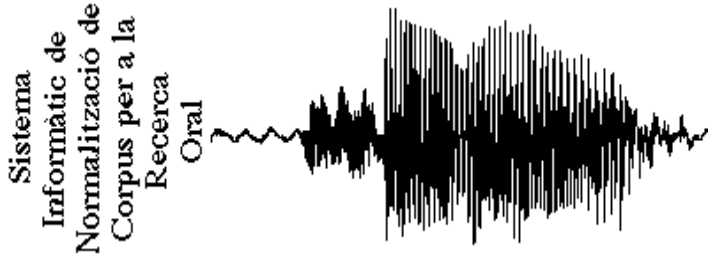
Institut Universitari de Lingüística Aplicada
Laboratori de Tecnologies Lingüístiques
Universitat **Pompeu Fabra**
iulalatel@upf.edu
935-421-344

Instruccions per transcriure ortogràficament des del SINCRO



Barcelona, 1 de febrer del 2007

Índex de continguts

1. Presentació

Aquest manual esta pensat per aquelles persones que hagin de transcriure ortogràficament, i sense sincronització, els segments d'una o més mostres i, per tant, el seu objectiu és donar les pautes necessàries per portar a terme aquesta tasca: es tracta d'escoltar cada segment, reproduir-lo ortogràficament en el quadre de diàleg que es visualitza a la part inferior esquerra de la pantalla i, eventualment, codificar el text en la mesura que es trobin incidències destacables (vegeu l'apartat 7**. Codificacions**) segons el criteri de la persona que dirigeixi la vostra feina.

A continuació donarem una sèrie de pautes per accedir als segments que cal transcriure ortogràficament: com se selecciona una mostra, què cal fer per escoltar-la, on s'ha d'introduir el text, com es pot corregir i què cal fer per validar-lo.

2. Què és el SINCRO?

El sistema SINCRO, desenvolupat a l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, ofereix un procediment per generar corpus orals sincronitzats. La sincronització es pot fer a partir de la grafia, del so o bé amb tècniques mixtes, però partim de la suposició que tenim ambdues coses: el text escrit i el text oral. L'objectiu és arribar a les unitats de sincronització més petites possible amb una intervenció humana mínima.

Les unitats bàsiques de subdivisió del corpus per a la sincronització seran diferents si partim de la grafia (frases o paraules) o del so (segments entre silencis). Totes dues haurien de coincidir amb una certa facilitat si hi hagués una correspondència total entre el text escrit i l'oral. Per raons d'economia productiva, el sistema SINCRO s'aplica als textos escrits previs a la gravació, és a dir, als textos que suposadament haurien d'haver llegit els locutors. La principal característica d'aquest procés és que les subdivisions naturals de l'oral i les de l'escrit no acostumen a coincidir gairebé mai. Quan no es disposa d'un text previ, hi ha menys facilitats però tot i així no deixa de ser més productiu que treballant amb sistemes de pedals o altres andròmines *infolítiques*.

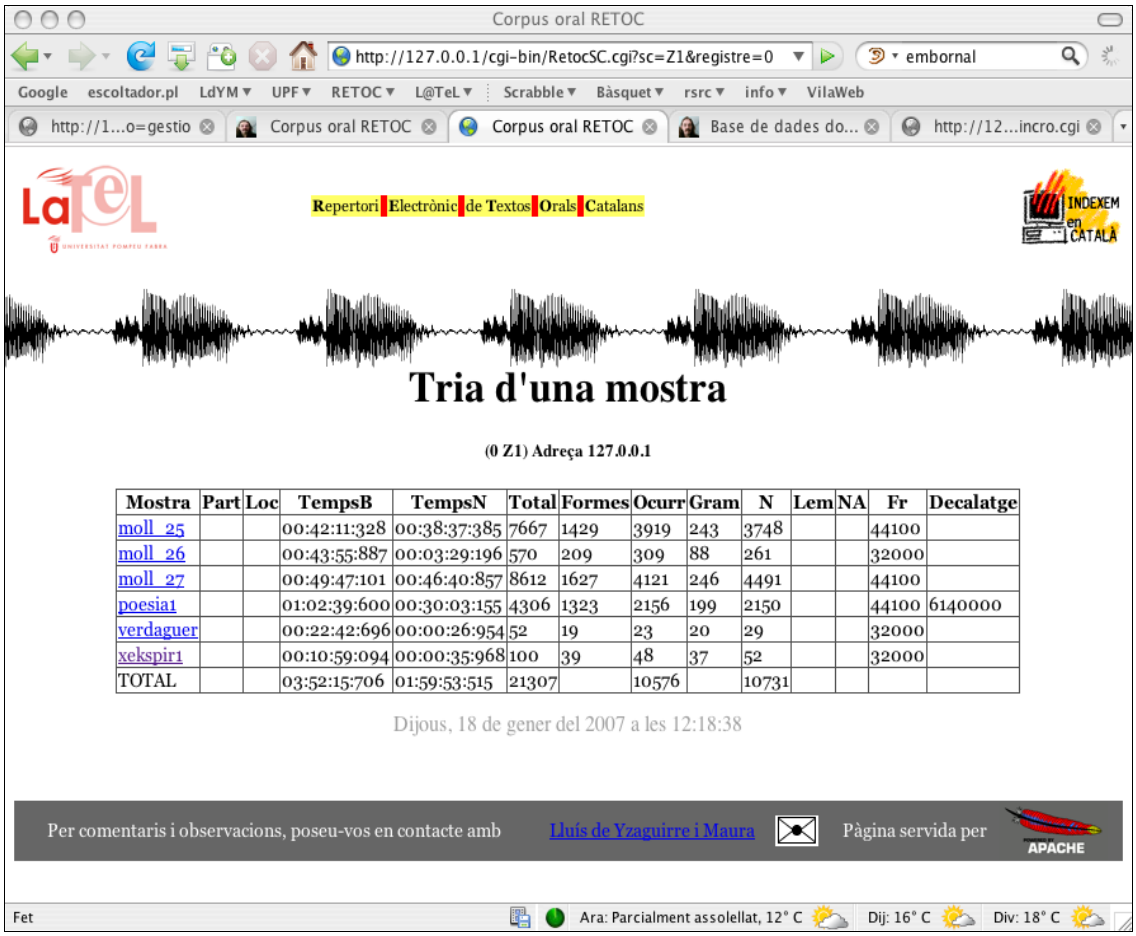
Per raons d'economia, la transcripció és ortogràfica; això no exclou que per a una determinada recerca s'incloguin marques que ajudin a tipificar incidències relacionades amb, per exemple, l'entonació, la prosòdia, l'ortologia o fins i tot qüestions pragmàtiques.

3. Procediment per seleccionar una mostra per transcriure-la ortogràficament (via lenta)

a) Accediu a la pàgina http://retoc.iula.upf.edu/html/ des d'Internet.

b) Trieu l'opció *Menú de SubCorpus* de la pàgina principal.

c) Seleccioneu un subcorpus com, per exemple, *Literari* i obtindreu el resultat que es mostra en la figura 1:



Mostra	Part	Loc	TempsB	TempsN	Total	Formes	Ocurr	Gram	N	Lem	NA	Fr	Decalatge
moll_25			00:42:11:328	00:38:37:385	7667	1429	3919	243	3748			44100	
moll_26			00:43:55:887	00:03:29:196	570	209	309	88	261			32000	
moll_27			00:49:47:101	00:46:40:857	8612	1627	4121	246	4491			44100	
poesia1			01:02:39:600	00:30:03:155	4306	1323	2156	199	2150			44100	6140000
verdaguer			00:22:42:696	00:00:26:954	52	19	23	20	29			32000	
xekspir1			00:10:59:094	00:00:35:968	100	39	48	37	52			32000	
TOTAL			03:52:15:706	01:59:53:515	21307		10576		10731				

Figura 1. Conjunt de mostres d'un subcorpus

Quan seleccioneu la mostra concreta, per exemple *xekspir1*, es visualitza una pantalla repartida en tres parts. La part superior esquerra té el menú que reproduïm en la figura 2. Si feu clic en l'última opció d'aquesta llista –*Transcriure*–, accedireu al SINCRO un cop us hàgiu autentificat escrivint el vostre codi i la contrasenya (sobre aquest tema vegeu l'annex que s'adjunta al final d'aquest document).

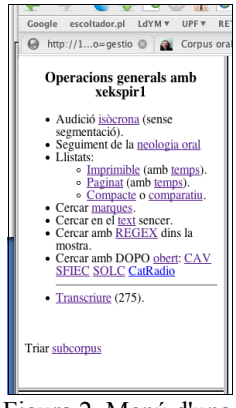


Figura 2. Menú d'una mostra

4. Procediment per seleccionar una mostra per transcriure-la ortogràficament (via ràpida)

Si sabeu el número de mostra (que, per la via lenta s'indica dins d'un parèntesi al costat de l'opció *Transcriure*) podeu accedir directament al SINCRO des de l'enllaç següent: http://retoc.iula.upf.edu/CGIs/sincro.cgi?operacio=previ&numMostra=275, bo i canviant la darrera xifra per la de la mostra a què voleu accedir. D'aquesta manera accedireu directament a la pàgina de validació d'accés de les aplicacions del LATEL i, un cop introduïu el vostre codi (o el nom i el cognom) i la vostra contrasenya, accedireu a la mostra que heu triat abans per transcriure-la ortogràficament.

Figura 3. Pàgina de validació d'accés

5. Operacions bàsiques

A continuació es mostra la pantalla de treball del SINCRO i les diferents operacions bàsiques que es poden fer per un segment (cada pantalla conté una llista amb set segments), representades amb icones.

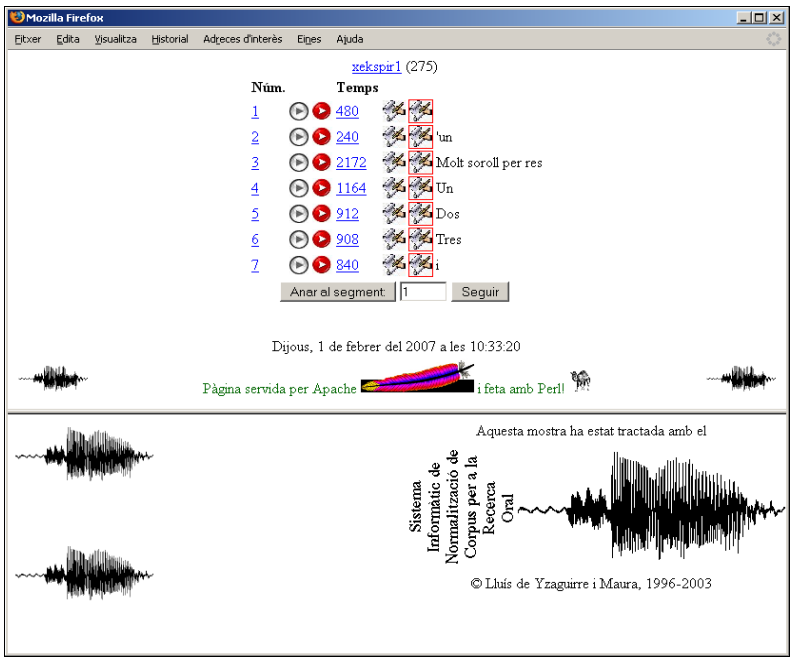


Figura 4. Pantalla de treball del SINCRO

El significat de les diferents icones que conté la pantalla de treball és el següent:

Permet escoltar el segment.

Permet escoltar un segment i el següent (aquesta opció és perillosa en la mesura que dos segments aparentment contigus poden tenir intercalats un lapse de música o altres realitzacions que s'hagin negligit o que el programa no hagi sabut segmentar; això podria crear problemes de memòria a l'ordinador si el segment resultant fos massa llarg...).

Dóna pas al quadre de diàleg en què s'ha de transcriure ortogràficament el segment bo i escoltant el segment.

Permet corregir una transcripció ortogràfica ja feta sense necessitat d'escoltar-la (és la mateixa icona que la de la transcripció, però sobre fons roig). Al costat de la icona es visualitzarà la transcripció ortogràfica d'aquest segment un cop introduït el text, i després de fer clic en el botó *Entrar el segment x.*

Quan heu escoltat un segment, n'heu transcrit el text dins del quadre de diàleg de la part inferior dreta de la pantalla i ja estigueu segurs que la transcripció ortogràfica és correcta, feu clic en el botó que hi ha a sobre d'aquest quadre de diàleg: *Entrar el segment x* per validar el text i fer possible que es visualitzi[1] al costat de la icona  , que permet corregir el text transcrit ortogràficament.

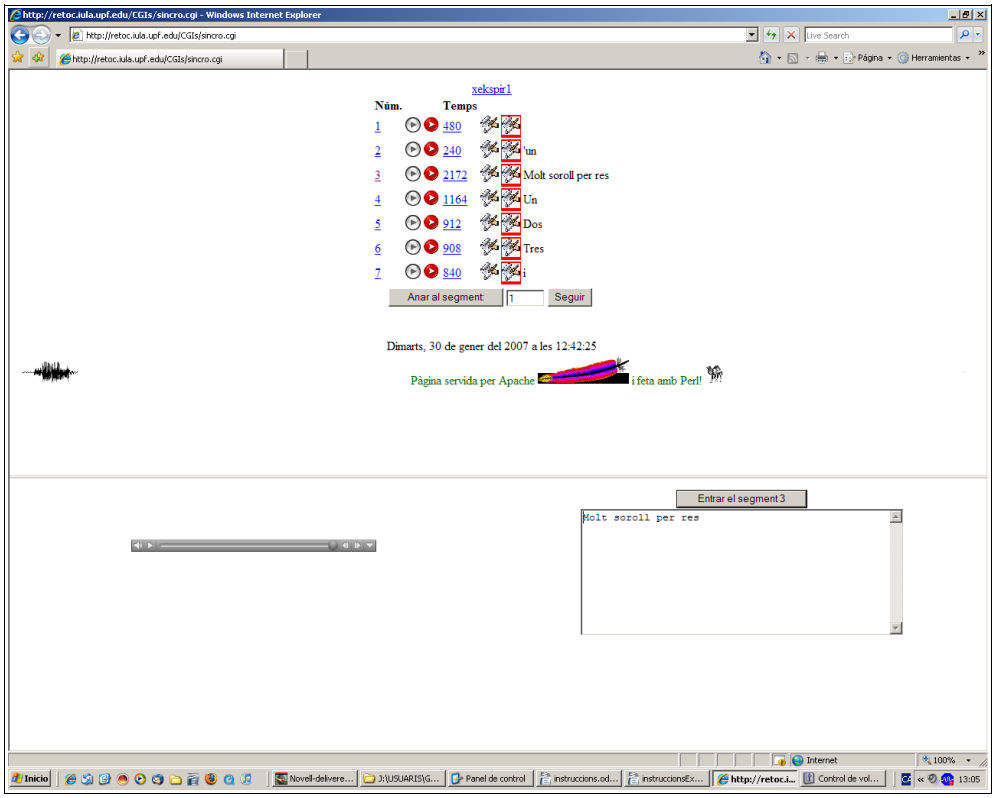


Figura 5. Pantalla de treball del SINCRO amb el quadre de diàleg on cal reproduir el text que heu escoltat

Feu servir el quadre de diàleg *Anar al segment*[2] per indicar el segment amb què voleu treballar, i el botó *Seguir* per accedir a la pantalla amb els set segments següents.

1 Aquesta visualització no és automàtica per no alentir el procés; si voleu tenir confirmació de com ha quedat, utilitzeu el quadre de diàleg *Anar al segment*.

2 Quan vulgueu tancar una sessió, convé que us anoteu el número de segment en què esteu treballant, d'aquesta manera hi podreu accedir directament si l"escriviu en el quadre de diàleg *Anar al segment* en el moment que us convingui continuar la sessió de treball.

6. Problemes més freqüents

A continuació es plantegen els problemes més usuals amb què us podeu trobar quan escolteu un segment i el vulgueu transcriure, i com els heu de resoldre.

6.1. Problemes amb el *plugin*

Els usuaris de Mozilla o Firefox o Netscape trobareu informació sobre el tema a:

http://plugindoc.mozdev.org/faqs/firefox-windows.html#install-quicktime

How do I install QuickTime? If you have QuickTime 5.0 or later installed, Mozilla Firefox will automatically use the plugin. You do not need to do anything. If you install QuickTime after Mozilla Firefox, the QuickTime installer will automatically install the plugin for Mozilla Firefox. [Download QuickTime for Windows -> http://www.apple.com/quicktime/download/win.html]

6.2. Salt entre dos segments (falta algun fragment d'audició)

El fet que manqui un fragment de veu entre dos segments es pot confirmar si es fa clic en el número que indica el temps de cada segment, ja que permet visualitzar el diàleg temporal següent:

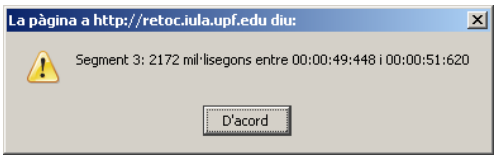


Figura 6. Diàleg temporal que evidencia l'interval del segment que manca

Tot i així, no cal fer-hi res perquè entre les operacions de manteniment està tipificada la de detectar automàticament aquesta mena de situacions, que solen tenir el seu origen en la dificultat de segmentar: soroll de fons, baixa qualitat (comunicació telefònica), encavalcament dels diversos interlocutors en un debat...

6.3. Segments que comencen o acaben amb un mot a mitges

Quan es doni aquesta circumstància, inicieu el segment amb punts suspensius (si el mot incomplet és al principi) o bé poseu-los al final (quan es tracta de l'últim mot).

6.4. Segments massa llargs

Es pot donar el cas que el segment que s'ha de transcriure ortogràficament sigui més llarg del que el programa té previst. Si és així, en validar el text amb el botó *Entrar el segment*, la pantalla es divideix en dues parts. En la part inferior –de color salmó– es pot veure el segment sencer que heu transcrit i a sota –precedit del signe +– com s'ha reduït el text (noteu que en la captura de pantalla següent s'ha marcat manualment la part del segment que ha quedat tallada[3]).

3 Convé crear un document de text que contingui aquestes parts dels segments que han quedat tallades, i que estiguin identificades pel número de segment, per tal de fer-lo arribar al LATEL un cop s'acabi la transcripció d'una mostra. Així el vostre esforç no haurà estat en va.

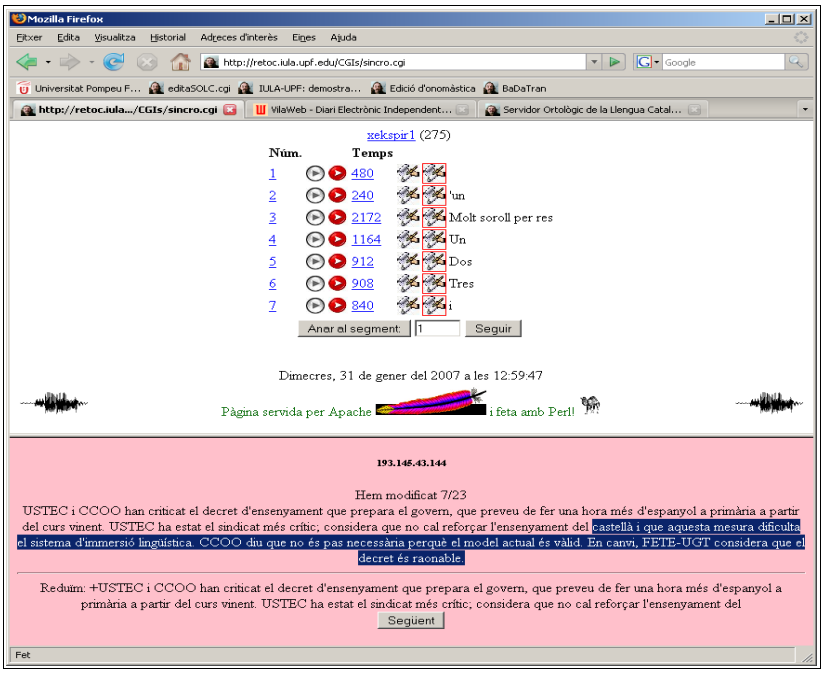


Figura 7. Pantalla que notifica que el segment és massa llarg

6.5 Segments amb un breu silenci, soroll o música

Quan el contingut d'un segment és un silenci, un soroll o música, indiqueu-ho amb la paraula 'silenci', 'soroll o 'música' continguda en un parèntesi. Tant en aquest cas com en el subapartat següent (6.6. Segment en una altra llengua), si tot un bloc de segments consecutius comparteixen la mateixa indicació, podeu limitar-vos a posar-ho en el primer i el darrer segment, deixant els intermedis en blanc.

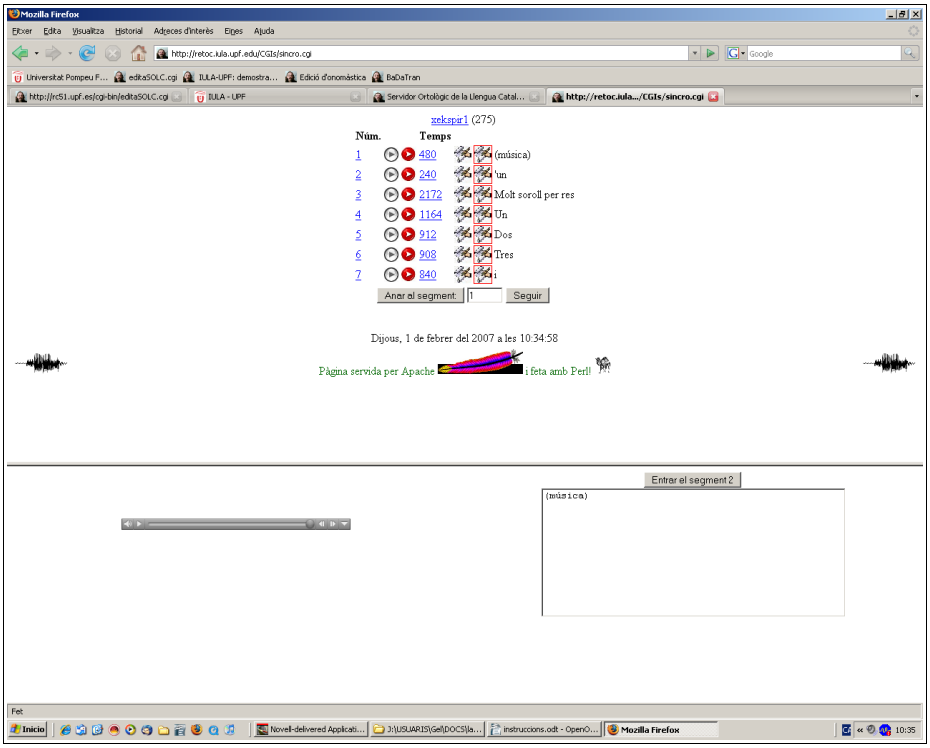


Figura 8. Exemple de marcatge d'un segment amb música

6.6. Segment en una altra llengua

De vegades us podeu trobar que una mostra contingui algun segment en una altra llengua diferent del català. En aquests casos indiqueu entre parèntesis de quina llengua es tracta.

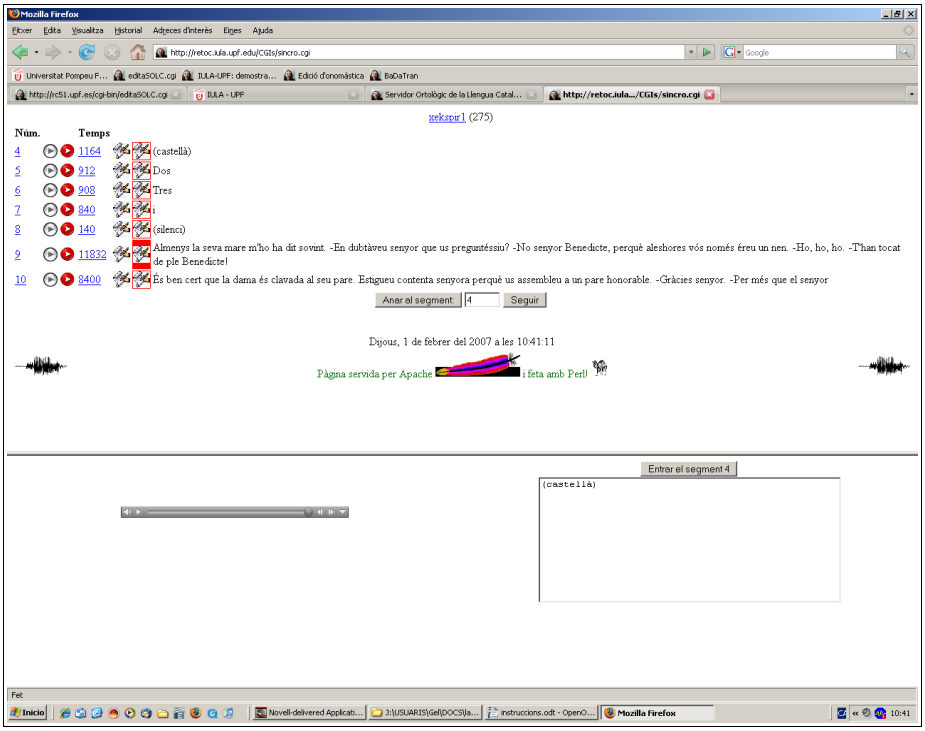


Figura 9. Exemple de marcatge d'un segment en castellà.

7. Codificacions

Els corpus orals són molt costosos de produir. Per aquest motiu cal donar les màximes facilitats per tal que els diferents equips de recerca amb necessitats de marcatge diferents puguin compatibilitzar estratègies de marcatge. En versions futures d'aquest manual podria ser que aquesta secció inclogui informacions adreçades a col·laboradors d'altres projectes, en la mesura en què estem oberts a qualsevol col·laboració en aquest camp. De moment, però, només hi trobareu les pautes que hem fet servir per indicar la presència de fenòmens significatius des d'una perspectiva ortològica.

8. Marcatge d'incidències ortològiques sobre la transcripció ortogràfica

Quan escolteu els segments sovint trobareu necessari fer algun comentari sobre mots conflictius, sigui perquè hi ha alguna cosa en la seva pronúncia que us resulta estranya, però no sabeu ben bé de què es tracta, sigui perquè voleu marcar que el conflicte es planteja explícitament en la *Proposta d'estàndard oral* de l'IEC. En tots els casos inseriu un asterisc (*) a l'esquerra del mot conflictiu i seguiu la codificació que es proposa a continuació per marcar el tipus de problema:

Codi	Tipus de problema	Exemple
* ?	Quan el mot és conflictiu, però no sabeu què dir	
* núm. PEOSFIEC	Quan la incidència es planteja explícitament a la PEOSFIEC (el número que heu d'indicar és el de l'apartat de la *Proposta* que en parla.	frase*18 (quan aquest mot s'ha pronunciat amb 'e' tancada).
*b	barbarisme que cal diferenciar d'un mot genuí	pues*b, meno*b
*mXX	ha pronunciat emfàticament la *essa* d'*aquest*m19*: la ema fa referència al quadern de morfologia i la xifra és la pàgina (tenim pendent fer-ne una versió més operativa)	

Codificació amb claus {esperat/observat}

Un altre tipus d'informació que cal afegir quan escriviu el text que heu escoltat és aquell que té a veure amb una pronúncia diferent a la prevista. Per marcar aquesta incidència, indiqueu entre claus –i separats per una barra inclinada– primer el que esperàveu sentir i després, el que heu escoltat (per exemple, la conjunció 'però' pronunciada sense 'e' es marcaria així: {e/}. Aquesta codificació permet expressar ortogràficament les diferències entre allò que esperàvem escoltar i el que efectivament trobem (sigui perquè tenim un guió escrit del que s'ha llegit, sigui perquè s'ha usat una forma col·loquial allunyada de l'estàndard). La importància d'aquesta codificació rau en l'eficàcia que ens proporciona a l'hora de cercar (doncs, l'estàndard aquí no pretén vehicular cap preferència per la llengua més artificiosa i elaborada sinó fornir un instrument d'equiparació entre variants formals que ens permeti indexar plegades les diverses variacions d'un mot, com naltros, nantros, natros, mosatros, nosatres...). Cal tenir clar que sempre que la PEOSFIEC preveu dues pronúncies vàlides no cal marcar-ho (jo com a ʒɔ́ | jɔ́) . També és important no marcar més del que és estrictament necessari, car es perdria la capacitat de generalitzar: per exemple, si codifiquem "brenar" i "vritat" així {berenar/brenar} i {veritat/vritat} ens apareixeran indexats com si no tinguessin res en comú; en canvi, si codifiquem b{e/}renar i v{e/}ritat, aquestes dues situacions totalment comparables ens apareixeran indexades sota la marca {e/} que indica que en la realització observada manca una e (o neutra) que esperàvem.

representació d'allò esperat en relació a a llò observat	primer esperat, després observat
ha pronunciat sons, síl·labes o mots que no s'esperaven	p{/er }els, on{/t}
ha omès sons, síl·labes o mots que s'esperaven	m{e/}itat, tipu{s/}
ha dit una cosa en comptes d'una altra	{la mort/l'amor}
ha monoftongat quaranta o quarantena	{qua/co}ranta, {qua/co}rantena

Annex. Codi d'accés i contrasenya

Per accedir a SINCRO us heu d'identificar escrivint un codi i una contrasenya en la pantalla de validació d'accés de les aplicacions del LATEL (Laboratori de Tecnologies Lingüístiques)[4]

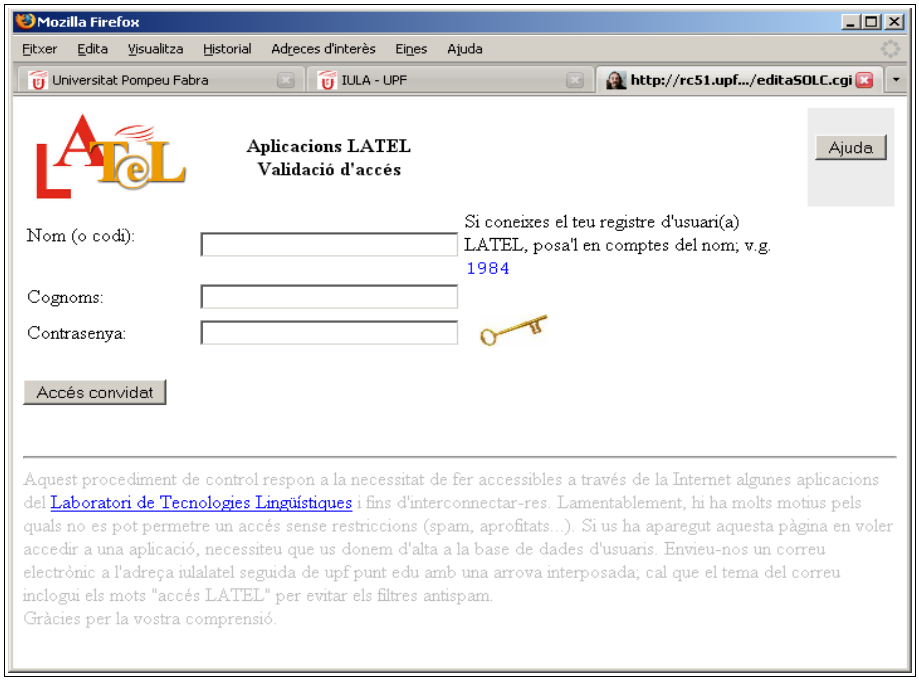


Figura 1. Pantalla d'accés a les aplicacions LATEL

La pantalla d'accés té tres camps: **Nom (o codi de l'usuari),** on cal escriure el vostre nom[5]; **Cognoms,** on cal escriure el vostre primer cognom, i **Contrasenya**, on la primera vegada que es vol accedir al SINCRO cal escriure una contrasenya que us donarem des del LATEL i que després cal canviar per una de personal.

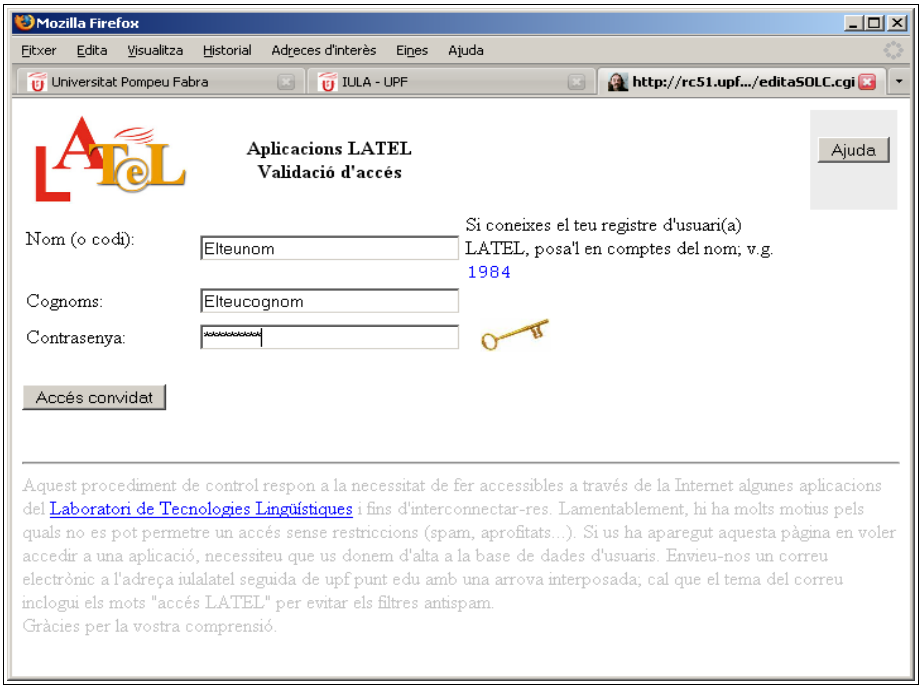


Figura 2. Pantalla d'accés a les aplicacions LATEL amb la informació necessària per a poder-hi accedir (nom, cognoms i contrasenya)

4 Aquest codi no serveix per totes les aplicacions del LATEL; només per aquelles que l'administrador ha cregut necessàries per portar a terme les vostres tasques.

5 Si en comptes del vostre nom escriviu el codi (que us donarà l'administrador), no farà falta que empleneu el camp **Cognoms**.

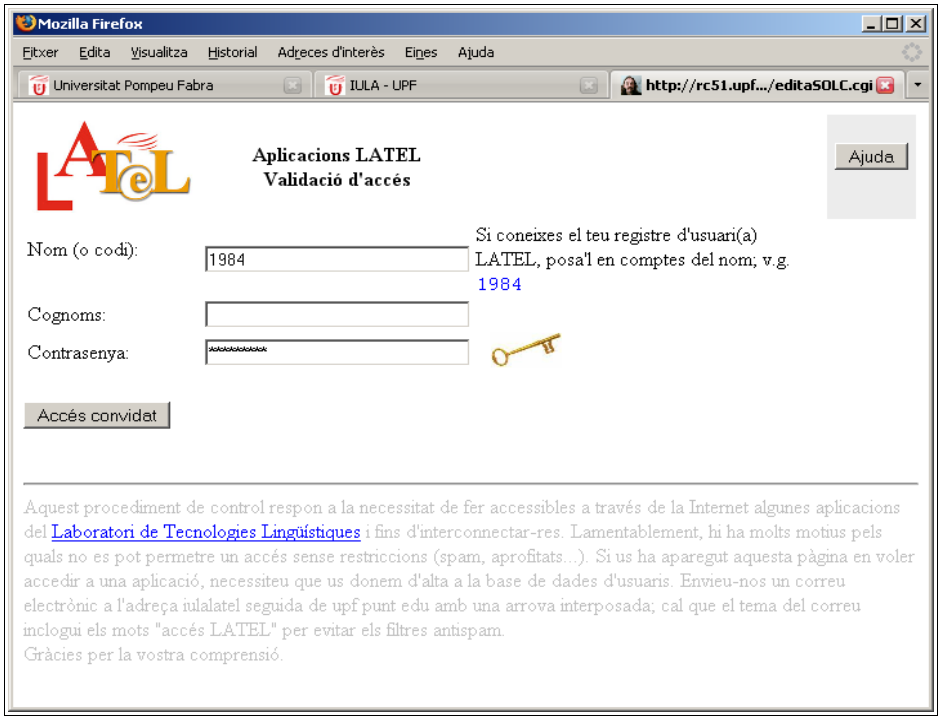


Figura 3. Pantalla d'accés a les aplicacions LATEL amb la informació necessària per a poder-hi accedir (codi i contrasenya)

Un cop introduïda la contrasenya, cal fer clic en la icona de la clau per accedir al SINCRO. Si és la primera vegada que es fa aquesta operació, es visualitza una pantalla com la següent que permet canviar la contrasenya proporcionada per l'administrador per una de més personal:

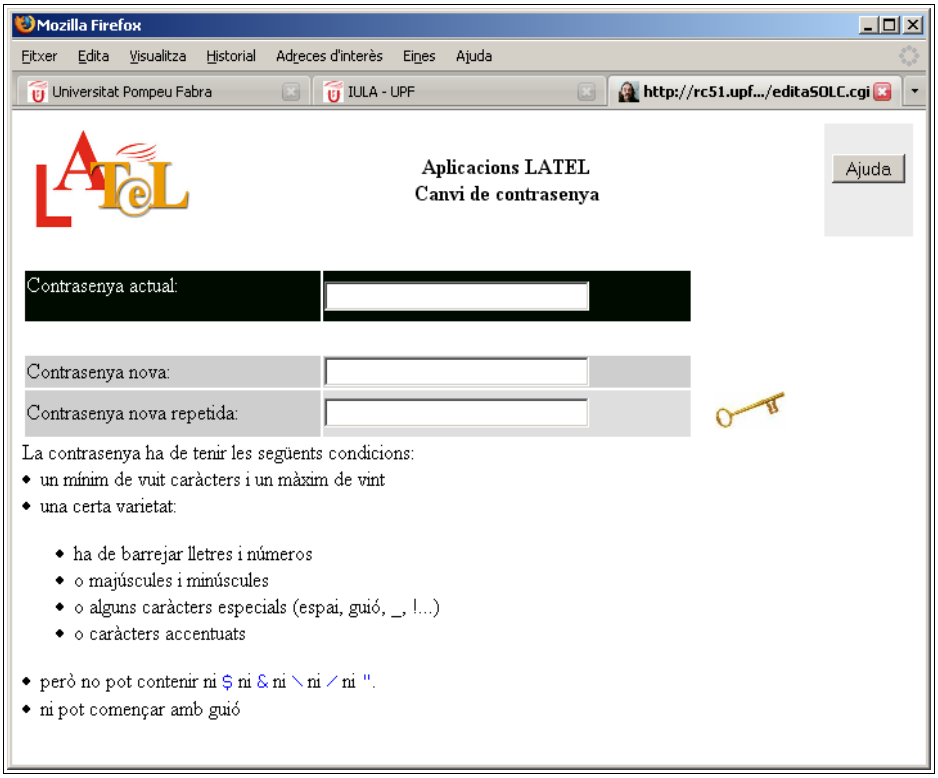


Figura 4. Pantalla per canviar la contrasenya

En cas que la contrasenya no s'adeqüés als requisits que es demanen tornarà a sortir la mateixa pantalla per tal de poder repetir l'operació. Si la contrasenya és correcta, es visualitza la pantalla principal del SINCRO[6].

6 Aquest procediment només cal seguir-lo el primer cop que s'accedeix a l'aplicació i cada vegada que oblideu la contrasenya i el LATEL l'hagi de reinicialitzar)