

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

## **CLUB-1, 1er Col.loqui Lingüístic de la Universitat de Barcelona: Corpus, corpora.**

**20 de desembre de 1993**

**Secció de Lingüística Catalana, Departament de Filologia Catalana, Universitat de Barcelona.**

### **ELS CORPUS LINGÜÍSTICS ORALS**

**Joaquim Llisterri**

**Departament de Filologia Espanyola, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona. Fax: 581.16.86 Correu electrònic: Joaquim.Llisterri@cc.uab.es**

#### **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).

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 confluïr 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 en 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:

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

	<b>Lingüística de corpus</b>	<b>Tecnologies de la parla i fonètica experimental</b>
Materials	Parla espontània, no preparada ( <i>unelicited speech</i> )	Corpus controlat ( <i>elicited speech</i> )
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 (1993:4) "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 (1991:3) "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, 1993:65-70).

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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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<sup>1</sup>**

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 aquest 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**

<sup>1</sup> 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 general dels projectes als Estats Units pot consultar-se Lamel (1992); un resum de les principals iniciatives japoneses ha estat publicat a NESCA - *The European Speech Communication Association Newsletter* 13 (1994): 11-15; Badia *et al.* (1994) recull corpus textuals i orals per al català i Arrarte i Llisterrri (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](mailto:esca@icp.grenet.fr); servidor WWW: <http://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.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 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 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 BDBSONS, 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).

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 BDBSONS - *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*<sup>2</sup> (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 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

<sup>2</sup> 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 80834, Barcelona. Fax: (93) 401.64.47. Correu electrònic: nadeu@tsc.upc.es

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 es centren en una llengua, el corpus conegut com EUROM.1 (Sherwood i Fuller, 1992) és de naturalesa multilingüe. Existeix actualment en alemany, anglès, castellà<sup>3</sup>, 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 cal diàlegs el més real 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 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.

<sup>3</sup> 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 80834, 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; Llisterra *et. al.*, 1993)

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

## 2.3.- Corpus per a l'estudi de la llengua oral<sup>4</sup>

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 es 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,

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*<sup>5</sup> (Marcos Marín, 1991; Marcos Marín, Ballester i Santamaría, 1993) és un recull de 1.100.000 paraules transcrits 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

<sup>4</sup> 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 mitjançant 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)*

<sup>5</sup> El corpus és accessible per ftp anònim a lola@llj.uam.es (150.244.8.2).

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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).

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

### 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*<sup>6</sup> (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 el qual els 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<sup>7</sup> en el marc del projecte ESPRIT 6819 SAM-A - *Speech Technology Assessment in Multilingual Applications* (Mariño i Llisterra, 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 alfabetes 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

<sup>6</sup> 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. Participen, entre d'altres, la Universitat Politècnica de Madrid i *Telefónica*.

<sup>7</sup> Cal esmentar aquí el treball previ d'A. Quilis i E. Enríquez en el projecte ESPRIT 2104 POLYGLOT 1.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 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 Llisterra, 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 que 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<sup>TM</sup>.

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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 entorns 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)

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 MULTTEXT - *Multilingual Text Tools and Corpora*<sup>8</sup>, 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, desambigüitzadors 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 definir el suport, queda encara organitzar una bona infraestructura de distribució i 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'estandardització i la difusió de corpus orals. En primer lloc es fa referència als projectes d'estandardització 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<sup>9</sup>.

<sup>8</sup> 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.

<sup>9</sup> 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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

## 4.1.- Iniciatives dedicades a corpus textuais i orals

### 4.1.1.- TEI, *Text Encoding Initiative*<sup>10</sup>

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 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

---

associacions, a bulletins de discussió i distribució per correu electrònic, i a arxius de textos.

<sup>10</sup> 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](mailto: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](mailto: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 missatge a [listserv@uicvm.uic.edu](mailto:listserv@uicvm.uic.edu); amb "get teip3 package" s'obté el document en 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.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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*<sup>11</sup>**

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 Màlaga, 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 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'inclougui una representació digital del senyal acompanyada d'una representació espectrogràfica i de la corba melòdica.

<sup>11</sup> Per a obtenir més informació sobre NERC cal adreçar-se a, 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](mailto:glottolo@icnucevm.cnuce.cnr.it)

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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*<sup>12</sup>**

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 Forschungszentrum für Künstliche Intelligenz*, Saarbrücken), 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 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 de EAGLES seran accessibles a finals de 1994.

<sup>12</sup> La informació sobre EAGLES pot obtenir-se adreçant-se a: Sr. G Arrarte o J. Llisterrri, Àrea de Investigación, Instituto Cervantes, Libreros 23, 28801 Alcalá de Henares, Madrid. Fax: (91) 883.50.10. Correu electrònic: [eagles@cervantes.es](mailto:eagles@cervantes.es). Vegeu també les presentacions publicades al *DGXIII Magazine* (Brinkhoff, 1993) i a *ELSNNews* (EAGLES, 1993). *ELSNNews* 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, CCS, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. Fax (44.31) 650.45.87. Correu electrònic: [elsnet@ed.ac.uk](mailto:elsnet@ed.ac.uk). Servidor WWW: <http://www.cogsci.ed.ac.uk/elsnet/home.html>

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

## 4.2.- Iniciatives dedicades a corpus orals

### 4.2.1.- SAM, *Multilingual Speech Input/Output Assessment, Methodology and Standardisation*<sup>13</sup>

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.

### 4.2.2.- COCOSDA, *Coordinating Committee for Speech Databases and Speech Input/Output Systems Assessment*<sup>14</sup>

COCOSDA és un comitè 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 esmentat 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

<sup>13</sup> 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](mailto:adrian@phonetics.ucl.ac.uk).

<sup>14</sup> 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](mailto:cocos@phonetics.ucl.ac.uk). o bé [euro@phon.ucl.ac.uk](mailto:euro@phon.ucl.ac.uk). Existeix un butlletí electrònic de discussió sobre temes generals ([cocosda@atr.co.jp](mailto:cocosda@atr.co.jp)) i un d'especificament dedicat a corpus ([cocosda\\_corp@atr.co.jp](mailto:cocosda_corp@atr.co.jp))

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

*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*<sup>15</sup>

El *Linguistic Data Consortium* és una agrupació americana d'universitats, empreses i centres del govern organitzada i finançada per l'*Advanced Research Projects Agency* (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<sup>16</sup>:

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
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

<sup>15</sup> 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](mailto:ldc@unagi.cis.upenn.edu)

<sup>16</sup> 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.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

ATIS3	8		Parla espontània	ARPA
YOHO	2		Autenticació 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, dígit, 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*<sup>17</sup>

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*<sup>18</sup>

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.

<sup>17</sup> 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>.

<sup>18</sup> 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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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 Forschungszentrum 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 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 situacions. 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.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

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.

## 6.- Referències

ALTENBERG, B. (1991) "A bibliography of publications relating to English computer corpora", in JOHANSSON, S.- STENSTRÖM, A. (Eds.) *English Computer Corpora. Selected Papers and Research Guide*. Berlin: Mouton de Gruyter. pp. 355-396

ALVAR EZQUERRA, M.- VILLENA PONSODA, J.A. (Coords.) (1994) *Estudios para un corpus del español*. Málaga: Universidad de Málaga (Anejo 7 de Analecta Malacitana, Revista de la Sección de Filología de la Facultad de Filosofía y Letras de Málaga)

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

ANDERSEN, O.- DALSGAARD, P. (1992) "DKISALA V1.1- Users Guide (SAM-IES-059)" in *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.- GREYTER, R.- OMOLOGO, M. (1993) "Automatic segmentation and labelling of English and Italian speech databases" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 653-656

ARRARTE, G.- LLISTERRI, J. (1994) *Informe sobre recursos lingüísticos para el español (I): Corpus escritos y orales disponibles y en desarrollo en España*. Madrid: Instituto Cervantes. [Per a més informació: G. Arrarte / J. Llisterri, Área de Investigación, Instituto Cervantes, Libreros 23, 28001 Alcalá de Henares, Madrid. Fax (91) 883.50.10. Correu electrònic: g.arrarte@cervantes.es - joaquim.llisterri@cervantes.es]

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

BADIA, T.- CABRÉ, M.T.- LLISTERRI, J.- 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. [ Per a més informació: M.T. Cabré, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Rambla de Santa Mònica 30, 08002 Barcelona. Fax (93) 542.24.49. Correu electrònic: cabre@upf.es]

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

BLOMBERG, M.- CARLSON, R. (1993) "Labelling of speech given its text representation" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 1775-1778

BNC (1991) *British National Corpus - Spoken Corpus Transcription Guide* TGCW 21, 18 December 1991

BONNEAU-MAYNARD, H.- GAUVAIN, J.-L. - GOODINE, D.- LAMEL, L.F.- POLIFRONI, J.- SENEFF, S. (1993) " A French version of the MIT-ATIS system: portability issues" in *Eurospeech'93*.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

*3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 2059-2062

BRINKHOFF, N. (1993) "Towards standards in language engineering: EAGLES", *DGXIII Magazine* pp. 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: 75-76

BRUCE, G. (1989) "Report from the IPA working group on suprasegmental categories", *Lund University Department of Linguistics, General Linguistics, Phonetics, Working Papers* 35: 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.- BARRY, B. (1992) "PTS Software V.4.40, user manual" in *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 ", in VIDAL BENEYTO, J. ( Dir.) *Las industrias de la lengua*. Trad. de M. Alvar *et al*. Salamanca / Madrid: Fundación Sánchez Ruipérez / Pirámide ( Biblioteca del Libro, 5 ). pp. 108-118

CARRÉ, R. (1992) "Speech Databases" in AINSWORTH, W.A. (Ed.) *Advances in Speech, Hearing and Language Processing. A Research Annual. Volume 2*. London: Jai Press. pp. 199-216.

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

CASTAGNERI, G. (Ed.) (1991) *Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech i/O Assessment Methods*. Chiavari 26-28 September 1991 ( Italy ). Organized by CSELT in cooperation with CEC DGXIII, ESCA, ESPRIT PROJECT 2589 (SAM)

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

CRIN-INRIA (1992) "ELSA-ESPRIT labelling system assessment software. User's guide for V2.4 (SAM-UCL/CRIN-042)" in *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.- THOMPSON, S.A. (1991) " Towards a new corpus of spoken American English", in AIJMER, K.- ALTENBERG, B. (Eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman. pp. 64-82

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

DE GINESTEL-MAILLAND, A.- DE CALMÈS, M.- PÉRENNOU, G. (1993) "Multi-Level Transcription of Speech Corpora from Orthographic Forms" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 2 pp. 1441-1444

DÍAZ, J.E.- RUBIO, A.J.- PEINADO, A.M.- SEGARRA, E.- PRIETO, N. - CASACUBERTA, F. (1993) "Development of task-oriented Spanish speech corpora", in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993.

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

DOLMAZON, J.M. (1994) "BDSONS", in *Language Engineering Convention, CNIT, La Défense, Paris, Abstracts*. Compiled by L. Jackson-Eve. July 6-7, 1994. pp. 89-90

DRAXLER, C.- TILLMANN, H.G.- EISEN, B. (1993) "Prolog Tools for Accessing the PhonDat Database of Spoken German" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 191-194

DU BOIS, J.W.- SCHUETZE-COBURN, S.-CUMMING, S.- PAOLINO, D. (1993) "Outline of discourse transcription", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 45-90

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

EDWARDS, J.A. (1992) "Design principles in the transcription of spoken discourse" in SVARTVIK, J. (Ed) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*. Stockholm, 4-8 August, 1991. Berlin: Mouton de Gruyter. pp. 129-147

EDWARDS, J.A. (1993a) "Principles and Contrasting Systems of Discourse Transcription", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 3-32

EDWARDS, J.A. (1993b) "Survey of Electronic Corpora and Related Resources for Language Researchers", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 263-310

EISEN, B. (1993) "Reliability of speech segmentation and labelling at different levels of transcription" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 673-676

ENGSTRAND, O. (1987) "The IRIS speech data base - a status report" in ENGSTRAND, O. (Ed) *Papers from the Swedish Phonetics Conference Held in Uppsala October 17-18, 1986 ( RULL, Reports from the Uppsala University Department of Linguistics, 17) : 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: 99-106

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

FOURCIN, A.- DOLMAZON, J.M. (presented on behalf of the SAM Project) (1991) "Speech knowledge, standards and assessment" in *Actes du XIIème Congrès International des Sciences Phonétiques*. 19-24 août 1991, Aix-en-Provence, France. Aix-en-Provence: Université de Provence, Service des Publications. Vol 5 pp. 430-433.

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

FRENCH, J.P. (1992) "Transcription proposals: multilevel system", Working paper, University of Birmingham, October 1992. NERC-WP 4-50

GAROFALO, J.S. - PALLET, D.S. (1989) "Use of CD-ROM for speech database storage and exchange" in TUBACH, J.P.- MARIANI, J.J. (Eds.) *Eurospeech 89. European Conference on Speech Communication and Technology*. Paris- September 1989. Edinburgh: CEP Consultants Ltd. pp. 309-312

GARSIDE, R.- LEECH, G.- SAMPSON, G. (Eds.) (1987) *The Computational Analysis of English: A Corpus-based Approach*. London: Longman

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora. Actes del 1er i 2on Col.loqui Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

GIBBON, D. (1989) Survey of Prosodic Labelling for EC Languages. SAM-UBI-1/90, 12 February 1989; Report e.6, in ESPRIT 2589 (SAM) *Interim Report, Year 1*. Ref. SAM-UCL G002. University College London, February 1990.

GLASS, J.- GOODINE, D.- PHILLIPS, M.- SAKAI, S.- SENEFF, S.- ZUE, V. (1993) "A bilingual Voyager system" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 2063-2065

GREENBAUM, S.- SVARTVIK, J. (1990) "The London-Lund Corpus of Spoken English" in SVARTVIK, J. (Ed) *The London-Lund Corpus of Spoken English. Description and Research*. Lund: Lund University Press. pp. 11-63

GREENBERG, J.A. -FERGUSON, C.A.- MORAVCSIK, E.A. (Eds.) *Universals of Human Language. Volume 2: Phonology*. Stanford: Stanford University Press.

GUMPERZ, J.J.- BERENZ, N. (1993) "Transcribing Conversational Exchanges", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 91-122

HERNÁEZ, I.- BARANDIARÁN, J.- MONTE, E. (1993) "A segmentation algorithm based on acoustical features using a self organizing neural network" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 661-664

HIRST, D. (1991) "Intonation models: towards a third generation", in *Actes du XIIème Congrès International des Sciences Phonétiques*, 19-24 août 1991, Aix-en-Provence, France. Aix-en-Provence, Université de Provence, Service des Publications, Vol 1 pp. 305-310

HIRST, D.- DI CRISTO, A. (en premsa) "A survey of intonation systems" in HIRST, D. - DI CRISTO, A. (Eds.) *Intonation Systems. A Survey of 20 Languages*. Cambridge: Cambridge University Press.

HIRST, D.- DI CRISTO, A.- LE BESNERAIS, M.- NAJIM, Z.- NICOLAS, P.- ROMÉAS, P. (1993) "Multilingual modelling of intonation patterns", in HOUSE, D.- TOUATI, P. (Eds.) *Proceedings of an ESCA Workshop on Prosody*. September 27-29, 1993, Lund, Sweden. Lund University Department of Linguistics and Phonetics, Working Papers 41. pp. 204-207

HIRST, D.J.- IDE, N. - VÉRONIS, J. (1994) "Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project", in *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*. September 12-15, 1994. Mohonk Mountain House, New Paltz, New York, USA. pp. 77-80

HIRST, D.J.- NICOLAS, P.- ESPESSER, R. (1991) "Coding the Fo of a continuous text in French: An experimental approach" in *Actes du XIIème Congrès International des Sciences Phonétiques*, 19-24 août 1991, Aix-en-Provence, France. Aix-en-Provence, Université de Provence, Service des Publications, Vol 5 pp. 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) in *SAM User Guide to ETR Tools*. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Ref, SAM-UCL-G007.

JONES, K. - MARIANI, J. (Eds.) (1992) *Proceedings of the 1992 Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment*. Monday, 12 October 1992. Banff Springs Hotel, Banff, Canada.

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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

KNOWLES, G. (1991) "Prosodic labelling: the problem of tone group boundaries", in JOHANSSON, S.- STENSTRÖM, A. (Eds.) *English Computer Corpora. Selected Papers and Research Guide*. Berlin: Mouton de Gruyter. pp. 149-163

KNOWLES, G.- LAWRENCE, L. (1987) "Automatic intonation assignment" in GARSIDE, R.- LEECH, G.- SAMPSON, G. (Eds.) *The Computational Analysis of English: A Corpus-based Approach*. London: Longman. pp. 139-148

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

KOHLER, K. (1991) "Phonetic data bases for German" in *Actes du XIIème Congrès International des Sciences Phonétiques*. 19-24 août 1991, Aix-en-Provence, France. 5 vols. Aix-en-Provence: Université de Provence, Service des Publications. Vol. 2 pp. 466-469.

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

LAMEL, L.F.- GAUVAIN, J.-L.-ESKÉNAZI, M. (1991) " BREF, a Large Vocabulary Spoken Corpus for French", in *Eurospeech 91. 2nd European Conference on Speech Communication and Technology*. Genova, Italy, 24-26 September 1991. vol 2. pp. 505-508

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

LEECH, G. (1991) " The State of the Art in Corpus Linguistics" in AIJMER, K.- ALTENBERG, B. (Eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman. pp. 8-29.

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

LLISTERRI, J. (1994b) *Prosody Encoding Survey*. WP 1 Specifications and Standards. T1.5. Markup Specifications. Deliverable 1.5.3. Final version, 15 September 1994. LRE project 62-050 MULTEXT.

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

MacKAY, I. (1994) "Review of The Oxford Acoustic Phonetic Database on Compact Disk", *Linguist List* Vol 5-256, Sun 06 March 1994; *NESCA, Newsletter of the European Speech Communication Association* 14: 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 : 11-25

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

MARCOS MARÍN, F. (1991) " Corpus oral de referencia de la lengua española contemporánea" in MARCOS MARÍN, F. (1991) *Archivos Digitales*. Sociedad Estatal del V Centenario. Area de Industrias de la Lengua. 3.07.1991. p. 1-25

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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

MARIÑO, J.B. - LLISTERRI, J. (1993) *Spanish adaptation of SAMPA and automatic phonetic transcription*. SAM-A/UPC/001/v1 20th April 1993. ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications).

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

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

MORENO, A.- POCH, D.- BONAFONTE, A.- LLEIDA, E.- LLISTERRI, J.- MARIÑO, J.B.- NADEU, C. (1993) "ALBAYZIN Speech Database: Design of the Phonetic Corpus" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 175-178

NERC (1994) *NERC-1 Network of European Reference Corpora. Final Report*. Pisa: Istituto di Linguistica Computazionale - CNR.

OCHS, E. (1979) "Transcription as Theory" in OCHS, E.- SCHIEFFELIN, B.B. (1979) *Developmental Pragmatics*. New York: Academic Press. pp. 43-72

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

PAUL, D.B.- BAKER, J.M. (1992) " The design for the Wall Street Journal - based CSR Corpus ", *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.- ROSNER, B.S (1993) *The Oxford Acoustic Phonetic Database on Compact Disk*. Oxford: Oxford University Press (2 CDs)

RENOUF, A. (1987) "Corpus development" in SINCLAIR, J. (Ed) *Looking Up, An Account of the COBUILD Project*. London: Collins. pp. 1-40

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

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.- JACK, M.A. (1993) "Phonetic transcription standards for European names (ONOMASTICA)" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 279-282

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

SHERWOOD, T.- FULLER, H. (1992) *Guide to EUROM.1 Speech Database*. Doc no. SAM-NPL-102, Final version 21 April 1992.

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

LLISTERRI, J. (1996) "Els corpus lingüístics orals", in PAYRATÓ, Ll.- BOIX, E.- LLORET, M.-R.- LORENTE, M. (Eds.) *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70. [http://liceu.uab.es/~joaquim/publicacions/UB\\_Corpus\\_96.pdf](http://liceu.uab.es/~joaquim/publicacions/UB_Corpus_96.pdf)

SINCLAIR, J. (1994) "Spoken Language", "Phonetic/Phonemic and Prosodic Annotation" in NERC (1994) *NERC-1 Network of European Reference Corpora, Final Report*. Pisa: Istituto di Linguistica Computazionale - CNR.

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

SVENDSEN, T.- KVALE, K. (1992) "ELABSEG V2.5, Users Manual" in *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.- FLIGELSTONE, S. (1991) "A survey of English machine-readable texts", in JOHANSSON, S.- STENSTRÖM, A. (Eds.) *English Computer Corpora. Selected Papers and Research Guide*. Berlin: Mouton de Gruyter. pp. 319-354

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

TILLMANN, H.G.- POMPINO-MARSCHALL, B. (1993) "Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems" in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology*. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 1691-1694

UCL (1992) "Speech acquisition and Annotation Protocols and Index of Mnemonics (SAM-UCL-018)-Section IV: SAMPA" in *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: 31-54

WELLS, J.C.- BARRY, W.- GRICE, M.- FOURCIN, A.- GIBBON, D. (1992) Standard Computer-Compatible Transcription. SAM Stage Report Sen.3 SAM UCL-037, 28 February 1992. In *SAM (1992) ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Final Report*. Year Three: 1.III.91-28.II.1992. London: University College London.

WICHMANN, A. (1991) "A study of up-arrows in the Lancaster/IBM Spoken English Corpus", in JOHANSSON, S.- STENSTRÖM, A. (Eds.) *English Computer Corpora. Selected Papers and Research Guide*. Berlin: Mouton de Gruyter. pp. 165-178

ZEILIGER, J.- SERIGNAT, J.F. (1991) "Europec software V.4.1 User's Guide (SAM-ICP-045)" in *SAM User Guide to ETR Tools*. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Ref, SAM-UCL-G007, 1992

ZUE, V.- GLASS, J.- GOODINE, D.- HIRSCHMAN, L.- LEUNG, H.- PHILLIPS, M.- POLIFRONI, J.- SENEFF, S. (1991) " The MIT ATIS system: Preliminary development, spontaneous speech data collection and performance evaluation" in *Eurospeech 91. 2nd European Conference on Speech Communication and Technology*. Genova, Italy, 24-26 September 1991. vol 2. pp. 537-540

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