Generació de recursos lingüístics per al desenvolupament de tecnologies de la parla en català

La generació de recursos lingüístics per al desenvolupament de tecnologies de la parla en català és el resultat d'un conveni signat per:

· *Departament d'Universitats, Recerca i Societat de la Informació, Generalitat de Catalunya*
· *Universitat Politècnica de Catalunya.*

Desenvolupat per:

· *Centre de Tecnologies i Aplicacions del Llenguatge I la Parla (TALP).*

Directora del projecte:

· *Dra. Asunción Moreno*

Generació de recursos lingüístics per al desenvolupament de tecnologies de la parla en català

El centre TALP de la Universitat Politècnica de Catalunya (UPC) i l'empresa *Applied Technologies on Language and Speech* (ATLAS) estan desenvolupant el projecte *Generació de recursos lingüístics per al desenvolupament de tecnologies de la parla en català*.

L'objectiu és fer disponibles per a empreses i centres d'investigació els recursos lingüístics necessaris per poder:

1. Posar el *català* al nivell d'altres llengües europees pel que fa a disponibilitat, accessibilitat i qualitat de recursos lingüístics.

2. Adaptar les tecnologies desenvolupades en altres llengües al *català.*

3. Atraure empreses tecnològiques per desenvolupar productes en *català.*

4. Fer accessible a la comunitat científica eines i recursos lingüístics.

5. Promoure la recerca en l'àmbit del processament de la parla i el llenguatge en *català*.

El projecte està centrat en la generació del Recursos Lingüístics necessaris per al reconeixement automàtic de la parla. Per aquest fi, es requereixen grans bases de dades orals per entrenar els sistemes, és a dir, perquè aprenguin diferents formes de pronunciació. Amb la tecnologia actual, les bases de dades consisteixen en els enregistraments d'un gran nombre de persones que llegeixen textos o frases curtes en diferents entorns. Posteriorment es fa una transcripció ortogràfica del que realment ha pronunciat el locutor i una transcripció fonètica. La tecnologia actual és molt sensible al canvi d'entorn, soroll i micròfons i, per aquesta raó, és necessari tenir molts enregistraments en un entorn concret per assegurar que el sistema de reconeixement pot treballar

adequadament en aquest entorn (per exemple, enregistraments en telefonia fixa no serveixen per reconèixer veu en un entorn d'oficina).

Amb aquest projecte es pretén assolir un nivell d'igualtat amb les bases de dades orals que existeixen actualment en la majoria dels idiomes parlats a Europa per aplicacions industrials. El projecte segueix els estàndards europeus en la generació de recursos lingüístics i finalitzarà durant l'any 2006.

Els recursos que es generaran són:

• Telefonia fixa i mòbil: S'enregistrarà a 4000 persones amb àmplia diversitat en dialectes, edats i entorns d'enregistrament. Cada persona llegeix uns 15 minuts de textos amb diversa informació. Aplicacions típiques d'aquestes dades són accés a centraletes telefòniques, punts d'informació telefònica, reserva d'hotels, trens, viatges, reserva d'entrades de cinema...

• Consum: Unes 550 persones llegeixen uns 60 minuts cadascuna amb un sistema equipat amb micròfons des de diferents entorns: casa, oficina, llocs públics. Les aplicacions d'aquestes dades són quioscs d'informació, control d'eines al treball (típicament ordinadors) i a casa, ajuda a gent gran, serveis de transcripció d'informes....

• Cotxes: al voltant de 300 persones llegeixen uns 60 minuts cadascuna dins de cotxes amb diferents situacions de trànsit i velocitat. Les aplicacions es situen en el comandament de sistemes de GPS, control dels dispositius del cotxe i, com que el missatge reconegut es pot transmetre a qualsevol lloc mitjançant un telèfon mòbil, també s'inclourien aplicacions tals com control a distància de fax, telèfon,... de qualsevol dispositiu que pugui ser operat per control i les típiques aplicacions de telefonia.

Totes les bases de dades s'enregistraran segons les especificacions generades en projectes europeus previs. Les especificacions inclouen, per a cada base de dades, el criteris per al disseny dels corpus, les especificacions de les plataformes d'enregistrament, la distribució dels informants per dialecte, sexe, grups d'edat i ambient d'enregistrament, la normativa per a la transcripció d'allò que realment s'ha pronunciat, la documentació de les bases de dades finals i els criteris per a la validació o homologació de les bases creades. Tota aquesta informació és accessible a la pàgina web del projecte

http://gps-tsc.upc.es/veu/projects/BDG

La validació de cadascuna de les bases de dades creades la farà un centre extern de validació.

A continuació es descriu breument el contingut i els criteris de distribució d'informants per a cada base de dades. Posteriorment es fa un resum de les plataformes d'enregistrament utilitzades en el projecte, els formats dels fitxers i els criteris de transcripció. Finalment, l'últim apartat es dedica a l'accessibilitat de les dades.

Definició del corpus i entorns d'enregistrament de les bases de dades de telefonia fixa i mòbil

Aquesta base de dades contindrà la veu de 4000 persones, la meitat dones i la meitat homes. Els enregistraments es fan mitjançant una interfície telefònica RDSI a 8KHz de freqüència de mostreig, 8 bits per mostra i codificada amb llei A.

El corpus està dissenyat per donar suport a la creació de teleserveis comandats per veu. Els informants (les persones a qui s'enregistra la veu) pronuncien 40 textos curts, que comprenen dígits aïllats i connectats, números naturals, quantitats de diners, lectura lletra per lletra, frases de dia i hora, frases de confirmació/rebuig, noms, cognoms, ciutats, empreses, paraules comunes d'aplicació, paraules d'aplicació inserides en frases i frases fonèticament riques. La majoria de les locucions són textos llegits i d'altres són respostes espontànies. Les bases de dades es lliuraran amb documentació extensa i estandarditzada. La veu es transcriu a nivell ortogràfic, i s'hi anoten també un seguit d'esdeveniments clarament audibles (sorolls, respiració,...). L'edat i la regió dialectal de procedència dels informants també queden reflectides en la base de dades. La documentació inclou un diccionari de pronunciació que conté totes les paraules aparegudes en el corpus amb la corresponent transcripció fonètica amb SAMPA. El fitxers de dades tenen el format SAM.

Entorns d'enregistrement

El conjunt d'entorns des d'on es fan les trucades és: casa, oficina, vehicle, lloc públic i cotxe utilitzant un sistema mans lliures. Es preveu la següent distribució d'informants en funció del tipus de telèfon, regió d'enregistrament i entorn:

Telèfon	Entorn	Distribució a la base de dades total	Distribució per regió d'enregistrament
Fixa (2000)	Casa/Oficina	100%	Entre 100 i 600 per regió
Mòbil (2000)	Vehicle	20% ± 5%	>20%
	Lloc Públic	25% ± 5%	
	Carrer	25% ± 5%	
	Casa/Oficina	25% ± 5%	>20%
	Kit de cotxe	5% ± 1%	Sense restricció

Taula 1. Distribució d'informants en les bases de dades de telefonia en funció del tipus de telèfon, regió de gravació i entorn

Definició del corpus i entorns d'enregistrament de la base de dades per aplicacions de consum

La base de dades contindrà la veu de 550 persones, cadascuna enregistrada en 1 sessió, de les quals aproximadament la meitat seran dones i la meitat homes. Una sessió consisteix en unes 291 locucions llegides i en un màxim de 30 més de parla espontània enregistrades amb 4 micròfons mitjançant una plataforma mòbil. En una sessió s'enregistra la següent informació:

• Informació de calibratge: consisteix en unes mesures acústiques del lloc on s'efectua l'enregistrament.

• Locucions de parla espontània lliure: 5 min de parla espontània lliure amb locucions de context ric (explicar una història).

• Locucions espontànies curtes: consisteixen en dates, hores, noms propis, noms de ciutats, seqüències de lletres, respostes de preguntes, números de telèfon, idioma.

• Paraules (llegides) bàsiques: consisteixen en 31 paraules i frases generals, com dígits aïllats, seqüència de dígits aïllats, seqüència de dígits connectats, número de telèfon, nombres naturals, quantitat de diners, frases de temps analògic i digital, data analògica, data relativa i digital, seqüències de lletres, noms propis, noms de ciutat o de carrer, preguntes amb resposta sí o no, caràcters especials del teclat, adreces web i adreces d'email. També s'inclouen 208 paraules d' aplicacions i frases específiques per comandament de dispositius com comandes bàsiques IVR, paraules per navegació, edició, control de sortida, missatges i navegació per internet, funcions de l'organitzador, encaminament, automoció, àudio i vídeo.

Condicions d'enregistrament

S'han definit 4 ambients:

• Oficina: Una oficina, és a dir, una habitació on la gent treballa amb escriptoris, normalment o probablement amb un ordinador. No haurien de tenir lloc reunions a l'oficina durant els enregistraments. Es faran 200 enregistraments des d'aquest entorn.

• Entreteniment (ambient domèstic): Sala d'estar, és a dir, una habitació amb alguns mobles i llocs on la gent pot seure. Hi hauria d'haver alguns mobles, una taula, un televisor o algun equip de so. També seria possible fer l'enregistrament en una habitació d'hotel. Es faran 75 enregistraments des d'aquest entorn.

• Cotxe: Vehicle per 4 o 5 passatgers. Es faran 75 enregistraments des d'aquest entorn.

• Lloc públic: Un vestíbul gran o espai obert. El vestíbul hauria de tenir almenys 3 parets i un sostre; amb gent més o menys ocupada però no massa silenciosa. Un espai obert no té parets i tampoc un sostre tancat. Evidentment, pot estar delimitat per les parets dels edificis del voltant. En aquest cas, un màxim de 2 parets poden estar a menys de 2 metres. Això permet fer enregistraments al racó format per 2 edificis. En tots els casos, els arbres, les botigues petites, un espai obert on prendre el cafè, el trànsit així com també una vorera, són possibles. Es faran 200 enregistraments des d'aquest entorn.

Definició del corpus i entorns d'enregistrament de la base de dades de cotxe

Aquesta base de dades comprèn els enregistraments de 600 sessions diferents fetes per 300 informants. Una sessió conté 119 locucions llegides. Les últimes 200 sessions també contenen locucions espontànies. Totes han estat enregistrades mitjançant quatre micròfons instal·lats en cotxes. Les locucions llegides consisteixen en paraules clau d'activació de sistemes per veu, diverses seqüències de dígits aïllats i connectats, dates espontànies i llegides, hores, frases amb paraules clau, paraules lletrejades, noms espontanis, ciutats, empreses, frases i paraules fonèticament riques i més de 70 paraules específiques d'aplicació per telèfon mòbil, funcions IVR i productes de cotxes.

Condicions d'enregistrament

Hi ha definides 7 condicions d'ambient. Cada ambient està igualment representat a la base de dades final.
1. cotxe aturat amb el motor en marxa
2. cotxe en trànsit urbà
3. cotxe en trànsit urbà, en condicions sorolloses
4. cotxe circulant a baixa velocitat en condicions de carretera rugosa
5. cotxe circulant a baixa velocitat en condicions de carretera rugosa en condicions sorolloses
6. cotxe circulant a alta velocitat en condicions de bona carretera
7. cotxe circulant a alta velocitat en condicions de bona carretera amb l'equip d'àudio *on.*

A més, s'ha recopilat la següent informació durant els enregistraments:
• condicions meteorològiques: pluja, cel clar, vent, etc.
• accessoris utilitzats durant els enregistraments: neteja parabrises, ventilació, ventilador, ràdio
• estat de funcionament del ventilador: apagat, baix, mitjà, alt

Informants

El català és la llengua parlada a Catalunya, València, Balears i Andorra. També es parla en altres llocs, encara que de forma minoritària, com el Rosselló i el Vallespir (sud de França), la franja fronterera entre Catalunya i l'Aragó, i l'Alguer a Sardenya.

Hi ha una divisió principal entre l'est i l'oest que creua Catalunya i també separa els dialectes de València i Balears. Hi ha una divisió secundària entre el nord i el sud que separa els dialectes de l'est i l'oest de Catalunya dels dialectes de València i Balears (Joan Veny, Els parlars catalans, Edit. Moll, Mallorca, 1993).

El mapa de la Figura 1 mostra las quatre zones dialectals a Catalunya objecte de la cerca d'informants. Amb el propòsit de forçar una recollida de mostres tan variada com sigui possible, s'ha contemplat una subdivisió al nord de Catalunya amb el Gironí i una altra al sud amb el Tortosí.

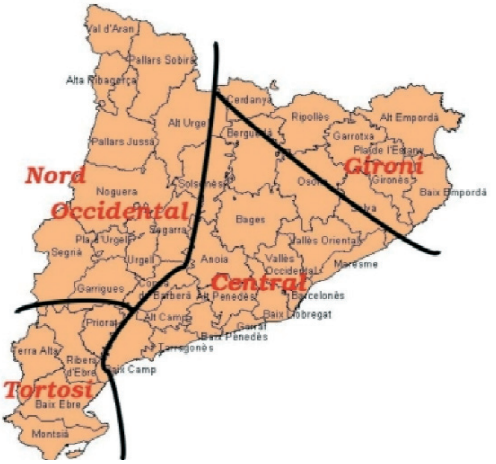

Figura 1. Dialectes de Catalunya

En totes les bases de dades, la meitat dels informants seran homes i l'altra meitat dones. També es tindrà en compte la distribució del informants per grups d'edat, que haurà de complir les especificacions de la Taula 2.

Grups d'edat	Nombre d'informants (en % sobre el total)
Menys de 16	0
16-30	>20%
31-45%	>20%
46-60	>15%
Més de 60	opcional

Taula 2. Distribució d'informants per grups d'edat

En les bases de telefonia fixa i mòbil, el nombre d'informants és de 2000 per cada tipus de telèfon. Es consideraran, en cada cas, 1500 locutors de Catalunya distribuïts en les regions dialectals marcades en el mapa de la Figura 1, més 300 locutors de València i 200 de Balears.

En la base de dades de consum, el nombre d'informants és de 550 i cada un enregistra una sessió. Els informants són seleccionats per assolir un equilibri entre dialectes. El mapa de la Figura 1 mostra les quatre regions dialectals definides per aquesta base de dades. Un mínim de 97 informants per regió és obligatori.

En la base de dades de cotxe, el nombre d'informants és de 300 i cada un enregistra dues sessions. Els informants són seleccionats per assolir un equilibri dialectal. El mapa de la Figura 1 mostra les quatre regions dialectals definides per aquesta base de dades. Són necessaris un mínim de 60 informants (o 120 sessions) de cada regió.

Lloc i plataformes d'enregistrament

Els enregistraments de les bases de dades de telefonia fixa i mòbil es fan a la Universitat Politècnica de Catalunya. Les característiques principals de la plataforma són:

• Interfície: RDSI d'accés bàsic
• Targeta: AVM-ISDN-A1.
• Ordinador: Pentium PC at 120 MHz, 32 MB RAM 4 GBytes SCSI Hard disk. PCI Network card
• DOS: Windows
• Interfície de programari: COMMON-ISDN-API Version 2.0 (CAPI 2.0)
• Programari: UPC ADA
• Línies: 2

La plataforma d'enregistrament de les bases de dades de consum i cotxe consisteix en un ordinador portàtil que usa un slot PCMCIA com a interfície per a

l'equipament d'àudio. El sistema operatiu és Windows XP. La UPC ha desenvolupat un programari d'enregistrament específic per a aquest projecte. És possible enregistrar fins a quatre micròfons de manera simultània.

Els micròfons utilitzats depenen de l'ambient on es realitza l'enregistrament: oficina, entreteniment, cotxe o lloc públic. Cadascun d'aquests ambients d'enregistrament té les seves característiques de soroll, nombre i tipus de micròfons a ser enregistrats simultàniament i posició de la plataforma d'enregistrament i dels micròfons.

Tots els informants porten dos micròfons per fer enregistraments de curta distància. En l'oficina i lloc d'entreteniment, es col·loca un micròfon a mitjana distància, sobre d'una taula, i un micròfon a llarga distància. En el llocs públics no s'utilitza el micròfon de llarga distància i, en el cotxe, els micròfons es col·loquen a prop del sostre en llocs adients.

La Figura 2 dóna una visió general de les posicions de muntatge. Els micròfons de *curta distància* es posen a 2 cm i a 10 cm de la boca, respectivament.

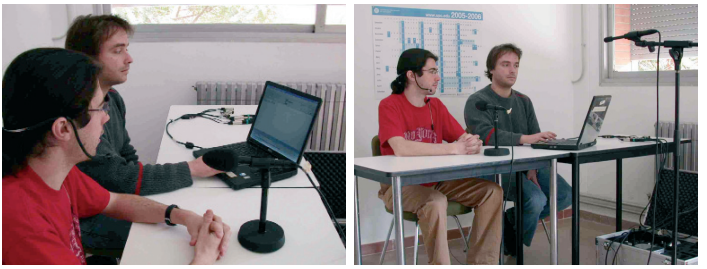

Figura 2. Posicions pel muntatge dels micròfons

En tots els escenaris d'enregistrament, els micròfons de *mitjana distància* i *llarga distància* estan encarats a l'informant. La persona enregistrada s'asseu en una cadira durant tota la sessió. Els dos micròfons de *curta distància* estan muntats sobre el cos de l'informant i els micròfons de *mitjana* i *llarga distància* estan situats a 1 metre i entre 2 i 3 metres del informant respectivament. Els micròfons es situen a una alçada mitjana de 1,2 metres, i permeten una desviació de 50 cm. Pel que fa a les propietats de reverberació d'un lloc, la posició dels informants relativa a objectes reflectors, com les parets, és important. Unes etiquetes de posició diferencien en categories aquestes posicions de forma genèrica. Per a cada lloc d'enregistrament i posició específica, es mesura la resposta impulsional de l'habitació. Per a cada sessió, es mesura un nivell de soroll.

El procediment d'enregistrament està completament supervisat per un operador.

Formats del fitxers de veu

Els fitxers de veu s'emmagatzemen com seqüències de 8 bits a 8 kHz en llei A sense compressió. Cada registre s'emmagatzema en un fitxer separat. Cada fitxer de veu té un fitxer d'etiquetes SAM associat.

En les bases de consum i cotxe s'enregistren quatre canals d'àudio d'alta qualitat mitjançant una plataforma mòbil. Les dades s'emmagatzemen en seqüències de 16 bits sense compressió i utilitzant una freqüència de mostreig de 16 kHz. Cada registre s'emmagatzema en un fitxer separat. Cada fitxer de veu té un fitxer d'etiquetes SAM associat on hi ha una descripció de la freqüència de mostreig, la quantificació i el nombre de bytes per mostra, entre d'altres. A més, també hi ha informació relativa al nivell de soroll ambiental en el moment de l'enregistrament i del valor de la relació senyal-soroll del fitxer de veu.

Transcripció de les bases de dades

La transcripció la duu a terme l'empresa ATLAS. Estarà inclosa en aquesta base de dades i la característica principal és que és ortogràfica i lèxica amb alguns detalls que representen sorolls audibles (veu i no-veu) presents en els corresponents senyals d'àudio. Les marques extres contingudes en la transcripció ajuden a interpretar el text de la frase. Les transcripcions es fan en dos passos: un primer pas en el qual es transcriuen les paraules i un segon pas on s'afegeixen els detalls addicionals.

Figura 3. Transcripció de dades

Les marques extres s'utilitzen per a males pronunciacions, paraules inintel·ligibles i sorolls. Els símbols pels sorolls són:

[fil]: Pausa sonora
Aquests sons es poden modelar bé en un model de pauses sonores en reconeixedors de veu. Alguns exemples son: uh, um, er, ah, mm.
[spk]: Soroll d'informant.
Tots els sorolls i sons fets per l'informant i que no formen part del text preparat com soroll de llavis, tossir, aclariment de la gola, clicks amb la llengua, respiració sorollosa, riures,...

[sta]: Soroll estacionari
Aquesta categoria conté sorolls de fons que no són intermitents i tenen un espectre d'amplitud més o menys estable. En són exemples el soroll de cotxe, soroll de carrer, soroll de canal, GSM, veus de fons, soroll de fons de llocs públics,...

[int]: Soroll intermitent
Aquesta categoria conté sorolls de naturalesa intermitent. Aquests sorolls típicament ocorren una vegada (cop de porta) o tenen pauses (ring del telèfon), o canvien el seu espectre amb el temps (música). En són exemples: música, veu de fons, nen plorant, telèfon sonant, cop de porta, campana timbre, paper arrugat, converses creuades.

[dit]: To (beep)
Aquest soroll és produït per la plataforma d'enregistrament per a indicar l'informant que pot començar a enregistrar. El sistema no ha d'enregistrar aquest soroll però, si ho fa, ha de ser anotat amb aquesta marca.

La base de dades es transcriu mitjançant el programari UPCRevBD.v1, desenvolupat a la UPC. Un 1% de les transcripcions es transcriu dues vegades per a detectar errors. La base de dades final serà supervisada i validada per un organisme extern independent.

Informació lèxica i fonètica

La documentació inclou un lexicó. El fitxer amb el lexicó és una llista ordenada alfabèticament de les diferents partícules lèxiques (essencialment paraules en el nostre cas) que ocorren en el corpus amb la corresponent informació de pronunciació. Cada paraula diferent té una entrada diferent. Com que el lexicó es deriva del corpus, usa la mateixa codificació alfabètica per a caràcters especials i accentuats com en les transcripcions (ISO-8859). El fitxer inclou també un recompte de freqüència d'aparició per a cada entrada en el lexicó.

Després de la fase de transcripció es genera un lexicó que conté totes les paraules que han aparegut ordenades alfabèticament, el seu nombre d'aparicions, i la seva transcripció fonètica. Les paraules apareixen en el lexicó exactament igual que en la transcripció. Les marques de sorolls, fragments i paraules mal pronunciades no apareixen en el lexicó.

El programari SEGRE, desenvolupat en l'àmbit del CREL, s'utilitza per a transcriure fonèticament les paraules amb la notació SAMPA. El lexicó es transcriu automàticament. El noms propis i noms d'empreses es faran manualment.

Control de qualitat

Totes les bases de dades del projecte seran validades per un centre de validació extern a aquell que les produeix. L'objectiu de la validació és assegurar que els recursos produïts compleixen les especificacions i puguin ser homologats. Per a cada base de dades existeixen uns criteris de validació que han estat produïts en diversos projectes europeus. El centre de validació aplica els criteris a cada base de dades i les desviacions, si existeixen, són documentades en un informe públic. L'informe és positiu si les desviacions estan en el marc d'uns marges prèviament establerts. En cas d'un informe negatiu, la base de dades ha de ser corregida i sotmesa a una nova validació. L'informe de validació s'afegeix a la documentació de la base de dades corresponent per a la seva distribució.

Seguiment del projecte

Els objectius, les especificacions i l'estat actual del projecte són accessibles a la pàgina web http://gps-tsc.upc.es/veu/projects/BDG.

En el projecte hi han col·laborat conjuntament enginyers i lingüistes per a la elaboració dels corpus d'enregistrament i les plataformes d'enregistrament. Estan col·laborant més de 40 estudiants de diverses universitats catalanes d'arreu de Catalunya en la recerca de locutors. Deu lingüistes graduats estan fent les tasques de transcripció.

Disponibilitat

Les bases de dades seran públiques i gratuïtes. Accessibles via ELRA *(European Language Resources Association).*

Informació

asuncion@gps.tsc.upc.edu