UN DICCIONARI ORTOLÒGIC CATALÀ

Presentació

En aquesta comunicació presentem el diccionari ortològic telemàtic que hem dissenyat i implementat a l'Institut de Lingüística Aplicada de la Universitat Pompeu Fabra, explicarem com el seu model de regles vehicula la *Proposta d'estàndard oral de la Secció Filològica de l'IEC* (*PEOSFIEC*) i valorarem els resultats obtinguts fins ara.

Fa pocs anys que tenim una primera aproximació a un model ortològic per al català, destinada fonamentalment als mitjans de comunicació oral (MCO), és a dir a professionals amb una formació lingüística limitada, que no sempre tenen personal de suport, depenen de poders públics amb actituds diverses davant l'autoritat lingüística o d'empreses privades que hi tenen un interès sovint conjuntural i que reben pressió de les seves audiències.

Aquesta proposta oblida altres col·lectius professionals (polítics, màrqueting, oradors, docents...), alguns dels quals tenen models propis (Església catòlica), d'altres tenen un panorama força complex, com el de la docència, que necessitaria diversos submodels i que té una situació relativament peculiar en els mestres desplaçats del seu àmbit dialectal i d'altres de rellevants, com els polítics, amb tot el seu impacte mediàtic, o l'àmbit del dret, etc.

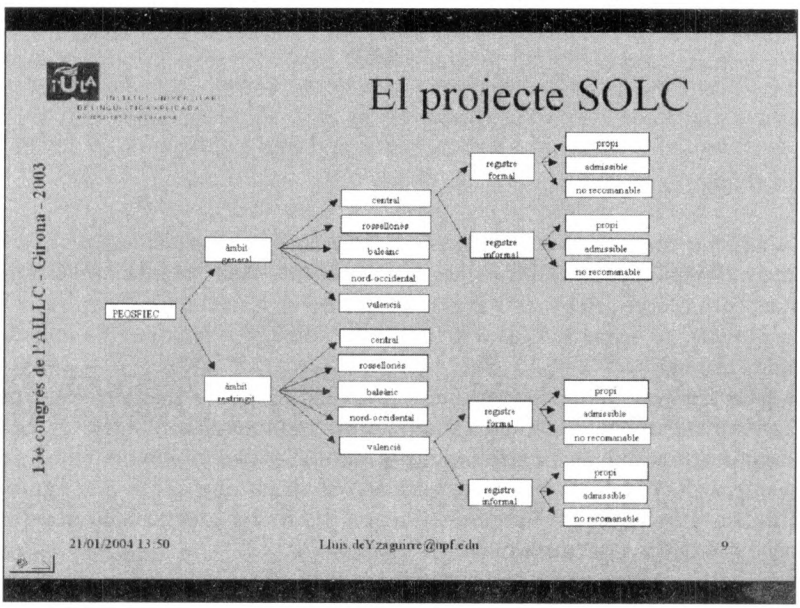
Característiques de la *PEOSFIEC*

Les principals característiques de la *PEOSFIEC* són a) la seva flexibilitat (no ofereix un model monolític de pronúncia, sinó que en possibilita una munió segons el dialecte del locutor, segons l'homogeneïtat dialectal de l'audiència, el nivell de formalitat i amb tolerància a certes opcions admissibles) i b) la incompleció (en la seva condició de proposta, és incompleta i assistemàtica).

Ens interessa ressaltar el fet que la riquesa de solucions de la *PEOSFIEC* es torna una dificultat quan el qui l'ha de vehicular és un professional dels MCO sense for-

mació filològica, per tal com la combinació de tots els valors possibles de totes les variables incloses a la *PEOSFIEC* dóna un sostre hipotètic de seixanta realitzacions per a cada fenomen contemplat.

Hem sintetitzat aquesta complexitat en la gràfica següent:



Pel que fa als fenòmens no descrits, podríem citar, per exemple, els grups inicials **#pt**, **#pn** o **#mn** en contraposició a **#ps** mantingut (tot i que *psalm* té *salm* i que seqüències **CCC#ps** resulten embarbussadores).

D'altres tenen una formulació que no contempla tota la casuística. Així, no és recomanable la «pronunciació de la *t* ortogràfica en mots com *setmana* o *cotna*. La pronunciació recomanable és *semmana*, *conna*» (*PEOSFIEC* 31):[1] hi ha el precedent de Vallverdú (1984: 27) que considera diferent aquest grup en els mots patrimonials com *setmana* dels cultismes com *aritmètica*. Atesos els destinataris principals de la *PEOSFIEC* i el rol de Vallverdú a la CCRTV, s'hauria d'haver assumit o descartat explícitament aquesta excepcionalitat i d'altres de semblants. Altres casos en què no es contempla tota la casuística serien els mots amb GL/BL (*PEOSFIEC* 25) o les excepcions al grup TLL.

1. Per a les referències a la PEOSFIEC, consulteu la versió de treball amb numeració d'apartats, citada a la fi.

Mentre es recomana una assimilació homoorgànica, se'n desaconsellen o ignoren d'altres; el cas del grup TM (*PEOSFIEC* 31) precedent contrasta amb el tracte que es dóna a la pronúncia balear de CT en mots com *doctor* (*PEOSFIEC* 32) o amb els grups DM o BM, transcrits al diccionari VOX (Badia 1990) com a *ammirar* i *summarí* que no es contemplen.

Altres mancances són les moltes llistes no exhaustives de «mots com...», l'ús d'un concepte com el de «cultisme», de mala aplicació en mans d'un llec.

Necessitat d'un diccionari ortològic

Molts dels reduccionismes en aquesta problemàtica vénen de la il·lusió de voler tenir la mínima distància entre llengua oral i llengua escrita. Moltes de les ingenuïtats que s'han dit sobre la conveniència que els MCO assumeixin la llengua del carrer o el català que ara es parla cerquen inconscientment aquesta immediatesa, sense adonar-se que tot el que acostem la llengua escrita a una modalitat concreta de l'oral l'allunyem de les altres i, inversament, tot el que acostéssim el model estàndard de llengua oral al model unitari de llengua escrita en mataria la vivacitat. Hem d'assumir com a normal que hi hagi una distància entre l'escrit i l'oral, que no tot sigui predictible (hi ha excepcions, hi ha neologismes amb diferents graus de facilitat d'integració, en el cas dels manlleus la distància entre la pronúncia endòfona i l'exòfona pot ser molt variable, com el nivell de coneixença entre els parlants de la llengua d'origen, etc.).

Tots aquests problemes es resolen fent un diccionari ortològic, però l'empresa presenta algunes dificultats, car ha de cobrir un lèxic que no para de créixer amb una multiplicitat de formes encara no tancada i entrades enciclopèdiques encara més nombroses; tampoc no és fàcil concretar els criteris: trobem diferències de transcripció fins i tot entre diccionaris amb un únic model de transcripció (que tenen destinataris diferents, en un marc sociopolític canviant...); també els diccionaris fonètics més recents ometen diferències de formalitat o de propietat. En resum, els fets demostren que un diccionari ortològic que resolgui els dubtes dels professionals de la llengua oral donant cobertura a totes les variants contemplades per la *PEOSFIEC* seria inviable econòmicament si s'havia de vehicular en paper, especialment pel seu caràcter efímer i la seva caducitat.

L'alternativa actual a un diccionari en paper és una base de dades accessible a través de l'Internet. Però un cop es té la solució tecnològica per oferir un diccionari en línia, és possible obtenir uns valors afegits per al producte. El primer valor afegit és el de respondre millor a la immediatesa del treball quotidià als MCO. El segon valor afegit és el de tenir una formalització computacional de la *PEOSFIEC* que, a

més de la finalitat primària d'alimentar el diccionari en línia, permeti als ortòlegs estudiar els límits, les incongruències, les assistemacitats que pugui tenir o no la *PEOSFIEC*.

Model per regles

Hem elaborat, doncs, un sistema de transcripció que formalitza fins on és possible la *PEOSFIEC*. Aquest model es basa en el que els anys 1986-88 va servir per generar la base de dades sil·làbica (2.418.000 registres) usada a la tesi de De Yzaguirre (1990) sobre l'estructura sil·làbica del català central, model presentat en demostració pública al congrés de la Sociedad Española de Procesamiento del Lenguaje Natural del 1989. Els objectius d'aquest model són:

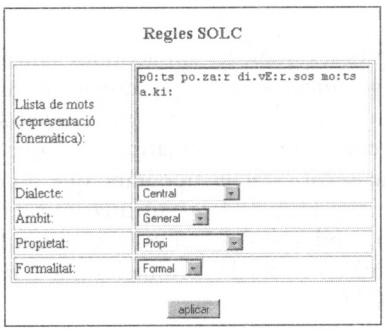
- Implementar la *PEOSFIEC*
- Sistematitzar-la i aplicar-la exhaustivament
- Modelitzar-la per preveure'n evolucions
- Recopilar tota la casuística que no s'avé a un tractament regular
- Separar l'ortologia sintàctica de la lèxica

Aquestes regles són sensibles a les variables *PEOSFIEC* (dialecte, àmbit, propietat i formalitat) i treballen a partir d'una representació comuna pseudofonemàtica (desortografiada), que pot generar-se automàticament, però que requereix intervenció humana allà on l'ortografia no és biunívoca (la E unitària de *menja* no es pot destriar de la E amb tres solucions de *cadena*, basant-se només en l'ortografia). Des d'un punt de vista tècnic se'n pot dir que han estat implementades en Perl fonamentalment amb expressions regulars.

Regles lèxiques

La majoria de les regles de la *PEOSFIEC* actuen entre els límits del mot, per la qual cosa les anomenem lèxiques, en oposició a les que actuen en situacions de contacte entre mots, que anomenem sintàctiques. Poden ser provades lliurement a http://retoc.iula.upf.edu/SOLC/marcSOLCmot.htm.

Com es pot veure en el diàleg home-màquina de la imatge següent, l'usuari introdueix els mots a transcriure en una mena de representació fonemàtica que pot consultar en línia, expressa una determinada combinació dels valors de les variables de la *PEOSFIEC*:

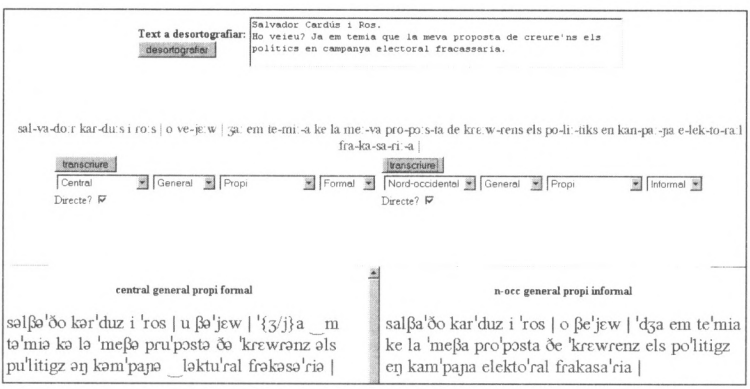


i obté els resultats següents, que corresponen respectivament al central i al valencià; primer ens mostra una taula amb els mnemònics de cada regla precedits d'un + o - que indica si aquella regla està activada o no en aquella combinació i després tenim cada mot en una línia que ens mostra el punt de partida, el d'arribada i la llista de regles que han intervingut en aquella transcripció, expressades amb llurs mnemònics.

Tota la interacció es fa usant un navegador de generació recent que suporti Unicode; entre la informació auxiliar l'usuari trobarà la que l'ajudarà, si cal, a instal·lar una tipografia fonètica gratuïta amb l'Alfabet Fonètic Internacional en Unicode.

Regles sintàctiques

A diferència de les regles lèxiques, les sintàctiques són majoritàriament alienes a la *PEOSFIEC*; ha calgut incorporar-les per donar uns resultats mínimament aprofitables, però necessiten encara força correccions. Aquest servei, a més, parteix de l'ortografia i no de la representació fonemàtica, cosa que el fa més accessible (i, ensems, vulnerable a qualsevol caràcter tipogràfic que no s'hagi previst). Amb tot, qualsevol investigador que hagi de fer avui una transcripció predictiva (per exemple, per a un manual), tindrà molta menys feina si usa la URL indicada tot seguit i després corregeix a mà els errors de l'ordinador que si la fa tota manualment http://retoc.uila.upf.edu/SOLC/.



Entre els formularis de consulta que oferim, n'hi ha algun que permet comparar transcripcions, com el de la pantalla precedent, que mostra, de dalt a baix, el text original (del diari *Avui*), la desortografiació i dues transcripcions.

La base de dades ortològica

Per tal que el model de regles ortològiques sigui solvent, cal depurar-lo i controlar-ne l'actuació cada vegada que es modifica una regla. La manera més àgil de fer això és tenir vocabulari controlat (mot de partida i resultat de l'aplicació de les regles) i que el propi ordinador monitoritzi després de cada canvi l'adequació dels resultats. Aquest vocabulari controlat és el punt de partida de la base de dades ortològica que, a mesura que anirà creixent, s'anirà convertint en una garantia que les regles produeixen realment el que pertoca. El projecte només ha creat un primer

nucli de vocabulari (de 333 paraules) que conté tota la casuística coberta fins avui per la *PEOSFIEC*. Sempre que ha estat possible, les transcripcions d'aquestes 333 paraules s'han confrontat amb els diccionaris publicats, amb la limitació que no n'hi ha cap que contingui més d'una modalitat (v.g. formal i informal). L'objectiu que ens hem fixat és que cada modalitat sigui revisada per estudiosos de l'àmbit corresponent.

Nuclis de vocabulari

Les etapes que estem preparats per cobrir (amb la selecció ja completada dels successius nuclis de vocabulari) es recullen a la taula següent:

1	333 mots crucials	(A)
2	2.000 mots del vocabulari bàsic	(A)
3	10.000 mots corrents	(B+C)
4	70.000 entrades del DIEC	(C)
5	Entrades enciclopèdiques catalanes	(C)
6	Lèxic i onomàstica forans	(D)

En cada etapa, el vocabulari cobert serà transcrit automàticament i la transcripció obtinguda serà revisada. Els mots revisats pels ortòlegs esdevenen el control de qualitat de les regles noves i de les modificacions de les antigues; tot el contingut del Servidor serà reprocessat cada cop que es modifiquin les regles fins a harmonitzar els mots revisats amb les regles. El codi de la taula anterior indica el procediment de revisió que pretenem fer:

A	revisió manual per ortòlegs de cada mot en cadascuna de les modalitats incloses
B	revisió aleatòria percentual per regles implicades: atès que el sistema de regles ens deixa constància de quines regles han intervingut en la transcripció de cada mot, es revisarà un percentatge dels mots modificats per cada regla; en el cas que una regla hagi actuat erròniament se'n revisarà un percentatge superior un cop corregida la regla. A mesura que s'avanci, la validació serà més àgil car si un mot és triat per a la validació de dues regles diferents només caldrà corregir-lo o confirmar-lo la primera vegada. Això passarà amb freqüència car es prioritzarà que els mots revisats per a cada regla cobreixin diferents interactuacions d'aquella regla amb totes les altres.
C	sota demanda dels professionals o per conflictes en el procés incremental: està previst que el material no supervisat no sigui accessible al públic en general sinó només als estudiosos i professionals de la llengua oral. Implica l'accessibilitat dels professionals de la llengua oral (UALs, estudis de doblatge, ràdios, formadors d'ortòlegs, de locutors i d'actors) als materials encara no revisats i que tinguin capacitat de deixar constància dels errors que han detectat
D	sota demanda de qualsevol usuari

Pel que fa als ortòlegs responsables de la revisió de la base de dades amb els nuclis de vocabulari, l'ideal fóra que comptessin amb alguna fórmula de suport

o supervisió per part de la Secció Filològica de l'IEC. Quan el projecte comptarà amb l'equip d'experts encarregats de la revisió, aquests hauran de definir el nivell de precisió o laxitud en la representació fonètica abans de començar; durant el procés, hauran de consensuar la representació fonemàtica única en cas de conflicte: és inherent a la concepció del Servidor Ortològic el fet que totes les regles han de treballar a partir d'una única representació d'entrada.

El públic en general haurà de poder accedir només als continguts revisats un cop assolits els primers 2000 mots; un cop tancats els primers 10000, l'accés es generalitzarà a tots els que no hagin requerit regles rares o combinacions infreqüents de regles

Revisió de regles sintàctiques

El SOLC ha de solucionar bàsicament problemes lèxics; les regles sintàctiques es podran revisar en la mesura en què s'utilitzin en projectes vinculats (com RETOC o DOPO); aquesta part està molt menys descrita en la *PEOSFIEC* i es realitza amb molta més variabilitat (pauses imprevistes, rectificacions, relaxacions...). Preveiem sol·licitar la col·laboració dels usuaris (que ens remetin transcripció original i còpia revisada, quan sigui possible, pot agilitar molt el procés de refinament de les regles sintàctiques, especialment per a les combinacions *PEOSFIEC* poc sol·licitades).

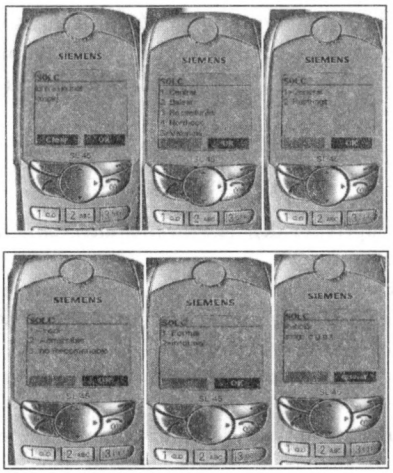
Beneficis col·laterals

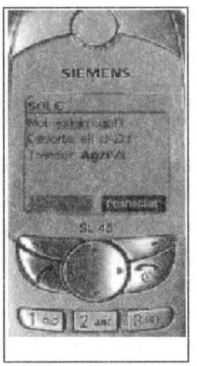
A més del fet que el conjunt del treball permetrà, un cop acabat, millorar la *PEOSFIEC*, el sistema de regles es podrà aplicar al corpus RETOC (i a qualsevol projecte no comercial), a l'àmbit de la fonètica forense, al de la tecnologia de la veu, a la sistematització de les representacions fonemàtiques comunes, i la recopilació de la casuística no regularitzable alimentaran la recerca fonològica i els manuals de llengua.

Opcions futures

En primer lloc, cal parlar de millores, com ho seria (per a les regles sintàctiques) el fet d'entroncar amb analitzadors morfològic i sintàctic amb desambiguació (per tractar ambigüitats com: *coure, circular, poder, fer, tens, sou, son, les, valent, calent...*;

cent homes vs. *cent o més*) o com incorporar regles exòfones per als exònims («eu» alemany = «oi») o compilar les representacions fonemàtiques que permetin aplicar les regles endòfones (s'ha de dir «fròjdià» sempre?)





En segon lloc, i per tal d'explorar la complexitat tècnica d'algunes de les vies d'obtenir valor afegit del SOLC, hem desenvolupat un sistema de consulta WAP, en protocol WML, que possibilita la consulta d'un mot, amb pantalles successives que permeten concretar els valors de les variables *PEOSFIEC* i que donen com a resultat l'aplicació de les regles *PEOSFIEC* a partir de l'ortografia. Tot i que el protocol WML integra Unicode, no hem trobat cap telèfon mòbil que incorpori la secció AFI de l'Unicode, de manera que hem optat per representar la transcripció pseudofonèticament. Aquest sistema de consulta WAP es podria posar en funcionament si hi havia alguna institució o espònsor interessat.

Conclusions

Tant pel nivell de coneixença de la llengua entre els professionals de la llengua oral com per la riquesa de possibilitats i relativa novetat de la *PEOSFIEC*, l'àmbit professional de la llengua oral necessita un servei com el que hem dissenyat i implementat. Ara cal generalitzar-ne els beneficis implicant-hi estudiosos, professionals

i empreses de tot el domini lingüístic. La feina que ja està feta es podria oferir al públic en general, però amb un marge d'errors davant dels quals els no iniciats probablement no sabrien protegir-se, de manera que sembla prudent restringir-ne de moment la difusió entre els professionals i els estudiosos.

FRANCESCA SALVÀ, LLUÍS DE YZAGUIRRE, MARIA TERESA CABRÉ
Laboratori de Tecnologies Lingüístiques, Institut de Lingüística Aplicada,
Universitat Pompeu Fabra

REFERÈNCIES BIBLIOGRÀFIQUES

BADIA 1990: Antoni Maria Badia i Margarit (supervisor de), *Vox essencial diccionario castellà-català, català-castellà*, Barcelona, Bibliograf.

CABRÉ *et alii* 1999: Maria Teresa Cabré, Lluís De Yzaguirre i Esteve Clua, «Diccionari ortològic català», a *Llengua i mitjans de comunicació*, ed. Imma Creus, Joan Julià i Sílvia Romero, Lleida, Pagès.

CAMPS *et alii* 2004: Oriol Camps, Lluís de Yzaguirre i Francesca Salvà, «Diagnòstic ortològic assistit», a *Actes AILLC 13*, II, **85-91**.

DE YZAGUIRRE 1990: Lluís de Yzaguirre, «L'estructura sil·làbica del català central», tesi doctoral, Universitat de Barcelona.

DE YZAGUIRRE *et alii* 2004: Lluís de Yzaguirre, Antoni Jaume Farriols i Jaume Martí, «El corpus RETOC: Un corpus oral per a la recerca i la docència», a *Actes AILLC 13*, II, **495-504**.

PEOSFIEC: IEC, Secció Filològica, *Proposta per a un estàndard oral de la llengua catalana I. Fascicle de fonètica*, Barcelona, IEC, 1990.

VALLVERDÚ 1986: Francesc Vallverdú, *Elocució i ortologia catalanes*, Barcelona, Jonc.

Enllaços

PEOSFIEC (versió treball)	http://retoc.iula.upf.edu/docs/ortol/*PEOSFIEC*.htm
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