

Projectes Orals

*Laboratori de Tecnologies Lingüístiques
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Lluís de Yzaguirre*

Sumari

- Presentació
- Antecedents
- Creació i manipulació de corpus orals
- Modelització ortològica per regles
- Aplicabilitat a d'altres llengües
- Col·laboració amb d'altres projectes
- Oralistes sense fronteres

Presentació

El LaTeL té oberta una línia de treball en el camp de la llengua oral. Aquesta línia està desenvolupant aplicacions per a la manipulació o/i modelització de la llengua oral.

Els treballs quotidians s'orienten al català, però des del punt de vista tecnològic el 95% del desenvolupament és universal.

Hi ha actualment tres tesis doctorals obertes que treballen en aquesta línia (Corrales, Salvà, Teixidor).

Antecedents

La tesi doctoral de LdYM sobre l'estructura sil·làbica del català (1990) es va beneficiar d'un sistema de transcripció per regles que és a la base d'un dels treballs que presentarem.

Els corpus orals de la UB, en què varen participar Cabré, Lorente i LdYM mentre foren membres del DFCUB, representen un altre antecedent del que s'està desenvolupant a l'IULA.

Corpus orals

Creació i manipulació de corpus orals

- SINCRO (Sistema informàtic de normalització de corpus per a la recerca oral)
- DOPO (Diagnòstic ortològic per ordinador)
- RETOC (Repertori electrònic de textos orals catalans)

http://terminotica.upf.es/membres/DE_YZA/PUBLI/retoc.pdf
http://terminotica.upf.es/membres/DE_YZA/PUBLI/Dopo.pdf

Corpus orals

El Repertori Electrònic de Textos Orals Catalans (RETOC) és un corpus de veu digitalitzada, parcialment transcrita i accessible universalment a través de l'Internet, actualment operatiu, però encara millorable, pel que fa a les funcions de consulta i a les de desenvolupament (tractament de les mostres, transcripció o/i sincronització). La versió actual és multiplataforma i funciona a partir d'eines públiques com el llenguatge de programació PERL i el laboratori fonètic virtual Praat; i 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.



Corpus orals

Oral

Fonamentalment, hi ha conferències de nivell universitari i enregistraments de mitjans de comunicació oral, organitzats en subcorpus a través dels quals es pot accedir a cada enregistrament. Pel que fa al nombre d'hores, totes accessibles, en podeu consultar l'índex general. La previsió és que la infraestructura actual ens permetrà acumular-ne fins a 1800h, amb una estimació de més de 15M de mots. Ara en tenim 275, de les quals n'hi ha més de 25 transcrits, amb més de 252.000 paraules.

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

7



Corpus orals

Oral

Aquest projecte embrionari s'ha pogut desenvolupar sobretot gràcies a la relació amb el projecte DOPO (Diagnòstic Ortològic per Ordinador) de col·laboració del Laboratori de Tecnologies Lingüístiques (LATEL) amb la Unitat d'Assessorament Lingüístic (UAL) de Catalunya Ràdio, projecte que n'ha finançat una part molt important.

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

8



SOLC

Oral

Modelització ortològica per regles

- SOLC (Servidor ortològic de la llengua catalana)

http://terminotica.upf.es/membres/DE_YZA/PUBLI/Solc.pdf

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

9



SOLC

Oral

El projecte del diccionari ortològic automatitzat de català (SOLC) està orientat a resoldre les necessitats de molts professionals, lingüistes i no lingüistes davant un estàndard oral molt plural i, en alguns casos, poc definit en relació a la normativa. Està sobretot enfocat als mitjans de comunicació, pel seu paper cabdal en la difusió d'un model de llengua estàndard, perquè doni solucions a locutors, ortòlegs, actors de doblatge, productors multimèdia, ensenyants, oradors o enginyers lingüístics.

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

10



SOLC

Oral

L'objectiu és convertir en regles les formulacions de la Proposta d'Estàndard Oral de l'IEC per fer-la més exhaustiva ja que, ara per ara, és incompleta i asistemàtica. L'eina inclourà, a part del vocabulari general (sempre en creixement), terminologia, informacions enciclopèdiques i lèxic i onomàstica forans, ja que els dubtes dels professionals no s'aturen al vocabulari compendiat en una eina lexicogràfica.

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

11



SOLC

Oral

A partir d'una representació comuna pseudofonèmica (desortografiada), l'eina extreu el model de pronunciació segons la combinació de les variants següents: dialecte/propietat/àmbit/formalitat. A més, també es poden consultar les regles que hi actuen i accedir a l'aplicació de les regles sintàctiques (encara en provatures).

11/10/2004 09:02

Laboratori de tecnologies lingüístiques iula.LaTel@upf.edu

12



Altres llengües

El SINCRO només té dos aspectes idiosincràtics: la partició sil·labica i la verbalització de xifres i altres elements. La transcripció és tributària del SOLC. Per la resta és transportable com ho demostren les mostres sobre Sango (Diki-Kidiri) i Black English (Sutcliffe).

DOPO és dependent del català quant a la interpretació pseudofonemàtica de l'ortografia, que es pot adaptar sense problemes a d'altres llengües (asumint-ne els límits)

SOLC és idiosincràtic quant a l'inventari de fonemes i al·lòfons i quant al banc de regles. Però ambdós aspectes es podrien externalitzar per fer-lo aplicable a d'altres llengües.



Col·laboració

Els diversos projectes orals col·laboren reciprocament.

Col·laborem amb l'ObNeo.

Ens convindria col·laboració de telecos per a la segmentació mot a mot.

Explotarem la possible ajuda en mètodes numèrics amb finalitats fonomètriques.



Oralistes sense fronteres

Davant les 45000 hores de transcripció pendents i moltes altres tasques factibles, qualsevol aportació serà benvinguda.



***Laboratori de Tecnologies Lingüístiques
Institut Universitari de Lingüística
Aplicada
Universitat Pompeu Fabra
Lluís de Yzaguirre
iulaLatel @ upf.edu***