DIAGNÒSTIC ORTOLÒGIC ASSISTIT

Presentació

El projecte Diagnòstic Ortològic Per Ordinador (DOPO) pretén posar les noves tecnologies al servei de: a) l'anàlisi ortològica amb finalitats de control de qualitat en els àmbits professionals de la llengua oral; b) la formació acadèmica d'aquests professionals; i c) la recerca en fonètica prescriptiva.

En aquest projecte hi han confluït els interessos de recerca del Laboratori de Tecnologies Lingüístiques amb els de control de qualitat de la Unitat d'Assessorament Lingüístic de Catalunya Ràdio. En aquest escrit presentarem la gènesi i evolució del projecte, exemples dels procediments d'utilització i dels resultats obtinguts així com les perspectives de futur que li veiem.[1]

Finalitat del *DOPO*

Creada el 1983 per una llei del Parlament de Catalunya, Catalunya Ràdio, té com un dels objectius principals contribuir a la normalització lingüística, el procés que té per finalitat posar la llengua catalana en condicions de sobreviure en el seu espai històric, que ha estat ocupat, especialment en els usos públics, pel castellà o el francès.

A l'època de la seva creació, Catalunya Ràdio obria un espai de comunicació en català, ja que en aquell moment no hi havia emissores de radio en català, fora d'algunes excepcions que no arribaven a cobrir tot el territori ni totes les hores del dia i de la setmana.

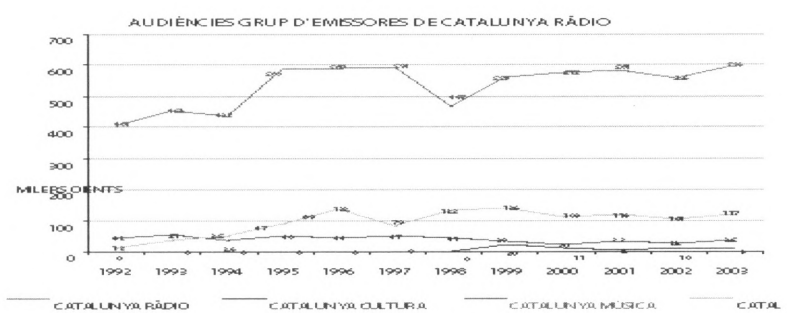
Actualment, la nostra primera emissora posseeix una audiència mitjana de 590.000 oients aproximadament, i és líder en la ràdio convencional a Catalunya, a més de 100.000 oients de distància de la segona classificada. Té també un canal exclusivament informatiu, Catalunya Informació, que també és líder en el seu gènere, amb 67.000 oients, i dues emissores més: Catalunya Música (música clàssica)

1. El projecte DOPO ha gaudit de finançament per part del Ministerio de Ciencia y Tecnología (FIT 150.200-2000-248).

i Catalunya Cultura, que explora les possibilitats de la ràdio en els temes de la cultura, evitant al mateix temps la política i els esports.

Al moment actual es pot dir que existeix un espai català de comunicació, almenys pel que fa a la ràdio, com ho prova el fet que des de fa uns anys diversos empresaris privats s'han llançat a la creació d'emissores de ràdio que parlen en català, primer especialitzades en música pop-rock i poc després (setembre del 2000) generalistes.

Una vegada establert aquest espai de comunicació, als mitjans de comunicació públics els queda la important missió de ser líders de la qualitat de les programacions, i especialment de la qualitat lingüística. Aquesta missió deriva del fet de ser pagats (ni que sigui parcialment) pel pressupost del govern autònom. En efecte, una de les funcions dels mitjans de comunicació de masses (MCM) és establir, conservar, eixamplar i difondre, a través de l'ús, la llengua estàndard.



Tanmateix, les condicions en què s'ha de mantenir la qualitat lingüística de les emissions no són pas favorables, tenint en compte la pressió que el castellà i altres llengües exerceixen sobre la nostra. S'ha de destacar que gairebé totes les entrades d'informació es fan en castellà, francès o anglès, i que només les notícies locals arriben als mitjans en català. En la feina dels periodistes i dels redactors es produeix, doncs, un incessant contacte de llengües, que deriva de la mateixa naturalesa de la tasca informativa, així com de la situació de doble oficialitat lingüística.

Això significa que no es pot pensar en el manteniment de la qualitat sense el treball d'uns quants professionals lingüistes (correctors, ortòlegs) que miren de compensar aquesta pressió de les altres llengües. Això vol dir que han de vigilar la introducció de mots estrangers i fins i tot de maneres de dir importades que no siguin necessàries, i que han de controlar també la pronunciació en antena, i no

tan sols en una varietat del català, sinó en les dues varietats principals (oriental i occidental) i respectant unes quantes subvarietats.

Això significa, a més, que malgrat el progrés del català a l'ensenyament (que pretén formar ciutadans en teoria perfectament bilingües), s'ha de posar en marxa una selecció lingüística rigorosa del personal que posarà veu a la ràdio.

Fins i tot si una part important del treball a favor de la qualitat lingüística es fa en la fase escrita dels textos difosos (correcció de textos abans que es llegeixin per antena), queda la feina de control de la qualitat de la llengua efectivament utilitzada en les emissions, i la necessària selecció del personal. Aquesta feina d'escoltar la ràdio, o un enregistrament, o l'actuació d'un candidat a periodista de ràdio amb la finalitat de detectar-hi errors de pronunciació s'ha de fer d'orella. Però això implica alguns problemes. Sobretot els que deriven de les diferències de sensibilitat —entenguem-hi nivells d'exigència— entre les diverses persones que l'exerceixen.

És per aquesta raó que la CCRTV ha decidit procurar-se una eina informàtica que pugui assegurar una més gran objectivitat de l'observació —els mateixos fenòmens per a tothom, amb adaptació a la fonètica utilitzada per cadascú, cosa que vol dir també una exigència igual per a tothom. De moment, això no és possible amb el tractament de fitxers de veu, en vista de les dificultats que es troben en l'explotació dels sistemes de reconeixement de la veu, i del gran nombre de veus a observar (més de tres centenars).

Deixeu-nos dir que el cervell humà encara és més ràpid que els sistemes electrònics per identificar les correspondències entre un so i la seva representació en l'ortografia. Però és possible de preveure, per als programes de ràdio basats en la lectura de textos —que, en les quatre emissores de Catalunya Ràdio, arriben a gairebé 35 hores per dia— una eina basada en el text llegit.

Pel que fa al Laboratori de Tecnologies Lingüístiques, ha concebut el DOPO com una manera d'aconseguir sinèrgies amb el món empresarial que permetin assumir els costos de creació de corpus orals sincronitzats i del disseny informàtic per tractar-los. És a dir, que la possibilitat d'aplicar el DOPO a un corpus oral en fa més econòmica la seva producció. A més la capacitat del DOPO de concretar cerques molt complexes en un procediment relativament senzill en fa molt interessant l'aplicació a la docència (Filologia, Interpretació, Comunicació Audiovisual...) i a d'altres projectes de recerca.

Versions del DOPO

Una primera generació del producte va ser presentada el 2000 al 1er. Freiburger Arbeitstagung zur Romanistischen Korpuslinguistik (Camps 2000). Les seves prin-

cipals limitacions eren que tot el tractament s'havia de fer per encàrrec a l'Institut Universitari de Lingüística Aplicada i que el procediment de treball generava pàgines HTML estàtiques: en cas de detectar un error de transcripció, calia corregir-lo, regenerar totes les pàgines i tornar-les a instal·lar al servidor.

Per aquest motiu, es va decidir dissenyar una versió més evolucionada de l'eina que fos multiplataforma i que es pogués transferir a la CCRTV. Aquesta eina és el DOPO-2, que presentem avui i que, tot i que millora la majoria d'aspectes del DOPO-1, encara no ha assolit la seva simplicitat d'ús, car s'hi han afegit moltes prestacions que fan difícil compaginar la potencia amb la usabilitat.

Es tracta d'una implementació en Perl del sistema de filtratge que permet actuar sobre qualsevol mostra a l'instant i generar un hipertext dinàmic de consulta. Cada investigador o equip de recerca pot definir els seus filtres amb la finalitat d'obtenir protocols d'observació dels fenòmens del seu interès. Aquests filtres s'instal·len al servidor i esdevenen d'ús general, encara que, si calgués, es podrien limitar a determinats usuaris.

Formalisme actual

En la pantalla següent es pot veure un exemple d'interrogació amb el DOPO:

Les minúscules indiquen la grafia literal, i les majúscules codifiquen grups associats a fonemes amb els límits imposats per l'ortografia; es poden agrupar condicions i combinar-les amb inclusions o exclusions de mots. Es poden demanar coses com:

Vx#V;Vs#V	mots acabats en *s* o *x* postvocàlica seguits de vocal (altern. V[sx]#V)
A#A	dues vocals tòniques amb frontera de mot
VNVRV	seqüència de vocal-sonora-vocal-sorda-vocal
nt#V; nd#V	*nt* o *nd* finals seguides de vocal (alternativament, n[td]#V)
Cx	qualsevol consonant seguida de *x*

Tot això amb la limitació que ens basem en l'ortografia i se'ns poden escapar, per exemple, bona part de les vocals tòniques, però les que trobem poden servir si tenim en compte aquesta manca d'exhaustivitat.

Es poden fer filtres més complexos combinant regles amb .OR. o amb .AND. i amb llistes d'exclusió o inclusió fixades en fitxers ad hoc. L'usuari tria el format de sortida com a protocol imprès, protocol d'Internet anònim (=enquesta) o nominal (=examen). També pot fer restriccions manuals a posteriori, per tal de restringir la casuística d'un filtre. Pot limitar el nombre d'exemples, fer-los triar aleatòriament i visualitzar-los en ordre cronològic o desordenadament.

A més de les codificacions específiques del DOPO, aquest «hereta» de Perl les expressions regulars, que li afegeixen una potència espectacular —que exigeix, però, un esforç suplementari d'aprenentatge, optatiu.

Exemples d'aplicació

La pantalla següent ens mostra un llistat de resultats obtingut en sol·licitar el filtre Vx#V;Vs#V:

Cerquem Vx#V;Vs#V per a 80.58.35.237 oblig=amb excl=quan

Num	Mot
oir	*a partir de l'estiu passat amb trobades entre*
oir	*I pel que fa a la previsió meteorològica hem de dir que El sol avui serà present a totes les comarques, amb temperatures altes al migdia.*
oir	*de divendres amb alguns núvols prime cap a l'extrem nord de Catalunya durant aquesta tarda*
oir	*Gràcies a Montse Mir. Aquesta comunicació d'urgència amb el Camp Nou on s'ha confirmat la baixa*
oir	*En només un mes, i comparant les dades amb un sondeig que va publicar el febrer el mateix diari,*
oir	*Segons l'enquesta preelectoral del CIS*, el Partit Popular guanyaria les pròximes eleccions generals amb el*
oir	*Cook va tenir una reunió amb el llavors ministre d'Afers Estrangers xilè Gabriel Valdés a Nova York.*
oir	*per beneficiar Catalunya amb més autogovern i amb un finançament just.*
oir	*ha denunciat amb paraules molt dures*
oir	*Segons el conservador Daily Telegraph, que des del primer dia ha exigit l'alliberament de Pinochet, el pacte per acabar amb el procés es va gestar*
oir	*L'Independent coincideix amb molts dels detalls del Telegraph i afegeix que el ministre Straw,*
oir	*així com l'atenció crítica amb aquestes unitats especials.*
oir	*El candidat socialista ha acusat el govern d'Aznar d'haver actuat en aquest cas amb hipocresia i mentides.*

Hem trobat 13 registres sobre 598.Hem trobat 13 registres entre el registre 0 i el 598

(mots acabats en *s* o *x* postvocàlica i seguits de vocal), supeditats a dues llistes, la d'inclusió que conté *amb* i la d'exclusió, que conté *quan*; la paraula *oir* (triada per la seva concisió) és un botó que permet escoltar el fragment en qüestió.

Aquest document hipermèdia pot ser conservat com a fitxer propi a l'ordinador del corrector o ortòleg; això li permetrà interactuar amb els locutors que supervisa d'una manera molt efectiva i àgil o acumular exemples per a la formació interna.

Alternativament, el DOPO pot ser aplicat al nivell de la paraula o grup de paraules consecutives, cosa que generalment evita la necessitat de tenir llistes d'exclusió o d'inclusió, però que és més lent perquè treballa amb més unitats. En qualsevol cas, es recupera el fragment sencer i s'escolta sencer.

Previsions d'evolució

Encara que la versió actual del DOPO es pot considerar un producte acabat, sempre se li poden fer millores. La primera és la utilització de les regles del SOLC per refinar les cerques a partir de la transcripció fonètica automàtica esperada. Una altra és la d'oferir la possibilitat de demanar buidatges estadístics, cosa que pot interessar a l'hora de valorar la importància d'una certa incidència ortològica en probabilitat d'aparició.

Conclusions

Hem establert un procediment pel qual es pot fer recerca amb finalitats ortològiques (formació, correcció, recerca...) sobre corpus digitalitzats i transcrits ortogràficament. L'hem refinat amb la Unitat d'Assessorament Lingüístic de la CCRTV i ara aquest organisme està en fase d'implementar-lo en el seu sistema informàtic. El procediment de filtratge s'ha incorporat al corpus RETOC.

Oriol Camps, Lluís De Yzaguirre i Francesca Salvà
Unitat d'Assessorament Lingüístic CCRTV /
Laboratori de Tecnologies Lingüístiques,
Institut Universitari de Lingüística Aplicada,
Universitat Pompeu Fabra

REFERÈNCIES BIBLIOGRÀFIQUES

CAMPS *et alii* 2002: Oriol Camps, Lluís De Yzaguirre i Anna Matamala, «DOPO, un outil d'analyse orthologique», a *Romanistische Korpuslinguistik. Korpora und gesprochene Sprache*, ed. C.D. Pusch i W. Raible, Tübingen.

DE YZAGUIRRE *et alii* 2005: Lluís De Yzaguirre, Antoni Jaume Farriols, Jaume Martí, «El corpus RETOC: Un corpus oral per a la recerca i la docència», a *Actes AILLC 13*, II, 495-504.

SALVÀ *et alii* 2005: Francesca Salvà, Lluís De Yzaguirre, i Maria Teresa Cabré, «Un diccionari ortològic català», a *Actes AILLC 13*, II, 409-418.

ENLLAÇOS

Catalunya Ràdio: http://www.catradio.com/
Aplicació DOPO al 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