Matxin:
developing sustainable MT for a less-resourced language

Kepa Sarasola

(Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Julen Ruiz, Aingeru Mayor)

Ixa taldea.

University of the Basque Country

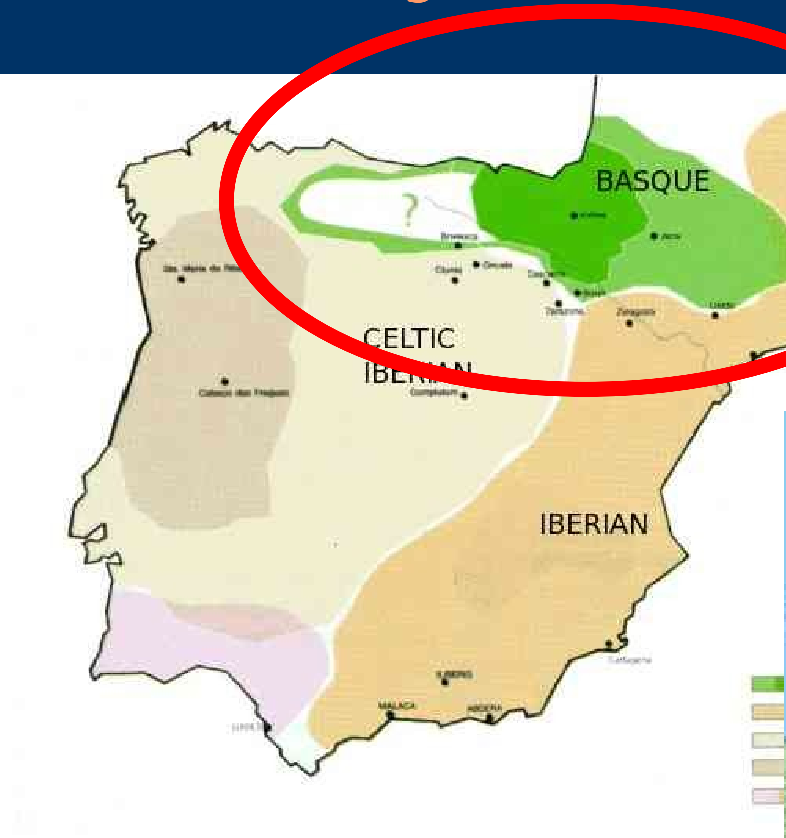
Universidad del País Vasco
Euskal Herriko Unibertsitatea

FreeRBMT 2009, Alacant

Outline

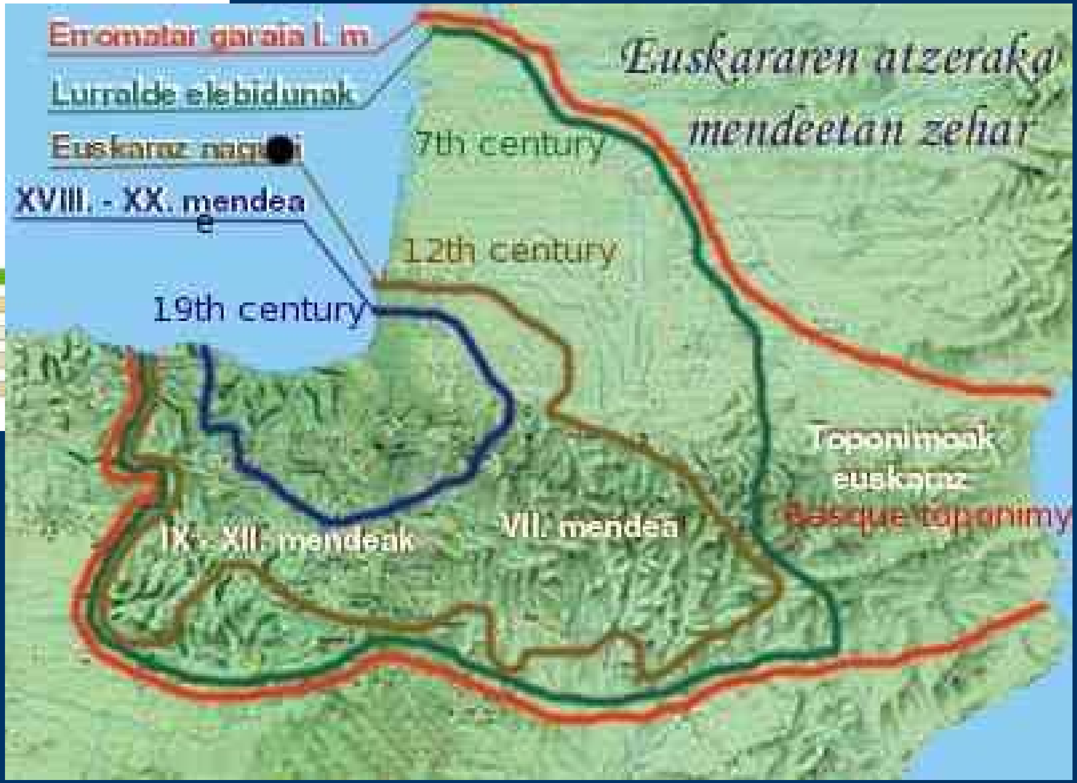
- Basque: a Less Resourced Language (LRL)
- Strategy for sustainable HLT (and MT) for Basque
- Machine Translation for Basque (Matxin)
- Evaluation of Matxin
- Future: Combining RBMT and Corpus-based MT
- Recent elements and conclusions

History of Basque



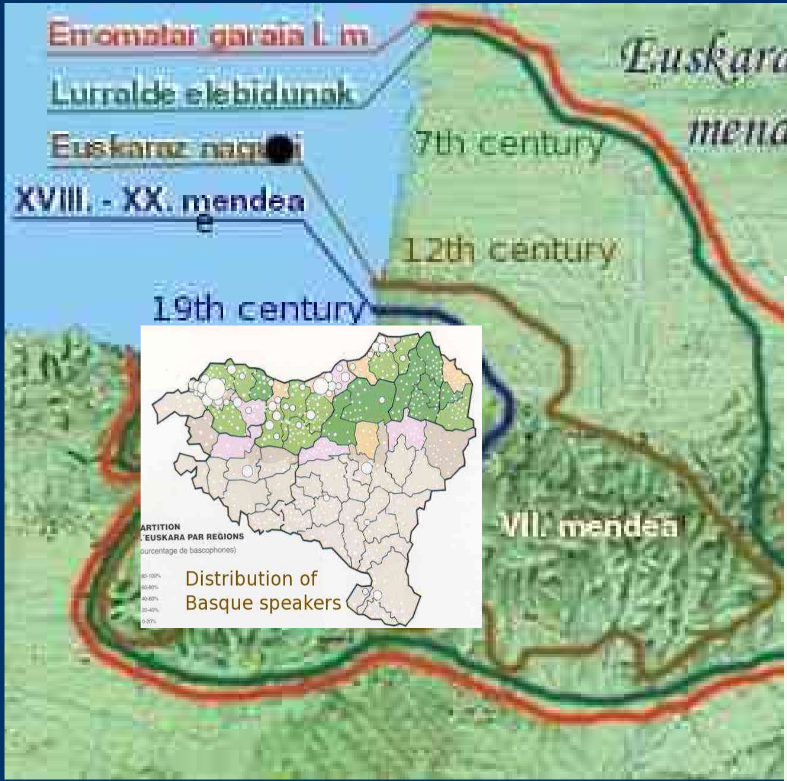
Prerromanic languages in Spain

Basque in 7th, 12th and 19th centuries



History of Basque

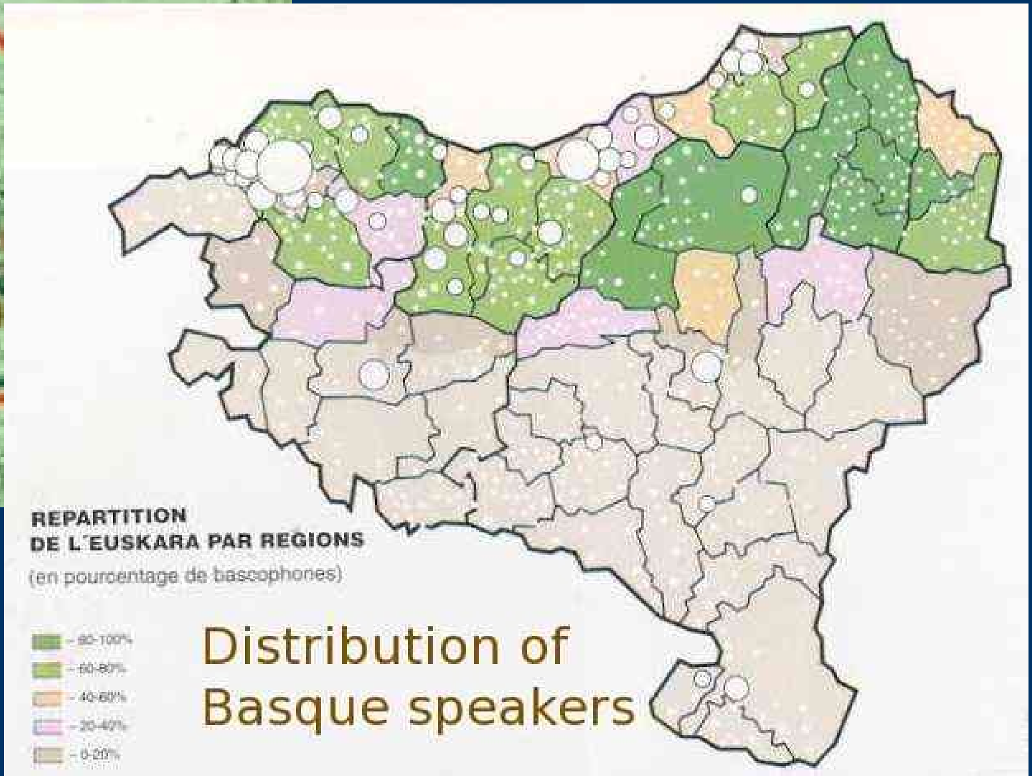
Basque in 7th, 12th and 19th centuries



1,033,900 Speakers

(First lang.: 700,000)

Non homogeneous distribution!

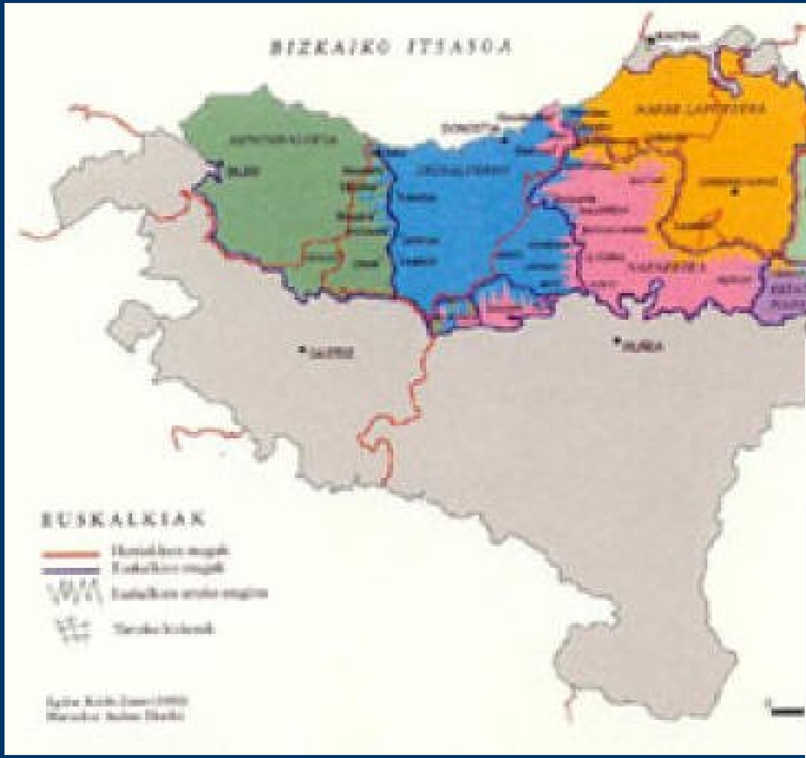

REPARTITION DE L'EUSKARA PAR REGIONS
(en pourcentage de bascophones)
80-100%
60-80%
40-60%
20-40%
0-20%
Distribution of Basque speakers

Basque nowadays

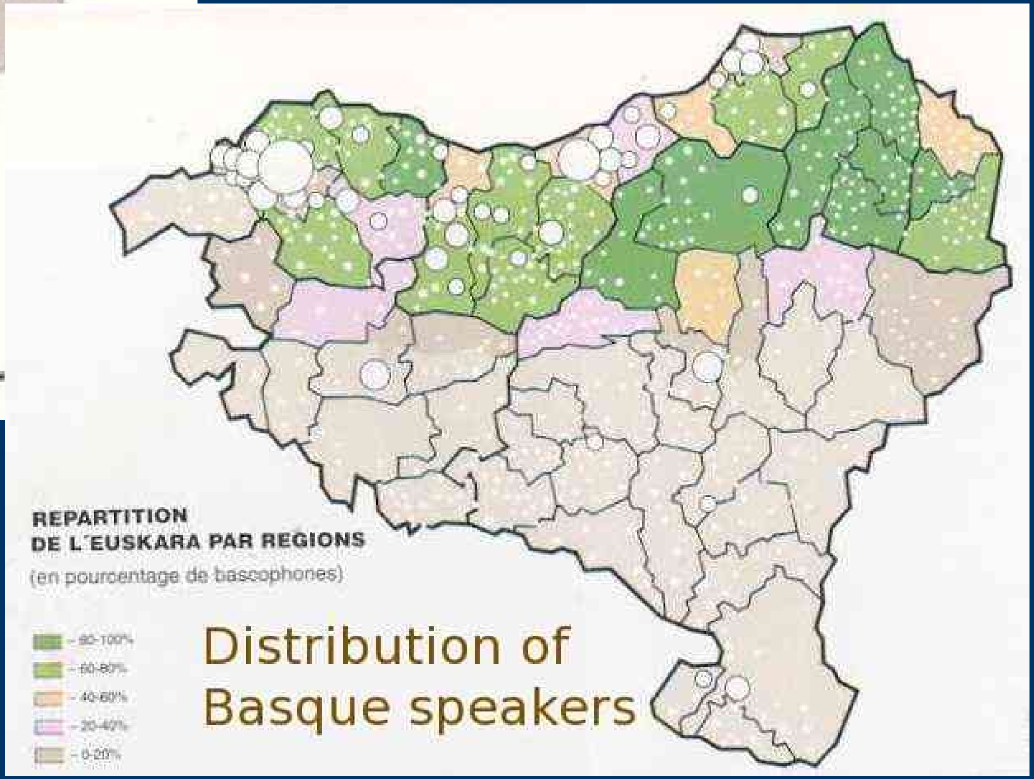
1,033,900 Speakers

(First lang.: 700,000)

Non homogeneous distribution !



Six different dialects !



Distribution of Basque speakers

Main reasons of Basque regression.

- No official language
- Out of the education system
- 6 dialects!
- Out of media
- Out of industry

Main reasons of Basque regression

		But since 1980...
• No official language	→	Coofficial language
• Out of the education system	→	Integrated in education (even at university)
• 6 dialects!	→	Unified Basque (1966)
• Out of media	→	TV, newspaper...
• Out of industry	→	Out of new ICTs ???

Basque. Linguistic features: Agglutinative language

Case	Undet.	Det.sing.	Det.Pl.	CloserPl.
Absolutive	*katu*	*katua*	*katuak*	*katuok*
Ergative	*katuk*	*katuak*	*katuek*	*katuok*
Dative	*katuri*	*katuari*	*katuei*	*katuoi*
Genitive1	*katuren*	*katuaren*	*katuen*	*katuon*
Associative	*katurekin*	*katuarekin*	*katuekin*	*katuokin*
…				
…				
…	***~with cat***	***with the cat***	***with the cats***	***~with these cats***

14 different cases

In fact, at least 360 possible word forms for each lemma

In theory, more than one million word forms are possible for each lemma

Basque. Linguistic features:

Case suffixes and free order of components

- Case suffixes and free order of sentence components

The dog brought the newspaper in his mouth

Txakur-rak	egunkari-a	aho-an	zekarren.
The-dog	the-newspaper	in-his-mouth	brought
ergative-3-s	absolutive-3-s	inessive-3-s	
Subject	Object	Modifier	Verb

Alternative possible orders:

Txakur-rak	aho-an	egunkari-a	zekarren.
Txakur-rak	aho-an	zekarren	egunkari-a.
Egunkari-a	txakur-rak	zekarren	aho-an.

...

Basque. Linguistic features:

Ergative language & multiple agreement

- Ergative case. Subject of transitive verbs

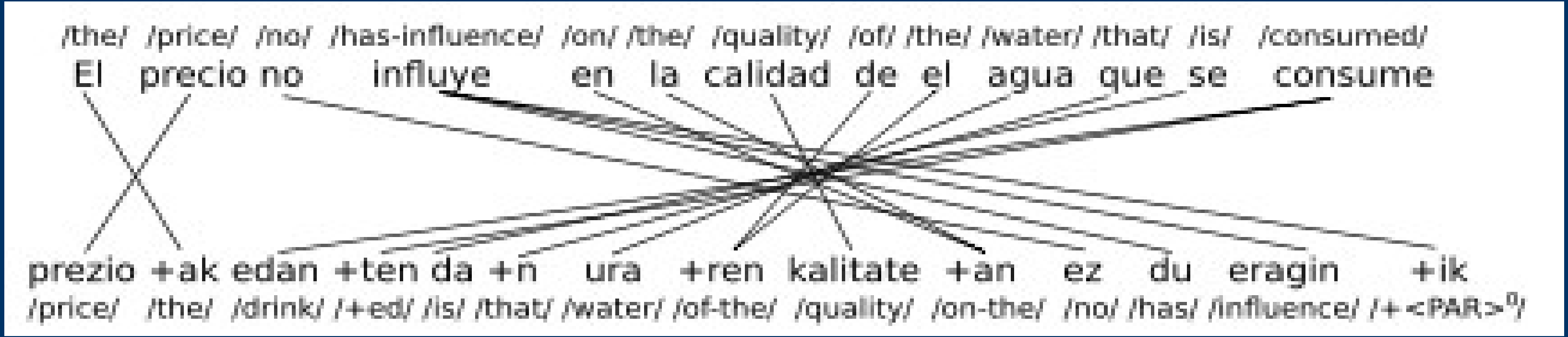
– I am	Ni naiz	(absolutive)
– I saw the cat	Nik katua ikusi nuen	(ergative)

- Agreement in number and person between verb and (subject, object and indirect object)

I saw the cat	Nik katua ikusi nuen
I saw the cats	Nik katuak ikusi nituen
I saw you	Nik zu ikusi zintudan

Basque. Linguistic features and MT

- Basque morphology and Syntax are very different comparing with Spanish, English, French, Catalan or Galician.
  - Rich morphology
  - Different component order at noun phrase level.


/the/ /price/ /no/ /has-influence/ /on/ /the/ /quality/ /of/ /the/ /water/ /that/ /is/ /consumed/
El precio no influye en la calidad de el agua que se consume
prezio +ak edan +ten da +n ura +ren kalitate +an ez du eragin +ik
/price/ /the/ /drink/ /+ed/ /is/ /that/ /water/ /of-the/ /quality/ /on-the/ /no/ /has/ /influence/ /+<PAR>0/

  - Free-order of components at sentence level.

=> Translating to Basque is more difficult!

Outline

- Basque : a Less Resourced Language (LRL)
- Strategy for sustainable HLT (and MT) for Basque
- Machine Translation for Basque (Matxin)
- Evaluation of Matxin
- Future: Combining RBMT and Corpus-based MT
- Recent elements and conclusions

Strategy to develop HLT in Basque

IXA Research Group

- 1986: 4-5 university lecturers (computer science)
- 2009: Interdisciplinary team
  - *32 computer scientists*
    - 19 lecturers (15 doctors)
    - 4 researchers
    - 9 PhD students (research grants)
  - *8 linguists*
    - 6 lecturers (4 doctors)
    - 2 PhD students (research grants)
  - 2 research assistants assigned to projects

http://ixa.si.ehu.es

IXA Group. Milestones

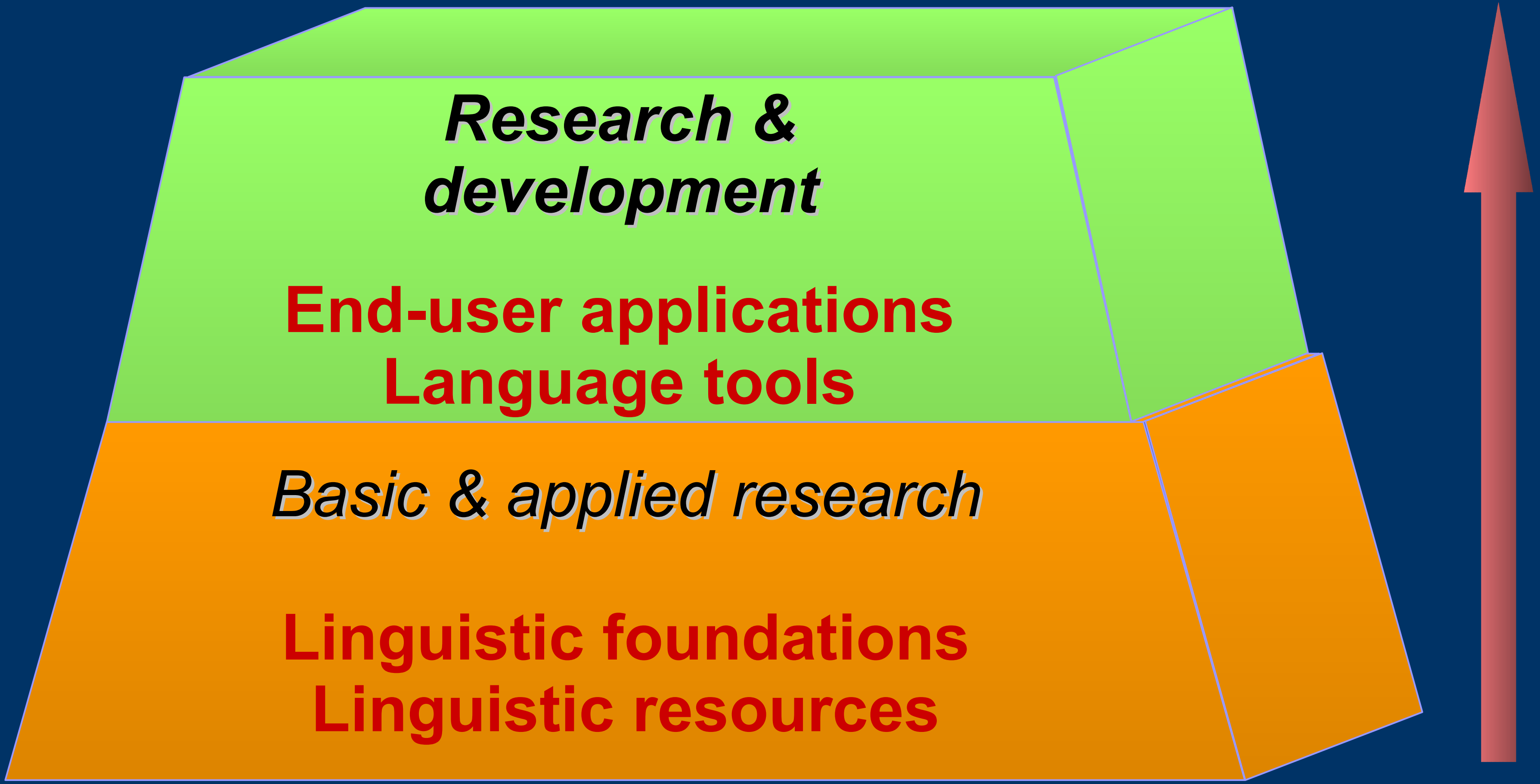
1987 1990 1995 2000 2007

Projects	Province Gov.	Basque Gov.	Madrid Cicyt	Europa (Meaning)	Basque G. Industry	Europe (IE-IR) Madrid (MT)	
Companies Basque C.	UZEI	Eusenor	Plazagune	Elhuyar	ASP	Diana Vicomtech Robotiker	ArgazkiPress
Companies abroad	Microsoft	Eatoni	Lexiquest	Irion	Scansoft	Prompsit	Imaxin
Spin-off companies	Eleka						
Products	Spelling checker	EDBL Lexical DB	Lemmatizer	Parser	BasqueIWordnet	MT-system	

Underlying strategy

- Need of standardization of resources to be useful:
  - in different researches
  - in different tools
  - in different applications
- Need of incremental design and development of language foundations, tools, and applications
  - in a parallel and coordinated way
  - in order to get the best benefit from them

Strategic priorities: from basic research to application development



Research & development

End-user applications
Language tools

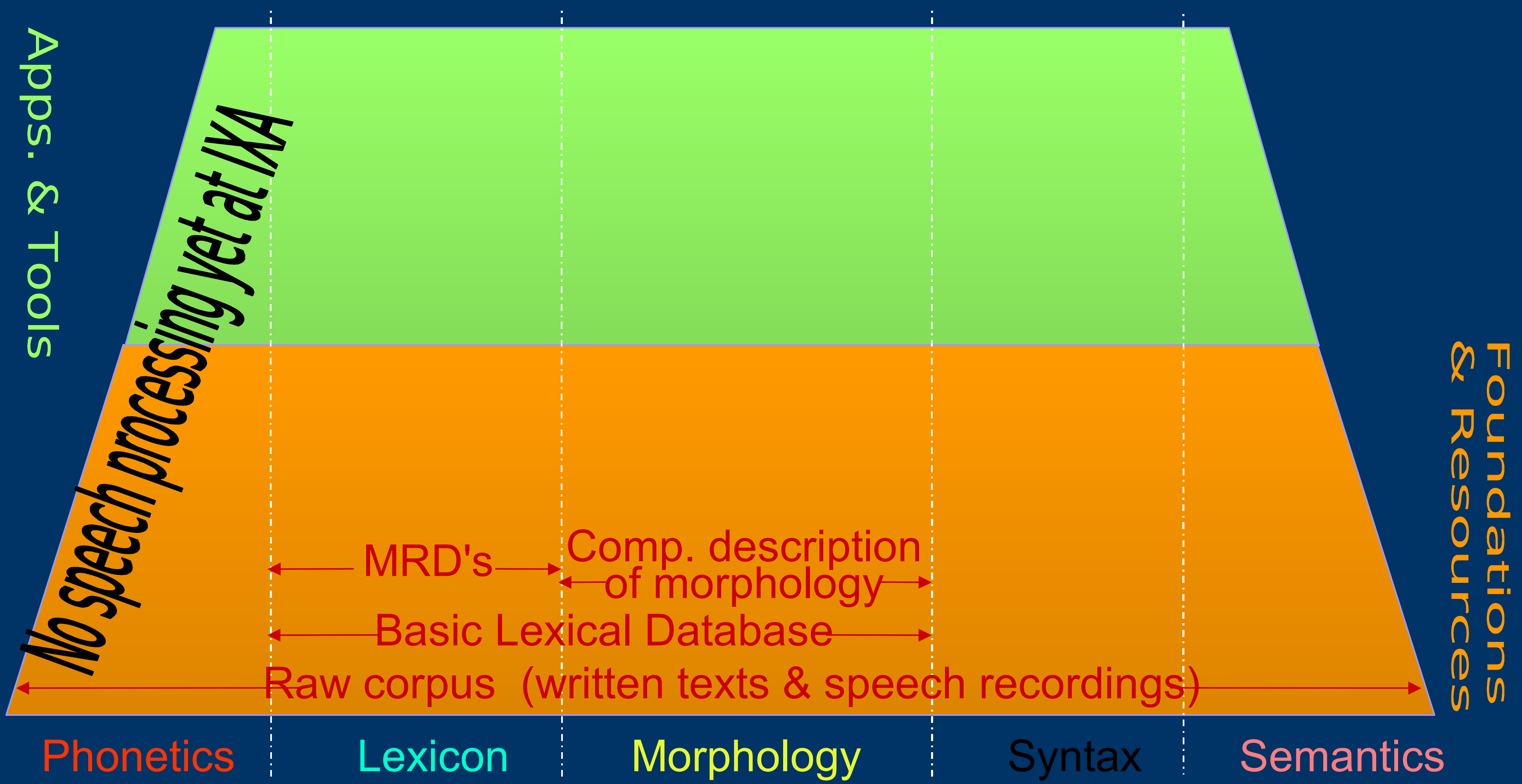
Basic & applied research

Linguistic foundations
Linguistic resources

Linguistic foundations & resources, tools and applications

- Linguistic foundations and resources: necessary infrastructure for the automatic processing of a language.
- Tools: mainly intended to application developers.
- Applications: commercial or non-commercial, for non-specialised end-users.

Phase I: laying foundations



Apps. & Tools

No speech processing yet at IXA

MRD's

Comp. description of morphology

Basic Lexical Database

Raw corpus (written texts & speech recordings)

Foundations & Resources

Phonetics

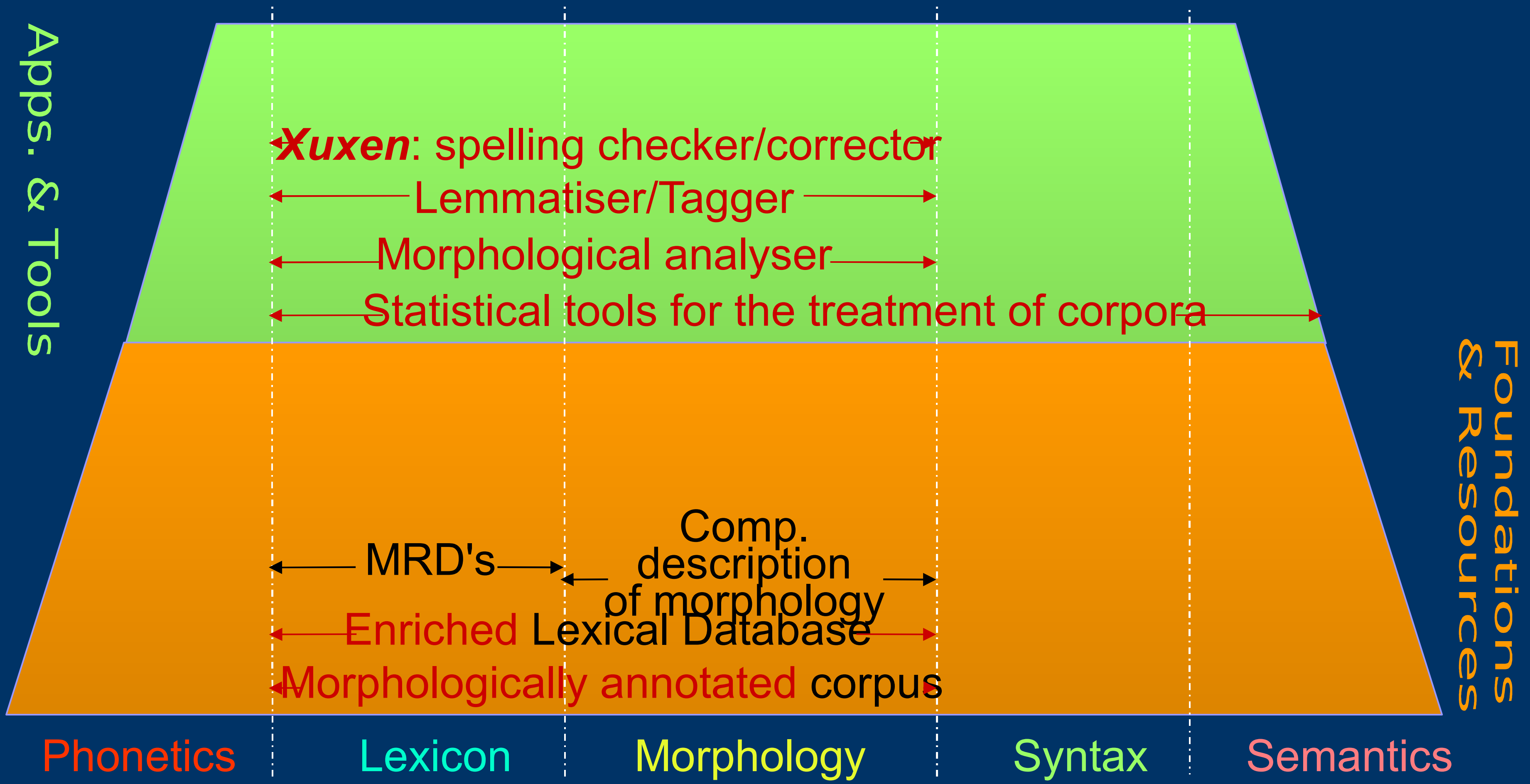
Lexicon

Morphology

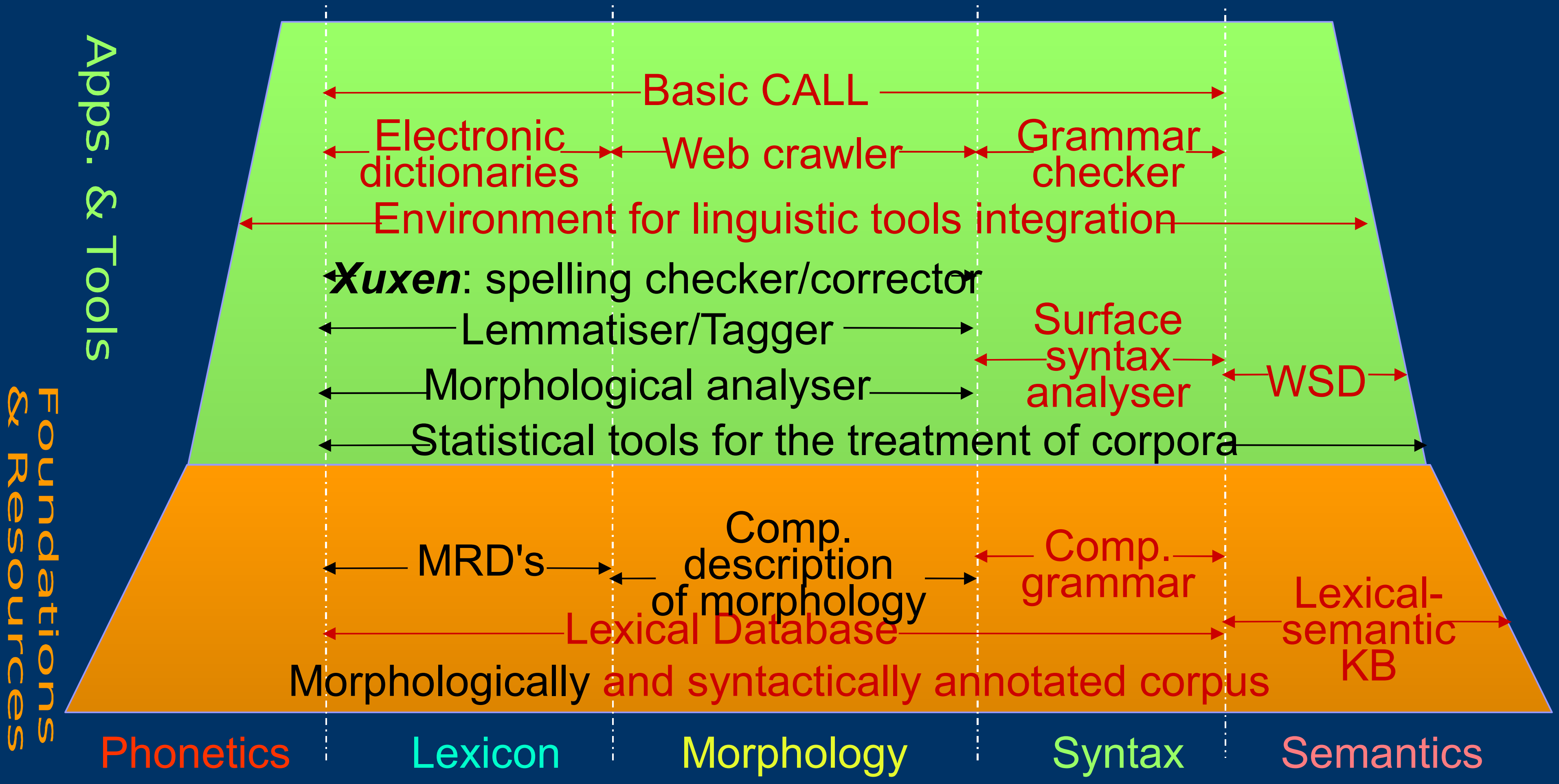
Syntax

Semantics

Phase II: *first basic tools and applications*


Apps. & Tools
Xuxen: spelling checker/corrector
Lemmatiser/Tagger
Morphological analyser
Statistical tools for the treatment of corpora
MRD's
Comp. description of morphology
Enriched Lexical Database
Morphologically annotated corpus
Foundations & Resources
Phonetics
Lexicon
Morphology
Syntax
Semantics

Phase III: more advanced tools and applications



Apps. & Tools

Basic CALL

Electronic dictionaries

Web crawler

Grammar checker

Environment for linguistic tools integration

Xuxen: spelling checker/corrector

Lemmatiser/Tagger

Surface syntax analyser

Morphological analyser

WSD

Statistical tools for the treatment of corpora

Foundations & Resources

MRD's

Comp. description of morphology

Comp. grammar

Lexical Database

Lexical-semantic KB

Morphologically and syntactically annotated corpus

Phonetics

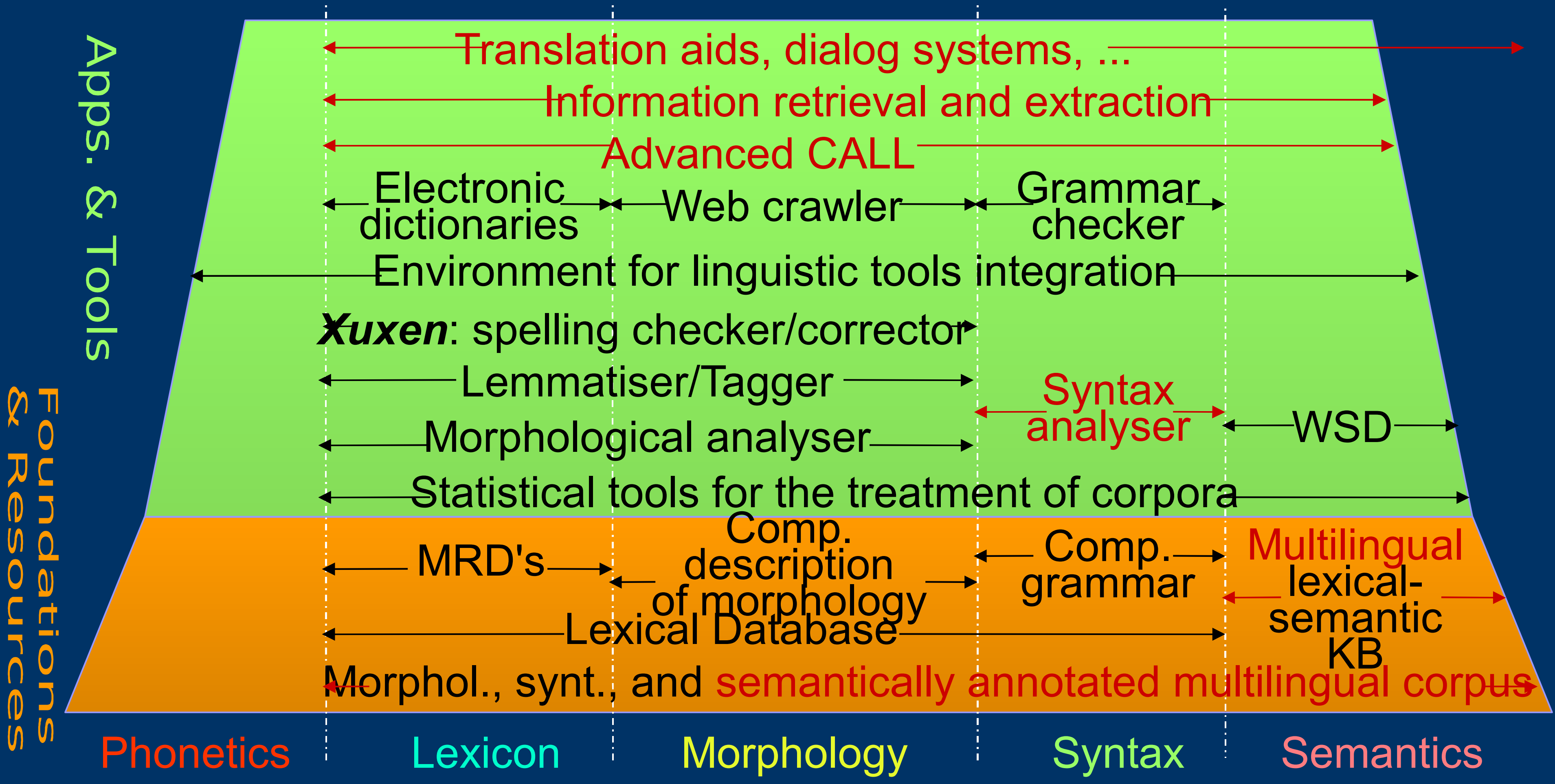
Lexicon

Morphology

Syntax

Semantics

Phase IV: multilinguality and general applications



Apps. & Tools

Translation aids, dialog systems, ...

Information retrieval and extraction

Advanced CALL

Electronic dictionaries

Web crawler

Grammar checker

Environment for linguistic tools integration

Xuxen: spelling checker/corrector

Lemmatiser/Tagger

Syntax analyser

Morphological analyser

WSD

Statistical tools for the treatment of corpora

Foundations & Resources

MRD's

Comp. description of morphology

Comp. grammar

Multilingual lexical-semantic KB

Lexical Database

Morphol., synt., and semantically annotated multilingual corpus

Phonetics

Lexicon

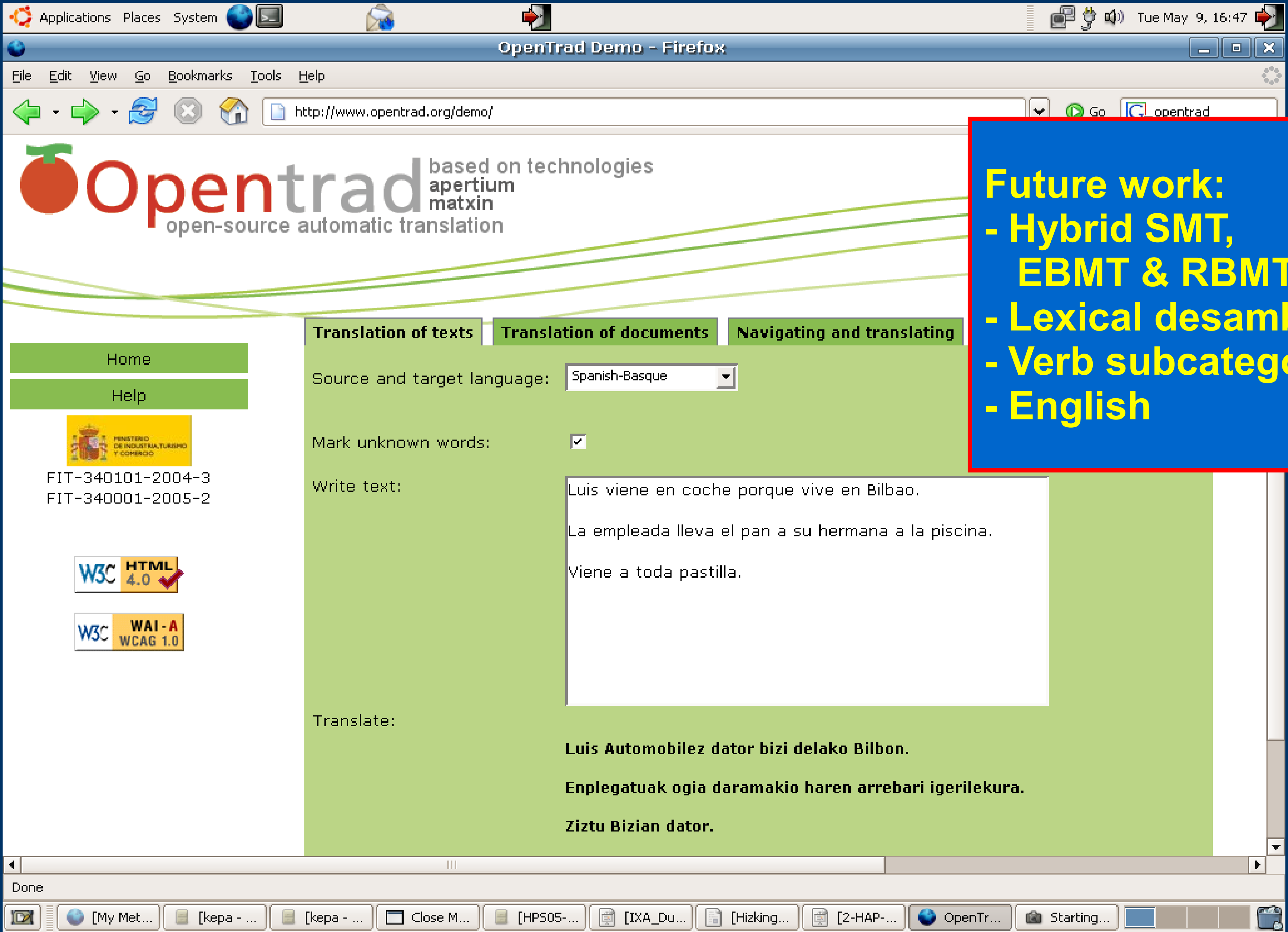
Morphology

Syntax

Semantics

Applications

- *Spelling checker/corrector*
- *3 lemmatization based on-line bilingual /monolingual dictionaries*
- *Lemmatization based on-line dictionary of synonyms*
- *Lemmatization based search machine*
- *Basque Wordnet*
- *Spanish-Basque transfer based MT system (Matxin)*



FIT-340101-2004-3
FIT-340001-2005-2

Source and target language: Spanish-Basque

Mark unknown words: ☑

Write text:

Luis viene en coche porque vive en Bilbao.

La empleada lleva el pan a su hermana a la piscina.

Viene a toda pastilla.

Translate:

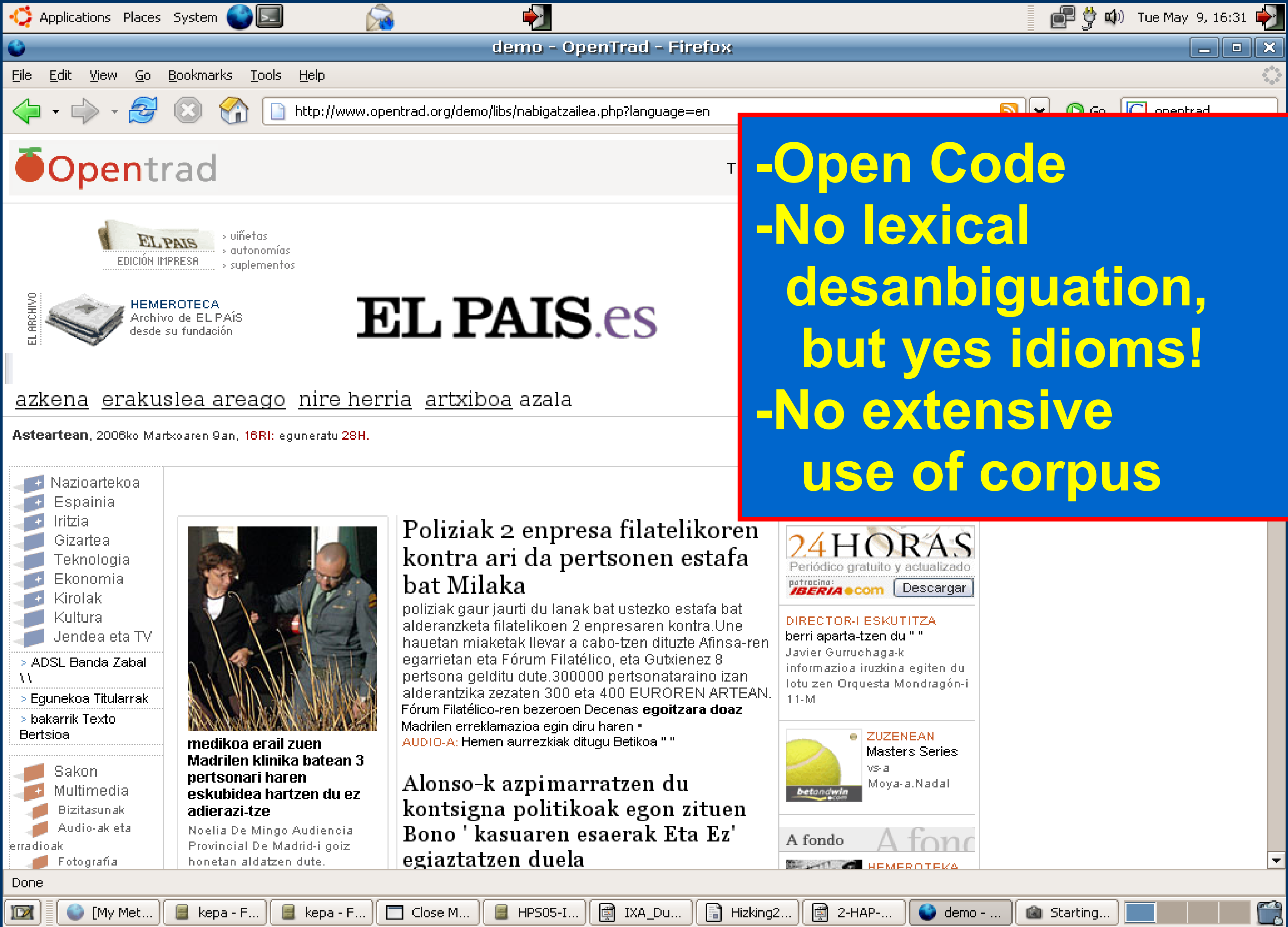
Luis Automobilez dator bizi delako Bilbon.

Enplegatuak ogia daramakio haren arrebari igerilekura.

Ziztu Bizian dator.

Future work:
- **Hybrid SMT,**
  EBMT & RBMT
- **Lexical desambiguation**
- **Verb subcategorization**
- **English**

Spanish-Basque transfer MT


Applications Places System
Tue May 9, 16:31
demo - OpenTrad - Firefox
File Edit View Go Bookmarks Tools Help
Go
opentrad
Done
[My Met... kepa - F... kepa - F... Close M... HPS05-I... IXA_Du... Hizking2... 2-HAP-... demo - ... Starting...

http://www.opentrad.org/demo/libs/nabigatzailea.php?language=en

Opentrad

EL PAIS EDICIÓN IMPRESA › viñetas › autonomías › suplementos

EL ARCHIVO HEMEROTECA Archivo de EL PAÍS desde su fundación

EL PAIS.es

azkena erakuslea areago nire herria artxiboa azala

Asteartean, 2006ko Martxoaren 9an, 16RI: eguneratu 28H.

Nazioartekoa
Espainia
Iritzia
Gizartea
Teknologia
Ekonomia
Kirolak
Kultura
Jendea eta TV

› ADSL Banda Zabal \ \

› Egunekoa Titularrak

› bakarrik Texto Bertsioa

Sakon
Multimedia
Bizitasunak
Audio-ak eta erradioak
Fotografia

medikoa erail zuen Madrilen klinika batean 3 pertsonari haren eskubidea hartzen du ez adierazi-tze

Noelia De Mingo Audiencia Provincial De Madrid-i goiz honetan aldatzen dute.

Poliziak 2 enpresa filatelikoren kontra ari da pertsonen estafa bat Milaka

poliziak gaur jaurti du lanak bat ustezko estafa bat alderanzketa filatelikoen 2 enpresaren kontra.Une hauetan miaketak llevar a cabo-tzen dituzte Afinsa-ren egarrietan eta Fórum Filatélico, eta Gutxienez 8 pertsona gelditu dute.300000 pertsonataraino izan alderantzika zezaten 300 eta 400 EUROREN ARTEAN.

Fórum Filatélico-ren bezeroen Decenas **egoitzara doaz**

Madrilen erreklamazioa egin diru haren •

AUDIO-A: Hemen aurrezkiak ditugu Betikoa " "

Alonso-k azpimarratzen du kontsigna politikoak egon zituen Bono ' kasuaren esaerak Eta Ez' egiaztatzen duela

24HORAS Periódico gratuito y actualizado patrocina: IBERIA.com Descargar

DIRECTOR-I ESKUTITZA

berri aparta-tzen du " "

Javier Gurruchaga-k informazioa iruzkina egiten du lotu zen Orquesta Mondragón-i 11-M

ZUZENEAN

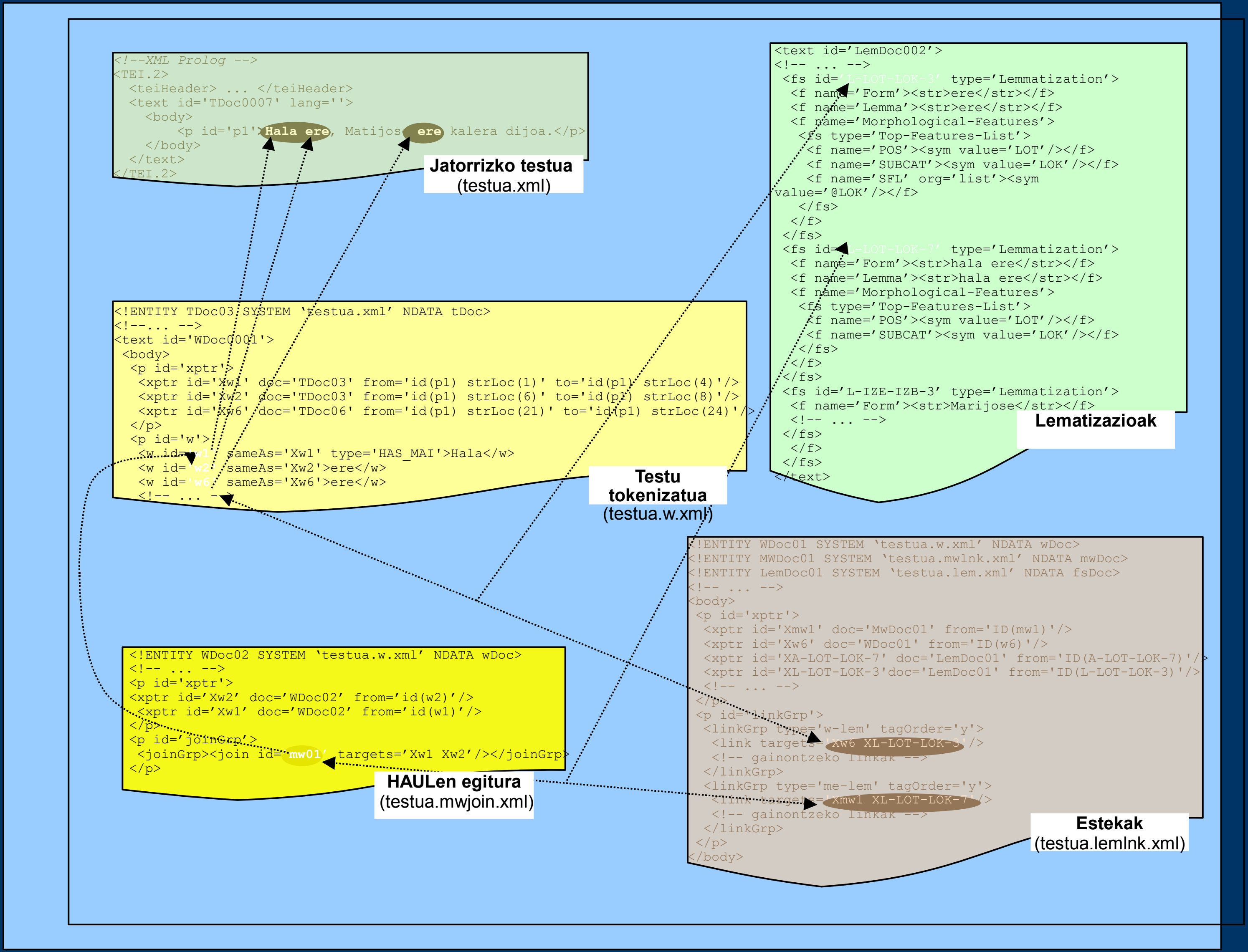
Masters Series vs-a Moya-a.Nadal

A fondo

HEMEROTEKA

-Open Code
-No lexical
  desanbiguation,
  but yes idioms!
-No extensive
  use of corpus

Spanish-Basque transfer MT



Jatorrizko testua
(testua.xml)

```
<!--XML Prolog -->
<TEI.2>
  <teiHeader> ... </teiHeader>
  <text id='TDoc0007' lang=''>
    <body>
        <p id='p1'>Hala ere, Matijose ere kalera dijoa.</p>
    </body>
  </text>
</TEI.2>
```

Testu tokenizatua
(testua.w.xml)

```
<!ENTITY TDoc03 SYSTEM ‘testua.xml’ NDATA tDoc>
<!--... -->
<text id='WDoc0001'>
 <body>
  <p id='xptr'>
   <xptr id='Xw1' doc='TDoc03' from='id(p1) strLoc(1)' to='id(p1) strLoc(4)'/>
   <xptr id='Xw2' doc='TDoc03' from='id(p1) strLoc(6)' to='id(p1) strLoc(8)'/>
   <xptr id='Xw6' doc='TDoc06' from='id(p1) strLoc(21)' to='id(p1) strLoc(24)'/>
  </p>
  <p id='w'>
   <w id='w1' sameAs='Xw1' type='HAS_MAI'>Hala</w>
   <w id='w2' sameAs='Xw2'>ere</w>
   <w id='w6' sameAs='Xw6'>ere</w>
   <!-- ... -->
```

Lematizazioak

```
<text id=’LemDoc002’>
<!-- ... -->
 <fs id=’L-LOT-LOK-3’ type=’Lemmatization’>
  <f name=’Form’><str>ere</str></f>
  <f name=’Lemma’><str>ere</str></f>
  <f name=’Morphological-Features’>
   <fs type=’Top-Features-List’>
    <f name=’POS’><sym value=’LOT’/></f>
    <f name=’SUBCAT’><sym value=’LOK’/></f>
    <f name=’SFL’ org=’list’><sym
value=’@LOK’/></f>
   </fs>
  </f>
 </fs>
 <fs id=’L-LOT-LOK-7’ type=’Lemmatization’>
  <f name=’Form’><str>hala ere</str></f>
  <f name=’Lemma’><str>hala ere</str></f>
  <f name=’Morphological-Features’>
   <fs type=’Top-Features-List’>
    <f name=’POS’><sym value=’LOT’/></f>
    <f name=’SUBCAT’><sym value=’LOK’/></f>
   </fs>
  </f>
 </fs>
 <fs id=’L-IZE-IZB-3’ type=’Lemmatization’>
  <f name=’Form’><str>Marijose</str></f>
  <!-- ... -->
 </fs>
  </f>
 </fs>
</text>
```

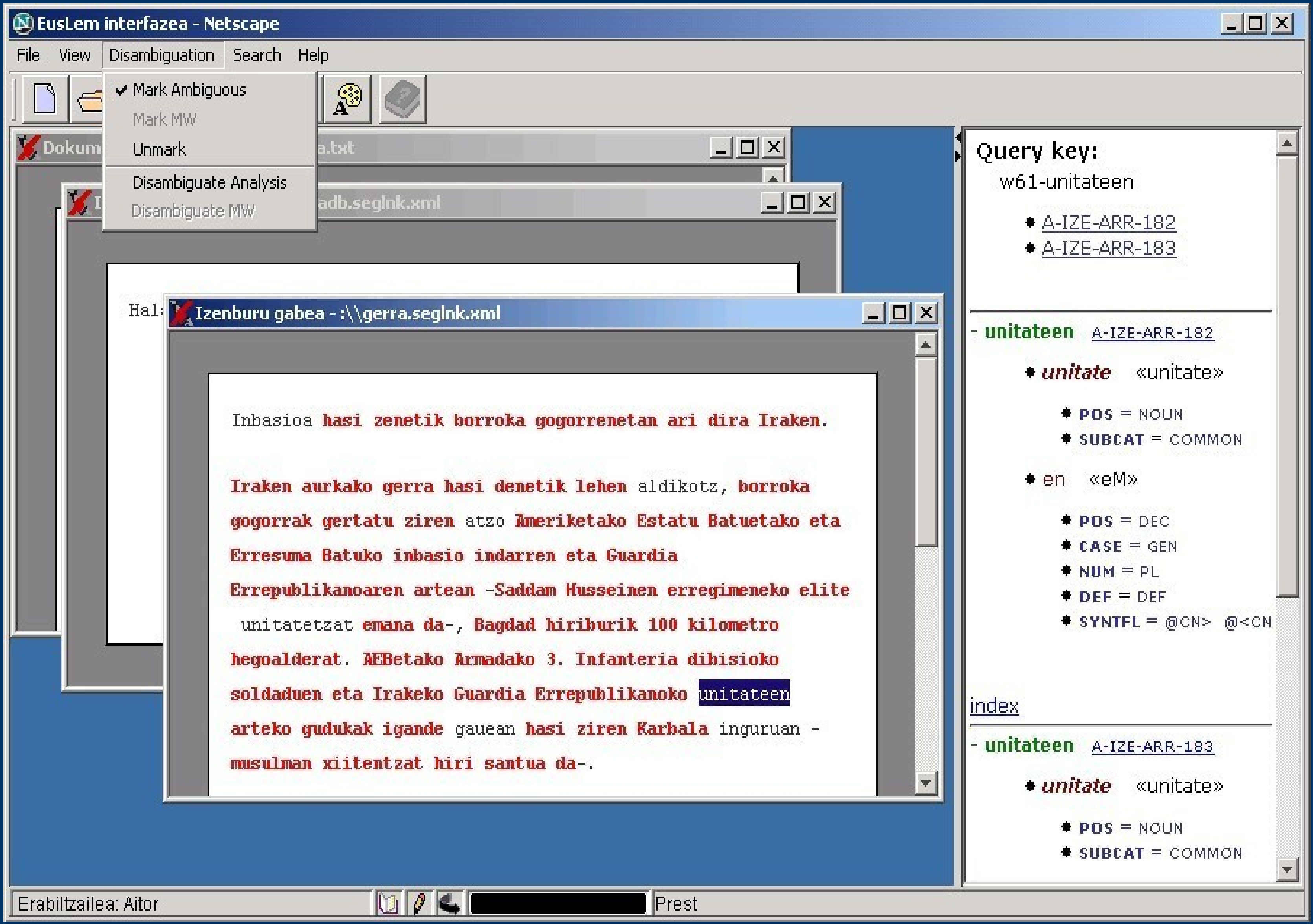
HAULen egitura
(testua.mwjoin.xml)

```
<!ENTITY WDoc02 SYSTEM ‘testua.w.xml’ NDATA wDoc>
<!-- ... -->
<p id='xptr'>
<xptr id=’Xw2’ doc=’WDoc02’ from=’id(w2)’/>
 <xptr id=’Xw1’ doc=’WDoc02’ from=’id(w1)’/>
</p>
<p id=’joinGrp’>
 <joinGrp><join id=’mw01’ targets=’Xw1 Xw2’/></joinGrp>
</p>
```

Estekak
(testua.lemlnk.xml)

```
<!ENTITY WDoc01 SYSTEM ‘testua.w.xml’ NDATA wDoc>
<!ENTITY MWDoc01 SYSTEM ‘testua.mwlnk.xml’ NDATA mwDoc>
<!ENTITY LemDoc01 SYSTEM ‘testua.lem.xml’ NDATA fsDoc>
<!-- ... -->
<body>
 <p id='xptr'>
  <xptr id='Xmw1' doc='MwDoc01' from='ID(mw1)'/>
  <xptr id='Xw6' doc='WDoc01' from='ID(w6)'/>
  <xptr id='XA-LOT-LOK-7' doc='LemDoc01' from='ID(A-LOT-LOK-7)'/>
  <xptr id='XL-LOT-LOK-3'doc='LemDoc01' from='ID(L-LOT-LOK-3)'/>
  <!-- ... -->
 </p>
 <p id='linkGrp'>
  <linkGrp type='w-lem' tagOrder='y'>
   <link targets='Xw6 XL-LOT-LOK-3'/>
   <!-- gainontzeko linkak -->
  </linkGrp>
  <linkGrp type='me-lem' tagOrder='y'>
   <link targets='Xmw1 XL-LOT-LOK-7'/>
   <!-- gainontzeko linkak -->
  </linkGrp>
 </p>
</body>
```

Methodology for stand-off corpus tagging (TEI, feature structures and XML)



Inbasioa hasi zenetik borroka gogorrenetan ari dira Iraken.

Iraken aurkako gerra hasi denetik lehen aldikotz, borroka gogorrak gertatu ziren atzo Ameriketako Estatu Batuetako eta Erresuma Batuko inbasio indarren eta Guardia Errepublikanoaren artean -Saddam Husseinen erregimeneko elite unitatetzat emana da-, Bagdad hiriburik 100 kilometro hegoalderat. AEBetako Armadako 3. Infanteria dibisioko soldaduen eta Irakeko Guardia Errepublikanoko unitateen arteko gudukak igande gauean hasi ziren Karbala inguruan -musulman xiitentzat hiri santua da-.

Query key:
w61-unitateen

- A-IZE-ARR-182
- A-IZE-ARR-183

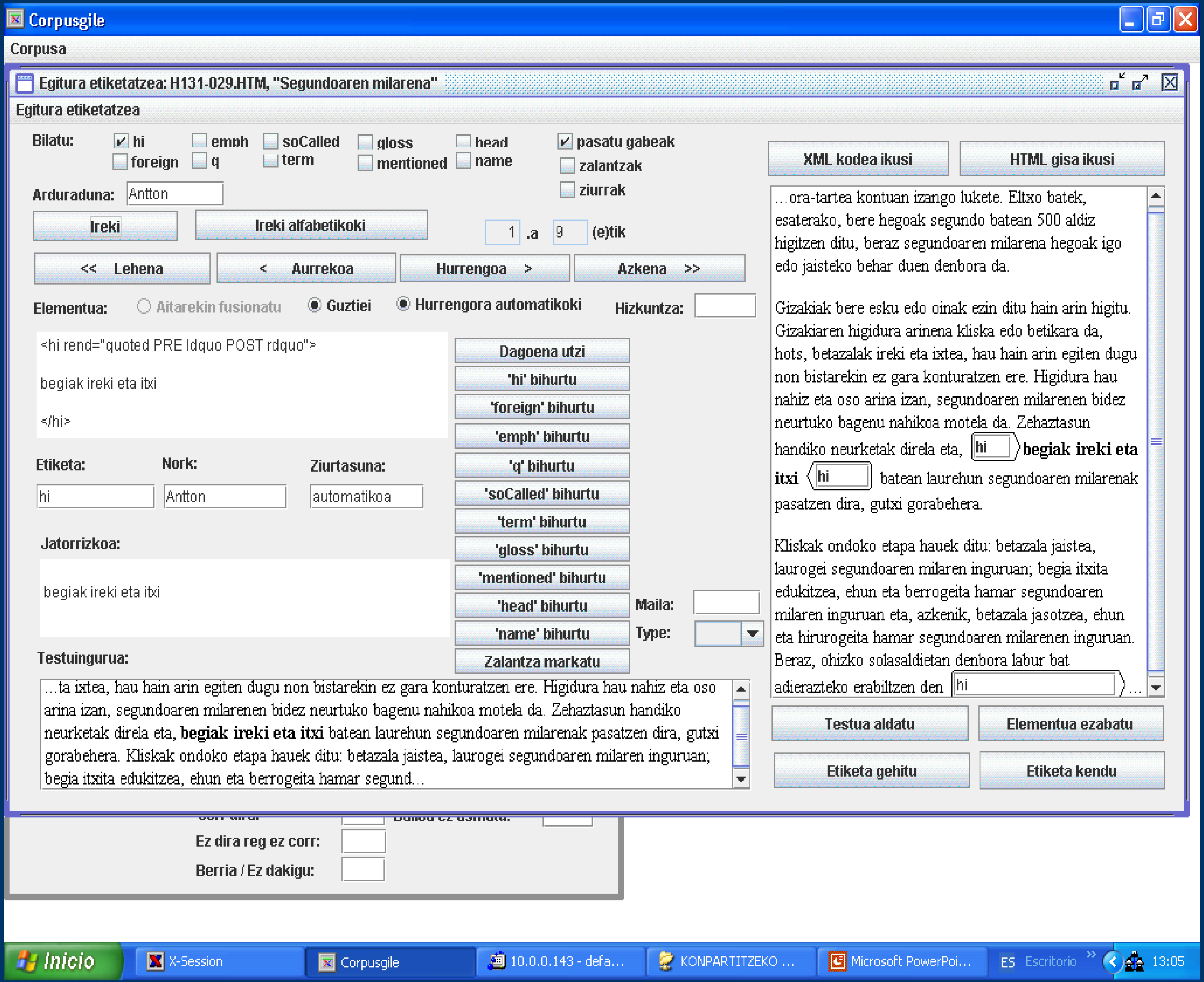
- unitateen A-IZE-ARR-182
  - *unitate* «unitate»
    - POS = NOUN
    - SUBCAT = COMMON
  - en «eM»
    - POS = DEC
    - CASE = GEN
    - NUM = PL
    - DEF = DEF
    - SYNTFL = @CN> @<CN

index

- unitateen A-IZE-ARR-183
  - *unitate* «unitate»
    - POS = NOUN
    - SUBCAT = COMMON

EULIA: tool for monolingual corpus tagging

(EULIBELTZ: tool for bilingual corpus tagging)



Bilatu: ☑ hi ☐ foreign ☐ emph ☐ q ☐ soCalled ☐ term ☐ gloss ☐ mentioned ☐ head ☐ name ☑ pasatu gabeak ☐ zalantzak ☐ ziurrak

Arduraduna: Antton

1 .a 9 (e)tik

Elementua: ○ Aitarekin fusionatu ● Guztiei ● Hurrengora automatikoki

Hizkuntza:

```
<hi rend="quoted PRE ldquo POST rdquo">

begiak ireki eta itxi

</hi>
```

Etiketa: hi

Nork: Antton

Ziurtasuna: automatikoa

Jatorrizkoa:

begiak ireki eta itxi

Maila:

Type:

Testuingurua:

...ta ixtea, hau hain arin egiten dugu non bistarekin ez gara konturatzen ere. Higidura hau nahiz eta oso arina izan, segundoaren milarenen bidez neurtuko bagenu nahikoa motela da. Zehaztasun handiko neurketak direla eta, **begiak ireki eta itxi** batean laurehun segundoaren milarenak pasatzen dira, gutxi gorabehera. Kliskak ondoko etapa hauek ditu: betazala jaistea, laurogei segundoaren milaren inguruan; begia itxita edukitzea, ehun eta berrogeita hamar segund...

...ora-tartea kontuan izango lukete. Eltxo batek, esaterako, bere hegoak segundo batean 500 aldiz higitzen ditu, beraz segundoaren milarena hegoak igo edo jaisteko behar duen denbora da.

Gizakiak bere esku edo oinak ezin ditu hain arin higitu. Gizakiaren higidura arinena kliska edo betikara da, hots, betazalak ireki eta ixtea, hau hain arin egiten dugu non bistarekin ez gara konturatzen ere. Higidura hau nahiz eta oso arina izan, segundoaren milarenen bidez neurtuko bagenu nahikoa motela da. Zehaztasun handiko neurketak direla eta, hi **begiak ireki eta itxi** hi batean laurehun segundoaren milarenak pasatzen dira, gutxi gorabehera.

Kliskak ondoko etapa hauek ditu: betazala jaistea, laurogei segundoaren milaren inguruan; begia itxita edukitzea, ehun eta berrogeita hamar segundoaren milaren inguruan eta, azkenik, betazala jasotzea, ehun eta hirurogeita hamar segundoaren milarenen inguruan. Beraz, ohizko solasaldietan denbora labur bat adierazteko erabiltzen den hi ...

Ez dira reg ez corr:

Berria / Ez dakigu:

CORPUSGILE:
tool for compiling and consulting corpus

Outline

- **Basque : a Less Resourced Language (LRL)**
- **Strategy for sustainable HLT (and MT) for Basque**
- **Machine Translation for Basque (RBMT:Matxin)**
- **Evaluation of Matxin**
- **Future: Combining RBMT and Corpus-based MT**
- **Recent elements and conclusions**

The RBMT approach

- Since 2000, after years working on basic resources and tools, we faced MT from Spanish or English to Basque.
- Design of the MT system:
  - Reusability of previous resources: lexical resources, morphology of Basque, parsing of Spanish and English.
  - Standardization and collaboration: General framework useful for other language pairs and groups. Spanish, Galician and Catalan.
  - Open-source: Anyone having the necessary computational and linguistic skills will be able to adapt or enhance our system.

The RBMT approach

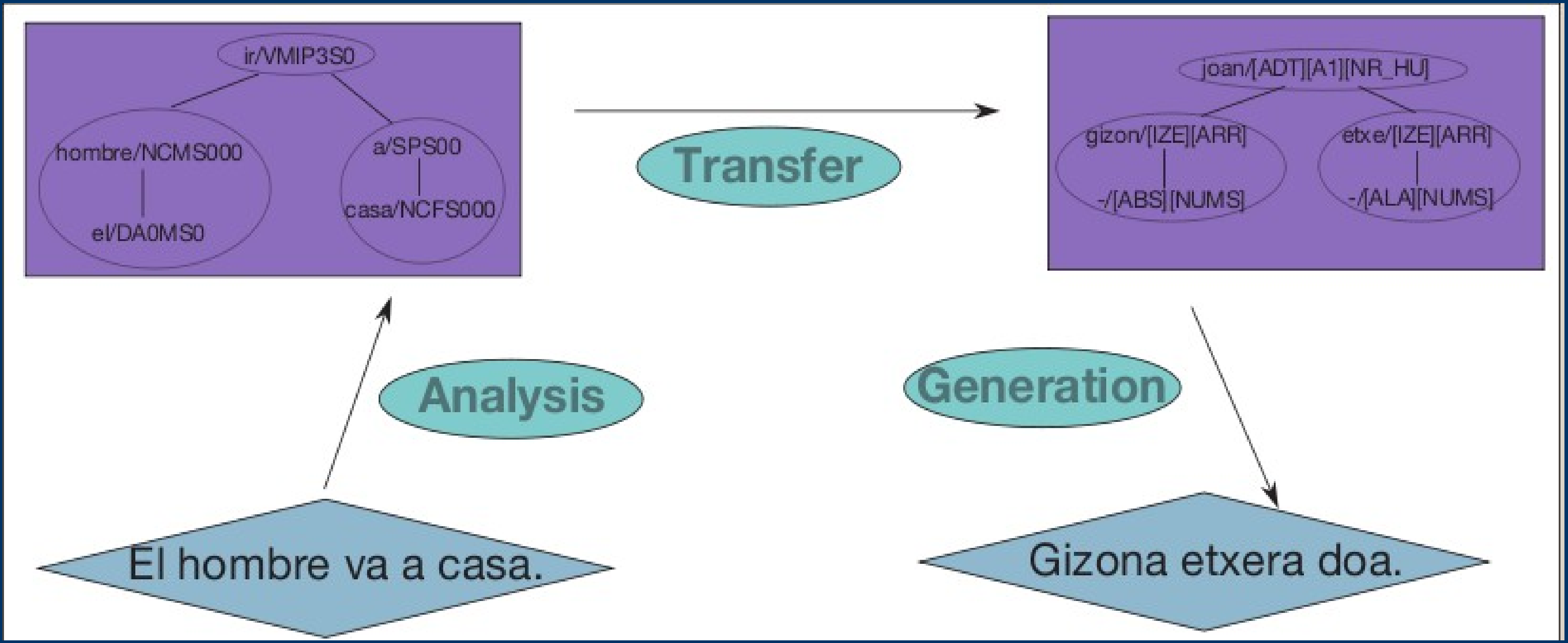
- Integrated in OpenTrad initiative (www.opentrad.com):
  - Open, reusable and interoperable framework.
  - Translation among the four main languages in Spain.
- Design and programs are language independent
  - Depending on the language pair it might be necessary to add, reorder and change some modules
  - but it will not be difficult because a unique XML format is used for the communication among all the modules.
  - Present work: New unified formalism to represent transfer and generation rules (Mayor &Tyers, 2009)

The RBMT approach: Opentrad-Matxin

Two different designs in OpenTrad

- Apertium (apertium.sourceforge.net)
  - Shallow-transfer MT engine for pairs of similar languages (Spanish, Catalan and Galician...).
  - The MT architecture uses
  - finite-state transducers for lexical processing,
  - hidden Markov models for part-of-speech tagging,
  - and finite-state based chunking for structural transfer
- Matxin (matxin.sourceforge.net)
  - A deeper-transfer engine for the Spanish-Basque pair.
  - Some modules, data formats and compilers from Apertium
  - The Spanish analysis module is FreeLing (Carreras et al., 2004). Another open source engine

The rule based approach.
Matxin design: Spanish-Basque



The RBMT approach: Spanish-Basque

- **Analysis:**
  - the Freeling toolkit to carry out the Spanish parsing
- **Tranfer**
  - lexical transfer: a bilingual dictionary is reused
  - syntactic transfer: tree transformation rules
- **Generation**
  - syntactical generation: the order of the dependency tree elements is redefined.
  - lexical generation: the word forms are generated, adding suffixes with morphological information to the lemmas. A previous morphological analyser is reused.

Outline

- Basque : a Less Resourced Language (LRL)
- Strategy for sustainable HLT (and MT) for Basque
- Machine Translation for Basque (Matxin)
- Evaluation of Matxin
- Future: Combining RBMT and Corpus-based MT
- Recent elements and conclusions

The RBMT approach
Evaluation of Matxin

The results for the Spanish-Basque RBMT system using *FreeLing* and *Matxin* are acceptables (Mayor, 2007)

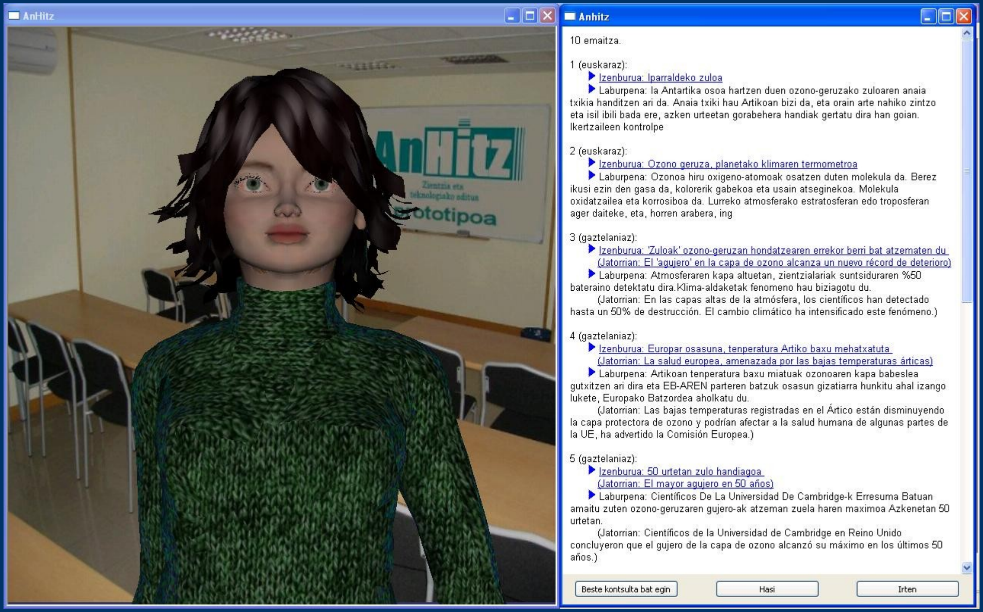
40.41 editing corrections are required for every 100 tokens.

	BLEU	Edit-distance TER
Corpus1 (newspapers)	9.30	40.41
Corpus2 (web magazine)	6.31	43.60

The RBMT approach
Evaluation in context *(IE-IR, MT, ASR, TTS)*

Matxin is integrated in AnHitz, a virtual expert person in scientific and technological themes.

- With Question Answering and Cross Lingual IR systems.
- The interaction in Basque and is speech-based (ASR &TTS)
- Matxin translates not-Basque results of the CLIR module



10 emaitza.

1 (euskaraz):
▶ Izenburua: Iparraldeko zuloa
▶ Laburpena: la Antartika osoa hartzen duen ozono-geruzako zuloaren anaia txikia handitzen ari da. Anaia txiki hau Artikoan bizi da, eta orain arte nahiko zintzo eta isil ibili bada ere, azken urteetan gorabehera handiak gertatu dira han goian. Ikertzaileen kontrolpe

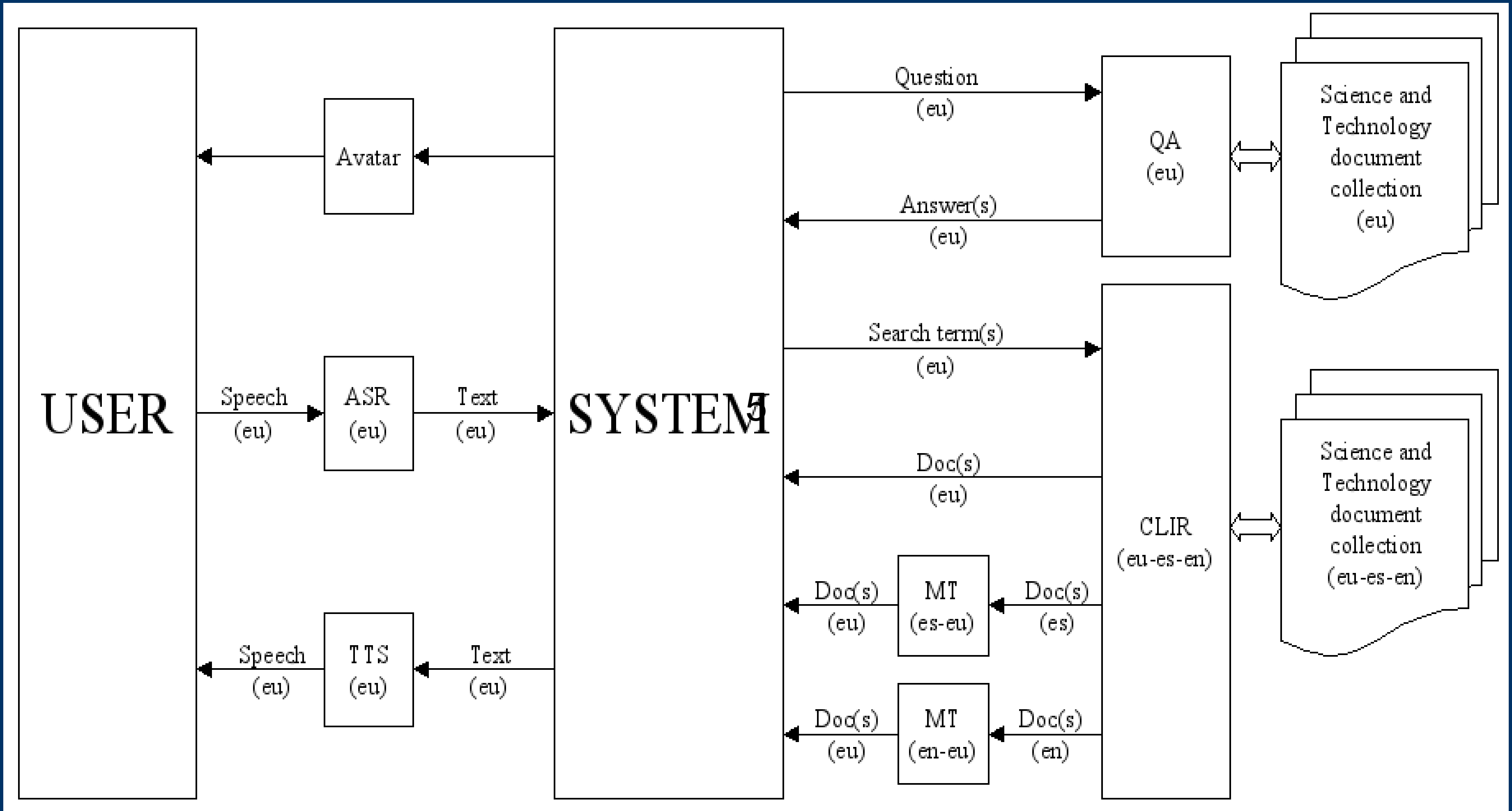
2 (euskaraz):
▶ Izenburua: Ozono geruza, planetako klimaren termometroa
▶ Laburpena: Ozonoa hiru oxigeno-atomoak osatzen duten molekula da. Berez ikusi ezin den gasa da, kolorerik gabekoa eta usain atseginekoa. Molekula oxidatzailea eta korrosiboa da. Lurreko atmosferako estratosferan edo troposferan ager daiteke, eta, horren arabera, ing

3 (gaztelaniaz):
▶ Izenburua: 'Zuloak' ozono-geruzan hondatzearen errekor berri bat atzematen du
(Jatorrian: El 'agujero' en la capa de ozono alcanza un nuevo récord de deterioro)
▶ Laburpena: Atmosferaren kapa altuetan, zientzialariak suntsiduraren %50 bateraino detektatu dira.Klima-aldaketak fenomeno hau biziagotu du.
(Jatorrian: En las capas altas de la atmósfera, los científicos han detectado hasta un 50% de destrucción. El cambio climático ha intensificado este fenómeno.)

4 (gaztelaniaz):
▶ Izenburua: Europar osasuna, tenperatura Artiko baxu mehatxatuta
(Jatorrian: La salud europea, amenazada por las bajas temperaturas árticas)
▶ Laburpena: Artikoan tenperatura baxu miatuak ozonoaren kapa babeslea gutxitzen ari dira eta EB-AREN parteren batzuk osasun gizatiarra hunkitu ahal izango lukete, Europako Batzordea aholkatu du.
(Jatorrian: Las bajas temperaturas registradas en el Ártico están disminuyendo la capa protectora de ozono y podrían afectar a la salud humana de algunas partes de la UE, ha advertido la Comisión Europea.)

5 (gaztelaniaz):
▶ Izenburua: 50 urtetan zulo handiagoa
(Jatorrian: El mayor agujero en 50 años)
▶ Laburpena: Científicos De La Universidad De Cambridge-k Erresuma Batuan amaitu zuten ozono-geruzaren gujero-ak atzeman zuela haren maximoa Azkenetan 50 urtetan.
(Jatorrian: Científicos de la Universidad de Cambridge en Reino Unido concluyeron que el gujero de la capa de ozono alcanzó su máximo en los últimos 50 años.)

The RBMT approach
Evaluation in context (IE-IR, MT, ASR, TTS)



The RBMT approach
Evaluation in context (IE-IR, MT, ASR, TTS)

Evaluation of Matxin integrated in AnHitz prototype
(Leturia et al., 2009)

50 users who have completed a total of 300 tests

- **30.00% : "very good", "good" or "quite good"**
- **38.89% : "comprehensible"**
- **31.11% : "quite bad", "bad" or "very bad"**

=> Matxin is useful in assimilation applications

AnHitz has good performance and acceptance

Outline

- Basque : a Less Resourced Language (LRL)
- Strategy for sustainable HLT (and MT) for Basque
- Machine Translation for Basque (Matxin)
- Evaluation of Matxin
- Future: Combining RBMT and Corpus-based MT
- New elements and conclusions

Milestones in MT

	1950	1960	1970	1980	1990	2000	2004	2006	2008	2010	2011
	1949, MT proposal		1970 Logos				2004, Apertium				
	1954,IBM	1966, ALPAC					2005, Matxin				
RBMT		1968 Systran						2007, Opentrad			
			1977, Meteo								
EBMT				1984, EBMT (Nagao)				2006, MaTrEx			
						2003, EBMT (Carl&Way)					
SMT					1991, SMT (IBM)		2004, Pharaoh		2008, GT		
						2001, Giza++		2006, Moses			
SMT Corpus							2005, Europarl (~30Mw per pair)				
SMT Corpus eu								2006, 1Mw es-eu	2009, 7Mw es-eu		
SMT Metrics						2001, BLEU		2006 BLEU? (Callisson-Burch)			
Hybrid systems								2007, (Multi Engine MEMT)			
								2007, Stat. post-edition (SPE)			
Postediting Tools									2009 GT's toolkit		
									2009, Firefox,WWL		

Combining RBMT and Corpus-based MT

Now we are working on two hybrid MT systems:

- MEMT : Multi-Engine MT
  - EBMT + SMT + RBMT
  - Needs: The three MT systems
    Confidence scores

- SPE: Statistical Post Edition
  - Statistical postediting of RBMT output
  - Needs: RBMT system
    Huge corpus: manual postediting of RBMT translations

MEMT : Multi-Engine MT

- We are working (Alegria eta al., 2008) on the construction of a MEMT system based on the different approaches to MT:

  EBMT + SMT + RBMT

- Specific domain: Labor Agreements
- Needs:
  - The three MT systems
  - Confidence scores

The corpus based approach. EBMT Translation Patterns

- Automatic extraction of translation patterns from the bilingual parallel corpus:

Aligned sentences	Aligned sentences with generalized units	Translation pattern
En Vitoria-Gasteiz, a 22 de Diciembre de 2003.	En <rs type=loc> Vitoria-Gasteiz </rs> , a <date date=22/12/2003> 22 de Diciembre de 2003</date> .	En <rs1> , a <date1>.
Vitoria-Gasteiz, 2003ko Abenduaren 22.	<rs type=loc> Vitoria-Gasteiz </rs> , <date date=22/12/2003> 2003ko Abenduaren 22</date>.	<rs1>, <date1>.

The corpus based approach. *EBMT Translation Patterns*

- Automatic extraction of translation patterns from the bilingual parallel corpus
  - 7,599 translation patterns
  - covering 35,450 sentence pairs

- Very high precision but quite low coverage

- Interesting to combine with the other engines
  - Specially in this kind of domain
    (formal and quite controlled language)

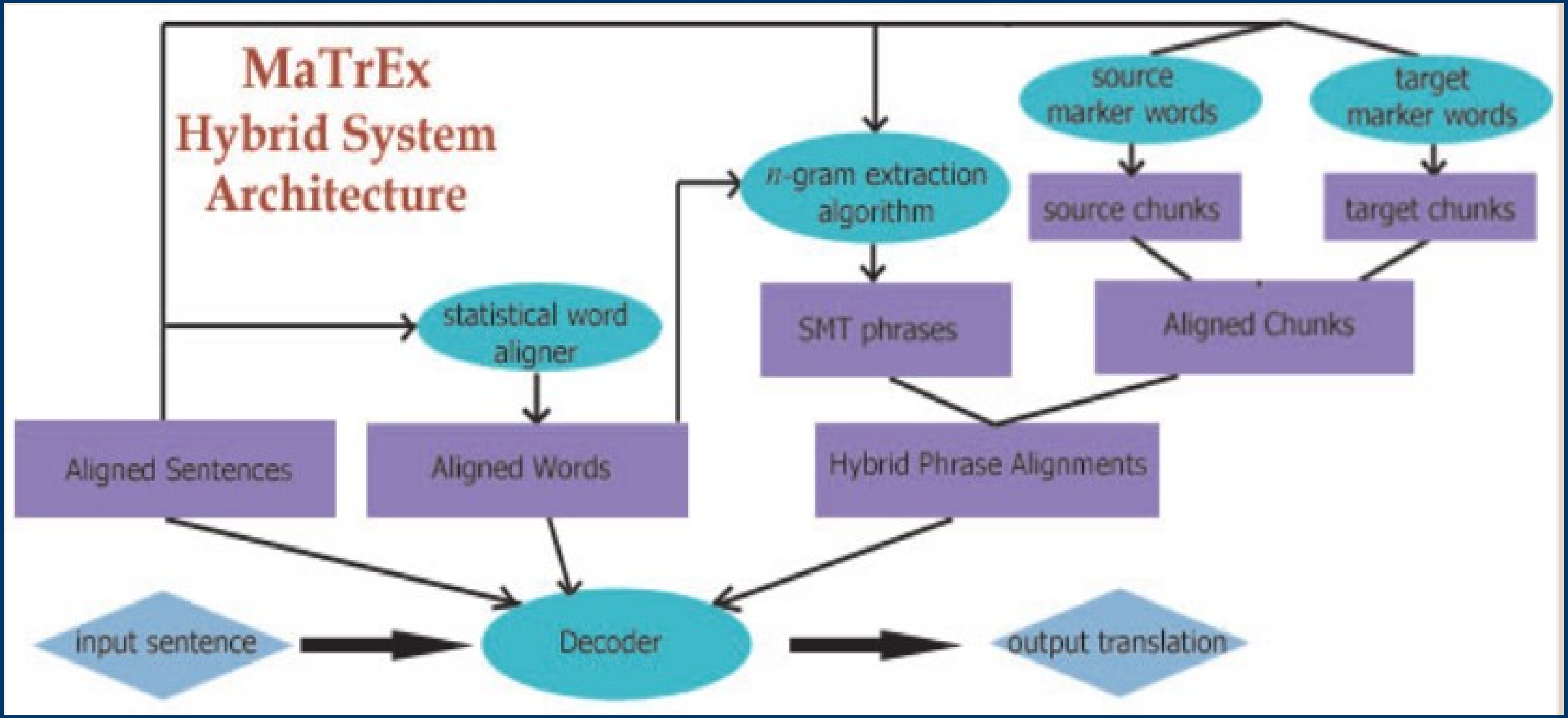
The SMT approach

- **Some tools have been reused for this purpose:**
  - GIZA++: For word/morpheme alignment (Och and Ney, 2003)
  - Moses decoder: the decoder is also a hybrid system which integrates EBMT and SMT. It is capable of retrieving already translated sentences and also provides a wrapper around the PHARAOH SMT decoder (Koehn, 2004).
  - MaTrEx: a data-driven MT engine, built following an extremely modular design. It consists of a number of extendible and re-implementable modules (Way and Gough, 2005).
  - Eusmg: a toolkit to chunk Basque sentences.

The SMT approach, Matrex

- Carried out in collaboration with the National Centre for Language Technology in Dublin
- The system exploits SMT technology to extract aligned chunks

The SMT approach



The SMT approach

- Three approaches:
  - Conventional SMT machine
  - Morpheme-based SMT machine

The SMT approach

- Both systems (conventional and morpheme-based) were optimized using Minimum Error Rate Training. Metric: BLEU
- Preliminary evaluation:

	BLEU	NIST	WER	PER
SMT	9.51	3.73	83.94	66.09
Morpheme-based SMT	8.98	3.87	80.18	63.88

MEMT: The RBMT approach

Adaptation to the domain (labor agreements)

- **Terminology.**
  - Semiautomatic extraction. Elexbi (Alegria *et al.*, 2006). 807 terms extracted
- **Lexical selection.**
  - New order for the possible translations calculated on the parallel corpus using GIZA++
- **Resolution of format and typographical variants frequents in the administrative domain**.

RBMT and SMT (preliminary evaluation)

Automatic evaluation (BLEU and NIST)

SMT performs better on the in-domain corpus
RBMT performs better on the out-domain corpus

Manual evaluation (HTER)

RBMT performs better, irrespective of the corpus

	BLEU RBMT	BLEU SMT	HTER RBMT	HTER SMT
EiTB corpus (news) Out-domain	**9.30**	9.02	**40.41**	71.87
Consumer (magazine) In-domain	6.31	**8.03**	**43.60**	57.97

Combination:
Multi-Engine MT for Basque

- Combining the different methods in a domain where translation memories were available.

  Text is divided into sentences,

  Each sentence is processed using each engine (parallel processing is possible).

  Finally one of the translations is selected.

- Facts to define this selection:

  EBMT: very high precision, but low coverage

  The SMT engine gives a confidence score.

  The RBMT engine does not give a confidence score.

  RBMT translations are more adequate for human post-edition

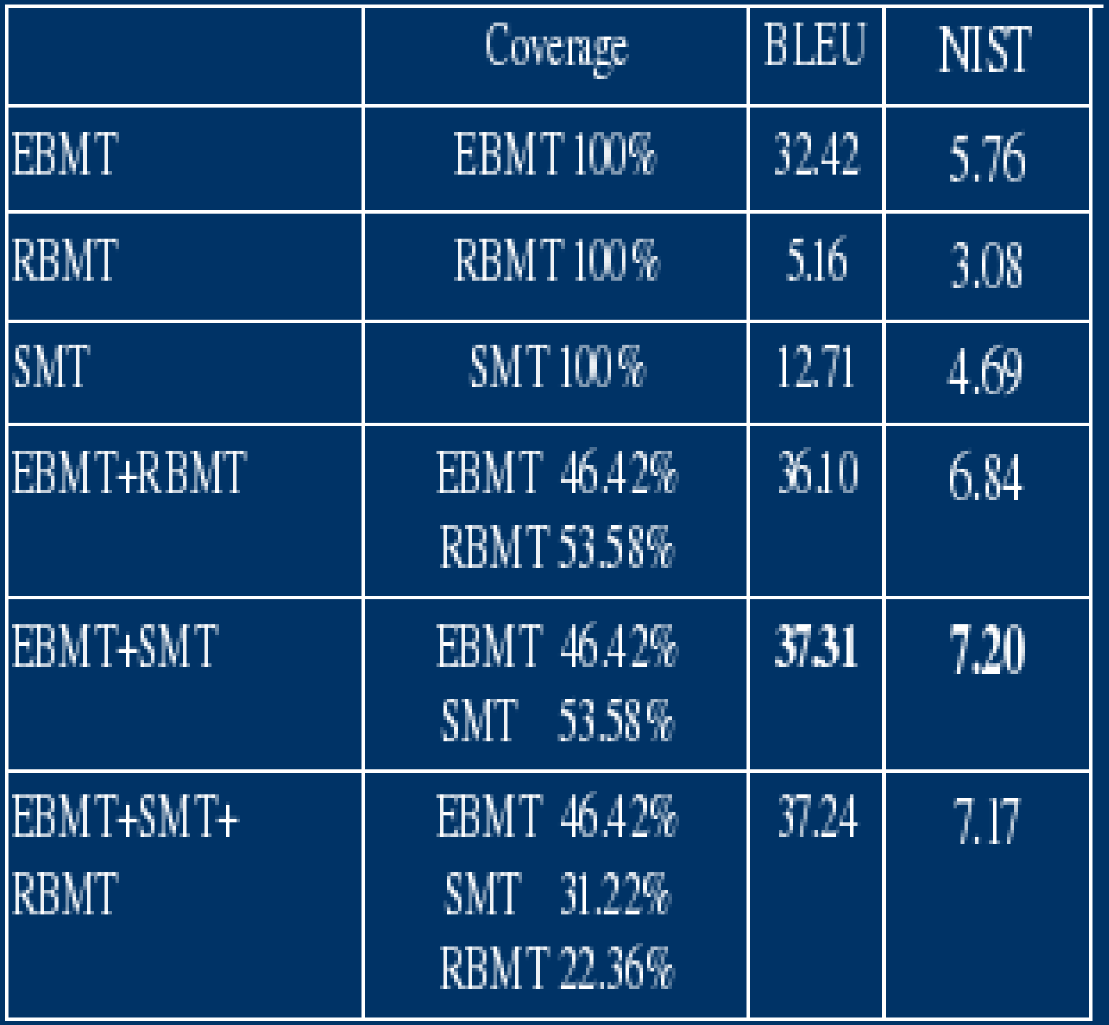
  SMT gets better scores when BLEU and NIST (only one reference)

Combining the approaches.
Multi-Engine MT for Basque

Combining three approaches in a simple hierarchical way:

```
if the EBMT engine covers the sentence
        EBMT translation is selected
else if   the SMT's confidence score  >  a given threshold
        SMT translation is selected
otherwise
        RBMT translation is selected
```

MEMT evaluation



	Coverage	BLEU	NIST
EBMT	EBMT 100%	32.42	5.76
RBMT	RBMT 100%	5.16	3.08
SMT	SMT 100%	12.71	4.69
EBMT+RBMT	EBMT 46.42% RBMT 53.58%	36.10	6.84
EBMT+SMT	EBMT 46.42% SMT 53.58%	**37.31**	**7.20**
EBMT+SMT+ RBMT	EBMT 46.42% SMT 31.22% RBMT 22.36%	37.24	7.17

- Very significant improvement
  193% relative increase for BLEU comparing EBMT+SMT+RBMT and SMT alone
- 15% relative increase comparing EBMT + SMT and EBMT alone.

but
a deeper evaluation was necessary.

Combination:
RBMT + Statistical Postedition

Sentence

→ RBMT system

→ Intermediate translation

→ SMT system (trained on corpus of posteditions)

→ Final translation

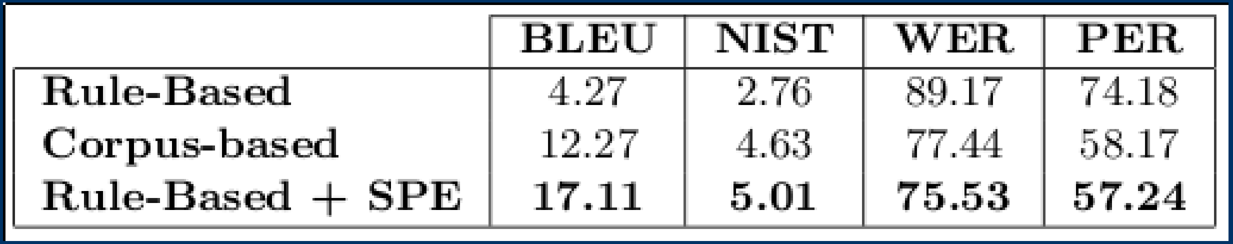
Combination:
RBMT + Statistical Postedition

Creation of a pseudocorpus of post editions

- We have first translated Spanish sentences in the parallel corpus using Matxin.
- Using automatically translated sentences and their corresponding Basque sentences in the parallel corpus, → parallel corpus to train our statistical post-editor

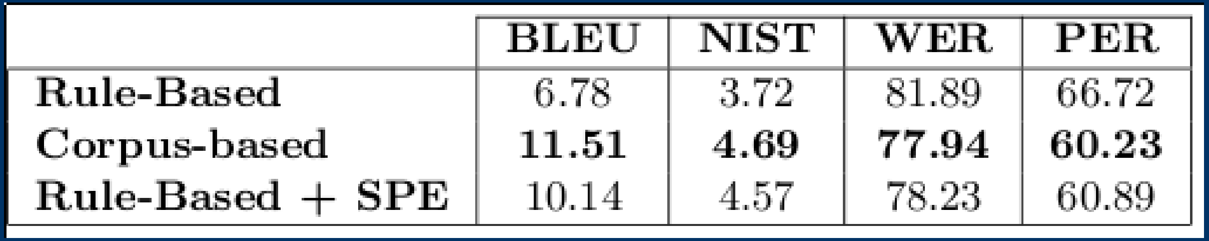
Of course, it would have worked better
using real Post-Editing parallel corpus
… but we had not postedited translations :-(

SPE evaluation



	BLEU	NIST	WER	PER
Rule-Based	4.27	2.76	89.17	74.18
Corpus-based	12.27	4.63	77.44	58.17
Rule-Based + SPE	**17.11**	**5.01**	**75.53**	**57.24**

Evaluation on domain specific corpus



	BLEU	NIST	WER	PER
Rule-Based	6.78	3.72	81.89	66.72
Corpus-based	**11.51**	**4.69**	**77.94**	**60.23**
Rule-Based + SPE	10.14	4.57	78.23	60.89

Evaluation on general domain corpus

Results on Labor Agreements Corpus

- RBMT gets a very low performance (not adapted to the restricted domain),
- RBMT+SPE gets 40% relative improvement with Corpus based system
- No improvement in general domain

but

a deeper evaluation was necessary.

Conclusion on combination of MT approaches

The results of combining RBMT with other MT paradigms are promising (Alegria et al., 2008)

But deeper evaluation is necessary:

- More than one reference with BLEU and NIST

or

- Human evaluation, postedition cost (HTER)

Outline

- Basque : a Less Resourced Language (LRL)
- Strategy for sustainable HLT (and MT) for Basque
- Machine Translation for Basque (Matxin)
- Evaluation of Matxin
- Future: Combining RBMT and Corpus-based MT
- New elements and conclusions
    - New parallel corpora
    - New evaluation
    - Firefox-WWL web publication and postedition tool

Collecting corpus

- Being Basque a less-resourced language, one of our main difficulties is getting a larger enough bilingual corpus.

- Up to now:
  - 1 million Basque words bilingual corpus (1.3 million words in Spanish)

- Labaka (2009)
  - 7 million Basque words bilingual corpus (9 million in Spanish).
  - 28 million words monolingual Basque text to be used for the training of the language model

Improving basic systems

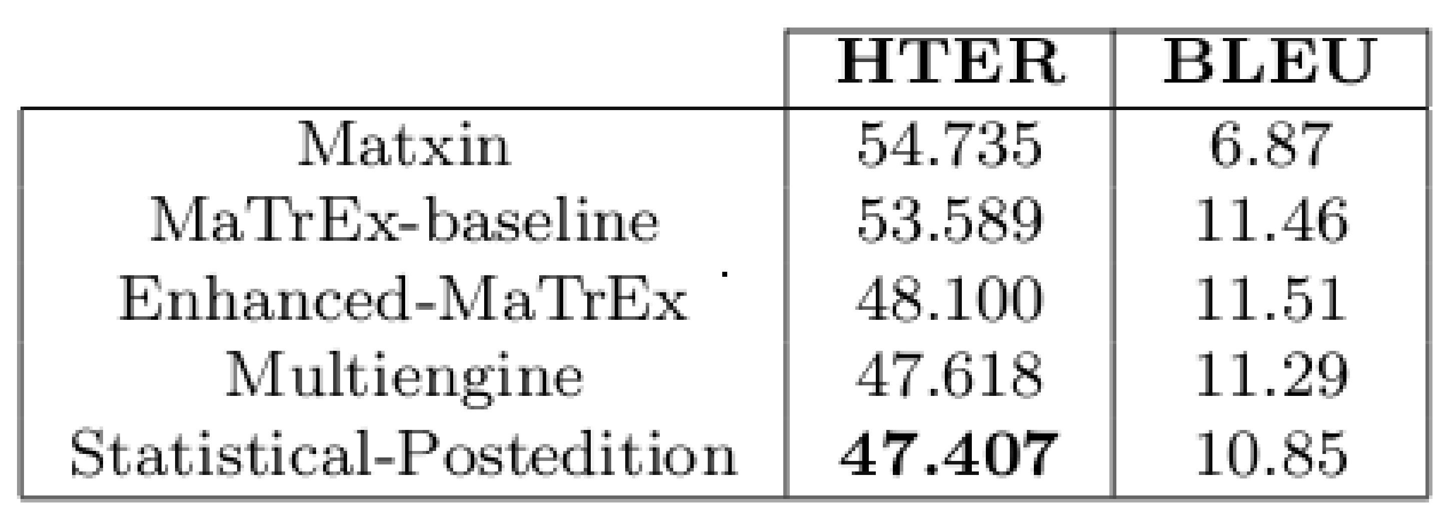
- SMT deeper architectures working
  - Morphological segmentation of words
  - Word reordering in source language text
- EBMT
  - Extraction of new patterns
- RBMT
  - Lexical enrichment...

Final evaluation: HTER

- HTER evaluation based on hand-made post-editions give us a more confident score,
  - It measures the real work a professional translator needs to achieve a correct translation starting from the output of the MT system.
  - Difficulty for interpreting the BLEU scores.
- HTER evaluation is expensive but cheaper than creating several references to get more accurate BLEU scores.

Final evaluation: HTER

- MEMT and SPE combinations are valuable.
- The RBMT system Matxin was not properly tuned when the evaluation was performed.
  - But we can observe that it helps in the MEMT's performance



	HTER	BLEU
Matxin	54.735	6.87
MaTrEx-baseline	53.589	11.46
Enhanced-MaTrEx	48.100	11.51
Multiengine	47.618	11.29
Statistical-Postedition	**47.407**	10.85

- There is still room for improving via MEMT +SPE (37.847 HTER for oracle MEMT +SPE)

Final evaluation: HTER Conclusions (Labaka, 2009)

- The usefulness of RBMT systems for assimilation is probed
  - 69% of the users found RBMT translation useful when integrated in a MultiLingual Information Retrieval system (Leturia et al., 2009).

- But if we were able to achieve the translation quality obtained by the oracle system (37.847 HTER score)...
  Spanish-Basque MT would be useful
  also for Computer-Aided Translation system
  - HTER <40%,
  - Post-editing a MT output would be definitely faster than creating a new translation.

Conclusions

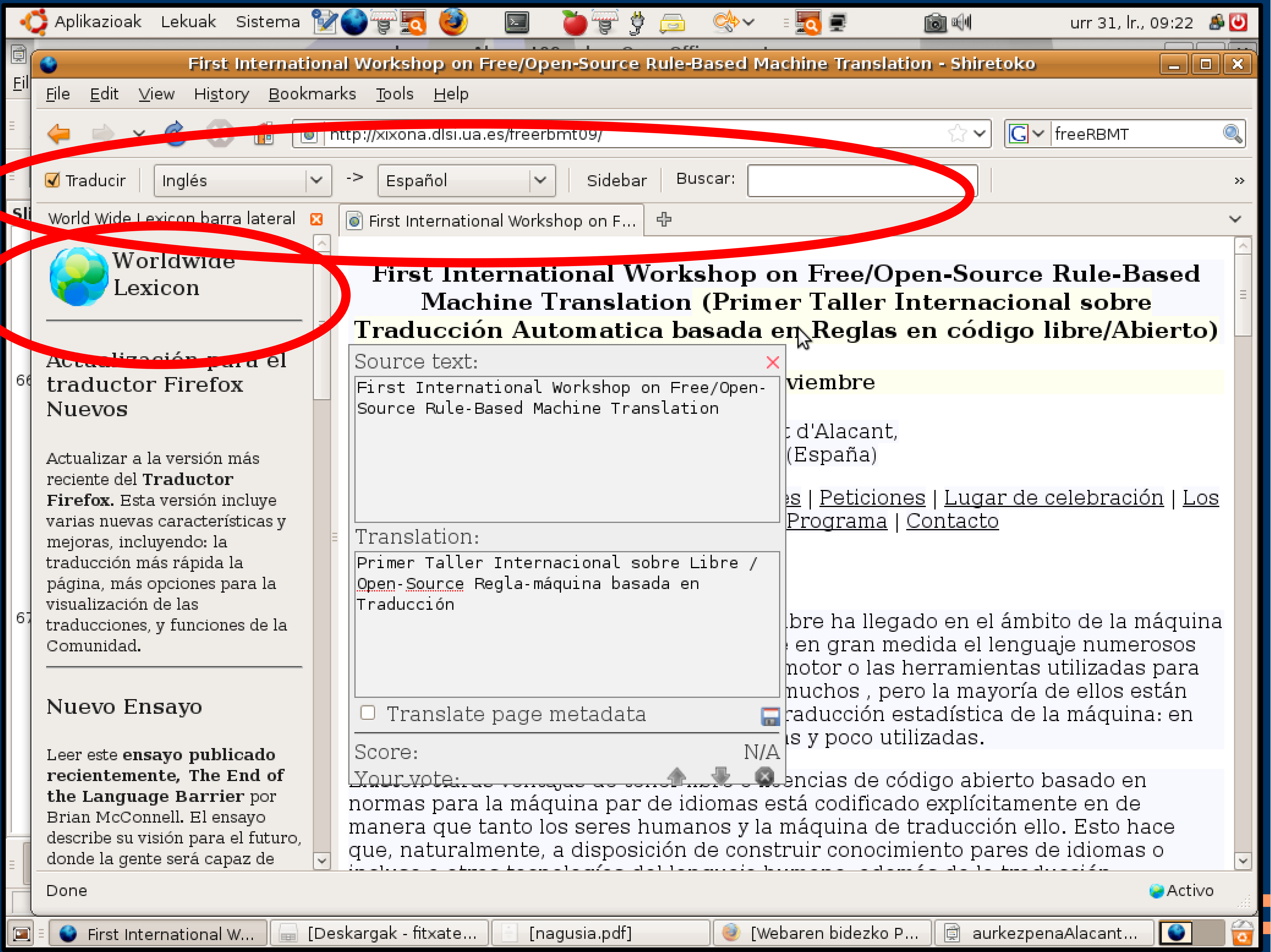
- Less privileged languages have to do a great effort to face language technology.
  - Need of high standardization
  - Reusing language foundations, tools, and applications
  - Incremental design and development of them
  - Open source
- Those guidelines seems to be trivial,
  but from our experience we know that they are not followed in many HLT projects related with these languages

Conclusions

- This strategy has been completely useful to create MT systems for Basque
    - Reusing of previous works for Basque (that were defined following XML and TEI standards)
    - Reusing other open-source tools (Opentrad and Freeling)
- Satisfactory results in a short time
- Two results publicly available:
    - free code for the es-eu RBMT system
      matxin.sourceforge.net
    - on-line demo:
      www.opentrad.org

Future Work

- New experiments
  - MEMT combination of the outputs based on a language model
  - Confidence scores for RBMT
    - penalties when suspicious or very complex syntactic structures are present in the analysis,
    - penalties for high proportion of highly polysemic words,
    - promoting translations that recognize multiword lexical units
    - …
- Collaboration with a web community (Basque Wikipedia)
  - to adapt web tools (Firefox Translator and WWL ?) for MT output postedition and web publication.
  - to collect corpus of translation posteditions



Worldwide Lexicon

Actualización para el traductor Firefox Nuevos

Actualizar a la versión más reciente del **Traductor Firefox.** Esta versión incluye varias nuevas características y mejoras, incluyendo: la traducción más rápida la página, más opciones para la visualización de las traducciones, y funciones de la Comunidad.

Nuevo Ensayo

Leer este **ensayo publicado recientemente, The End of the Language Barrier** por Brian McConnell. El ensayo describe su visión para el futuro, donde la gente será capaz de

First International Workshop on Free/Open-Source Rule-Based Machine Translation (Primer Taller Internacional sobre Traducción Automatica basada en Reglas en código libre/Abierto)

Source text:

First International Workshop on Free/Open-Source Rule-Based Machine Translation

Translation:

Primer Taller Internacional sobre Libre / Open-Source Regla-máquina basada en Traducción

□ Translate page metadata

Score: N/A

Your vote:

viembre

d'Alacant, (España)

es | Peticiones | Lugar de celebración | Los Programa | Contacto

bre ha llegado en el ámbito de la máquina en gran medida el lenguaje numerosos motor o las herramientas utilizadas para muchos , pero la mayoría de ellos están raducción estadística de la máquina: en s y poco utilizadas.

encias de código abierto basado en normas para la máquina par de idiomas está codificado explícitamente en de manera que tanto los seres humanos y la máquina de traducción ello. Esto hace que, naturalmente, a disposición de construir conocimiento pares de idiomas o

Thank you very much!

ixa.si.ehu.es

www.opentrad.es

ixa.si.ehu.es/openmt

Matxin: developing sustainable MT for a less-resourced language

Kepa Sarasola

(Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Julen Ruiz, Aingeru Mayor)

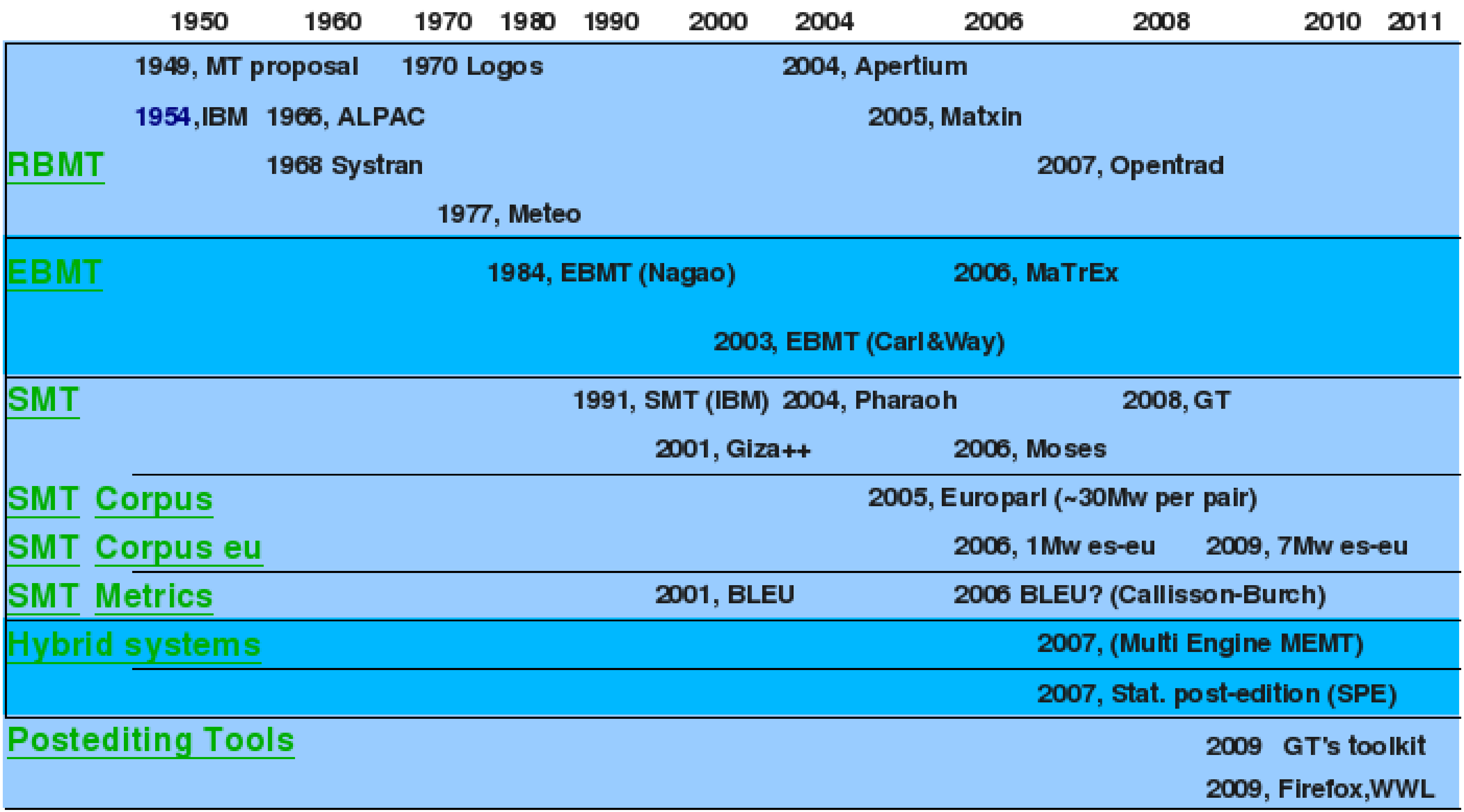
Ixa taldea.

University of the Basque Country

Universidad del País Vasco Euskal Herriko Unibertsitatea

FreeRBMT 2009, Alacant

Milestones in MT



	1950	1960	1970	1980	1990	2000	2004	2006	2008	2010	2011
	1949, MT proposal		1970 Logos				2004, Apertium				
	1954,IBM	1966, ALPAC					2005, Matxin				
RBMT		1968 Systran						2007, Opentrad			
			1977, Meteo								
EBMT				1984, EBMT (Nagao)				2006, MaTrEx			
						2003, EBMT (Carl&Way)					
SMT					1991, SMT (IBM)		2004, Pharaoh		2008, GT		
						2001, Giza++		2006, Moses			
SMT Corpus							2005, Europarl (~30Mw per pair)				
SMT Corpus eu								2006, 1Mw es-eu	2009, 7Mw es-eu		
SMT Metrics						2001, BLEU		2006 BLEU? (Callisson-Burch)			
Hybrid systems								2007, (Multi Engine MEMT)			
								2007, Stat. post-edition (SPE)			
Posteditting Tools									2009 GT's toolkit		
									2009, Firefox,WWL		