sinó que ha afectat tota la lingüística. Per a veure algun dels elements que es consideren bàsics en els processos de reutilització dels recursos creats, podeu veure Sadler i Markantonatou (1994).

al voltant del lèxic, de manera que se'n maximitzi les possibilitats de transport i reutilització).

Bibliografia

BADIA, T. (1994), *Lexicografia i models lingüístics: Les teories lingüístiques i el lèxic*, "Caplletra", 17, pàg. 15-46.

BLANCHE-BENVENISTE, C. i TEMPLE, L. (1986), *Décrire le français parlé*. "Le français dans le monde", pàg. 26-33.

BRESNAN, J. (ed.) (1982), *The mental representation of grammatical relations*, MIT Press, Cambridge, Mass.

BRISCOE, T. i CAROLL, J. (1993), *Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars*, "Computational Linguistics", 19, 1, pàg. 25-60.

BROWN, P. F., DELLA PIETRA, S. A., DELLA PIETRA, V. J. i MERCER, R. L. (1993), *The Mathematics of Statistical Machine Translation Parameter Estimation*, "Computational Linguistics", 19, 2, pàg. 263-312.

CARPENTER, B. (1992), *The logic of typed feature structures*, Cambridge University Press, Cambridge.

CHANDIOUX, J. (1976), *Météo: un système opérationel pour la traduction automatique des bulletins météorologiques destinés au grand public*, "Meta", 21, pàg. 127-133.

CHURCH, K.W. (1988), *A stochastic parts program and noun phrase parser for unrestricted text*, dins *Second Conference on Applied Natural Language Processing*, Austin, Texas.

CHURCH, K. W. i MERCER, R. L. (1993), *Introduction to the special issue on computational linguistics using large corpora*, "Computational Linguistics", 19, 1, pàg. 1-24.

GARSIDE, R., LEECH G., i SAMPSON G. (eds.), (1987), *The computational analysis of English: A corpus based approach*, Longman, Londres.

GAZDAR, G. i MELLISH, Ch. (1989), *Natural Language Processing in Prolog*, Addison, Wesley.

GAZDAR, G., KLEIN, E., PULLUM, G. K. i SAG, I. (1985), *Generalised Phrase Structure Grammar*, Blackwell, Oxford.

VAN HERWIJNEN, E. (1994). *Practical SGML*, Kluwer, Boston.

HINDLE, D. i ROOTH, M. (1993), *Structural Ambiguity and Lexical Relations*, "Computational Linguistics". 19, 1, pàg. 103-120.

KARLSSON, F. (1990), *Constraint Grammar as a Framework for Parsing Running Text*, "Coling-92", pàg. 168-173.

KAY, M. (1984), *Functional unification grammar: a formalism for machine translation*, "Coling-84", pàg. 75-78.

KOSKENNIEMI, K. (1983), *Two-level model for morphological analysis*, "IJCAI-83", pàg. 683-685.

PÉREZ, R., TROTZIG, D. i LLORÉ, X. (1994), *Morfeo: analizador morfológico y* tagger *del español*, dins *Actas del X Congreso SEPLN*, Còrdova.

POLLARD, C. J. i SAG, I. (1987), *Information-based syntax and semantics*, CSLI, Chicago University Press, Stanford.

_____ (1994), *Head-driven phrase structure grammar*, CSLI, Chicago University Press, Stanford.

PUSTEJOVSKY, J. (1991), *The generative lexicon*, "Computational Linguistics", 17, 4, pàg. 409-441.

PUSTEJOVSKY, J., BERGLER, S. i ANICK, P. (1993), *Lexical Semantic Techniques for Corpus Analysis*, "Computational Linguistics", 19, 2, pàg. 331-359.

SADLER, L. i MARKANTONATOU, S. (eds.) (1994), *Grammatical Formalisms: Issues in Migration*. CEC, Luxemburg.

SHANK, R. (1980), *Language and memory*, "Cognitive Science", 4, 3, pàg. 243-284.

SHANNON, C. (1948), *The mathematical theory of communication*, "Bell System Technical Journal", 27, pàg. 398-403.

SHIEBER, S. M. (1986), *An introduction to unification-based approaches to grammar*, CSLI, Chicago University Press, Chicago.

El corpus de conversación coloquial del grupo *Val.Es.Co.*

Antonio Briz (coord.)
(Universidad de Valencia)

Val. Es. Co, (Valencia. Español Coloquial) es un grupo de investigación surgido en el seno del Departamento de Filología Española de la Universidad de Valencia en 1990 que tiene como objeto de estudio el español coloquial. El grupo está constituido por profesores y becarios de investigación de los Departamentos de Filología Española y Teoría de los Lenguajes: José Ramón Gómez Molina, Enrique Serra, Beatriz Gallardo, Antonio Hidalgo, Salvador Pons, Leonor Ruiz, Juan Gómez, Julia Sanmartín, Ana Portela, José Padilla, Begoña Gómez, Marcial Terrádez y Immaculada Baixaulí.

1. Proyecto de investigación[1]

1.1. *Objeto de estudio e hipótesis de partida*

El objeto del proyecto de investigación del grupo *Val.Es.Co.* es, como ya se ha señalado, abordar el estudio del español coloquial en sus diferentes niveles de análisis a

[1] Vid. BRIZ, A. y GÓMEZ MOLINA, J. R. (1992), *Scheme of Study of Colloquial Spanish: Some Methodological Considerations*, "LynX. A Monographic Series in Linguistics and World Perception", 3, pág. 111-124.

partir de un corpus básicamente oral, extraído directamente de la conversación espontánea y de otro tipo de discursos.

Un trabajo ya concluido ha sido la recopilación y selección de un corpus de español coloquial, concretamente de conversación coloquial,[2] a partir fundamentalmente de los datos que ofrece la conversación corriente y natural en las relaciones cotidianas. Todo el material se ha obtenido por medio de grabaciones secretas y no secretas, con participación o no del investigador; conversación semidirigida o libre.

Su análisis en varias fases nos permitirá más tarde:

a) la caracterización formal del registro coloquial: descripción y explicación de los aspectos lingüísticos (frecuencias lingüísticas) que caracterizan este registro de habla,

b) el estudio de hechos relacionados con la competencia y práctica comunicativas: la estructura conversacional, la alternancia de turnos, la configuración secuencial de los mismos, la organización temática, el desarrollo y progresión del coloquio.

c) será posible de este modo abordar el estudio de la interacción entre las formas lingüísticas y sus funciones pragmático-comunicativas, lo que, en nuestra opinión, permitirá explicar el carácter sistemático de muchos fenómenos del registro coloquial.

d) por último, nuestro análisis se detendrá en los fenómenos de covariación sociolingüística, si los hubiera.

Partimos de la hipótesis de que es posible establecer un modelo de base pragmalingüística, basado en la posibilidad de integrar el análisis del discurso y el análisis conversacional, que explique la "gramática" del coloquio, no como ruptura o transgresión de la "otra gramática", sino como un conjunto de estructuras que llegan a constituirse como tales en el proceso mismo de interacción pragmática.

1.2. *Los inicios de la investigación*

Cuatro años de intensa actividad del grupo, de largas discusiones cada semana en interminables tardes de reunión, de desesperantes inicios, porque poco o nada parecía avanzar, han dado sus frutos, no sólo en el primero de los objetivos previstos, la elaboración de un amplio corpus de referencia del español coloquial, del que aquí se

[2] Más adelante insistimos en la necesidad de separar, por un lado, el registro "español coloquial" y, por otro, la conversación coloquial, como tipo de discurso que emplea dicho registro.

extrae una muestra, sino de algunos trabajos publicados o en fase de publicación.[3]

[3] Sirvan de botón de muestra los siguientes:

BRIZ, A. (1993), *Notas de español coloquial para extranjeros*, en *Actas del Simposio sobre* El español de España y el español de América. Virginia 1991, Valencia-Virginia UP, Virginia.

_____ (1993), *Los conectores pragmáticos en español coloquial: su papel argumentativo*, "Contextos", XI/21-22, pág. 145-188.

_____ (1993), *El papel metadiscursivo del conector pragmático en español*, "Español Actual", 59, pág. 39-56.

_____ (1994), *Hacia un análisis argumentativo de un texto coloquial*, "Verba", 21, pág. 369-395.

_____ (1995), *La atenuación en la conversación coloquial: una categoría pragmática*, en *Actas del I Simposio sobre el español coloquial*, Universidad de Almería.

BRIZ, A. et al. (1994), *La elaboración de un corpus de español coloquial. Problemas metodológicos previos*, en *Cahiers du Centre Interdisciplinaire des Sciences du Langage. Actes du Colloque* Le Dialogue en question. *Lagrasse, 1993*, Université de Toulouse-Le Mirail, pág. 103-109; revisado en *Actas del I Congreso de Lingüística General, Valencia, 1994*, (en prensa).

GALLARDO PAÚLS, B. (1992), *El dinamismo conversacional: subsunción y feed-back. Carácter acumulativo de los elementos conversacionales*, "Comunicación y sociedad", V, 1 y 2, pág. 51-75.

_____ (1993), *Lingüística perceptiva y conversación: secuencias*, "LynX", Annexa 4.

_____ (1993), *La transición entre turnos conversacionales: silencios, interrupciones y solapamientos*, "Contextos", XI/21-22, pág. 189-220.

_____ (1994), *Conversación y sintaxis*, en *Actes del I Congrés de Lingüística General*, Valencia.

GÓMEZ, B. (1994), *Dequeísmo y queísmo en el español hablado de Valencia*, Tesis de Licenciatura (inédita), Valencia.

GÓMEZ, J. R. (en prensa), *Dequeísmo y queísmo en el habla de Valencia: aproximación sociolingüística*, en *Actas del VI Colloque de Linguistique Hispanique (Toulouse-LeMirail), 1994*.

_____ (1995 (en prensa)), *Observaciones sobre la función de los extranjerismos en español coloquial: valores estilísticos, semánticos y pragmáticos*, en *II Simposio de Pragmática y Gramática del Español Hablado. El Español coloquial*.

GÓMEZ CAPUZ, J. (1994), *Observaciones sobre las variedades y registros de una lengua histórica; ensayo de delimitación y extracción de materiales*, en *Actes del I Congrés de Lingüística General*, Universidad de Valencia, Valencia.

HIDALGO NAVARRO, A. (en prensa), *Sobre los mecanismos de impersonalización en la conversación coloquial: el TÚ impersonal*, "Estudios de Lingüística", 11.

_____ (1994 (en prensa)), *Entonació i conversa: Aproximació als mecanismes prosòdics demarcatius d'unitats sintàctiques en la parla col·loquial*, "Caplletra".

_____ (1996) *Entonación y conversación coloquial. Sobre el funcionamiento demarcativo-integrador de los rasgos suprasegmentales*, Tesis doctoral, Universidad de Valencia.

HIDALGO NAVARRO, A. y PONS BORDERÍA S. (1991), *Algunas consideraciones sobre la paradoja del observador*, en *Actas del I Simposio de Lingüística Aplicada y Tecnología*, Valencia.

Grabar era fácil, lo difícil era dónde, cómo, a quién(es) grabar y, sobre todo, de qué modo tener la seguridad de que la grabación obtenida era representativa de aquello que llamábamos español coloquial, y que todos intuitivamente de forma previa creíamos conocer. Las tres primeras cuestiones hacían referencia a la técnica de grabación, al método y a su representatividad sociológica; la cuarta al propio reconocimiento del objeto de estudio.

Para obtener una grabación técnicamente buena el dónde y el cómo eran fundamentales; metodológicamente, por otro lado, la situación de grabación y la técnica de grabación (secreta, no secreta) eran básicas para obtener el mayor grado de espontaneidad.

Partiendo de la idea de que cuanto más familiar (menos marcado) fuera el lugar de grabación, más fácil sería obtener una muestra del llamado español coloquial, las primeras grabaciones se realizaron en espacios públicos cotidianos: bares, tiendas de barrio,

PADILLA, J. (1994), *Aproximación al estudio del orden de palabras en español coloquial*. Tesis de licenciatura (inédita), Valencia.

PONS, S. (en prensa), *La presencia de los enlaces extraoracionales en la tradición gramatical española (I): La clasificación de las conjunciones ilativas y continuativas*, "Anuario de Lingüística".

_____ (en prensa), *La presencia de los enlaces extraoracionales en la tradición gramatical española (II): La figura de Andrés Bello*.

_____ (1994) *Las perífrasis de relativo, la Teoría de la Polifonía y una discrepancia entre Bello y Cuervo, "LEA"*, XVI, 1.

PORTELA, A. (1994), *La negación semántico-pragmática en español coloquial*, Tesis de Licenciatura (inédita), Universidad de Valencia.

RUIZ GURILLO, L. (1992 (en prensa)), *Sobre la fraseología coloquial: corpus e investigación* en *I Congreso Internacional de AESLA: el Español: Lengua Internacional, 1492-1992*, Granada.

_____ (1994) *Algunas consideraciones sobre las estrategias de aprendizaje de la fraseología del español como lengua extranjera*", en Sánchez Lobato y I. Santos (ed.) *Problemas y métodos en la enseñanza del español como lengua extranjera. Actas del IV Congreso Internacional de ASELE*, SGEL, Madrid, pág. 141-151.

SANMARTÍN, J. (1994), *El argot común en el ámbito europeo*, "Studia Lingüística. Cuadernos de Filología".

_____ (1994), *De la parla marginal a la llengua estándard*, en *Actes del I Congrés de Lingüística General*, Valencia.

autobuses, etc. Además para asegurar el máximo de espontaneidad se recurrió de forma exclusiva en aquellos momentos iniciales a la grabación secreta sin intervención participante del investigador. No obstante lo anterior, las restricciones medioambientales y, quizá también, el tipo de magnetófono impedían muchas veces una perfecta grabación. En el autobús, el ruido del motor y la abertura y cierre de las puertas hacían imposible más tarde la transcripción de la cinta; no eran menos las dificultades que planteaba la grabación en un bar: si había muchos clientes, se superponían varias conversaciones, y si, por el contrario, sólo estaba ocupada una mesa, el investigador no podía acercarse lo suficiente para obtener una buena grabación; todo ello sin contar con el ruido de las máquinas tragaperras, que punteaban la conversación con su música de verbena.

Por ello se recurrió también a la grabación secreta con observación participante del investigador, dado que esta técnica nos permitía ampliar los espacios de grabación. a la vez que facilitaba posteriormente la tarea de transcripción que el mismo investigador llevaba a cabo.

Los informantes, no sólo en número, sino en características sociológicas, debían ser representativos del repertorio lingüístico del área estudiada, en nuestro caso, Valencia y área metropolitana.[4]

Superados estos obstáculos teóricos y prácticos iniciales (*vid.* §2), debíamos resolver el problema fundamental al que ya apuntaban las cuestiones anteriores: el reconocimiento del español coloquial. Es decir, antes de grabar, incluso una vez grabada la conversación, había que determinar si era representativa de eso que llamábamos "español coloquial".

Comprobada la insuficiencia de las definiciones dadas hasta ahora para reconocer el llamado español coloquial, antes que otra definición de características similares, había que proponer un modo operativo de reconocimiento (*vid.* §3). Así pues, ya que había acuerdo en la identificación del español coloquial como registro de uso en una situación determinada de comunicación, la cuestión era precisar las características pragmáticas y situacionales que propiciaban el desarrollo de una conversación coloquial y que nos permitirían reconocer de forma previa dicho registro. Varias preguntas que aludían al supuesto carácter coloquial de una hipotética grabación en determinadas circunstancias fueron guiando la reflexión teórica: ¿es coloquial la conversación entre dos amigos en tal

[4] Se ha considerado comunidad de habla de Valencia y su área metropolitana a la integrada por 44 municipios y una población de 1.321.197 habitantes, según datos del padrón de 1991.

o cual situación?, ¿es coloquial la conversación entre dos personas cuya relación vivencial es nula?, ¿podemos negar la coloquialidad de una interacción si existe relación de +poder y -solidaridad (superior solidario) entre los interlocutores?, ¿se puede usar este registro con alguien que no se conoce?, ¿es posible hablar coloquialmente de pragmática? etc. Se trataba así de observar la posible influencia de la situación, de las relaciones vivenciales, de las relaciones de poder, de la temática..., en la consideración y empleo del registro coloquial.

Al estudio de los rasgos pragmáticos -y no antes- seguiría el estudio lingüístico y el de las correlaciones lingüísticas y sociológicas, objetivo principal que guiaba la obtención de dicho corpus. De estas y otras cuestiones nos ocupamos más ampliamente a continuación.

2. La elaboración del corpus general

Para la obtención del corpus hemos considerado necesario aplicar criterios de metodología sociolingüística, acordes con algunos objetivos de la investigación, por dos motivos.

El primero, de carácter general, por la utilidad y ventajas que ofrece esta disciplina en sus planteamientos; entre ellos, el estudio del habla viva en el contexto social real, su carácter empírico, que la convierte en disciplina no normativa,[5] y su interpretación de los hechos de lengua dentro del contexto social en que se producen, analizados desde una perspectiva interactiva.

Es cierto que los estudios sociolingüísticos se han ocupado en su mayor parte de los aspectos fónicos y léxicos, ya que en éstos se comprueba más fácilmente la estratificación sociocultural, así como por razones de comodidad en la recopilación de datos, y han investigado menos las variables sintácticas porque, entre otras razones, se supone que apenas cuentan como índice de conciencia lingüística.[6] Ahora bien, esto no quiere decir que los hechos sintácticos sean menos relevantes, sino que, por ser los más desatendidos, deben constituirse en uno de los principales objetos de estudio de la lengua hablada y, en nuestro caso, de la conversación coloquial.

---

[5] *Vid.* ROTAETXE, K. (1988), *Sociolingüística*, Síntesis, Madrid, pág. 18-19.

[6] *Vid.* NARBONA, A. (1989), *Sintaxis española: nuevos y viejos enfoques*, Ariel, Barcelona, pág. 158.

La actuación discursiva concreta queda condicionada fundamentalmente por dos factores que operan de manera simultánea y que pueden originar diversos registros dentro de un *continuum* en cada acto comunicativo. Son, de una parte, el hablante-oyente (con sus factores psico-sociales) y, de otra, la relación de éstos con el contexto situacional. Todo hablante posee y domina, según su nivel sociocultural, una gama más o menos amplia de registros lingüísticos funcionalmente diferenciados, que son complementarios y se adaptan a las conveniencias situacionales del momento.

El segundo motivo, porque en el análisis del español coloquial -discurso conversacional- la Sociolingüística (como lingüística diastrático-diafásica que es) nos ayudará a descifrar si los mecanismos discursivos y la gramática de este registro covarían con factores sociales, es decir, si podemos establecer, siempre que se constate dicha variación, categorías o variables sociolingüísticas.

2.1. *Tamaño de la muestra y técnica de muestreo*

Es aún cuestión discutida entre lingüistas y sociolingüistas el número óptimo de individuos que debe componer la muestra de hablantes seleccionada. Por nuestra parte, intentamos aplicar aquellos procedimientos que han demostrado ser fiables y válidos y que luego nos permitirán obtener inferencia estadística. La muestra de informantes que proponemos pretende ser una reproducción exhaustiva y representativa de la comunidad de habla que investigamos, de acuerdo con los requisitos sociológicos. Se trata así de conseguir informantes cuyos idiolectos sean representativos del repertorio lingüístico de la comunidad de habla.

El plan de muestreo diseñado para una muestra aleatoria simple supone un nivel de confianza del 90% en los resultados obtenidos, con un margen de error en esa confianza del 6%. Este nivel de precisión[7] exige una muestra de 189 informantes, valor que nos permitirá un alto grado de seguridad en las inferencias obtenidas.[8]

---

[7] *Cfr.* ALBERDI J. et al. (1969), *Metodología de investigación por muestreo*, Fundación FOESSA, Madrid; COCHRAM, G. (1971), *Técnicas de muestreo*, Continental, México; SELLTIZ C. et al. (1980), *Métodos de investigación en las relaciones sociales*, Rialp, Madrid.

[8] Si bien es cierto que desde la perspectiva sociolingüística existe un interés mayor en el aspecto cualitativo que cuantitativo de los datos, nosotros deseamos mantener además un rigor sociológico. Este tamaño de muestra y su estratificación pudiera ser en opinión de algunos sociolingüistas elevado, pero estimamos que deben ser los procedimientos estadísticos, sobre todo el análisis multivariable, los que determinen cuál es el umbral de significación a partir del cual las conductas lingüísticas comienzan a ser

Determinado el tamaño de la muestra, la técnica de muestreo es otro de los pilares básicos. En nuestro caso, el método elegido es un muestreo estratificado por cuotas donde se asigna a cada celdilla una muestra según las variables sociológicas elegidas, tal como se refleja más abajo en el cuadro.[9] Los informantes de cada celda se han conseguido utilizando distintos procedimientos: elección intencionada (utilizando como criterio la opinión del investigador); aleatoria o accidental (cuando se ha presentado la oportunidad de grabar una conversación de diálogo libre en el mercado, en el parque, ...), etc.

2.2. *Variables sociológicas. Cuadro-resumen de la muestra*

Los factores de estratificación social que hemos establecido para esta investigación son los siguientes:

---

redundantes.

[9] Presentamos los valores de la muestra seleccionada de informantes, pues ocurre con frecuencia que los distintos componentes de la muestra no están en la misma proporción que en los estratos correspondientes de la población. Las fuentes han sido: para Valencia ciudad, el Boletín de Estadística Municipal referido al padrón de 1991, y para el Área Metropolitana, el Anuario estadístico del Consell Metropolità de l'Horta, también referido a 1991.

Valencia y Área metropolitana. Población (de hecho): 1.321.197 hab.

Variables sociológicas		Población real (%)	Muestra (%)
Sexo:	Varón	48.4 %	48.7 % (92)
	Mujer	51.6 %	51.3 % (97)
Demografía:	≤ 25 años	36.9 %	29.6 % (56)
	26-55 años	40.1 %	43.4 % (82)
	> 55 años	23 %	27 % (51)
Nivel sociocultural:	Alto	9 %	10.6 % (20)
	Medio	60 %	60.3 % (114)
	Bajo	31 %	29.1 % (55)
Lengua habitual:			
	Monolingües castellano	49.2 %	56.6 % (107)
	Bilingües	50.8%	43.4 % (82)

1) nivel sociocultural, determinado fundamentalmente por el nivel de estudios:[10]

Alto	(Código X)
Medio	(Y)
Bajo	(Z)

2) estrato generacional:

≤ 25	(I)
26 - 55	(O)
> 55	(U)

3) sexo:

Varón	(V)
Mujer	(M)

4) lengua habitual:[11]

monolingüe castellano	(E)
bilingüe	(B)

En cuanto a la procedencia u origen de los informantes, todos son nacidos o residentes, con un mínimo de quince años, en la ciudad de Valencia o su área metropolitana.

---

[10] Algunos estudios sociolingüísticos recientes (SAMPER, J. A. (1991), *El estudio de la disponibilidad léxica en Gran Canaria: datos iniciales y variación sociolingüística*, en *Actas del I Congreso Internacional* El estudio del español, Salamanca.) utilizan para determinar la estratificación sociocultural de los informantes el nivel educativo y la ocupación profesional, ya que son parámetros más simples y menos controvertidos que la clase social. Para la determinación de los tres estratos se ha establecido una baremación del nivel educativo (1 a 10) y de la ocupación profesional (1 a 6), quedando adscritos al nivel bajo o popular, el intervalo 2-5; al medio, el intervalo 6-13; y al nivel alto o culto, el intervalo 14-16. Aunque somos conscientes de la importancia que supone tomar en consideración la escala socioeconómica (renta) para observar fenómenos de hipercorrección como interpretación de prestigio lingüístico, dificultades de diversa índole, como la propia técnica empleada de grabación secreta, han impedido que pudiésemos controlar ese parámetro.

[11] La estratificación de esta variable, relacionada con la adquisición y desarrollo de las capacidades lingüísticas, no puede extraerse de los datos estadísticos municipales o comarcales. Nos hemos basado para ello en los datos de la Dirección General de Política Lingüística de la Conselleria de Cultura, Educación y Ciencia de la Generalitat Valenciana y en otros estudios sociolingüísticos sobre Valencia (vid. BLAS, J. L. (1993), *La interferencia lingüística en Valencia (dirección: catalán --> castellano)*, Universitat Jaume I, Castellón, Madrid.

CUADRO-RESUMEN DE LA MUESTRA

Edad	Nivel sociocultural						TOTAL	
	Alto		Medio		Bajo			
≤25	V 3 M 3 E 4 B 15	6	V 16 M 18 E 19 B 7	34	V 8 M 8 E 9 B 24	16	V 27 M 29 E 32 B 2	56
26-55	V 4 M 4 E 4 B 4	8	V 24 M 25 E 28 B 21	49	V 12 M 13 E 14 B 11	25	V 40 M 42 E 46 B 36	82
>55	V 3 M 3 E 3 B 13	6	V 15 M 16 E 18 B 6	31	V 7 M 13 E 14 B 11	25	V 25 M 26 E 29 B 3	51
TOTAL	V 10 M 10 E 11 B 9	20	V 55 M 59 E 65 B 49	114	V 27 M 28 E 31 B 24	55	V 92 M 97 E 107 B 82	189

2.3. *La recogida de datos. Modelo de ficha técnica*

Para llevar a cabo la recogida de datos se ha recurrido a las siguientes técnicas:

a) Grabación de una conversación o diálogo libre entre informantes,

-con observación participativa del investigador

-sin observación participativa

-grabación secreta (los hablantes ignoran que se les está grabando)

-grabación ordinaria (los hablantes son conscientes de la grabación)

b) Grabación de una conversación semidirigida (grabación ordinaria), encauzada hacia unos temas o propósitos concretos; intentamos evitar la identificación de ésta con una entrevista formal por parte del informante.[12]

La grabación secreta, con observación participativa o sin ella, ha sido la técnica más empleada en la recogida de datos, ya que constituye la forma más eficaz de recogida del llamado español coloquial y permite soslayar inconvenientes teóricos como la llamada paradoja del observador.[13] Otro tipo de técnicas, como la conversación semidirigida, sirven como instrumento auxiliar para la comprobación de determinados fenómenos.

Cada grabación va acompañada de una ficha técnica en la que se recogen estas y otras informaciones relevantes para el objeto de la investigación. (vid. al principio de cada una de las transcripciones de este volumen):

FICHA TÉCNICA:

a) Investigador:

Clave de la conversación:

b) Datos identificadores de la grabación

- Fecha de la grabación:
- Tiempo de la grabación:
- Lugar de grabación (zona, calle, distrito, mercado, parque, hospital, etc.):

c) Situación comunicativa:

- Tema o materia:
- Propósito o tenor funcional predominante:

  Interpersonal

  Transaccional

- Tono:

  Informal

  Formal

- Modo o canal:

---

[12] Una relación de las diversas técnicas para conseguir la naturalidad de los hablantes nos la muestran C. SILVA-CORVALÁN, ((1989), *Sociolingüística Teoría y análisis*, Alhambra, Madrid, pág. 30-35) y F. MORENO ((1990), *Metodología sociolingüística*, Gredos, Madrid, pág. 90-105).

[13] Vid. HIDALGO, A. y PONS, S. (1991), *Algunas consideraciones sobre la paradoja del observador*, en *Actas del I Simposio de Lingüística Aplicada y Tecnología*, Valencia.

d) Tipo de discurso registrado:

e) Técnica de grabación:

- Conversación libre

  Observador participante:

  Observador no participante:

  Grabación secreta:

  Grabación ordinaria:

- Conversación semidirigida (grabación ordinaria):

f) Descripción de los participantes

- Número de participantes: clave:

  activos:

  pasivos:

- Tipo de relación que los une:

- Sexo:

- Edad:

  ≤ 25

  26-55

  > 55

- Nivel de estudios

  analfabetos:

  primarios:

  secundarios:

  medios:

  superiores:

- Profesión:

-Residencia o domicilio habitual:

- Nivel sociocultural[14]

  alto:

  medio:

  bajo:

[14] *Vid.* n. 9 y 10.

- Lengua habitual

monolingüe en castellano:

bilingüe:

g) grado de prototipicidad coloquial (*vid.* § 3)

- coloquial prototípico:
- coloquial periférico:

h) Notas de campo:

El número de interlocutores en cada grabación oscila entre un mínimo de dos, el mínimo necesario para hablar de conversación, y un máximo aconsejable de cuatro, ya que un número superior de informantes dificulta una correcta identificación de éstos a la hora de transcribir. La presencia de cuatro informantes basta para que puedan producirse las típicas escisiones conversacionales (conversaciones paralelas). En las seleccionadas para esta publicación sólo dos conversaciones superan ese máximo, aunque no es relevante porque algunos de los participantes apenas intervienen. Todas las conversaciones recogidas se han realizado cara a cara y su duración oscila entre los cinco y los sesenta minutos. No obstante, el corpus de referencia de que disponemos incluye además grabaciones telefónicas, de radio y de televisión, y se irá ampliando sucesivamente a varios tipos de discurso hablado.

3. El reconocimiento de la conversación coloquial[15]

3. 1. *Planteamiento previo*

Como ya adelantábamos en la n. 2, conviene no confundir el "español coloquial" y la conversación coloquial. El español coloquial es el registro que aparece en muchas conversaciones que reciben por ello el nombre de conversaciones coloquiales. Ahora bien, como tal registro puede aparecer también en otros tipos de discurso hablado que no son estrictamente conversaciones; por ejemplo, un debate, un concurso. Pensemos

[15] El tema en cuestión ha sido tratado previamente por el grupo *Val. Es .Co* en tres reuniones científicas: en el Coloquio de la Asociación Internacional para el Estudio del Diálogo (IADA), Toulousse 1993, en el Simposio de la Sociedad Española de Lingüística, Lérida 1994, y en el I Congreso de Lingüística General, Valencia, 1994.

además que hay textos escritos que reproducen el registro coloquial de forma natural (las cartas de un soldado a sus familiares), que lo imitan artificialmente (algunas novelas u obras de teatro) o que recurren a éste para captar la atención del lector.

En cualquier caso, para avanzar en el estudio del español coloquial en general o en el de la conversación coloquial en particular es necesaria la elaboración de un corpus de referencia válido. Y esta validez viene determinada inicialmente por su grado de representatividad, que pasa a su vez por el reconocimiento exacto del objeto de estudio. Nótese que decimos reconocer y no definir.

Aunque algunos trabajos se han ocupado ampliamente del llamado español coloquial, las definiciones dadas hasta ahora son intuitivas y poco precisas, puesto que no llegan a caracterizarlo de forma exclusiva. Los rasgos propuestos aparecen atomizados; no se propone una tipología de los mismos y, en consecuencia, no se distinguen los rasgos que definen propiamente el tipo de discurso de los que caracterizan el registro; se confunde conversación con conversación coloquial; no se separan tampoco convenientemente los rasgos situacionales pragmáticos de los estrictamente lingüísticos; se opera frecuentemente sin corpus de referencia; etc.

Desde la definición pionera de W. Beinhauer han ido sucediéndose en el ámbito hispánico otras muchas que, si bien aportan datos de interés, no son suficientes, tal y como han sido planteadas, para delimitar nuestro objeto de estudio.[16] Todas ellas giran en

---

16 BEINHAUER, W. (1929), *Spanische Umgangssprache*, Dümmlers, Berlín, [versió cast. *El español coloquial*,.Gredos, Madrid, 1978]) define el español coloquial como "el habla tal como brota, natural y espontáneamente en la conversación diaria, a diferencia de las manifestaciones lingüísticas conscientemente formuladas, y por tanto más cerebrales, de oradores, predicadores, abogados, conferenciantes, etc., o las artísticamente moldeadas y engalanadas de escritores, periodistas o poetas" (9).

En esta línea, B. STEEL ((1985), *A Textbook of Colloquial Spanish*, SGEL, Madrid; M. CRIADO DE VAL, (1980), *Estructura general del coloquio*, SGEL, Madrid, señala: "The term colloquial (...) is commonly felt -albeit often pejoratively- to refer to particular informal (often racy or poppular) spoken usage, especially that usage which differs in some way from formal lenguage".)

A. M. VIGARA ((1992), *Morfosintaxis del español coloquial. Esbozo estilístico*, Gredos, Madrid.) afirma: "lo que designamos con el sintagma español coloquial es el empleo común que hacen de un determinado sistema lingüístico los hablantes de una determinada sociedad (la española) en sus actos cotidianos de comunicación" (35), haciendo además sinónimos conversación y español coloquial: "La conversación (o coloquio) no es, en suma, sino una forma de interacción verbal puntual, determinada por tres características que le son consustanciales: la actualización oral, su inmediatez y la interdependencia dinámica de todos los elementos en el proceso de la comunicación" (38).

Las definiciones se amplían con rasgos pragmático-comunicativos en trabajos como los de E. LORENZO (*Consideraciones sobre la lengua coloquial (Constantes y variables)*), en R. LAPESA (coord.)

torno a varios rasgos generales: el carácter cotidiano, oral y espontáneo, interactivo e informal, que pueden encuadrarse dentro de una clasificación bien conocida en *campo, modo, tenor y tono.*[17]

Aparte de las observaciones, entre otros, de Labov sobre el *casual style*, de Ventola o de T. de Mauro,[18] en el ámbito general tampoco encontramos respuestas satisfactorias a los problemas aludidos.

La solución inmediata no parece que sea proponer una nueva definición, sino un modo operativo de reconocimiento que asegure la representatividad del corpus recogido de acuerdo con el objeto de estudio, que no es otro en nuestro caso que la conversación coloquial.

La propuesta que planteamos a continuación parte de una serie de rasgos, en su mayoría pragmáticos y situacionales, algunos ya señalados en las definiciones anteriores, mediante los que se define el prototipo de la conversación coloquial, y de un mecanismo capaz de identificar el grado de *coloquialidad* de una conversación. Este mecanismo de reconocimiento, denominado coloquialización, permite determinar si una conversación a priori no estrictamente coloquial puede llegar a ser considerada como tal, aunque no se dé en ella alguno de los rasgos prototípicos, o aunque presente otros rasgos no propios de este tipo de conversaciones.

(1977), *Comunicación y lenguaje*, Madrid, Karpos, pág. 161-180) y M. CRIADO DE VAL ((1980). *Estructura general del coloquio*, Madrid, SGEL). Véase la revisión bibliográfica de J. POLO (*El español coloquial y zonas afines (ensayo bibliográfico)*, Yelmo, números del 1 al 28).

[17] *Vid.*, GREGORY y CARROL, *Language and Situation. Language Varieties and their Social Contexts*, Routledge, London; y Kegan P. (1978), [versió. cast. *Lenguaje y Situación. Variedades del lenguaje y sus contextos sociales*, FCE., México; Ll. Payrató ((1992), *Pragmática y lenguaje cotidiano. Apuntes sobre el catalán coloquial*, "Revista de Filología Románica", 9, pág. 143-153), a partir de la citada clasificación, define el registro coloquial como: CAMPO: cotidianidad; MODO: oral espontáneo; TENOR: interactivo; TONO: informal; y en relación con esos aspectos propone un decálogo de rasgos lingüísticos:
1) complementación del canal verbal con códigos paralingüísticos, 2) escaso control o conciencia de la producción lingüística, 3) aplicación de múltiples recursos entonativos, 4) alto grado de referencia exofórica, 5) vocabulario específico, 6) estructura gramatical específica, 7) abundante presencia de regularizaciones, simplificaciones y analogías gramaticales, 8) grado muy elevado de redundancia, 9) alta frecuencia de marcadores discursivos interactivos, 10) alta frecuencia de recursos expresivos.

[18] LABOV, W. (1983), *El estudio del lenguaje en su contexto social*, Cátedra, Madrid; VENTOLA, E. (1979), *The structure of casual conversation in English*, "Journal of Pragmatics", 3, pág. 267-298; de MAURO, T. (1993), *Lessico di frequenza dell'italiano parlato.*

3.2. *La influencia de las variables sociolingüísticas en el grado de coloquialidad de una conversación*

Las distintas variables que sirven para la extracción de un corpus homogéneo estratificacionalmente y, a su vez, adecuado en cuanto a su representatividad sociolingüística y a su validez, al menos extensional, tienen una influencia desigual sobre el empleo del registro coloquial en una conversación o sobre la coloquialidad de la misma. A priori,

a) el sexo: no es determinante en la consideración y empleo del registro coloquial; no obstante, hay que tener en cuenta que en aquellas conversaciones en que la participación respecto a esta variable no es equilibrada (por ejemplo, una mujer entre hombres o un hombre entre mujeres) el comportamiento (lingüístico o de otro tipo) puede variar y presentarse alterado;

b) la edad: aunque son los jóvenes los mayores exponentes del proceso de coloquialización que sufre la lengua española actualmente en su uso general, la edad no es tampoco determinante del llamado español coloquial, a pesar de que es factor importante, por ejemplo, para estudiar un tipo de habla modalizante (léxico especial, sufijos, muletillas...);

c) la clase social: el habla coloquial no es propiedad de ninguna clase social, si bien el nivel sociocultural puede influir en el mayor o menor dominio de registros de habla. En efecto, las clases populares (media-baja y baja) se desenvuelven lingüísticamente en el único registro que dominan, el coloquial;[19] por el contrario, el hablante culto (de nivel sociocultural alto) es capaz de usar adecuadamente según la situación de comunicación, tanto el registro formal como el coloquial.[20] El nivel de lengua (adquirido por la educación) actúa de forma proporcional sobre el dominio de registros de habla (adaptado a la situación comunicativa). A mayor nivel de lengua (competencia comunicativa), más capacidad para dominar varios registros; a menor nivel, menos capacidad. Es evidente,

[19] Ello no niega el posible uso en ciertas situaciones de lo que podrían llamarse registros intermedios. Piénsese, por ejemplo, en los intentos de acomodación del propio registro al de otra persona: un paciente ante el médico.

[20] *Vid.* SECO, M. (1973), *La lengua coloquial: Entre visillos de C. Martín Gaite*, en *AAVV, Comentario de textos I*, Castalia, Madrid, y nuestra n. 19.

sin embargo, que cuando la clase social se manifiesta como relación de solidaridad o de poder en una conversación el uso del registro queda afectado.[21]

Así pues, aunque a priori ni el sexo, ni la edad, ni tan siquiera el nivel sociocultural de cada uno de los participantes en una conversación (al menos, en todos sus estratos) nos permiten de *forma aislada* reconocer el llamado español coloquial y, sobre todo, no son suficientes para negar el carácter coloquial de una conversación, sí pueden ser relevantes en su consideración dinámica dentro del proceso comunicativo, es decir, en el comportamiento de los interlocutores desde el momento en que inician una conversación (por ejemplo, en situaciones de desigualdad social consciente entre éstos, puede notarse cómo algunos interlocutores intentan acomodar su expresión a la de otros, llegándose incluso a la alternancia de registros en una misma conversación).

Los sociolectos o niveles diastráticos pueden manifestarse lingüísticamente en el registro coloquial (lo que habrá que corroborar con un estudio posterior), y de hecho pueden afectar al modo de expresión coloquial; en efecto, podemos suponer que el español coloquial de hablantes de nivel sociocultural culto presente una serie de rasgos y frecuencias lingüísticas no coincidentes con los de hablantes de nivel bajo. Ahora bien, lo relevante para determinar de forma previa el carácter más o menos coloquial de una conversación es el *marco de interacción*, entendido como la relación que se establece entre los participantes y, sobre todo, la relación de éstos con la situación comunicativa concreta.

Estas consideraciones se pueden ilustrar con un ejemplo: la relación asimétrica médico-paciente en un espacio marcado como un hospital no favorece la coloquialidad. Si el marco de interacción cambia (médico y paciente son ahora enfermos ingresados en el mismo hospital y en la misma habitación y, como tales, son tratados por todo el personal del hospital), aunque el espacio (hospital) y las variables sociológicas sean los mismos, el cambio producido en el marco de interacción aumenta el grado de coloquialidad de la conversación.

[21] El sexo y la edad pueden influir también en la relación de poder o de solidaridad, pero, normalmente, actúan en conjunción con otros factores.

3.3. *Los rasgos primarios y las rasgos coloquializadores*

Para reconocer y delimitar previamente el carácter coloquial de una conversación hemos manejado dos tipos de rasgos:

a. *los rasgos primarios*,

b. *los rasgos coloquializadores*.

3.3.1. Los rasgos primarios

Dejando a un lado el rasgo oral, que afecta al canal de comunicación y es una característica de toda conversación y de otros modos de discurso hablado,[22] existe una serie de rasgos constantes interrelacionados que afectan al modo en que se produce y se desarrolla la conversación y, sobre todo, al fin que persigue. En la bibliografía sobre el tema[23] se hallan ampliamente documentados los siguientes:

---

[22] Hay que distinguir, por un lado, entre lo oral y lo escrito y, por otro, entre la reproducción oral de lo escrito (*cf.* el informativo de un telediario) y la reproducción escrita de lo oral (*cf.* una novela que intenta imitar el habla coloquial). A pesar de ello, la distinción oral/escrito no es tajante, sino un continuo gradual; de hecho, algunos textos escritos evidencian claros signos de oralidad: en la primera página de un periódico valenciano aparecían los dos titulares siguientes: *Viver: Muere la anciana que corneó el toro que se coló en su domicilio. Condenan a Vicente Pons a siete meses por estafa y un año por falsedad.*, Comp. BUSTOS, J., (1995), *Oralidad y escritura*, en *Actas del I Congreso sobre el español coloquial*. Almería; LAMÍQUIZ, V. (1994), *El enunciado textual. Análisis lingüístico del discurso*, Ariel, Barcelona; TANNEN, D. (1982), *The Oral/Literate Continuum in Discourse, Spoken and Written Language*, Georgetown University, Washintong, y BIBER B. (1988), *Variation across speech and writing*, Cambridge University Press.

[23] Algunas referencias bibiliográficas indicativas, además de las citadas hasta aquí (especialmente en núm. 16, 17 y núm. 22, son las siguientes: en el ámbito general:

GIVÓN, T. (1979), *From discourse to syntax: grammar as processing strategy*, en T. Givón (ed.) *Syntax and Semantics. Discourse and Syntax*, 12, Academic Press, New York.

OCHS, E. (1979),*Planned and unplanned discourse*, en T. Givón, (ed.) *Syntax and Semantics. Discourse and Syntax*, 12, Academic Press. New York.

SORNICOLA, R. (1981), *Sur parlato*. Il Mulino, Bologna.

En el ámbito hispánico:

GALLARDO, B. (1993), *Lingüística perceptiva y conversación: secuencias*, "Anejo de Lynx".

_____ (1994), *Conversación y conversación cotidiana: sobre una confusión de niveles*. en *Pragmalingüística*, Universidad de Cádiz.

MORENO, F. (1985), *Hacia una sociolingüística automatizada del coloquio*, en *Actas del III Congreso Nacional de Lingüística Aplicada*. Valencia.

NARBONA, A. (1989), *Sintaxis coloquial: problemas y métodos*, en *Sintaxis española: Nuevos y viejos*

en cuanto al *modo*:

- *interlocución en presencia*: conversación cara a cara;

- *inmediatez*: hace referencia al carácter actual de la conversación, aquí y ahora, e influye de modo decisivo sobre los rasgos siguientes;

respecto a *cómo se desarrolla*:

- toma de turno no predeterminada;

- *ausencia de planificación*, que favorece la espontaneidad;

- *dinamismo conversacional entre E y R*: tensión dialógica, alternancia de turno; la relación hablante-oyente es simultánea y/o sucesiva, es decir, supone una conversación más o menos prolongada, y no pares mínimos de intervenciones (rituales);

- *retroalimentación*, proyectada sobre el tema de la conversación: vuelta atrás; conducta cooperativa de un interlocutor respecto a la intervención de otro;

en relación con el *fin que persigue*:

- *interpersonal:* la comunicación por la comunicación; finalidad comunicativa socializadora;[24]

por el *tono empleado* (consecuencia de los rasgos citados y de los que apuntamos más abajo):

- informal.

En un intento de determinar una tipología dentro de los rasgos mencionados, cabe notar que tres de éstos, la *inmediatez* o *carácter actual*, *la toma de turno no predeterminada* y el *dinamismo conversacional entre E y R*, poseen un carácter más general porque lo que definen no es el español coloquial, ni tan siquiera la conversación coloquial, sino simplemente la conversación. Tales rasgos son una condición necesaria para que pueda hablarse de conversación. En efecto, si no existe inmediatez, no hay conversación, sino una simple sucesión de mensajes; si la toma de turno está predeterminada, estamos ante

enfoques, Ariel, Barcelona.

_____ (1989), *Problemas de sintaxis coloquial andaluza*, en *Sintaxis española: Nuevos y viejos enfoques*. Ariel, Barcelona.

_____ *Sintaxis coloquial y Análisis del discurso*, "REL", 21, 2, pág. 187-204.

[24] La conversación coloquial es aquella en la que no parecen existir derechos y obligaciones, excepto, quizás, los derivados de la cortesía (estratégica). El yo y el tú intercambian sus posiciones.

otro tipo de discursos como el debate, la entrevista, la clase magistral, el juicio oral,[25] etc.; y si no hay dinamismo conversacional, nos encontramos con acontecimientos comunicativos monológicos o, a lo sumo, con rituales (*cf.* saludo-despedida).

Otros rasgos, como la *interlocución en presencia*, determinan un tipo particular de conversación, la conversación cara a cara, pero no son suficientes ni necesarios para calificar un discurso como tal. Una interacción a través del teléfono es también una conversación, con la particularidad de que ha cambiado el canal de comunicación.

Por otro lado, la *ausencia de planificación,* la finalidad *interpersonal* y el *tono informal*, esta última como característica gestáltica,[26] hacen referencia al registro coloquial. Es decir, estos rasgos determinan ahora un tipo de conversación en función del registro, la conversación coloquial.[27]

En suma, todos estos rasgos se han denominado primarios porque su presencia es necesaria para la caracterización de un tipo de discurso, en primer lugar, como conversacional y, en segundo lugar, como coloquial. Es decir, en una conversación coloquial se combinan dos tipos de rasgos primarios, los que caracterizan propiamente la conversación y los rasgos que determinan la ubicación de esa conversación dentro del registro coloquial. Entre los conversacionales, hay unos que son propios de cualquier conversación, tales como la inmediatez o carácter actual, la toma de turno no predeterminada, el dinamismo o tensión dialógica y la retroalimentación;[28] y otros que determinan una cierta tipología conversacional, como el rasgo interlocución en presencia, cara a cara. Son rasgos típicamente coloquiales la ausencia de planificación, la finalidad comunicativa socializadora (el predominio de la función fática) y el tono informal.

La ausencia de alguno de estos rasgos primarios implicará, respectivamente, la negación de la marca [+conversación] y la de un tipo conversacional y registro concretos.

---

[25] Nótese que en los juicios orales existe inmediatez, pero la toma de turno está predeterminada, por lo que es imposible hablar en estos casos de conversación en sentido estricto.

[26] Una conversación se puede caracterizar intuitivamente como informal, pero la especificación de los rasgos que la hacen informal se revela como una tarea difícil, ya que éstos se captan en bloque, se perciben como un todo.

[27] Según puede notarse, a partir de esta clasificación y añadiendo otros rasgos podemos llegar a establecer una tipología discursiva y, más concretamente, una tipología conversacional.

[28] Éstos actúan de forma gradual según el tipo de discurso; así, por ejemplo, mientras la toma de turno no predeterminada es un rasgo básico y necesario en la conversación (coloquial), admite una gradación en otro tipo de discursos más formalizados (asamblea de vecinos, tertulia, debate...)

Pero, a pesar de su valor constante y de su carácter primario, los rasgos citados no son suficientes para reconocer una conversación como coloquial, más aún cuando algunos de ellos son comunes también a otros registros.

3.3.2. Los rasgos coloquializadores y el proceso de coloquialización

Junto a tales rasgos primarios proponemos otros que, en nuestra opinión, caracterizan lo que podríamos llamar la conversación coloquial prototípica, a la vez que precisan sociolingüística y pragmáticamente dicho registro de habla. Son los siguientes:

-relación de igualdad entre los interlocutores, ya sea en cuanto a los papeles sociales (determinados por el estrato sociocultural, la profesión, etc.) o funcionales (provocados por la situación; por ejemplo, un catedrático y un peón de albañil ingresados en el hospital en la misma habitación son funcionalmente enfermos): la relación entre iguales favorece la coloquialidad. Siguiendo a Brown y Gillman,[29] una situación en la que existan relaciones de [-poder] y [+solidaridad] (iguales solidarios)[30] propicia la conversación coloquial:

-relación vivencial de proximidad: conocimiento mutuo y experiencia común compartida entre los interlocutores (presuposiciones comunes).

-marco de interacción familiar (no marcado): viene determinado por el espacio físico (lugar) y por la relación concreta de los participantes con éste (entorno de la enunciación). Lo que se pretende resaltar es que cuanto más cotidiano sea el espacio interaccional para los hablantes, más probabilidades existen de que se desarrolle en él una conversación coloquial.

-temática no especializada: el contenido enunciativo lo constituyen temas al alcance de cualquier individuo.

De los rasgos mencionados, los dos primeros hacen referencia a la relación *dinámica* entre los participantes en una conversación; el tercero señala la relación de los participantes con la situación comunicativa, y el cuarto tiene que ver con el tema conversacional.

---

[29] BROWN A. y GILLMAN, A. *The pronouns of power and solidarity*, en T. Sebeok (ed.), *Style in Language*, Cambridge Mass.

[30] En el caso de individuos de extracción sociocultural alta, esta igualdad sólo actúa como determinante de la conversación coloquial cuando es consciente por parte de los interlocutores y se halla fuertemente imbricada con el rasgo de proximidad vivencial. En el caso de interlocutores de extracción sociocultural baja esto no ocurre porque su nivel de lengua se limita casi exclusivamente a una forma de habla coloquial, a un registro informal-coloquial.

La afinidad en cuanto a su relación social, vivencial y situacional y el tema no especializado favorece la coloquialidad.

Éstos, a los que hemos denominado *rasgos coloquializadores*, caracterizan, junto a los rasgos primarios, las conversaciones coloquiales. La presencia de todos ellos en una conversación permite identificarla y reconocerla previamente como prototípicamente coloquial. Aunque, teóricamente, de forma aislada son incapaces de *caracterizar* este registro de habla, estos rasgos coloquializadores pueden actuar por separado o asociados entre sí, haciendo posible que una conversación que no comparte alguno de los restantes rasgos prototípicos pueda derivar en una forma de habla coloquial. Llamaremos, así pues, *coloquialización* al proceso mediante el cual se puede hablar de registro coloquial en una conversación no prototípicamente coloquial (bien por la ausencia de algún rasgo coloquializador, bien por la presencia de otros rasgos no coloquiales) gracias a la acción niveladora de alguno de los rasgos coloquializadores presentes.[31]

A pesar de esto es preciso reconocer la mayor fuerza niveladora de alguno de estos rasgos en ciertas situaciones. Por ejemplo, en ocasiones los hablantes desconocen la posición social de sus interlocutores, pero, aun con conocimiento (real o por intuición) de la misma, ésta pierde relevancia por la proximidad entre los participantes (tanto vivencial como de espacio).

3.3.3. Ejemplos del proceso de coloquialización

El siguiente paso consiste en comprobar si efectivamente la intervención de estos rasgos permite soslayar la ausencia de otros rasgos prototípicos, posibilitando así que una conversación no identificada con la coloquial prototípica resulte ser también coloquial y, por tanto, sea válida su inclusión dentro del corpus. Para observar cómo actúa este proceso nivelador en tales casos, enfrentamos sucesivamente cada uno de los rasgos prototípicos que hemos considerado como niveladores, es decir, *coloquializadores*, con el rasgo opuesto al prototípico. Como ya señalábamos, estos rasgos pueden actuar en el proceso de coloquialización por separado o a la vez, existiendo en este caso con más

[31] Del mismo modo, podría reconocerse un proceso inverso de descoloquialización o, quizá más exactamente, de formalización. Por ejemplo, una conversación coloquial puede "tecnificarse" o dejar de tener una función predominantemente interpersonal para convertirse en *transaccional* (informativa, comercial, etc). Es posible que este proceso inverso deba tenerse en cuenta para estudiar el grado de alejamiento respecto al prototipo coloquial y así mismo el de acercamiento a las periferias del registro formal. (Agradecemos los comentarios al respecto del profesor Ll. Payrató).

frecuencia acción prioritaria de alguno de ellos. En efecto, cada situación prioriza uno u otro de los rasgos niveladores, por lo que no es posible jerarquizarlos previamente.

I. *Relación de desigualdad (consciente).*

Los rasgos niveladores (coloquializadores) en este caso son:

a) Relación vivencial de proximidad

b) Marco de interacción familiar

c) Temática no especializada

Ejemplo de la acción niveladora de (a) y (b): un patrón almuerza habitualmente con sus empleados en el bar próximo a la empresa en que trabajan; un tema tan popular como el del fútbol (acción de (c))[32] nivela si cabe aún más las diferencias sociales. Los espacios de ocio y diversión son claramente niveladores de las desigualdades sociales conscientes y marcos propicios para el uso del registro coloquial; cf. nuestra conversación [RB. 37. B.1], en la que tales rasgos actúan nivelando la desigualdad consciente que existe entre las interlocutoras (B y C son las dueñas de la casa y A es la limpiadora).

II. *Relación vivencial de no proximidad (desconocimiento mutuo).*

Los rasgos niveladores (coloquializadores) en este caso son:

a) Relación de igualdad entre los participantes

b) Marco de interacción familiar

c) Temática no especializada

Ejemplos de la acción niveladora de (a) (b) y (c): dos desconocidos, que viajan en asientos contiguos en un autobús, entablan una conversación sobre los jóvenes de hoy; en su relación social son ahora sólo *viajeros* en un marco cotidiano, un autobús, propicio para hablar por hablar y que nivela además la falta de relación vivencial. También en un taxi la relación vivencial es escasa, pero el marco favorece la conversación y que ésta pueda reconocerse como coloquial. Lo mismo puede decirse de la conversación entre dos personas en la sala de espera en la consulta de un médico.

Los lugares de esparcimiento, como espacios cotidianos y familiares, convierten en irrelevante el rasgo negativo no prototípico de no proximidad vivencial.

---

[32] Es obvio que un tema puede convertirse en más o menos especializado, más o menos técnico, según los interlocutores. El fútbol ha creado su propia terminología y para algunos conversadores se convierte así en un lenguaje especializado o "técnico". No obstante, como tema de conversación, el fútbol y el contexto que lo rodea es cotidiano. En cualquier caso, no es lo mismo *hablar* de fútbol que *decir* o *comentar* un partido de fútbol.

La temática no especializada que se da normalmente en esos espacios cotidianos, la conocida por todos, es al mismo tiempo otro rasgo coloquializador.

III. *Espacios y momentos no cotidianos, ni familiares (para todos los interlocutores o para alguno de ellos).*

Rasgos niveladores:

a) Relación de igualdad entre los participantes

b) Relación vivencial de proximidad

c) Temática no especializada

Estos son los rasgos niveladores que permiten que en un marco de interacción transaccional (marcado, en principio), sea el caso de nuestra conversación [H. 25. A.1], se desarrolle (en España),[33] simultánea o paralelamente a la conversación con fin transaccional una conversación coloquial, cuyo fin es la comunicación por la comunicación.

Ejemplo de la acción niveladora de b) y c): en una carnicería de barrio, comprador y vendedor se olvidan de sus papeles respectivos y hablan sin otro motivo que el de la comunicación por la comunicación. Por otro lado, no cabe duda de que una consabida relación de igualdad entre ambos, es decir el rasgo (a), favorece la presencia de la conversación no transaccional en tales intercambios comerciales. Un suceso acaecido en el barrio el día anterior (la acción de (c)) provoca el comentario sobre el mismo.

Dos españoles se encuentran en Alemania en un área de servicio de la autopista; no se conocen, pero ahora son turistas en un país extranjero y su misma nacionalidad les sirve ahora como relación vivencial de proximidad.

IV. *Temática especializada.*

Rasgos niveladores:

a) Relación de igualdad entre los participantes

b) Relación vivencial de proximidad

c) Marco de interacción familiar

Se puede hablar de las interjecciones, de Quevedo y del español coloquial coloquialmente si existe una relación vivencial entre los conversadores y el marco de interacción lo permite; por ejemplo, dos estudiantes que comparten una experiencia

[33] En otros países es inusual esta conversación paralela a la de carácter transaccional.

común (la clase de literatura y la clase de lengua) juegan con sus contenidos fuera del aula.

3.4. *Conclusión*

Puede afirmarse que los rasgos primarios y coloquializadores, de forma conjunta, permiten reconocer y calificar previamente una conversación como prototípicamente coloquial.

El *grado de prototipicidad* coloquial variará según la mayor o menor presencia de tales rasgos, los cuales actúan como medida del carácter coloquial. Una *conversación coloquial prototípica*, al menos desde el punto de vista teórico, es aquella en la que pueden reconocerse los rasgos del prototipo "español coloquial". Una *conversación coloquial periférica* es aquella que presenta un índice de prototipicidad menor porque le faltan rasgos o posee otros no prototípicamente coloquiales. De este modo, la coloquialización permite identificar como coloquial la conversación periférica y además hace patente su grado mayor o menor de alejamiento respecto al prototipo.

Por otro lado, nótese cómo el concepto de coloquialización evita binarizar la categorización del español coloquial en el sentido de tener que decidir si una conversación es o no coloquial por la presencia o ausencia de alguno o algunos de los rasgos mencionados. Dicho concepto hace posible diferenciar dentro de una categoría como es el registro coloquial la conversación prototípica coloquial de la conversación periférica, sin por ello negar el carácter coloquial de ambas.

Como señalábamos al principio de esta introducción, el español coloquial es un registro de uso según la situación comunicativa. Hasta aquí sólo hemos intentado delimitar dicho registro en el tipo de discurso más auténtico, la conversación coloquial, aquella que se aproxima al llamado modo pragmático[34] de la comunicación, a partir de una serie de rasgos que fundamentalmente precisan y determinan el concepto de situación comunicativa. La caracterización lingüística y paralingüística de tal registro, así como el estudio de los posibles correlatos pragmáticos y lingüísticos (p. ej., entre una forma lingüística como el tuteo y la relación de igualdad), lo que constituiría propiamente la

[34] GIVÓN, T. (1979), *From Discourse to Syntax: Grammar as Processing Strategy*, en T. Givón (ed.), *Syntax and Semantics. Discourse and Syntax, 12*, Academic Press, Inc, Londres, pág. 81-112 (esp. 101 y ss.).

definición integral del *español coloquial*, incluso superando el ámbito de la conversación, es una tarea posterior.

4. Características de las conversaciones coloquiales seleccionadas en esta publicación. Cuadro-resumen de la muestra transcrita

El corpus que aquí presentamos es una muestra de conversación coloquial. La componen nueve conversaciones, siete de las cuales, según los rasgos señalados, son prototípicas (responden a las claves, [H. 38. A.1]; [AP. 80. A.1]; [ML. 84. A.1]: [L. 15.A.2]; [J. 82. A.1]; [S. 65. A.1]; [G. 68. B.1 + G. 69. A.1]) y dos periféricas ([H. 25. A.1], dado que el marco en que se desarrolla es transaccional y no existe relación de proximidad entre todos los participantes, y [RB. 37. A.1], por la presencia del rasgo no prototípico [+desigualdad entre los interlocutores]).

Su duración oscila entre los diez y los treinta minutos; aproximadamente, tres horas y media de grabación. Para asegurar el mayor grado de espontaneidad, todas las grabaciones son secretas, con o sin observador participante; sólo al final de [G. 68. B1 + G.69.A.1] los interlocutores son conscientes de que les están grabando.

La muestra cubre, como puede observarse en el siguiente cuadro, todos los estratos sociales, de edad y de sexo. No obstante, en nuestra selección, hemos intentado, por el propósito de la muestra, que en unas conversaciones los participantes presentaran características socioculturales, de edad y de sexo idénticas y en otras que hubiera variaciones al respecto. No siempre el número de participantes en la conversación reflejado en la ficha técnica coincide con el tenido en cuenta en la estratificación, ya que los participantes pasivos no se han contabilizado. Son participantes pasivos aquellos que apenas intervienen o cuya participación se reduce al empleo de reguladores fáticos, los que sólo aparecen ocasionalmente durante el desarrollo y los investigadores, siempre que sean sólo observadores o intervengan sólo de forma ocasional (los investigadores de [H. 38. A.1] y [L. 15. A.2] aparecen reflejados en el cuadro, ya que, como demuestra la actuación de los mismos en la conversación, son participantes activos).

CUADRO-RESUMEN DE LA MUESTRA TRANSCRITA: VARIABLES SOCIOLINGÜÍSTICAS Y PROTOTIPICIDAD.

a) Prototípicas:

	Clave Part.	Sexo	Edad	Nivel Sociocultural
[H. 38.A1]	1 A	V	<25	alto
	1 B	V	<25	medio
	1 C	V	<25	medio
	1 D		<25	medio
[ML.84.A.1]	1 A	V	<25	medio
	1 B	M	<25	medio
[L. 15.A.2]	1 L	M	<25	medio
	1 E	M	<25	medio
	1 G	V	<25	medio
[S. 65.A.1]	1 A	M	>55	medio
	1 M	M	>55	
[A.P.80.A. 1]	1 S	V	26-55	bajo
	1 J	V	26-55	bajo
	1 G	V	26-55	bajo
	1 A	M	26-55	bajo
	1 L	M	26-55	bajo
[J.82.A.1]	1 A	V	26-55	alto
	1 G	V	26-55	alto
	1 V	V	26-55	alto
	1 S	V	26-55	alto
[G.68.B.1	1 P	M	>55	bajo
+G.69.a.1]	1 C	M	>55	medio

b) Periféricas:

[RB.37.B.1]	1 A	M	>55	bajo
	1 B	M	<25	medio
	1 C	M	<25	medio
[H.25.A.1]	1 A	V	>55	bajo
	1 B	M	26-55	bajo
	1 C	M	26-55	bajo

5. La transcripción. Signos y convenciones

La transcripción intenta reproducir lo más fielmente posible la conversación y al mismo tiempo facilitar la labor del lector. El sistema de signos y convenciones empleado es lo suficientemente estrecho, por tanto, para conseguir que el lector pueda reproducir aproximadamente la conversación original.

Cualquier sistema de signos puede ser reducido o ampliado en función de los objetivos marcados. Puede reducirse a una transcripción ancha, por ejemplo, cuando sólo se pretende un estudio léxico; ha de ampliarse, si el estudio es fonético. Nuestra propuesta atiende a fenómenos relacionados con la alternancia de turnos, la sucesión inmediata de emisiones, solapamientos, reinicios y autointerrupciones, escisiones conversacionales, pausas y silencios, entonación (inflexiones finales que influían notablemente en el curso de la conversación y que introducían novedades respecto a la prosodia normativa), fenómenos de énfasis, problemas relacionados con emisiones dudosas o indescifrables, de fonosintaxis, alargamientos fonéticos, preguntas retóricas, estilo directo, referencias contextuales, etc.

Un sistema será adecuado si está regido por principios de exhaustividad y pertinencia de los signos. El que aquí proponemos cumple ambos requisitos. Por un lado, es exhaustivo en relación con los objetivos marcados; por otro, es pertinente, ya que cada

signo representa un fenómeno y cada uno de los fenómenos aparece codificado mediante una única convención.[35]

5.1. *Sistema de transcripción*

La transcripción combina el sistema ortográfico con ciertas convenciones del *Análisis Conversacional*,[36] ambos ampliados o modificados en algunos signos por problemas de transliteración y de adecuación a nuestros programas informáticos. Los signos fundamentales son los siguientes:

:	Turno de palabra.
A:	Turno de palabra de un hablante identificado como A.
?:	Interlocutor no reconocido.
§	Sucesión inmediata, sin pausa apreciable, entre dos emisiones de distintos hablantes.
=	Mantenimiento del turno de un participante en un solapamiento.
[	Lugar donde se inicia un solapamiento o superposición.
]	Final del habla simultánea.
-	Reinicios y autointerrupciones sin pausa.
/	Pausa corta, inferior al medio segundo.
//	Pausa entre medio segundo y un segundo.
///	Pausa de un segundo o más.
(5")	Silencio (lapso o intervalo) de 5 segundos; se indica el nº de segundos en las pausas de más de un segundo, cuando sea especialmente significativo.

[35] El debate sobre sistemas de transcripción que tuvo lugar en la Universidad de Almería con motivo del I Simposio de español coloquial nos aportó algunos datos interesantes para completar y mejorar el conjunto de convenciones. Nuestro agradecimiento desde aquí a los colegas y amigos que participaron en el mismo. Agradecemos, en particular, los comentarios que, tras la revisión del texto, han realizado los profesores A. Narbona, Ll. Payrató, J. Polo y J. Portolés.

[36] *Cfr.* JEFFERSON, G. (1974), *Error correction as an interactional ressource*, "Language in Society", 3, pág. 181-201; SCHENKEIN, J. (1978), *Studies in the organization of conversational interaction*, Academic Press, New York; LEVINSON, S. C. (1983), *Pragmática*, Teide, Barcelona.

↑	Entonación ascendente.
↓	Entonación descendente.
→	Entonación mantenida o suspendida.
EN COCHE	Pronunciación marcada o enfática.
(( ))	Fragmento indescifrable.
((siempre))	Transcripción dudosa.
((...))	Interrupciones de la grabación o de la transcripción.
(en)tonces	Reconstrucción de una unidad léxica que se ha pronunciado incompleta, cuando pueda perturbar la comprensión.
pa'l	Fenómenos de fonética sintáctica entre palabras, especialmente marcados.
°( )°	Fragmento pronunciado en un tono de voz más bajo, próximo al susurro.
h	Aspiración de "s" implosiva.
l·l	Asimilación fonética.
(RISAS)	Cuando aparecen al margen de los enunciados. Si acompañan a lo dicho, se transcribe el enunciado y en nota al pie se indica "entre risas".
aaa	Alargamientos vocálicos.
nn	Alargamientos consonánticos.
¿¡ !?	Preguntas o exclamaciones retóricas (por ejemplo, las interrogaciones exclamativas: preguntas que no preguntan).
¿?	Interrogaciones. También para los apéndices del tipo "¿no?, ¿eh?, ¿sabes?"
¡!	Exclamaciones y enunciados irónicos.

Letra cursiva: Reproducción e imitación de emisiones. Estilo directo, característico de los denominados relatos conversacionales.

Notas a pie de página: Anotaciones pragmáticas que ofrecen información sobre las circunstancias de la enunciación. Rasgos complementarios del canal verbal. Añaden informaciones necesarias para la correcta

interpretación de determinadas palabras (por ejemplo. la correspondencia extranjera de la palabra transcrita en el texto de acuerdo con la pronunciación real), enunciados o secuencias del texto, de algunas onomatopeyas, etc.

Sangrados a la derecha: Escisiones conversacionales.

és que se pareix a mosatros: Fragmento de conversación en valenciano.[37]

* Las incorrecciones gramaticales (fónicas, morfosintácticas y léxicas) no aparecen marcadas por lo general. Así pues, según el usuario del corpus (por ejemplo, si este es utilizado por un estudiante de español como segunda lengua), puede ser recomendable el soporte explicativo del profesor.
* La mayoría de los antropónimos y topónimos que aparecen en las transcripciones son inventados.

5.2. *Precisiones sobre los signos prosódicos: pausas y entonación*[38]

Considerando nuestro criterio de partida, es decir, que la transcripción debe servir para facilitar la lectura y no para perturbarla, hemos optado por transcribir los índices prosódicos (pausas, inflexiones tonales...) que influían notablemente en el curso de la conversación o introducían alguna modificación en la prosodia "normativa".

Un ejercicio útil para observar la adecuación de nuestro mecanismo de transcripción ha sido comprobar, a través de la lectura del texto transcrito por alguien ajeno al trabajo, su grado de identidad con la conversación real. A partir de aquí fueron incorporándose ciertos signos y eliminándose, por el contrario, otros de carácter redundante.

5.2.1. Tonemas

La labor de transcripción en este sentido tiene un carácter provisional hasta que un estudio posterior preciso, que se está realizando, nos dé claves más precisas para codificar los fenómenos entonativos adecuadamente. Por este motivo, los signos

[37] En tales fragmentos la transcripción se adapta para reflejar algunos rasgos de dicha variante dialectal.
[38] Nos hemos servido de un *Visipitch 6097* y de una tarjeta de sonido Sound Blaster, incorporados a una computadora *IBM, PC 486*, así como del programa de sonido *Sound Edit-Pro*, incorporado a un *MACINTOSH QUADRA 840 AV*.

mínimos empleados han de entenderse como marcas indicativas de una curva melódica peculiar antes que como codificaciones definitivas de la misma.

A) *Curvas con valor modal (fuerza ilocutiva)*:

- Interrogación-pregunta: ¿?

Añadimos el tonema final siempre que no corresponda al patrón normativo (interrogativas absolutas- ascenso; parciales- descenso); indicamos, en su caso, los tonemas interiores de grupo:

C: ¿de qué marca te lo has compraoꜛ? [H. 38. A.1]

B:¿cómo quiés decirꜜ de vacacionesꜛ? [RB. 37. B.1]

A: ¿aquí- al volver la esquinaꜛ no hay un poyeteꜛ en una ventana/ de mármol?

[RB. 37. B.1]

En ocasiones las interrogaciones -preguntas aparecen cortadas, bien porque éstas se suspenden, bien porque existe un reinicio. No obstante, en ambos casos se ha marcado gráficamente el cierre (?) para facilitar la lectura:

C: ¿no dijo que co- que te conocía a ti y a uun? [H. 38. A.1]

G: ¿cómo te vas-? si acabas de empezar [L. 15. A.2]

Exclamación: ¡!

Cuando en las exclamativas no se señalan los tonemas hay que entender que la inflexión tonal corresponde a los esquemas normativos (descenso pronunciado). Indicamos, en su caso, los tonemas interiores de grupo:

JM: ¡qué divertido es todo! [ML. 84. A.1]

C: ¡joder el del helicópteroꜜ tío! [H. 38. A.1]

Aseveración:

Puesto que en general el tonema de la aseveración es descendente, sólo hemos marcado los tonemas interiores de grupo cuando éstos segmentan partes potencialmente informativas dentro del texto. En caso de duda se ha comprobado la dirección de la inflexión tonal mediante los aparatos (*vid.* n. 38).

A: a setiembreꜜ se le acababaꜜ la primera semana [RB. 37. B.1]

Interrogaciones exclamativas (*cf.* preguntas que no preguntan): ¿¡ !?

Estructuras lingüísticamente interrogativas, aunque sin valor modal de pregunta (no se solicita información al receptor).

A: ¡buf!// mira C./ es que/ es que/ no lo sé/ es que/ yo ¿¡quÉ quieres que haga!? [ML. 84. A.1]

Entonación dialectal:

Existe una estructura interrogativa peculiar del habla de Valencia (que suponemos general en el valenciano) consistente en iniciar una estructura interrogativa mediante una partícula *que* (a modo de marcador de pregunta).[39] Hemos detectado una diferencia prosódica para este elemento según introduzca interrogativas absolutas o parciales. En el primer caso se articula más enfáticamente, por lo que lo transcribiremos con mayúsculas y acento; de este modo, queda diferenciado, por un lado, del simple *qué* interrogativo y, por otro, del *que* únicamente enfático:

A: ¿QUÉ vale poco verdá? [RB. 37. B.1]

En el segundo caso se suele articular con menor intensidad, por lo que se ha transcrito con mayúsculas, pero sin acento:

A: ¿QUE qué es lo que le pasa? [RB. 37. B.1]

B) *Función demarcativa-segmentadora*:

- Dislocaciones sintácticas a la derecha o a la izquierda.

La entonación en estos casos es capaz por sí sola de individualizar segmentos de frase por diversos motivos (énfasis, realce, focalización, etc.). La dirección de los tonemas demarcativos de segmentos puede variar: ascendente o descendente (↑↓):

M: estaba preparando la comida↓ y eso [ML. 84. A.1]

Sin embargo, el mantenimiento de nuestro criterio de economía de rasgos redundantes ha hecho que no transcribamos las inflexiones demarcativas de grupo sintáctico, como, por ejemplo, los tonemas ascendentes (semianticadencia) que delimitan los dos miembros de un período coordinado o subordinado.

[39] Para una interpretación a partir del sistema del castellano, *vid.* BELLO, A. (1988), *Gramática española.1847*, Arco, Madrid, § 995.

C) *Función fática*:

- Tonemas demarcativo-continuativos.

Cuando la segmentación del enunciado se manifiesta mediante el ascenso tonal. con un valor adicional de mantenimiento del hilo discursivo, hemos empleado el mismo signo que para la inflexión ascendente ordinaria: ↑

C: te veo todo el rato↓ igual estás superbién conmigo↑ ahí superbién↑ ¿no? y con todo el mundo↓ y de repente te encierras↑ tío↑ yo no sé qué te pasa↑ si es que tienes algún problema en casa o algo↑ tío↑ [ML. 84. A.1]

En ocasiones se indica el tonema continuativo, aunque la lectura no lo precise. para marcar el énfasis o expresividad del que habla:

C: se fue a una reunión d'estas↑/ no compró ningún libro↑ y mira qué carterita [G. 68. A. 1 + G. 69. B. 1]

En los alargamientos vocálicos motivados por la vacilación del hablante (a modo de pausas oralizadas) resulta difícil determinar en muchos casos hasta qué punto aparece o no un tonema suspensivo (→) continuativo añadido. Hemos optado normalmente por transcribir sólo el alargamiento:

C: ¿pero él- pero él entendía↑ dee de reLOJES↑ ooo? [RB. 37. B.1]

- *Marcadores metadiscursivos de control del contacto*.[40]

Existen asimismo una serie de fórmulas de control del contacto que son formalmente interrogativas, aunque con un valor no siempre apelativo, del tipo *¿no?, ¿sabes?, ¿verdad?, ¿eh?* etc. Se representan mediante signos de interrogación (¿?), puesto que poseen en su mayor parte una curva tonal ascendente, propia de una interrogativa absoluta. En consecuencia, no se indica el tonema final ascendente característico de tales fórmulas, si bien es preciso señalar que el ascenso que manifiestan en posición interior de intervención es menos marcado (presenta en estos casos un valor expresivo-fático) que cuando se sitúa al final con el carácter de pregunta y de apelación al oyente.

- *Marcadores metadiscursivos del control del mensaje (demarcativos)*:

Aunque la mayoría de marcadores metadiscursivos poseen un contorno melódico propio, no siempre se ha marcado el tonema correspondiente. Así, el verbo *decir* es un

[40] Vid. BRIZ, A. (en prensa), *El papel metadiscursivo del conector pragmático en español*, "EA", 1994.

introductor de estilo directo en los *relatos* conversacionales y funciona como marca de frontera, de límite entre algo precedente y algo consecuente, a modo de transición de habla necesaria en la actividad formulativa del hablante; actúa, pues, como pausa léxica (rellenando el silencio propio de toda pausa). Puesto que ha aparecido siempre con tonema suspendido (→), éste no queda representado.[41]

A: y dice *nooo* dice *no tiene el mismo paso* [RB. 37. B.1]

Por supuesto los ascensos, descensos o suspensiones tonales tienen un valor relativo y no todos poseen la misma función (demarcativa, interactiva, expresiva...), pero la especificación de los parámetros acústicos y el análisis de su comportamiento específico corresponderá a trabajos posteriores. No cabe duda de que tales estudios entonativos harán necesaria, como ya señalábamos, la corrección y ampliación de estos signos.

5.2.2. Pausas

Con carácter provisional, dado que su comportamiento en la conversación precisa todavía de un estudio particular, se distinguen tres tipos de pausas en virtud de su duración:

/ : pausa corta, menos de 0'5 segundos (medio segundo).

Con frecuencia indica discontinuidad sintáctica ("incompletitud"). En ocasiones la pausa corta es imperceptible, pero aparece señalada de algún modo por un tonema de descenso precedente. Señalaremos, pues, la variación tonal correspondiente o la pausa, si ésta llega a percibirse:

A: no↓ bastante [RB. 37. B.1]

// : más de medio segundo, hasta un segundo

Marca, por ejemplo, un contorno entonativo completo

C: oye pues está bien ¿eh?// ¡qué tranquila! ¿eh? [RB. 37. B.1]

/// : más de un segundo

Puede señalar, por ejemplo, un cambio de tópico o reformulación del anterior:

A: eso parece/// EEEs que a mi marido lo han hecho fijo [RB. 37. B.1]

[41] Los mismos comentarios válidos para DECIR (conjugado) metadiscursivo se hacen extensivos a Y + DECIR (conjugado) metadiscursivo.

Al final de turno sólo aparece representada la duración de la pausa en el caso de que sea especialmente significativa (2", 3" 4", etc.).

5.3. *Los filtros*

Una tarea previa y necesaria ha sido el aprendizaje y adiestramiento de todo el grupo en la transcripción del material con el fin de unificar el empleo de los signos. Los ejercicios prácticos de transcripción y de corrección en común han hecho mucho más fácil y llevadera la tarea posterior de filtrado.

Las conversaciones, una vez transcritas, han sido sometidas a varios filtros. Cada transcripción ha pasado por

1º) el filtro del propio investigador que realizó la transcripción,

2º) el filtro de, al menos, dos investigadores del grupo por separado,

3º) el filtro llevado a cabo conjuntamente por otros dos investigadores, a través de los cuales han pasado todas las conversaciones.

Los nuevos aparatos (ordenadores, Visi Pitch, programas de sonido...) adquiridos este último año nos han facilitado las tareas de transcripción en los últimos momentos y al mismo tiempo nos han permitido en este tercer filtrado subsanar algunos errores y deficiencias de carácter suprasegmental (duración de las pausas, ciertos tonemas, etc...), que de otro modo nos hubieran pasado desapercibidos.

4º) el filtro final (a modo de corrección de pruebas) llevado a cabo por todos los miembros del grupo.

Pasar la lengua hablada y, concretamente, la conversación coloquial a un formato escrito es tarea ingrata y difícil, no exenta de ciertos subjetivismos y errores (a pesar de los varios filtrados). Somos conscientes de algunas de las limitaciones e inadecuaciones de carácter teórico que presenta nuestra propuesta de sistema de transcripción, aunque, creemos, se ajusta a la premisa ya señalada de que el lector pueda reproducir lo más fielmente posible la conversación, sin hacer uso de la cinta grabada. Y, sobre todo, se ajusta a los objetivos concretos señalados en varios apartados de esta introducción.

6. Un ejemplo de la transcripción de materiales. Fragmento incial de la conversación G.68.A.1 + 69.B.1.

C: no/ no/ pero yo ahora/ ¿qué estaba diciendo?/ ya no m'acuerdo→

P: lo de la →// los libros§

C: §ah sí// pues eso// mee- una compañera me dice *¡hala!/ Carmen/ ¿te vienes a-*

a esto de los libros que nos regalan una cosita de sumar →/ o de no sé qué HISTORIA?§

P: §sí§

C: §y yo digo

no no/ que si nos yo soy muy tonta y pico// no/ mujer/ no dona[42]*/ va//* y ya→ con tan(( ))- ((ella)) se apunta a un bombardeo// nos vamos las dos↑// empiezan a explicar aquello↑/ a mi me gusta/ la ENCICLOPEDIA/ la no sé cuantas/ pico yo// y me cascan lo menos sesenta mil cucas↓ aunque sean (( ))- las pagué ↑// en lugar de- de en un año/ tal/ pues dije/ en cuatro plazos y entonces te rebajan y TOdo/ pero aun así y todo/ tuve que cascar ese dinero que a mí me venía bien para otra cosa/ [en cambio aquél[43]=]

P: [((pero luego→ claro))]

C: = na(da) más se llevó la carterita y no// y en cambio él[44] fue más listo que yo§

P: § mm§

C: § se fue a una reunión d'estas↑/ no compró ningún libro↑/ y mira qué carterita/ [(RISAS=)]

P: [(RISAS)]

C: = más bonita/ dice [((he sido→))=]

P: [esta ((gratis))]

C: = gratis/ claro/ lo que daban de regalo§

J: § sí/ sí/ [y va muy bien][45]

P: [((de regalo))]

---

[42] Valenciano, "mujer".

[43] Se refiere a J.

[44] Se trata nuevamente de J.

[45] J lo dice con ironía, pues oculta una grabadora en dicha cartera.

C: [claro/ ((claro))] mira/ mira qué bonita es§

P: § hombre/ [mira si te ((cabe))→ claro (( ))=]

J: [y lo que cabe→]

C: [(RISAS)]

P: = mira si te [va bien (( ))]

C: [y además] además yo la he visto por ahí bastante/ porque se ve que gente que ha ido↑/ a cosa de los libros↑§

J: § (( ))§

C: §↑pues- pues tiene la cartera/ y la he visto varias [(( ))=]

P: [(( )) claro]

C: = a varios por ahí§

J: § sí sí/ yo también§

P: § [además que (( ))]

C: [y yo el otro] día cuando vi que la cogió↑/ oye ¿para qué [((la querría gastar)?]

P: [que puede meter much (( ))] oye/ déjalo[46] ahí/ que a [mí no=]

J: [mm]

P: = me molesta→// se puede poner mucha cosa ahí [dentro ((¿eh?))]

C: [claro/ claro]// pues/ ¿qué me estabas diciendo del chiquillo?

P: nada/ quee lo operaron/ lo tuvieron que operar↑/ porque tenía una hernia en un testículo§

C: § PO[BRECITO]

P: [y- y] le dijeron// lo llevó Mari Ángeles a un ciiirujano→y le dijo dice *bueno/ esto puede pasar*// dice/ *porque→/ si fuese mayor↑/ aún aún/ pero aún es pequeñito*// pero luego lo he llevao a éste y dice *NOO/ si fuese de ombligo↑/ le dejaríamos que el niño→* §

C: § se fuera desarrollando§

P: §*se fuese desarrollando* dice *pero esTO/ YA*// dice *porque el niño se le puede estrangular*/// bueno/ así [que ((lo))=]

C: [al pequeñín de→]

[46] Se refiere a una bolsa de mano.

P: = al chiquitín [de Mari Ángeles]

C: [de- de] Mari Ángeles y Jesús// lo han ope[rao]↑

J: [¿a/ a] Alejandro?§

C: § Ale[jandro]

P: [sí] y entonces pues [((nada))]

C: [eso me] estaba empezando a contar cuando tú has lla[mado]

P [eent] entonces pues lo prepararon/ y→ ((le)) dice *entonces ¿cuándo lo tengo que llevar?* y el cirujano dice/ *pues mira/// HOY tengo quirófano/ si quieres ahora→*/ pero ella está de baja casi dos meses/ con lo de las cervicales§

C: § ¿Mari Ángeles?§

P: § Mari Ángeles/ en[ton(ces)]

C: [¿de cuál]? ¿de estar tanto en la caja y coger o eso↑/ o→?§

P: § bueno/ ella ya lo lleva eso→/ no saben si de un golpe que se dio/ o de nacimiento/ o de qué// tiene como un esguince§

C: § YA

P: entonces/ le- le dijo el cirujano/ *hoy tengo quirófano/ si quieresÆ/ Mari Ángeles//* dice *noo/ hoy no porque tengo yo que ir/ a hacerme unas placas*

C: YA

P: entonces// le dice *bueno/ pues el martes siguie(nte)→* al martes siguiente creo que tuvimos que ir// y nada/ dice *te estás aquí a- a las ocho dee- de la noche ↑/- de siete y media a ocho ↑/ y dice y a las nueve lo operamos//* y así hicimos/ fueron a recogerme a mí al trabajo↑§

C: §((cierra))§

P: § me recogieron↑/ y fuimos§

C: § como es tan CHIQUITÍN

P: loo- lo prepararon↑/ porque había/ nos dijo el- el anestesista/ dice/ *mira* dice *hay otro niño/* dice/ *el que- sea más chiquitín se opera antes//* dice por *los líquidos↑/ porque luego se pueden deshidratar/ por si devuelven o algo//* total que/ el otro niño vino↑/ le tomaron la tempetarura↑ tenía fiebre y no lo pudieron operar/ así que pasó él el primero// pasó él y nos dijo el cirujano/ *no os asustéis/ va a llorar // porque va a llorar↓* - cuan- cuando se lo llevaron↑ no/ porque empezaron a gastarle bromas↑/ [y nada=]

C: [(RISAS)]

P: = pero una camilla como en una persona mayor// y se lo llevaron↑/ y y y nada/ (y) dice *pero cuando lo pinchemos/ lo más seguro→/* así fue/ UNOS gritos/ unos gri[tos [((por fuera))]

C: [(RISAS)] me acuerdo cuando pinchaban a éste§[47]

P: § sí§

C: § que le tenían que sacar la vena≠/ ¿sabes de dónde?§

P: § sí§

C: §¿la sangre?// d'ESA VENA DE AQUÍ§

P: § pues un§

C: § pero no podían con él/ dos enfermeras/ dos monjas/ y- y- y- y yo qué sé// y encima le dio la monja caramelos/ digo *¿caramelos?*§

J: § caramelo§

C: § digo *yo lo tiraba por la ventana*§

P: § pues unos gritos que pa qué/ y se iba oyendo ya el grito con me[nos fuerza(( )) con menos fuerz(a)=]

C: [claro/ que lo- ya- lo- iba durmiéndose]

P: = pero estuvieron casi una hora en la operación/// y na[da]

C: [¿y eso] qué lo tenía?/ ¿más altito o estrangulao casi?

P: no lo (sé)↓ se le ponía como moradito§

C: § ¡ah!/ ya

P: entonces// [cuando=]

C: [((ves))]

P: = salió el cirujano→/ dice *todo ha salido estupendo↓ Mari Ángeles/ pero va a salir igual que ha hecho/ llorando y chillando*// claro/ dice *y de momento no te va a conocer/ porque como está con l'anestesia*↑

C: no- no gilan[48] bien [o sea no (( ))]

P: [y eso (( ))] así que cuando salíaÆ/ chillando y llorando/ *buáa*[49]/ y venga a

---

47 Se refiere a J.

48 Con el sentido de "ver".

49 Imitación del llanto de un niño.

llorar/ UNAS LÁgrimas// y claro↓ se acercó Mari Ángeles↑/ y ¡CARIÑO!/ y ¡CARIÑO!/ y él/ se abrazó a su madre↑/ acercó a la cara así↑/ [así (( )) y no la desapegó]

ELS LINGÜICIALS[1] DEL CORPUS UB

Lluís de Yzaguirre Maura
(Universitat Pompeu Fabra)

1. Presentació

En el moment d'iniciar el Corpus UB, teníem clar que l'objectiu immediat era la recol·lecció dels materials i que progressivament anirien creixent les necessitats de manipulació i processament d'aquests materials. Érem conscients que disposàvem d'un conjunt d'eines molt limitat, però coneixíem l'existència de nombroses iniciatives similars més avançades que la nostra, de les quals esperàvem que produïssin o fessin produir aviat lingüicials com els que podíem necessitar.

Però no podíem esperar que programes estàndard resolguessin tots els nostres requeriments, especialment aquells específics del català, com els relacionats amb la lematització. En la mesura que les seves modestes possibilitats li ho permetien, el DFCUB va iniciar el desenvolupament de lingüicials ajustats a les seves necessitats, tasca que em fou encomanada i que es va traduir en els programes, que presentarem suara, destinats a la manipulació i recuperació de veu digitalitzada, realitzats entre 1990 i gener de 1995; un segon paquet de lingüicials ha estat desenvolupat des del febrer de 1995, en què vaig integrar-me a l'Institut de Lingüística Aplicada de la Universitat "Pompeu Fabra"

[1] L'anglès "lingware" s'usa per designar les aplicacions informàtiques de tractament del llenguatge natural; els francesos en diuen "linguiciels" o "linguisticiels"; m'he permès l'estalvi de no dir cada vegada "aplicació/ons informàtica/ques de tractament del llenguatge natural" adoptant "lingüicial" per "lingware".

(IULA), bo i mantenint-me com a membre de l'equip de recerca del Corpus UB: es tracta, bàsicament, dels lingüicials que lematitzen, etiqueten i, ara per ara parcialment, desambigüen, produïts amb l'ajut de tot l'equip de l'àrea de Corpus de l'IULA a partir dels programes i dades generats entre 1980 i 1990 per a la meva tesi doctoral sobre l'estructura sil·làbica del català central.

Per tal com el DFCUB es troba integrat a la Xarxa temàtica de lingüística aplicada pilotada per l'IULA i aquest a la Xarxa temàtica sobre la variació pilotada pel DFCUB, està garantida la continuïtat del suport informàtic al Corpus UB per part de l'IULA, a través del seu Grup d'Enginyeria Lingüística, de recent estructuració, que tinc el gaudi de coordinar.

L'esperit que ha presidit l'elaboració dels lingüicials del Corpus UB ha estat sempre el de l'eficàcia a curt termini dins la subsidiarietat davant de qualsevol aplicació general o específica que pugui fer una part o el tot d'alguna de les feines necessàries. O sia que si alguna operació es podia resoldre usant un gestor de bases de dades o fent una macro del tractament de textos, no s'han esmerçat esforços en programar ad hoc. Hem pretès que, el dia que hi hagi lingüicials distribuïts públicament en el marc de les iniciatives europees d'estandardització de corpus i recursos lingüístics, les dades del Corpus UB puguin beneficiar-se'n immediatament, ço que vol dir 1) tenir les dades 2) en un format flexible 3) fàcilment transferible als estàndards que s'adoptin. Els recursos informàtics que presentarem pretenen contribuir als tres objectius anteriors.

Una altra circumstància que ha condicionat el treball fet ha estat l'heterogeneïtat dels materials lingüístics que integren el Corpus UB, causada per la multiplicitat d'objectius de recerca que es pretén cobrir. A l'hora de dissenyar el Corpus, l'equip d'investigadors que el promou va acceptar aquesta heterogeneïtat en el convenciment que, tot i les seves peculiaritats diferencials, cadascun dels subcorpus podria ser explotat també contrastivament amb els altres gràcies a la generació de productes homogenis derivats de tots els subcorpus com índexs de freqüències, concordances, estadístics textuals (sobre paraules gramaticals, distància entre relatius, mitjana de mots per frase, proporció de subjuntius...), lemaris...

2. El paquet "ASCII850"

Un desenvolupament de tipus general, causat pel desig de facilitar la portabilitat de les dades del Corpus, ha estat el paquet "ASCII850", al voltant d'una tipografia per a Macintosh amb els caràcters ordenats segons el lloc que ocupen a la pàgina de codis 850 del sistema operatiu MS-DOS. Encara que aquest sistema operatiu és una relíquia del passat i que avui la portabilitat de les dades entre plataformes està totalment resolta, vàrem voler simplificar les operacions de transferència entre Mac (que havíem adoptat per les facilitats de manipulació de veu digitalitzada) i compatibles IBM que en el futur poguessin necessitar usuaris del Corpus externs al DFCUB.

El primer pas fou el disseny pròpiament de la tipografia "Courier850", feta amb l'ajut del Sr. Jordi Domènech. Aquesta tipografia és monoespaiada, per facilitar el seu ús en llistats i concordances, i ens permet que tant els fitxers ASCII com les bases de dades es puguin llegir i manipular indistintament i simultània des de les plataformes Mac i MS-DOS.

Com a conseqüència de l'adopció de la tipografia "Courier850" per manipular els materials del Corpus UB, va caldre crear dos fitxers de configuració de teclat "Català ASCII850" i "Català ISO ASCII850". El paquet es completa amb dos programes de conversió entre ASCII850 i ASCII de Mac (un per a fitxers ASCII i l'altre per a fitxers compatibles dBase) i un fitxer de criteris d'ordenació per al programa Le Concordeur,[2] que usem per al tractament lexicomètric del Corpus. El DFCUB ha cedit el paquet ASCII850 al domini públic; es pot aconseguir accedint via Internet a "http://lincat.fil.ub.es".

3. Gestió de veu digitalitzada

L'equip del Corpus UB va avaluar els costos de dos plantejaments alternatius del tractament del material oral: emmagatzemar-lo en alguna forma d'àudio convencional o digitalitzar-lo. En tots dos casos, el text s'hauria d'introduir a l'ordinador i hi hauria la possibilitat de recuperar qualsevol informació expressada textualment; la diferència rau en

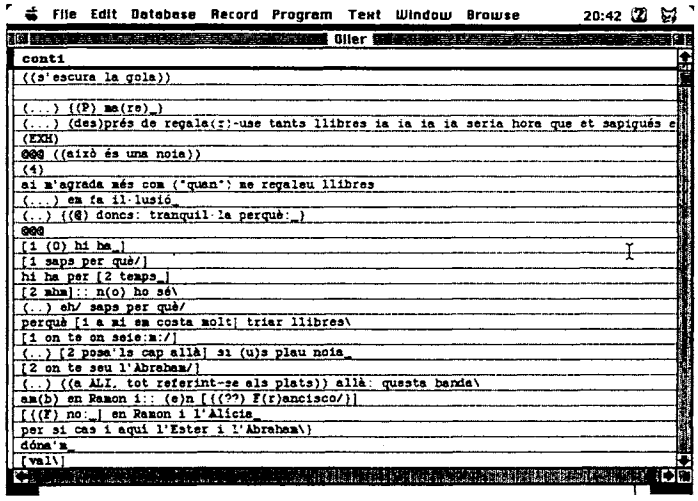
[2] Rand@ERE.Umontreal.CA

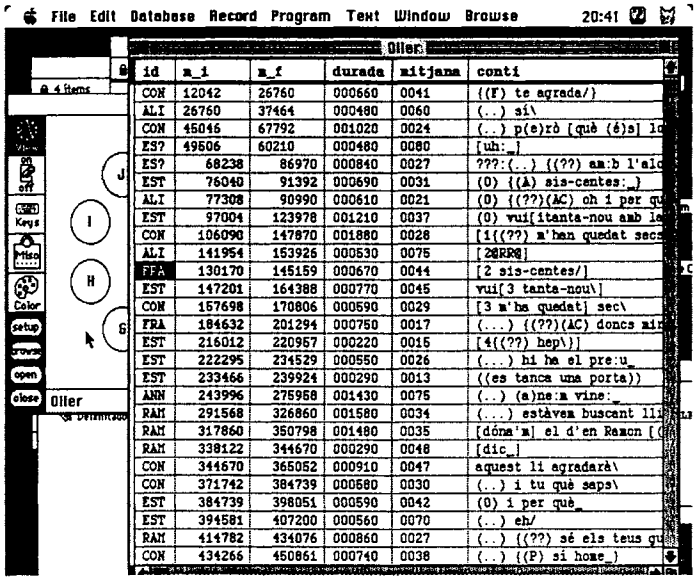
el fet que per contrastar el text amb l'àudio en un cas cal recuperar el senyal d'entre un conjunt voluminós de cintes mentre que en el segon cas es pot confiar a l'ordinador la gestió del senyal. En el primer cas, el cost humà es concentra en la recuperació de la informació; en el segon cas, el cost humà es concentra en l'operació de sincronitzar text i senyal, que anomenem "paral·lelització". Després de visitar diverses Universitats europees i del Quebec, l'equip del Corpus UB va considerar provat que el cost humà de paral·lelitzar el text dels materials orals amb el senyal digitalitzat dels enregistraments corresponents quedava àmpliament compensat pels beneficis molt superiors de les possibilitats de recuperació i per là reducció dràstica dels costos humans d'aquesta recuperació.

Com a conseqüència d'aquesta decisió, es varen elaborar un seguit de programes que permeten la integració del text transcrit dels materials orals amb la veu digitalitzada. El primer d'aquests programes segmenta un enregistrament digitalitzat en unitats menors entre silencis. En deixo constància, encara que ja no l'usem: la tecnologia informàtica evoluciona de manera vertiginosa; això explica que en alguns aspectes els condicionaments que teníem en començar han canviat; llavors vàrem creure que era millor disposar de moltes unitats petites que de poques unitats grans, mentre que, posteriorment, amb la generalització dels lectors de CD-ROM, l'aparició d'estampadores molt econòmiques i l'abaratiment brutal del cost de les oblies, vàrem preferir exactament el contrari, abandonant l'ús del programa segmentador de fitxers de veu digitalitzada, que, de tota manera, funciona i està a disposició, com tots els altres si no es diu el contrari, de qualsevol equip de recerca.

El segon programa gestor de veu permet millorar el procés de digitalització fornint dades que serveixen per avaluar-ne la qualitat comparant el nivell de quantització i el grau de saturació.

El tercer programa serveix per convertir automàticament un text al format de base de dades usat en la paral·lelització; a les imatges següents es poden veure detalls d'una base de dades, en format dBase IV, gestionada des d'un Mac, però amb el joc de caràcters 850.

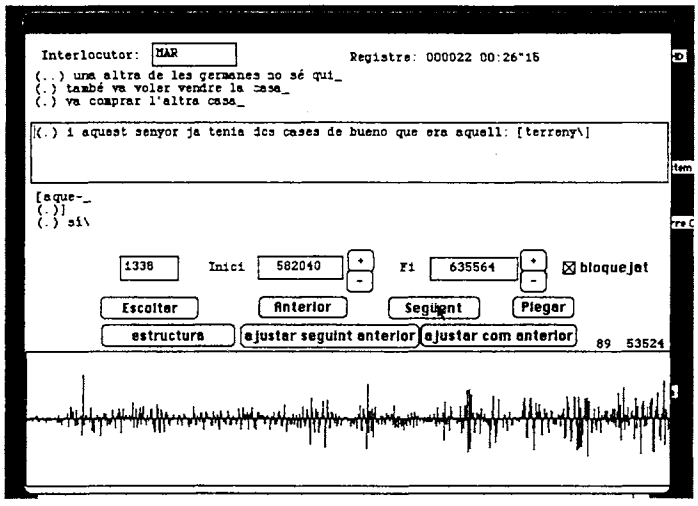




id	m_i	m_f	durada	mitjana	conti
CON	12042	26760	000660	0041	((F) te agrada/)
ALI	26760	37464	000480	0060	(..) sí\
CON	45046	67792	001020	0024	(..) p(e)rò [què (é)s] lo
ES?	49506	60210	000480	0080	[uh:_]
ES?	68238	86970	000840	0027	???:(..) ((??) am:b l'alc
EST	76040	91392	000690	0031	(0) ((A) sis-centes:_)
ALI	77308	90990	000610	0021	(0) ((??)(AC) oh i per qu
EST	97004	123978	001210	0037	(0) vui[itanta-nou amb le
CON	106090	147870	001880	0028	[i((??) m'han quedat secs
ALI	141954	153926	000530	0075	[2@RR@]
FRA	130170	145159	000670	0044	[2 sis-centes/]
EST	147201	164388	000770	0045	vui[3 tanta-nou\]
CON	157698	170806	000590	0029	[3 m'ha quedat] sec\
FRA	184632	201294	000750	0017	(...) ((??)(AC) doncs mir
EST	216012	220957	000220	0015	[4((??) hep\)]
EST	222295	234529	000550	0026	(...) hi ha el pre:u_
EST	233466	239924	000290	0013	((es tanca una porta))
ANN	243996	275958	001430	0075	(..) (a)ns:m vine:_
RAM	291568	326860	001580	0034	(...) estàvem buscant lli
RAM	317860	350798	001480	0035	[dóna'm] el d'en Ramon [(
RAM	338122	344670	000290	0048	[dic_]
CON	344670	365052	000910	0047	aquest li agradarà\
CON	371742	384739	000580	0030	(..) i tu què saps\
EST	384739	398051	000590	0042	(0) i per què_
EST	394581	407200	000560	0070	(..) eh/
RAM	414782	434076	000860	0027	(..) ((??) sé els teus gu
CON	434266	450861	000740	0038	(..) ((P) si home_)

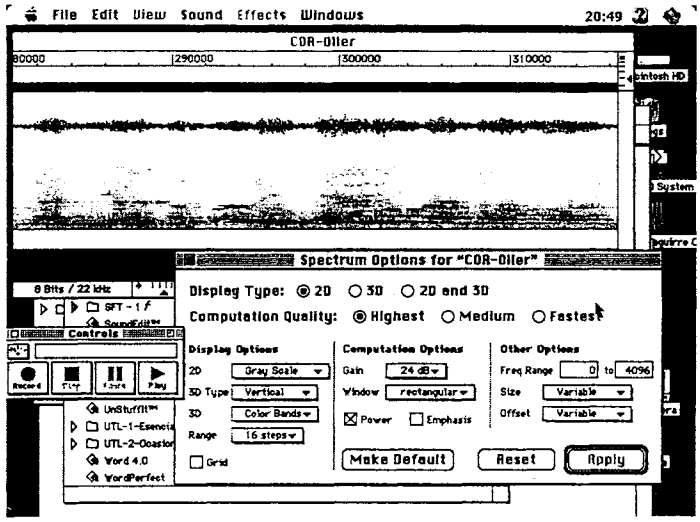
La segona d'aquestes imatges mostra les dades de la paral·lelització (mostra inicial=m_i i mostra final=m_f) i d'altres.

L'operació de paral·lelització, la fa un humà[3] amb ajut d'un quart programa (el paral·lelicial) que permet ajustar les marques de sincronia tot escoltant el senyal i visualitzant el text, que pot ser corregit en funció de les evidències que aquest procés forneix. El paral·lelicial extreu les mitjanes de durada de cada caràcter i millora progressivament la seva capacitat d'ajustar automàticament la durada del segment a sincronitzar, capacitat que gaudeix d'una rutina de detecció de silencis que, en funció del soroll de fons, sol millorar el procés:



Un cinquè programa permet explotar ("audiovisualitzar") els materials paral·lelitzats, que són exportats per un sisè programa a diversos formats usats per lingüicials estàndard i pels de manipulació de textos que esmentem a l'apartat següent. Cal destacar el fet que la veu digitalitzada està emmagatzemada en un fitxer en format SoundEdit, que es pot considerar un estàndard en la seva plataforma, cosa que permet usar diversos programes per manipular la veu o fer-hi comprovacions:

[3] La paral·lelització no podrà ser automàtica fins que no disposarem de sistemes eficients de reconeixement de la parla, tant robustos com per processar converses amb diversos interlocutors simultanis.



4. Etiquetatge gramatical

Es tracta de dos programes, un que lematitza i etiqueta i un altre que desambigua.

L'article *El projecte CECA (Corpus Escrit del Català)*, que trobareu en aquesta mateixa publicació, inclou dues mostres del resultat d'aquests programes. El primer d'aquests programes es pot considerar acabat provisionalment (en el sentit que sempre serà susceptible de millora, mentre la llengua evolucioni) i és l'únic que no podem deixar al domini públic per mor dels interessos comercials (d'editorials lexicogràfiques) que se'n podrien ressentir; però està a disposició de qualsevol equip de recerca la possibilitat de transferir-nos textos que els retornaríem lematitzats i/o etiquetats. El segon programa, destinat a desambiguar els textos etiquetats per l'anterior, està resolt tècnicament però requereix ara mateix un treball (permeteu-me redundar) laboriós de recopilació de regles

de base lingüística que s'iniciarà poc després del tancament d'aquest redactat. Durant la recopilació de les regles, el programa serà objecte de millores per incrementar-ne el rendiment en funció dels colls d'ampolla que les regles provoquin.

Aquesta informació és del dia 12 de novembre de 1995; tenint en compte l'entusiasme amb què treballem totes les implicades i també tots els implicats, us suggerim que ens seguiu el rastre a través del Web del DFCUB ("http://lincat.fil.ub.es"). També ho podeu fer a través del de l'IULA ("http://www.iula.upf.es"). També hi trobareu altra informació sobre el projecte i d'altres programes "menors" que aquí no s'han esmentat.

Bibliografia

BUTLER, C. S. (ed.) (1992), *Computers and written texts*, Basil Blackwell, Oxford.

SINCLAIR, J. (1991), *Corpus, concordance, collocation*, Oxford University Press, Oxford.

UNIVERSITAT DE BARCELONA
Secció de Lingüística Catalana
Departament de Filologia Catalana

COL·LECCIÓ
LINGÜÍSTICA
CATALANA
1