



Centre
d'Innovació

 Barcelona
Media

Grup de Veu i Llenguatge



El grup de Veu i Llenguatge es dedica a la investigació i la innovació en el camp del tractament del llenguatge natural:

- tant en les manifestacions escrites,
- com en les produccions orals



Centre
d'Innovació

22 Barcelona
Media

àrees de recerca

processament del llenguatge natural

- anàlisi
- generació

diverses dimensions del processament del llenguatge

- diferents llengües
- diversos nivells de descripció lingüística (morfosintaxi, sintaxi, semàntica, pragmàtica)
- diferents àrees temàtiques (vocabulari, estructures sintàctiques, significat...)



escalabilitat / adaptabilitat de les tècniques / recursos

- aproximacions robustes al processament del llenguatge
 - tècniques simbòliques de baix nivell (basades en autòmats)
 - tècniques estadístiques
- augmentar el poder dels models estadístics del llenguatge
 - tot incloent-hi informació lingüística, i
 - investigant en els algoritmes que treballen amb models de llenguatge complexos
- aplicar tècniques de baix nivell per modelar relacions lingüístiques complexes (sintàctiques, semàntiques, de discurs...)
- usar aprenentatge automàtic per adquirir informació lèxica o gramatical



Centre
d'Innovació

22 Barcelona
Media

àrees de recerca

aquests plantejaments es poden aplicar en un munt de contextos:

traducció automàtica

detecció d'errors ortogràfics i gramaticals

detecció de tema

extracció d'informació

resum

síntesi de veu

reconeixement de veu

. . .



processament de la veu

- síntesi
 - selecció de les unitats de descomposició (normalment difons)
 - recomposició de la parla (a partir de difons)
 - inclusió d'informació prosòdica
- reconeixement
 - augment del model acústic
 - per tractar la varietat de parlants (i de dialectes)
 - augment del model de llenguatge
 - per millorar la capacitat predictiva
 - per interactuar amb altres mòduls de processament del llenguatge
 - per automatitzar les tècniques d'entrenament per subllenguatges
 - millora dels algorismes de cerca
 - detecció de les emocions



Centre
d'Innovació

22 Barcelona
Media

línies d'investigació

Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



objectiu:

- tractament lingüístic de textos escrits
actualment, en català, espanyol, anglès
- amb la finalitat de:
 - tenir les paraules dels textos decorades amb informació
addicional que faciliten les tasques posteriors

investigacions:

- en lingüística
- en tècniques
- en arquitectura



anàlisi massiva de textos escrits

La fi de la guerra va suposar la fi de la lluita contra el règim

la	el	Det	AFS	DN>
fi	fi	Nom	N5-6S	CD_Subj
de	de	Prep	P	<NA
la	el	Det	AFS	DN>
guerra	guerra	Nom	N5-FS	<P
va	anar	Verb	VDR3S-	VAux>
suposar	suposar	Verb	VI----	VPrin
la	el	Det	AFS	DN>
fi	fi	Nom	N5-6S	CD_Subj
de	de	Prep	P	<NA
la	el	Det	AFS	DN>
lluita	lluita	Nom	N5-FS	<P
contra	contra	Prep	P	<NA_Advl
el	el	Det	AMS	DN>
règim	règim	Nom	N5-MS	<P



Centre
d'Innovació

22 Barcelona
Media

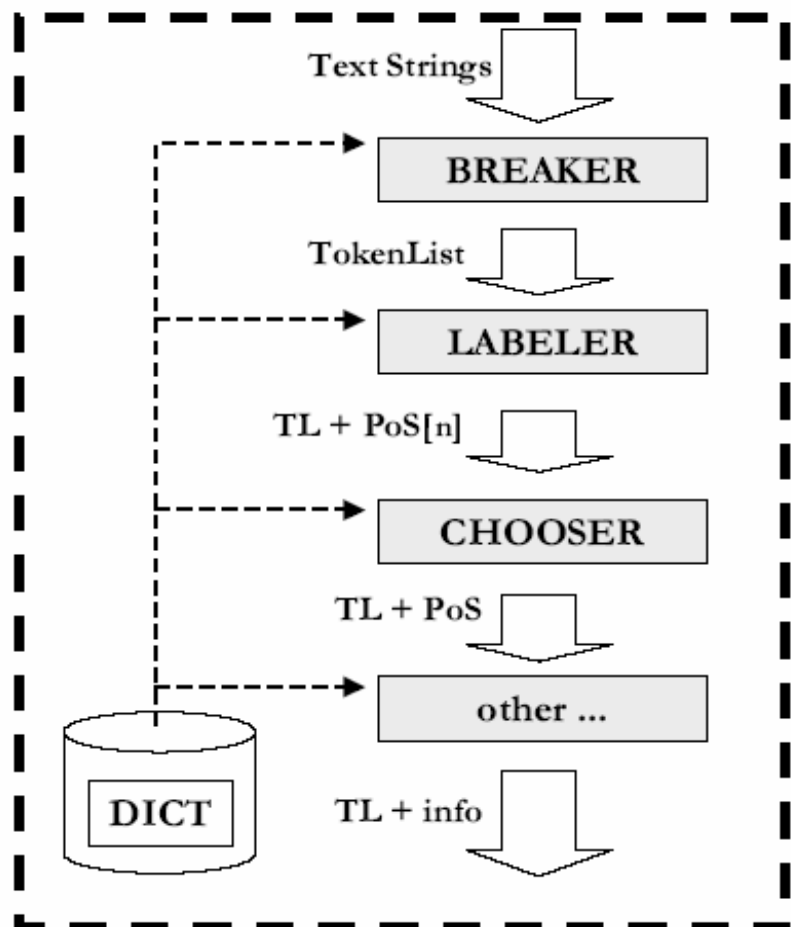
anàlisi massiva de textos escrits

recursos:

- **CATCG** (implementada sobre la Constraint Grammar® – Conexor oy)
 - demo: <http://mutis.upf.es/cgi-bin/catcg/demo.pl>
- arquitectura de processament **LINLaP**
feta en part, dins i3media
- diccionaris
- corpus de textos

2.1 Diagram

- Modular:
 - Flexible
 - Customizable
- Linear:
 - Segmentation
 - Dict. Lookup
 - PoS Tagging
 - Other...
- Progressive enrichment





CUCWeb: Corpus d'Ús del Català a la Web

- corpus extret i constituït a partir de les pàgines en català de les webs hostatjades a l'Estat Espanyol (domini .es i altres)
- consultable permanentment a la web:
<http://www.catedratelefonica.upf.es/>

fet en col·laboració amb la Càtedra Telefònica



Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



Centre
d'Innovació

22 Barcelona
Media

correcció de textos

objectiu:

correcció (lingüística) de textos:

- escrits per parlants nadius
 - correcció ortogràfica / gramatical / d'estil
- escrits per aprenents d'una segona llengua



Centre
d'Innovació

22 Barcelona
Media

correcció de textos

investigació:

- detecció dels errors
- classificació dels errors
- generació de propostes de correcció

tècniques:

- quantitatives
- qualitatives

corrector ortogràfic i gramatical per al català:

- codi obert
- multiplataforma (Windows, Linux, Mac)
- en múltiples entorns d'edició i sistemes operatius
 - Microsoft Office, Open Office, Mozilla...
- normatiu
- dialectal
- implementat en LINLaP
- flexible quant a funcionalitats
 - permet escollir opcions de correcció i dialecte

consultable ara a <http://parles.upf.es/corrector/index.aspx>

desenvolupat per encàrrec de la Generalitat



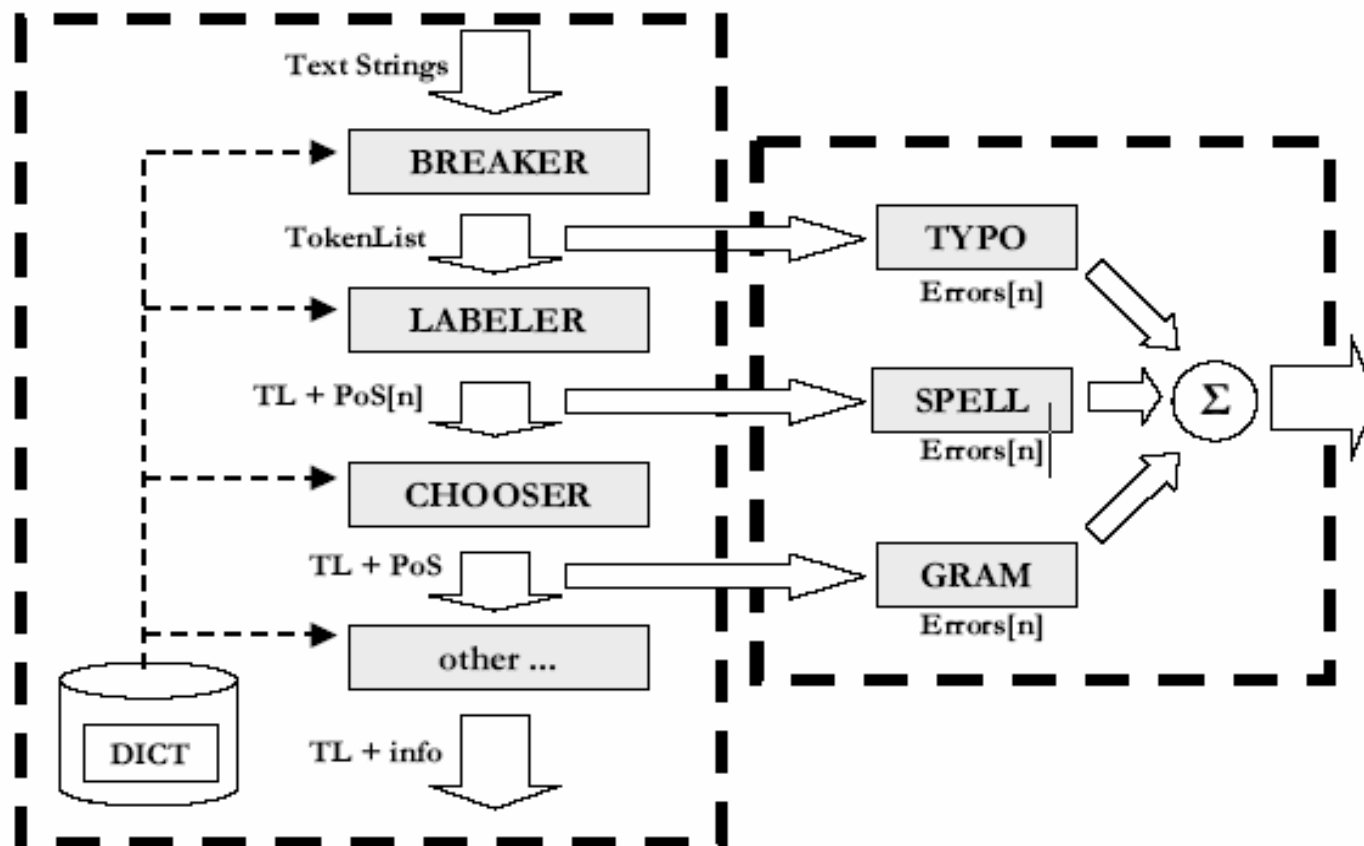
creació de correctors ortogràfics / gramaticals / d'estil

- adaptats a les necessitats d'una empresa o corporació
- sobre l'arquitectura LINLaP
 - usant diccionaris i corpus específics del client
 - personalitzant
 - les funcionalitats de correcció
 - els connectors als entorns d'edició

en contracte actiu amb CCMA



correcció de textos





Centre
d'Innovació

22 Barcelona
Media

correcció de textos

correcció de textos fets per aprenents de segones llengües

- aplicació de tècniques de correcció a textos escrits per aprenents
- aplicat a alemany, anglès, català, espanyol
 - en cooperació amb partners europeus
 - projectes ALLES i AUTOLEARN



- qüestions de recerca:
 - podem modelar el comportament del professor que corregeix?
 - quins errors es marquen? quan es marquen?
 - classificació dels errors detectats
 - tècniques automàtiques de detecció
 - comparació de textos d'aprenents amb textos escrits per nadius

http://217.91.104.155 - ALLES: Business English for Advanced Learners - Dev 1.01 - Mozilla Firefox

Customer Service and International Communication

Customer Service: IV

Subtask 0
Subtask 1
Subtask 2
Subtask 3
Subtask 4
Subtask 5
Subtask 6

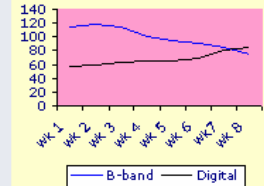
Use the charts and graphs on the right to help you to complete the writing exercise below. (Use the scroll bars)

1a. Look at graph 1 (Weekly sales of two Stanley products). Write a sentence to give an overall description of sales figures for Stanley Broadband over the 8 week period.

1b. Now, look at graph 1 (Weekly sales of two Stanley products) again and write a sentence to give an overall description of sales figures for Stanley Digital over the same period.

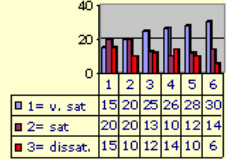
2. Add a sentence to give a general description of customer satisfaction with the Stanley Digital camera.

1) Weekly sales of two Stanley products



Week	B-band	Digital
Wk 1	120	60
Wk 2	115	65
Wk 3	110	70
Wk 4	105	75
Wk 5	100	80
Wk 6	95	80
Wk 7	90	80
Wk 8	80	80

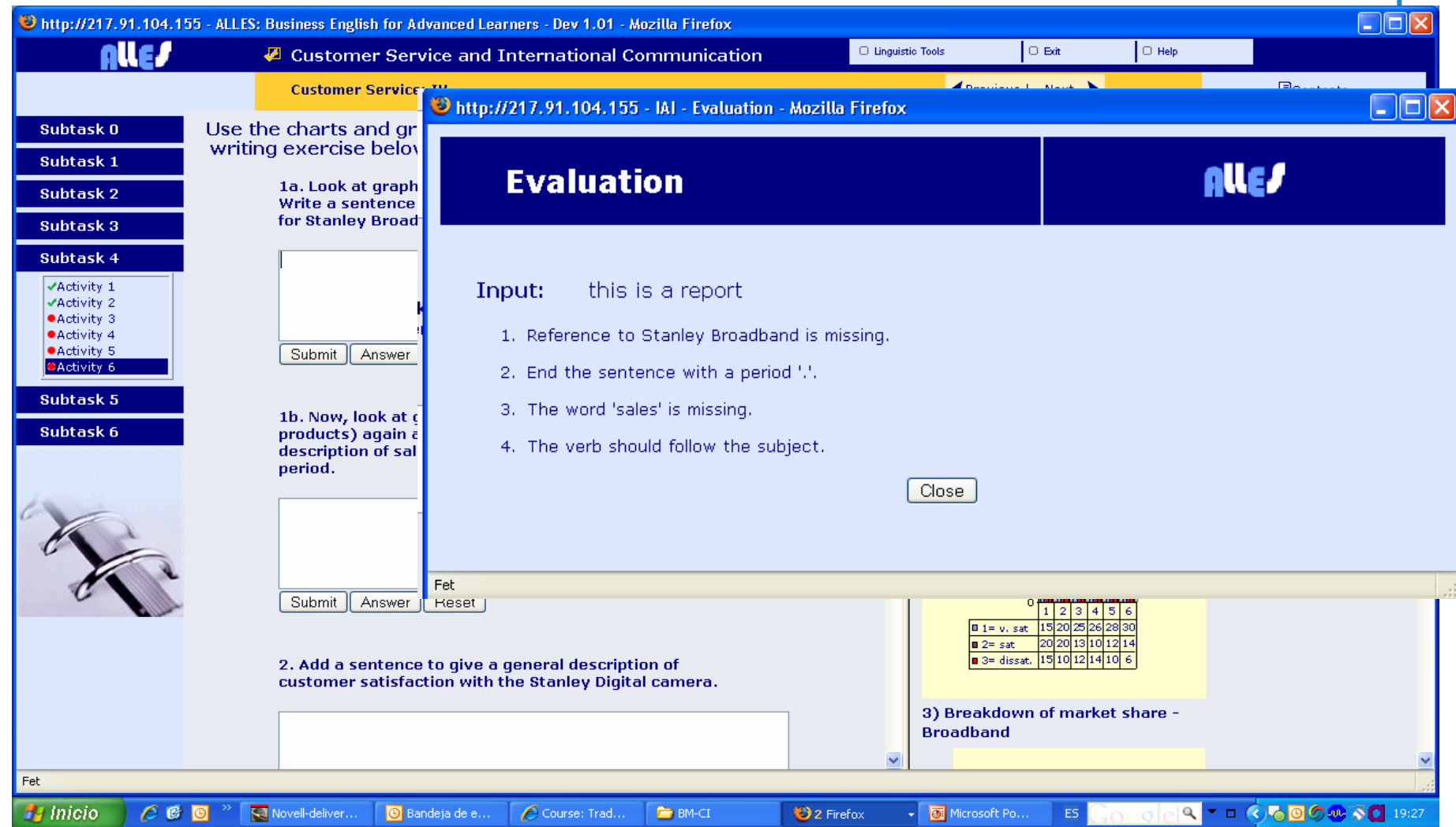
2) Weekly levels of customer satisfaction with Stanley Digital



	1	2	3	4	5	6
1= v. sat	15	20	25	26	28	30
2= sat	20	20	13	10	12	14
3= dissat.	15	10	12	14	10	6

3) Breakdown of market share - Broadband

Firefox taskbar: Inicio, Novell-deliver..., Bandeja de e..., Course: Trad..., BM-CI, 2 Firefox, Microsoft Po..., ES, 19:27



The screenshot shows a web browser window with two overlapping pages. The background page is a course interface for 'Customer Service and International Communication'. It features a sidebar with a task list and a main content area with writing exercises. The foreground window is an 'Evaluation' dialog box with a list of corrections.

Course Interface (Background):

- URL: <http://217.91.104.155> - ALLES: Business English for Advanced Learners - Dev 1.01 - Mozilla Firefox
- Page Title: Customer Service and International Communication
- Navigation: Linguistic Tools, Exit, Help
- Task List (Left Sidebar):
 - Subtask 0
 - Subtask 1
 - Subtask 2
 - Subtask 3
 - Subtask 4
 - Activity 1 (checked)
 - Activity 2 (checked)
 - Activity 3 (red dot)
 - Activity 4 (red dot)
 - Activity 5 (red dot)
 - Activity 6 (red dot)
 - Subtask 5
 - Subtask 6
- Main Content:
 - Use the charts and graph writing exercise below
 - 1a. Look at graph Write a sentence for Stanley Broad
 - 1b. Now, look at (products) again e description of sal period.
 - 2. Add a sentence to give a general description of customer satisfaction with the Stanley Digital camera.

Evaluation Window (Foreground):

- URL: <http://217.91.104.155> - IAI - Evaluation - Mozilla Firefox
- Title: Evaluation
- Input: this is a report
- Corrections:
 - Reference to Stanley Broadband is missing.
 - End the sentence with a period '.'.
 - The word 'sales' is missing.
 - The verb should follow the subject.
- Buttons: Submit, Answer, Close

Table (Bottom Right):

	1	2	3	4	5	6
1= v. sat	15	20	25	26	28	30
2= sat	20	20	13	10	12	14
3= dissat.	15	10	12	14	10	6

3) Breakdown of market share - Broadband

Windows Taskbar: Inicio, Novell-deliver..., Bandeja de e..., Course: Trad..., BM-CI, 2 Firefox, Microsoft Po..., ES, Google, 19:27



Centre
d'Innovació

22 Barcelona
Media

correcció de textos

- ara (AUTOLEARN)
 - estem fent una migració massiva a Moodle
 - amb l'objectiu d'avaluar les estratègies docents i de correcció amb estudiants en contextos educatius reals (a Alemanya, Escòcia, Turquia i Catalunya)



Centre
d'Innovació

22 Barcelona
Media

línies d'investigació

Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



objectiu:

- extreure informació de textos
 - informació sobre el contingut dels textos
 - informació lingüística
 - sobre l'estructura lingüística dels textos

investigacions:

- en lingüística
- en tècniques



a partir de l'arquitectura de LINLaP:

- afegir mòduls de processament
 - amb tècniques de baix nivell lingüístic

per extreure informació de base lingüística:

- entitats referides
- esdeveniments
- participants en els esdeveniments
- estats d'opinió



aquesta informació pot ser:

- sobre el contingut dels textos
 - útil per a empreses i clients
- rellevant per a l'obtenció d'informació lingüística
 - útil per a:
 - lingüistes
 - desenvolupadors de tecnologia lingüística

activitat dins i3media

en part, en col·laboració amb Grup de Recuperació d'Informació



sobre el contingut dels textos

- determinació del tema o temes,
- obtenció automàtica de metadades,
- extracció de paràfrasis
- extracció dels elements essencials del text
 - 1r pas cap al resum



rellevant per a l'obtenció d'informació lingüística

- classes de complements que tenen els verbs o predicats
- distingir entre sentits de les paraules / grups de paraules
- usos de construccions / paraules
- adquisició de terminologies i construcció d'ontologies



Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*

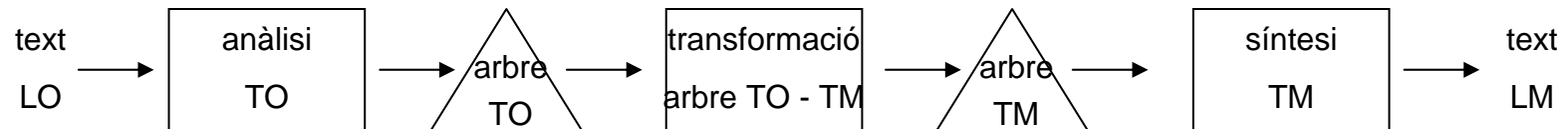
objectiu:

investigar en tècniques per a la traducció automàtica

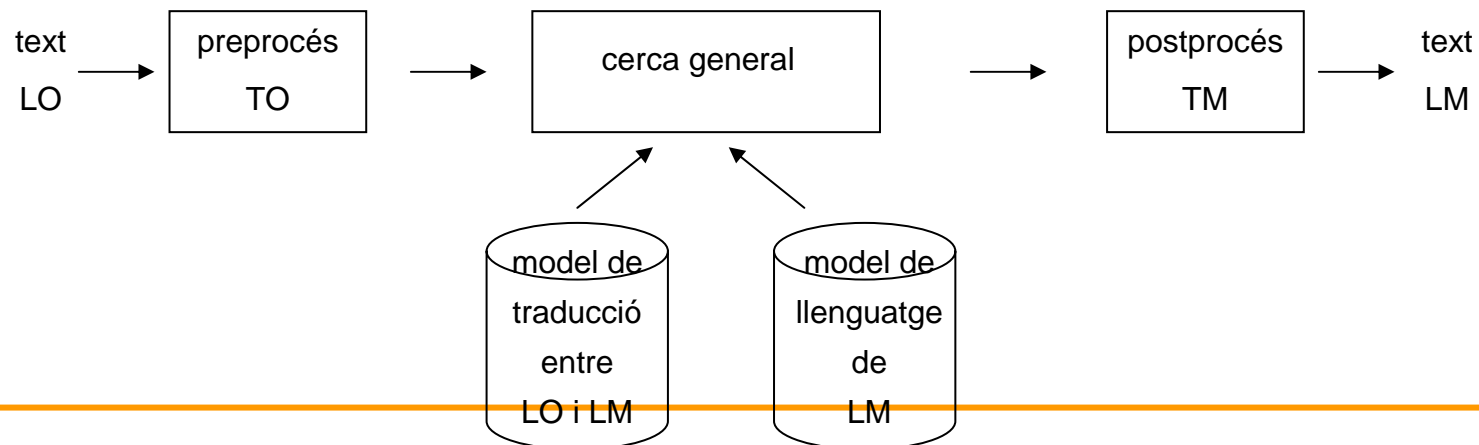
- que superin els colls d'ampolla actuals:
 - sistemes de base simbòlica
 - cost de desenvolupament
 - dificultats de manteniment
 - impossibilitat d'adaptació a nous entorns / tipus de textos
 - sistemes de base empírica
 - sostre en actuació
 - dificultat d'incorporar informació lingüística en els sistemes estadístics clàssics
- que puguin aprendre de textos traduïts correctes

mètodes clàssics de TA:

- basat en anàlisi lingüística

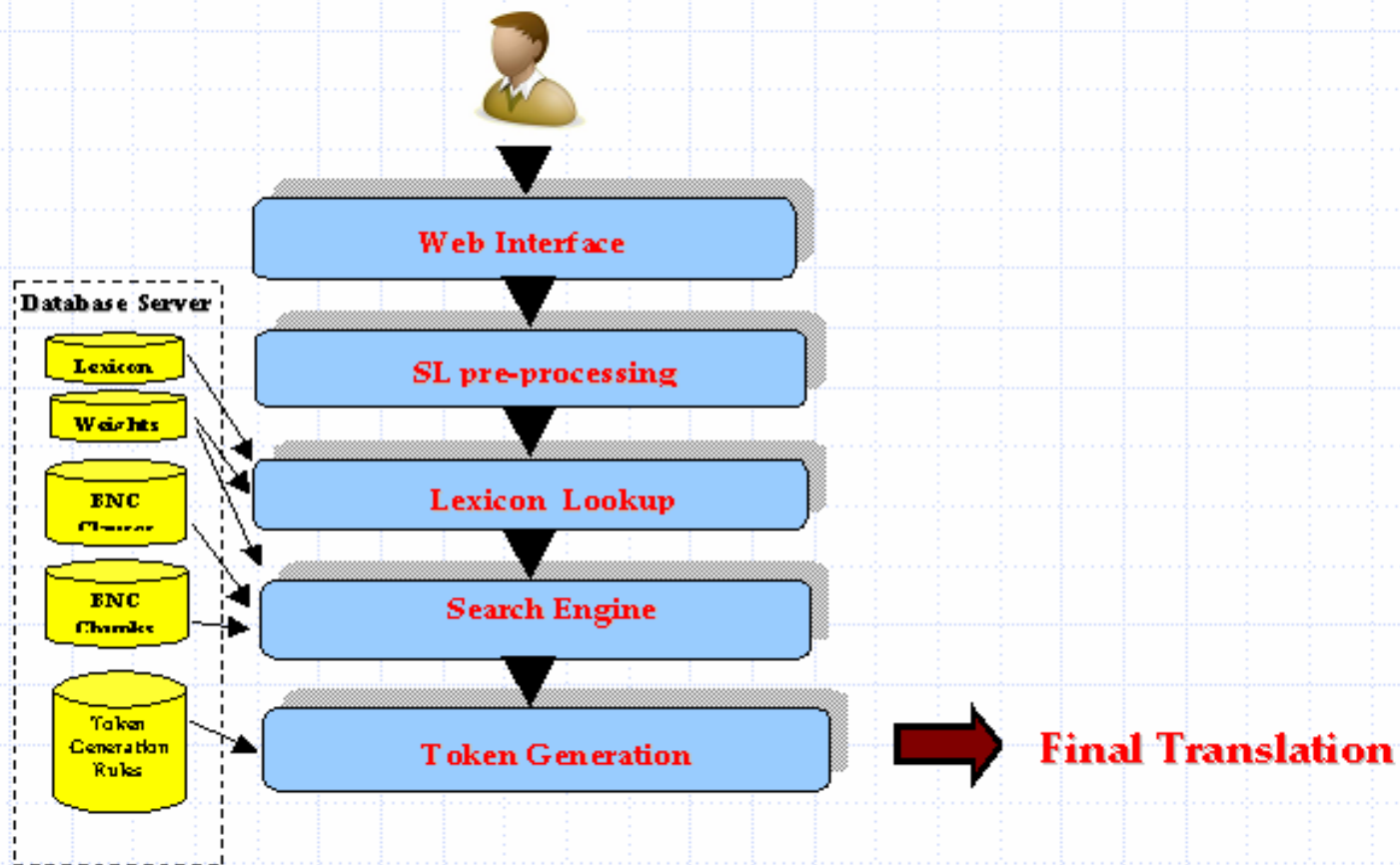


- estadístic





METIS II architecture





Centre
d'Innovació

22 Barcelona
Media

traducció automàtica

la nostra investigació:

- minimitzar els errors deguts a la mala qualitat dels textos paral·lels dels quals s'extreu el model de traducció
- introduir informació lingüística de forma natural
 - a partir de diccionaris bilingües
 - i d'esquemes de projecció d'estructures entre llengües
- maximitzar el model de llenguatge
 - “generation driven MT”

ara, activitat dins i3media
fins fa ben poc: METIS-II



Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



objectiu:

- poder tractar la llengua de signes catalana (LSC) com les llengües orals
- en concret:
 - codificar la LSC (i altres LS) de forma adequada
 - construir un prototipus de traductor automàtic de llengua oral a LS



- investigacions:
 - descripció de les LS
 - sistema de codificació de les LS
 - recopilació i anotació de corpus
 - integració amb motor d'animació de l'avatar

en part, dins Ulises
en col·laboració amb:

- grup de lingüística (UPF)
 - grups de Gràfics



Actualment, la nostra investigació se centra en sis línies principals:

- *anàlisi massiva de textos no restringits*
- *correcció de textos*
- *extracció d'informació de textos*
- *traducció automàtica*
- *tractament de la llengua de signes*
- *síntesi de veu*



Centre
d'Innovació

22 Barcelona
Media

síntesi de veu

objectiu

- crear
 - una veu catalana
 - una veu castellana
 - una veu bilingüe català – castellà
- introduir entonació natural (prosòdia)
- facilitar la creació de locutors especialitzats



investigacions

- lingüística sobre la prosòdia
 - en parla espontània
 - en diàleg
- recopilació de “corpus de prosòdia”
- agilitzar el procés de creació de veus i locutors
- mètodes d'introducció de prosòdia en les locucions sintètiques
 - a través dels corpus de gravació i de l'augment de l'espai de difons
 - a través de la manipulació del senyal
 - combinant els dos mètodes

en col·laboració amb una empresa escocesa (Cereproc)
investigació feta a i3media



Centre
d'Innovació

22 Barcelona
Media

síntesi de veu

millora del procés de creació de veus:

- en la selecció i processament del corpus de gravació
- en el mateix procés de gravació
- en el postprocés



Exemples en català:

1. “Barcelona Media-Centre d'Innovació és un centre tecnològic que es dedica a la recerca aplicada en l'àmbit de la comunicació o dels media, i a la transferència de coneixements i de tecnologia a la indústria d'aquest sector.”

[àudio](#)

2. “Jurídicament, Barcelona Media és una fundació sense ànim de lucre (Fundació Barcelona Media Universitat Pompeu Fabra), el patronat de la qual està format per representants d'empreses del sector de la comunicació, d'universitats i de l'administració pública.”

[àudio](#)



Exemples en castellà:

1. “Barcelona Media-Centro de Innovación es un centro tecnológico que se dedica a la investigación aplicada en el ámbito de la comunicación o de los media, y a la transferencia de conocimiento y de tecnología en la industria de este sector.”

[àudio](#)

2. “Jurídicamente, Barcelona Media es una fundación sin ánimo de lucro, el patronato de la cual está formado por representantes de empresas del sector de la comunicación, de universidades y de la administración pública.”

[àudio](#)



ara:

Beto Boullosa, Roberto Carllini, Judith Domingo, Juanma Garrido, Marc González, Bernat Grau, Yesika Laplaza, Montse Marquina, Guillem Massó, Maite Melero, Martí Quixal, Carlos Rodríguez, Julia Sidorova, Teresa Suñol, Oriol Valentín,

anteriorment:

Francesc Benavent, Eva Bofías, Gemma Boleda, Steffan Bott, Marta Carulla, Ariadna Font, Àngel Gil, Jana Kunova, Rosa Lucha, Araceli Martínez, Laia Mayol, Auke Oosterhof, Òscar Puente, Marc Riera, Anna Ruggia, Roser Saurí

col-laboradors UPF

Àlex Alsina, Carme Colominas, Joan Costa, Josep M Fontana, Louise McNally, Enric Vallduví



Centre
d'Innovació

22 Barcelona
Media

Grup de Veu i Llenguatge
