El corpus RETOC: Un corpus oral per a la recerca i la docència

Laboratori de Tecnologies Lingüístiques
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

Lluís **de Yzaguirre**
Antoni J. **Farriols**
Jaume **Martí**

Presentació

El Repertori Electrònic de Textos Orals Catalans (RETOC) és un corpus de veu digitalitzada, parcialment transcrita i accessible universalment a través de l'Internet. En el seu estat actual es pot considerar **operatiu** pel que fa als aspectes tecnològics i **oportunista** pel que fa als continguts:

- **Operatiu** perquè és plenament accessible tant en les funcions de consulta (millorables) com en les de desenvolupament (tractament de les mostres, transcripció o/i sincronització)

- **Oportunista** perquè s'ha alimentat bàsicament d'altres projectes (DOPO Camps 2004, Corpus d'Informatius Orals-UB) o de materials ja recollits (conferències, mostres de la Direcció General de Política Lingüística de la Illes Balears)

Les sinergies entre RETOC i DOPO han provocat la necessitat de transferir a la CCRTV el procediment de creació de corpus, per la qual cosa hem desenvolupat una segona generació de RETOC, que és la que presentem avui. Aquesta versió és multiplataforma i funciona a partir d'eines públiques com el llenguatge de programació PERL i el laboratori fonètic virtual Praat; té, a més, l'avantatge que el 95% del treball de constitució dels corpus es pot fer de manera cooperativa a través de l'Internet.

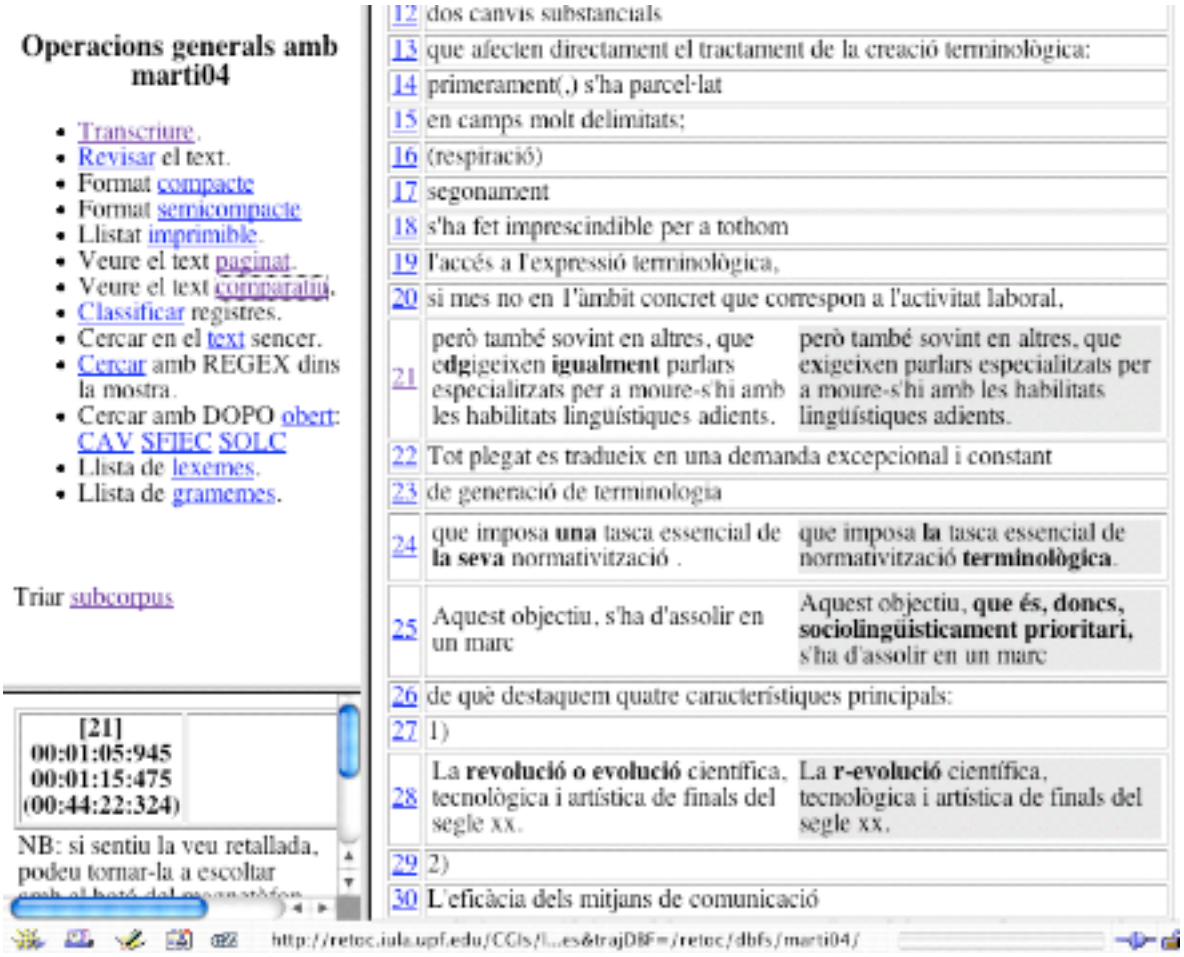
Repertori i dades

Actualment (setembre del 2003) conté:

◇ 253 mostres que duren 249h43'
◇ 20h07' estan transcrites, que dóna
◇ 205104 mots

Alguns dels materials no transcrits acabaran desapareixent perquè incompleixen els requeriments de qualitat que hem anat incrementant. Fonamentalment hi ha conferències de nivell universitari i enregistraments de mitjans de comunicació oral. Totes 249h són accessibles i la infraestructura actual ens permetrà acumular-ne fins a 1800h, amb una estimació de més de 15M de mots, que requeririen 45.000 hores de transcripció.

Procediment de consulta



Tota la consulta del RETOC es fa a través de l'Internet. Tenim en l'actualitat tot un ventall de possibilitats de consulta i cerca, que és preferible que els lectors ho explorin personalment a http://retoc.iula.upf.edu. A tall d'exemple, la imatge precedent recull un estat de consulta en què la pantalla està dividida en tres subfinestres:

- ◇ La més gran conté un llistat comparatiu entre la realització esperada (a la dreta) i l'observada (en aquest cas, car disposàvem d'un text escrit de referència que hem fet servir per transcriure més ràpidament, en una operació que anomenem sincronitzar) però que cal retocar (vegeu el recurs gràfic "r-evolució", que el conferenciant va pronunciar "revolució o evolució") per tal de facilitar una cerca efectiva de, per exemple, la seqüència "ó#o#e" que uneix aquests tres mots, operació que anomenem "ajustar" i que només podem aplicar als textos llegits. L'ajustament també ens permet afinar mesures de correspondència entre temps i nombre de caràcters o de síl·labes per registre. Un darrer detall a assenyalar d'aquesta subfinestra és que el número de registre és un botó hipertextual que ens fa escoltar el fragment corresponent.

- ◇ La subfinestra mitjana és un menú contextual que ofereix diverses modalitats de consulta o procediments de modificació. És contextual en un doble sentit: a) les possibilitats que ofereix depenen de l'estat de tractament en què es trobi una mostra i b) els procediments de modificació estan supeditats a la prioritat que tingui l'usuari que

consulta en funció de diversos procediments de control.

◇ La subfinestra més menuda s'activa quan l'usuari demana escoltar un fragment i permet interactuar amb la veu, reiterant-ne l'audició o aturant-la.

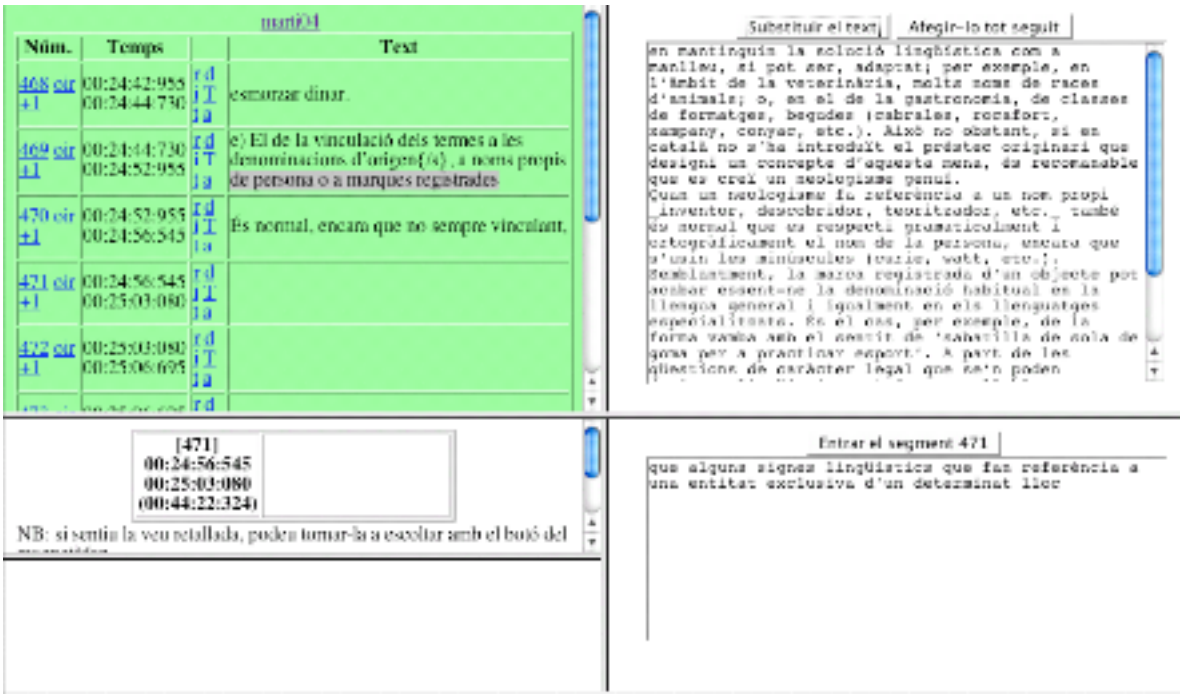
El que hem presentat suara és només una pinzellada dels procediments de consulta que ofereix un corpus hipermèdia i està subjecte a revisió en funció de les millores que ens suggereixen els usuaris.

Procediment de transcripció

Diposem d'un bon equipament d'àudio que permet digitalitzar des dels formats més habituals o directament des de la ràdio i quatre ordinadors amb accés a totes les plataformes

Un cop digitalitzada (convertida en un fitxer d'ordinador que pot ser escoltat en el propi ordinador o que podria ser transferida a un disc compacte, per exemple) una mostra a 16 bits i 32kHz, s'incorpora al corpus i pot ser objecte de transcripció ortogràfica.

En primer lloc, el fitxer es sotmet a una segmentació entre silencis, que es fa amb "scripts" de l'aplicació Praat. El resultat de la segmentació es transfereix a una base de dades que té un registre per a cada segment entre pauses. El procés de transcripció es farà a través de l'Internet, accedint a un diàleg que ens permet escoltar el segment delimitat entre pauses i entrar el text que li correspon.



La imatge precedent ens mostra un estadi del procés de transcripció (mostres espontànies o de les quals no tenim guió o text previ) o de sincronització/ajustament (mostres llegides, encara que sovint amb improvisacions).

◇ La subfinestra de fons verd o fosc (principal o de navegació) ens mostra el resultat de la transcripció, amb expressió dels límits temporals, i amb un conjunt de botons que ens

permeten escoltar el fragment delimitat sol o agrupat amb el següent o escoltar una extensió temporal diferent de la segmentada automàticament i un altre conjunt de botons que permeten modificar el text o el rang temporal (com subdividir el segment o agrupar-lo amb el precedent). La subfinestra té una barra de desplaçament que permet accedir a la part inferior de la mateixa subfinestra, que conté dispositius hipertextuals per navegar per la mostra.

- ◇ La subfinestra superior dreta (de sincronització) conté el text de referència que ens permet sincronitzar. Després de situar en el text una marca que correspon al final del text corresponent al fragment que s'està transcrivint, un botó facilita la transferència de la part delimitada a la subfinestra de transcripció/ajustament, tot eliminant-la del text a sincronitzar.

- ◇ La subfinestra inferior esquerra (de transcripció ajustament) rep el text sincronitzat o permet que el transcriptor teclegi el text. Un cop enviat al servidor, un diàleg permet saltar al fragment següent o subdividir el registre i el servidor en rebutja una part perquè supera els 250 caràcters màxims per registre. En aquest cas, és possible preservar per al registre següent el text corresponent a la segona meitat en què dividim el registre.

- ◇ La subfinestra restant és la d'audició, amb les mateixes funcionalitats que quan es consulta el corpus, però amb l'afegit que té dues subdivisions per tal de poder constrastar més fàcilment dues audicions (p.e. el fragment actual vs la suma de l'actual i el següent).

Convé destacar un aspecte tècnic del procés de transcripció: com que està concebut per facilitar el teletreball distribuït i es basa en la constant confrontació del text amb la veu, cal que el transcriptor disposi d'una bona connexió a l'Internet. Per aquest motiu, amb l'objectiu de permetre el treball de transcripció en desconnexió (com fan, per exemple, alguns col·laboradors de l'ObNeo) hem incorporat al servidor la possibilitat de generar hipertextos estàtics (HTML sense capacitats dinàmiques) que es vehiculen en cedé juntament amb una copia esmicolada de la veu en tants fitxers com fragments conté l'hipertext. La transcripció, en aquest cas, es fa en un document del tractament de textos que després s'importa automàticament al servidor. La feina resulta un xic més feixuga pel canvi constant d'aplicació entre el tractament de textos i el navegador, i a més no es poden modificar els rangs temporals dels fragments, però obvia el problema de la velocitat d'accés a l'Internet.

Experiències d'explotació: IES Cristòfol Rovira

En aquest IES s'han fet experiències d'implicar estudiants de secundària en el procés de transcripció. Presentem tres exemples, alguns ja experimentats amb èxit els darrers cursos a l'IES Cristòfol Ferrer.

a. **El treball de recerca de batxillerat**: És un treball obligatori per a tots els estudiants. Té el seu valor en crèdits del currículum. Cada estudiant l'ha de fer individualment i és tutorat també individualment per un professor del centre d'especialitat afí al tema del treball. Els estudiants escullen el tema i sovint el títol del seu treball a les acaballes del primer curs de batxillerat i el realitzen durant el segon. Ha de ser un treball de recerca, com el seu nom indica. Cal, doncs, que tingui apartats teòrics i empírics. A la vista d'aquests condicionaments acadèmics, RETOC ofereix un excel·lent treball de camp en la transcripció en si mateixa per a treballs de l'àrea de llengua catalana. Disposa d'un objectiu pràctic: assolir el material transcrit; i d'un aprenentatge evident en termes de llengua pel treball que s'hi fa (audició, transcripció, correcció del text escrit, adopció de sistemes i constants de notació, etc.). La sola presentació del material transcrit, amb la memòria de la seva elaboració (exposició de les

tècniques usades, codis i anotacions, descripció de les dificultats superades, dels aprenentatges assolits, etc.), segurament és suficient per a justificar-lo des del punt de vista acadèmic. Malgrat això, si es vol, de la transcripció, entesa com a pretext de recerca, es poden derivar fàcilment explotacions i ampliacions diverses que permeten augmentar l'ambició acadèmica (teòrica o experimental) del treball. Per exemple, l'aprofitament del contingut dels textos per a una ampliació teòrica, i és per això que s'ha escollit inicialment el corpus de conferències: textos monogràfics. Primer, pensant en candidats proclius a l'àrea de llengua, s'usen textos de lingüística, cosa que compacta el treball: praxi i teoria en l'àmbit de l'estudi lingüístic. Tot i això, no es pot descartar diversificar els àmbits temàtics dels textos per tal de motivar estudiants amb altres interessos (literatura, tecnologia, ciències diverses). O la proposta de teoritzar, investigar, ampliar sobre la transcripció en si mateixa: tècniques, sistemes de notació, etc. O la proposta de teoritzar, investigar, ampliar sobre els mecanismes informàtics i telemàtics que donen suport al projecte en les seves fases posteriors, etc.

b. **Exercici intensiu en agrupaments flexibles:** Sovint en algunes matèries comunes, com ara les llengües, els estudiants s'agrupen per nivells i els ensenyaments que s'hi imparteixen s'intenten adaptar a aquests nivells. Una possibilitat és la de muntar una activitat intensiva inserida en el currículum, per exemple de dues setmanes, en què un d'aquests agrupaments treballi algun aspecte relacionat amb la llengua (normatiu, descriptiu, d'estructura, d'expressió, de tipologia textual, d'oralitat, etc.) sobre la base de textos que ells mateixos hagin de transcriure, amb la qual cosa participen en la materialització del seu objecte d'estudi. No cal dir que, en aquest cas també, la temàtica dels textos pot esdevenir un pretext, per exemple, per aprofundir temes de literatura (història o tècnica) sobre la base de transcripcions de rondalles populars o conferències monogràfiques sobre algun autor o període literari que sigui objecte d'estudi en la programació d'aquell curs.

c. **Crèdit variable a l'ESO** (o matèries optatives de batxillerat): Amb un plantejament de continguts i explotació similar a la descrita anteriorment, aquesta mena de crèdits, solts, d'un trimestre, amb grups reduïts d'alumnes, optatius, en part alliberats del jou de la programació "troncal" del les matèries als crèdits comuns (obligatoris), permet muntar un treball monogràfic en grup sobre la base de les transcripcions. En aquest cas, el treball pot perllongar-se durant tot el trimestre, i per tant, les explotacions pedagògiques s'han de programar coherentment amb aquest nou escenari.

Experiències d'explotació: Comunicació Audiovisual

Durant el curs 2002-03 vam intentar explotar per a la docència la informació emmagatzemada a RETOC i les possibilitats de tractar-la que ens oferien les eines informàtiques de què disposàvem.

L'oportunitat va sorgir per a l'assignatura «Català I. L'estàndard oral», assignatura de quatre crèdits troncal de primer cicle de la llicenciatura en comunicació audiovisual de la UPF.

Val a dir que les característiques d'aquest material informatitzat, procedent dels informatius de la ràdio nacional de Catalunya, difícilment podia adequar-se més als objectius docents de la matèria, un dels quals és que els estudiants assoleixin el domini de la fonètica del català estàndard.

I això en un doble vessant: la capacitat d'articular-lo correctament i la capacitat d'analitzar críticament en relació a l'estàndard els trets fonètics en els discursos sentits.

És en aquest segon vessant que es va treballar amb el material i els recursos de què acabem de parlar; i ho vam fer, en aquesta fase, dissenyant uns exercicis que permetessin treballar en

l'observació dels trets fonètics conflictius; és a dir, els que, per l'experiència amb els mateixos estudiants, hem constatat que sovint es realitzen erròniament, ja sigui per desconeixement de l'estàndard o de la pronúncia genuïna, ja sigui per l'existència de dificultats articulatòries personals.

Exemples d'aquests trets (vocàlics i consonàntics), en són els que corresponen a les grafies subratllades que segueixen: serà / acord / Vallès / cuina / avions / que alguns / actual / client / trasllat / presentar / grans autors / localitzar / exigir / gestió / xilè / platges / govern / nord.

El resultat dels exercicis havia de ser sempre la indicació, per a cada cas, de si la pronúncia escoltada era correcta o no. I el mètode consistia en l'audició, captada a través de la xarxa, tantes vegades com fos necessari, d'unes seqüències dels informatius, prèviament triades per mitjans automàtics, que contenien els trets a observar.

Aquestes seqüències estaven coherentment agrupades en diferents unitats que constituïen els exercicis, objecte d'avaluació.

Abans d'acabar no voldríem deixar de subratllar un inconvenient i una virtut d'aquest novedós recurs que hem utilitzat. L'inconvenient, de moment encara superable –desafortunadament!-, és la dificultat de trobar en els professionals de la ràdio en qüestió determinats defectes fonètics que interessa de fer observar als estudiants; això fa més àrdua la feina de qui prepara els exercicis.

La virtut és la nitidesa de les seqüències que s'acaben proposant, deguda al fet que els problemes que s'han de detectar apareixen prou aïllats; això permet a l'estudiant centrar-se en aspectes concrets i treure'n profit, cosa que no succeiria si la proposta partís d'altres tipus de discursos orals.

Experiències d'explotació: Tesis doctorals IULA

Tres tesis doctorals en curs a l'IULA (Anna Corrales, Jordi Cicres i Francesca Salvà) es beneficien del corpus (dades o tècniques), al mateix temps que hi aporten millores o enriquiment pels seus propis resultats, per les seves dades o en tant que "testadors".

En el marc d'un seminari orientat a les necessitats ortològiques d'un traductor audiovisual o d'un intèrpret, estudiants de Traducció varen usar un filtre DOPO dissenyat per detectar la casuística plantejada a la Proposta d'estàndard oral de la SFIEC.

Col·laboracions cercades i possibilitats futures

Un corpus ambiciós com el que estem construint necessitarà molta col·laboració per poder-se completar, especialment perquè no el considerem tancat amb les aprox. 250 hores actuals. Atès el caràcter públic del nostre corpus, tenim esperances raonables de rebre col·laboracions externes, tant per part de simples parlants que puguin fer una primera transcripció ortogràfica que agiliti el treball posterior dels experts com per part d'especialistes, bàsicament de dos col·lectius: lingüistes i tecnòlegs.

Pel que fa als lingüistes, ens interessen:

- ◇ aportació de materials d'altres tipologies
- ◇ transcripció o sincronització de mostres

◇ revisió de transcripcions
◇ marcatges específics

Pel que fa als tecnòlegs, ens serien beneficioses aportacions que ajudin a:

◇ dividir automàticament a nivells inferiors al del segment entre pauses
◇ detectar els paràmetres òptims de segmentació automàtica per a enregistraments en entorns sorollosos o amb fons musical
◇ detectar canvis de parlant o encavalcaments
◇ separar segments nets dels altres
◇ detectar falques o mots freqüents
◇ identificar els locutors
◇ i, en un futur tan de bo no massa llunyà, reconèixer text sense ensinistrament

També esperem obtenir i proporcionar, doncs, intercanviar ajuts amb altres projectes o camps de recerca amb interessos pròxims:

Projecte	**benefici que obté de RETOC**	**benefici que aporta al RETOC**
DOPO	procediments de creació i consulta de corpus orals	millora i seguirà millorant els procediments de filtratge per localitzar casuística rellevant
SOLC	possibilitat de constrastar pronúncia prescriptiva amb casos reals	regles de transcripció per a implementar (aviat) cerques de base fonètica
cadena d'eines LIC	l'obliga a flexibilitzar-se per tal de tractar representacions escrites de llengua oral	possibilitat (futura) de marcar morfològicament el text per millorar les opcions de cerca
qualsevol tesi lingüística sobre llengua oral	una forma molt àgil i potent de tractament de materials orals	incorporació de mostres i ampliació de tipologies de mostres o revisió exhaustiva de mostres ja incorporades
ObNeo	utilitza els procediments de creació i consulta de corpus orals per detectar neologia oral	els materials que genera es poden incorporar com a mostres
SLCUB	enregistrament de materials compartits que el SLCUB empra en projectes de formació telemàtica	transcripció dels materials compartits
fonètica forense	tècnica de transcripció fonètica semiautomatitzada aproximativa amb les regles SOLC	enriquir la futura col·lecció de regles SOLC descriptives
tecnologia de la veu	bancs de proves	vegeu supra

A més del treball actual, estem arrodonint idees que poden millorar en el futur la qualitat dels materials transcrits o optimitzar els procediments de tractament, algunes de les quals ja hem apuntat:

◇ transcripció fonètica automàtica, que pugui ser corregida contrastivament per més d'un investigador i que arribi a ser autocorrectiva (respecte als errors no detectats habitualment per un determinat transcriptor) gràcies a tècniques de "machine learning"

- ◇ manteniment automàtic de la base de dades de locutors
- ◇ regles SOLC descriptives de subdialectes o modalitats amb un interès específic (v.g. per a la fonètica forense)
- ◇ reconeixement automàtic del canvi de locutor
- ◇ reconeixement automàtic de punts de vacil·lació o de canvi d'estil

En resum, estem davant d'una empresa gegantina, però creiem que si treballem adequadament les sinèrgies amb d'altres projectes o investigadors individuals, es podrà avançar molt més ràpidament.

Conclusions

Tenim un corpus accessible públicament, limitat a uns pocs tipus de llengua oral, però que ja ha començat a servir en projectes de recerca i de docència

Hem desenvolupat una tecnologia pròpia per a la generació i codificació de corpus que ens permet "servir-los" a l'Internet amb molta agilitat

Amb les nostres dades, la llengua catalana passa a ser la llengua romànica amb més dades orals accessibles a l'Internet i la primera entre les llengües sense estat.

Hem connectat amb d'altres projectes, sigui de recerca, sigui de docència, que ens potencien i que potenciem

Confiem que arribarem a ser un suport important per als projectes de tecnologia de la veu en català

Referències

Salvà, Francesca; de Yzaguirre, Lluís; Cabré, M. Teresa (2004): "Un diccionari ortològic català", dins de "*Actes del 13è Col·loqui de l'AILLC (Girona 2003)*", ed. Martí, Sadurní; Cabré, Miriam; Feliu, Francesc; Iglesias, Narcís i Prats, David. Barcelona, PAM, 2004.

Camps, Oriol; de Yzaguirre, Lluís; Salvà, Francesca (2004): "Diagnòstic ortològic assistit", dins de "*Actes del 13è Col·loqui de l'AILLC (Girona 2003)*", ed. Martí, Sadurní; Cabré, Miriam; Feliu, Francesc; Iglesias, Narcís i Prats, David. Barcelona, PAM, 2004.

Enllaços

Praat	http://www.praat.org/
ObNeo	http://www.iula.upf.es/obneo/
exemple d'aplicació DOPO al corpus RETOC:	http://retoc.iula.upf.edu/CGIs/escoltador.pl?operacio=previText&numMostra=12
SOLC	http://retoc.iula.upf.edu/SOLC/
LATEL	http://www.iula.upf.es/latel/lpresca.htm
Publicacions	http://terminotica.upf.es/membres/DE_YZA/PUBLI/PUBLIC.HTM