Valuable Language Resources and Applications Supporting the Use of Basque

Iñaki Alegria, Maxux Aranzabe, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Aingeru Mayor and Kepa Sarasola

Ixa Group

University of the Basque Country

LTC 2009, Poznan

Outline

- Basque: a Less Resourced Language
- Strategy to develop HLT for Basque
- Useful applications and resources
- Conclusions

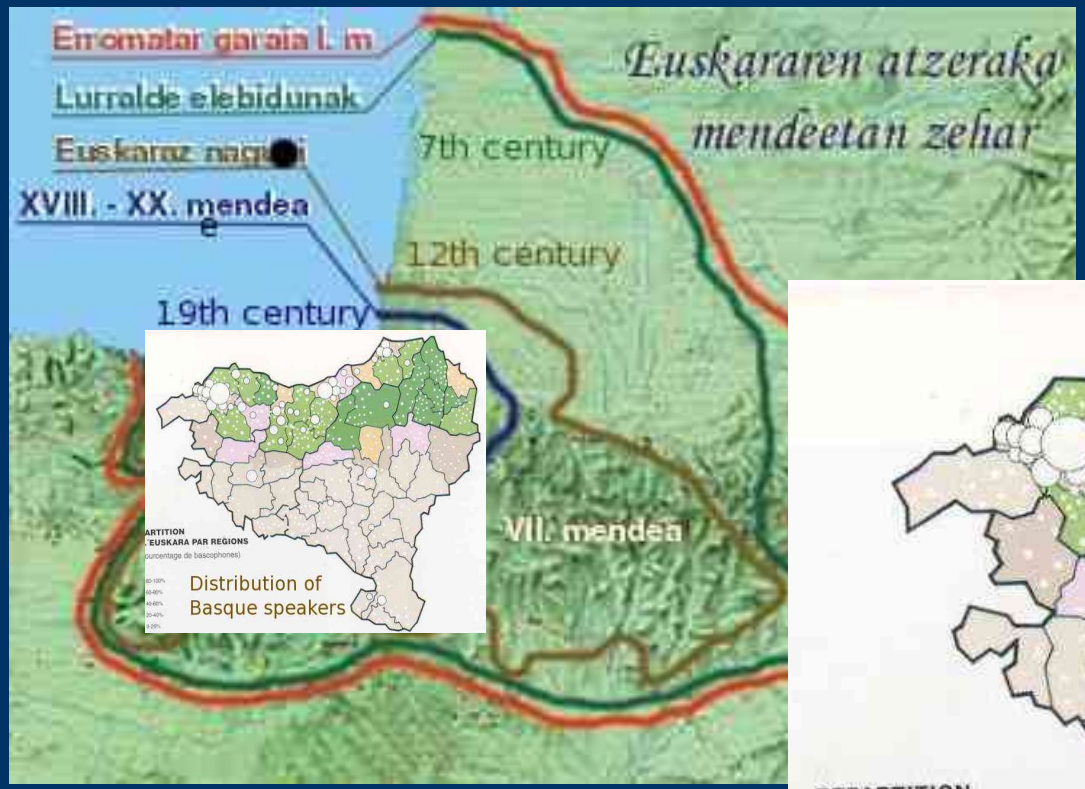
History of Basque

Prerromanic languages in Spain

Basque in 7th, 12th and 19th centuries

History of Basque

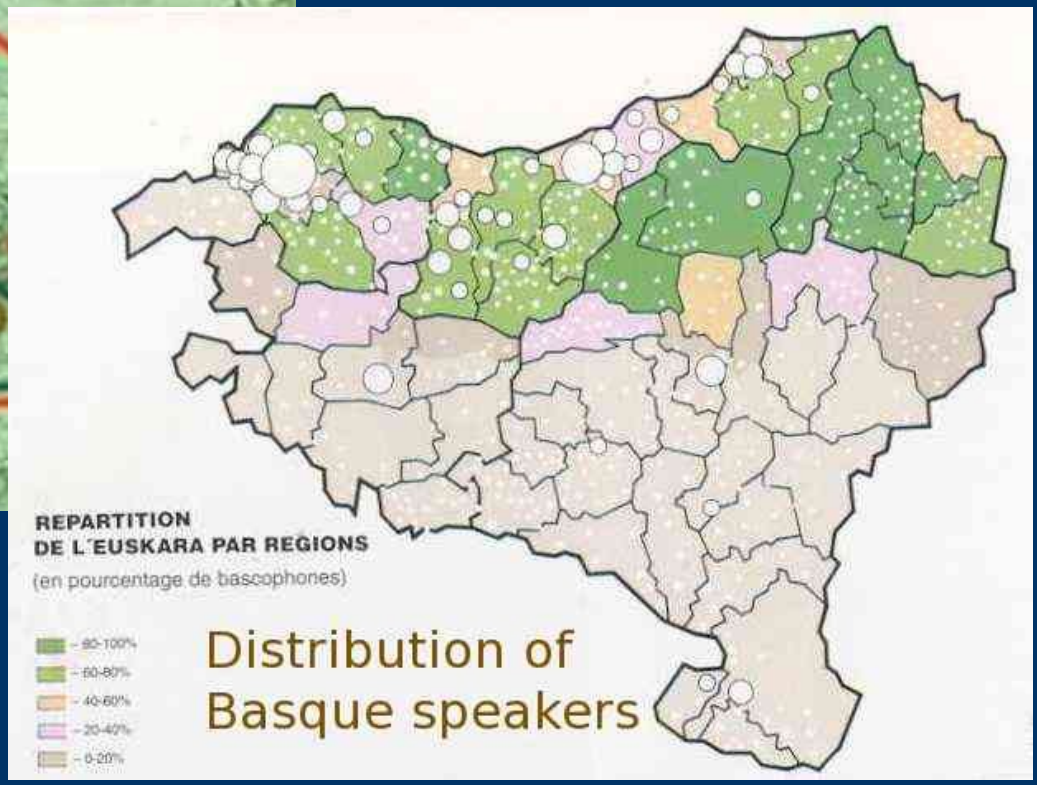
Basque in 7th, 12th and 19th centuries



1,033,900 Speakers

(First lang.: 700,000)

Non homogeneous distribution!

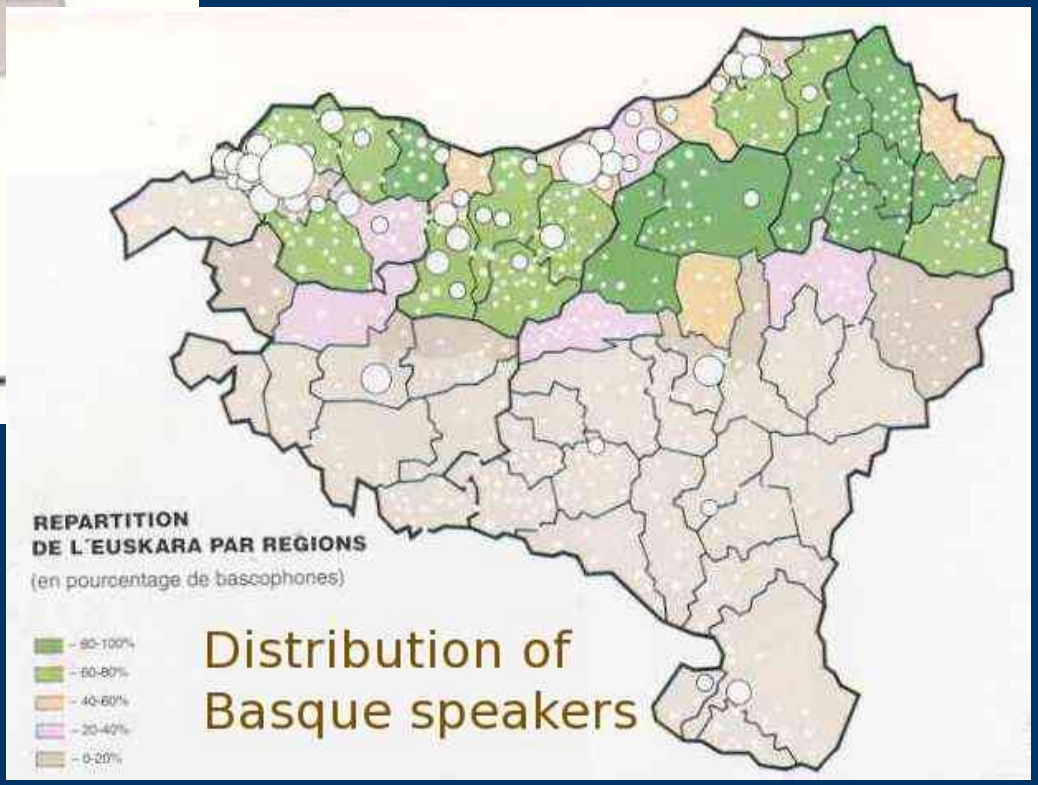


Basque nowadays

Six different dialects !

1,033,900 Speakers
(First lang.: 700,000)

Non homogeneous distribution !



Main reasons of Basque regression

- No official language
- Out of the education systems
- 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*
...				
...				
...				

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

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:

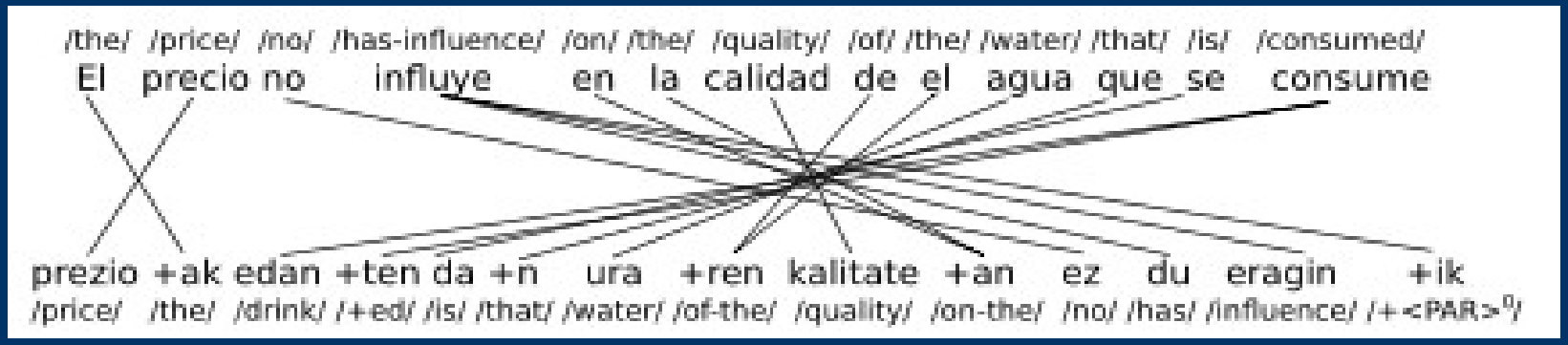
Ergative language & multiple agreement

- 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
  - Free-order of components at sentence level.
  - Different component order at noun phrase level.



=> Language technology for Basque is both:

- Real need to revitalize Basque
- Test bed for our strategy for developing language tools

Basque. Linguistic features and MT

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

=> Language technology for Basque is both:

- Real need to revitalize the language
- Test bed for our strategy for developing language tools

Outline

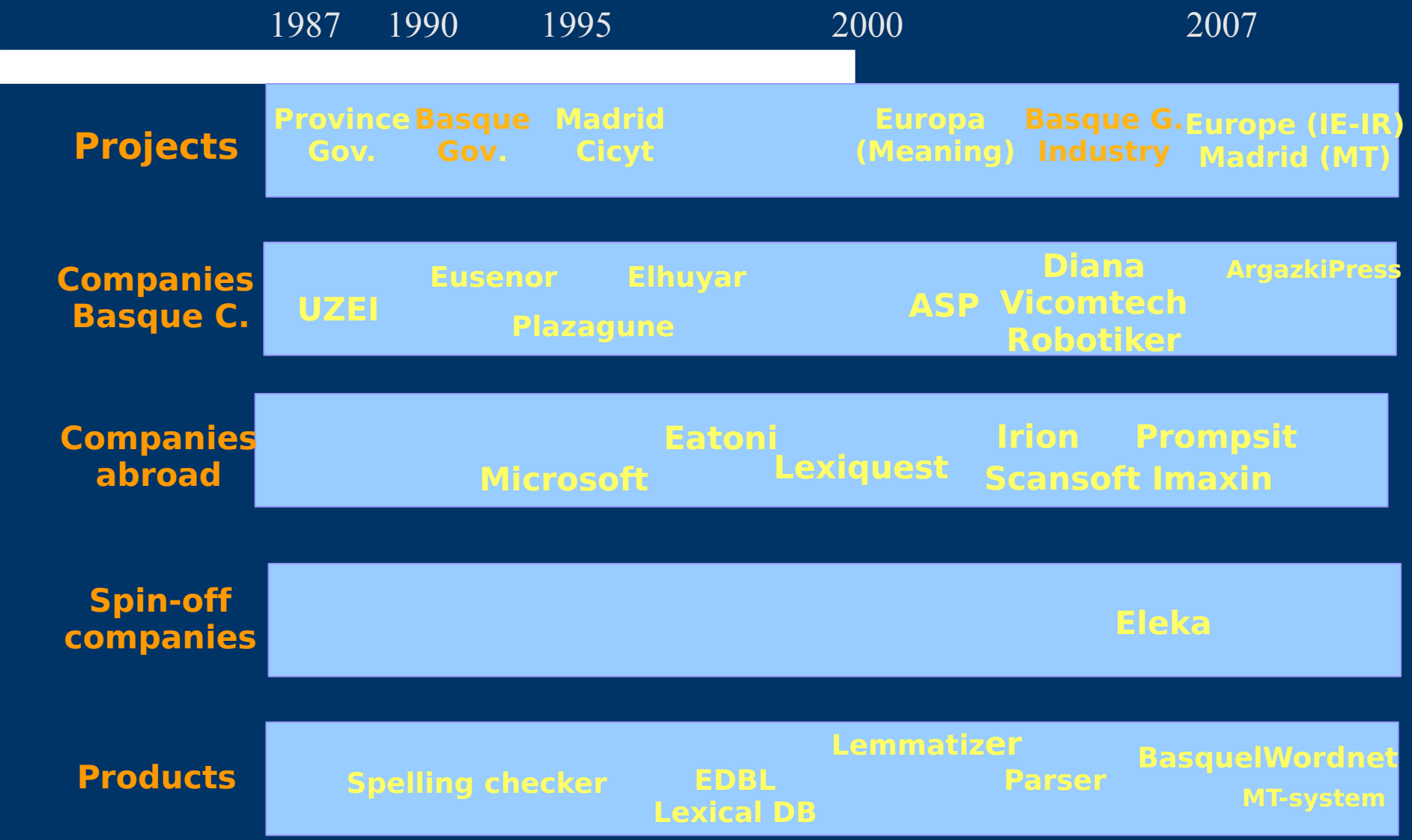
- Basque: a Less Resourced Language
- Strategy to develop HLT for Basque
- Useful applications and resources
- Conclusions

Strategy to develop HLT in Basque
IXA Research Group

- 1986: 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



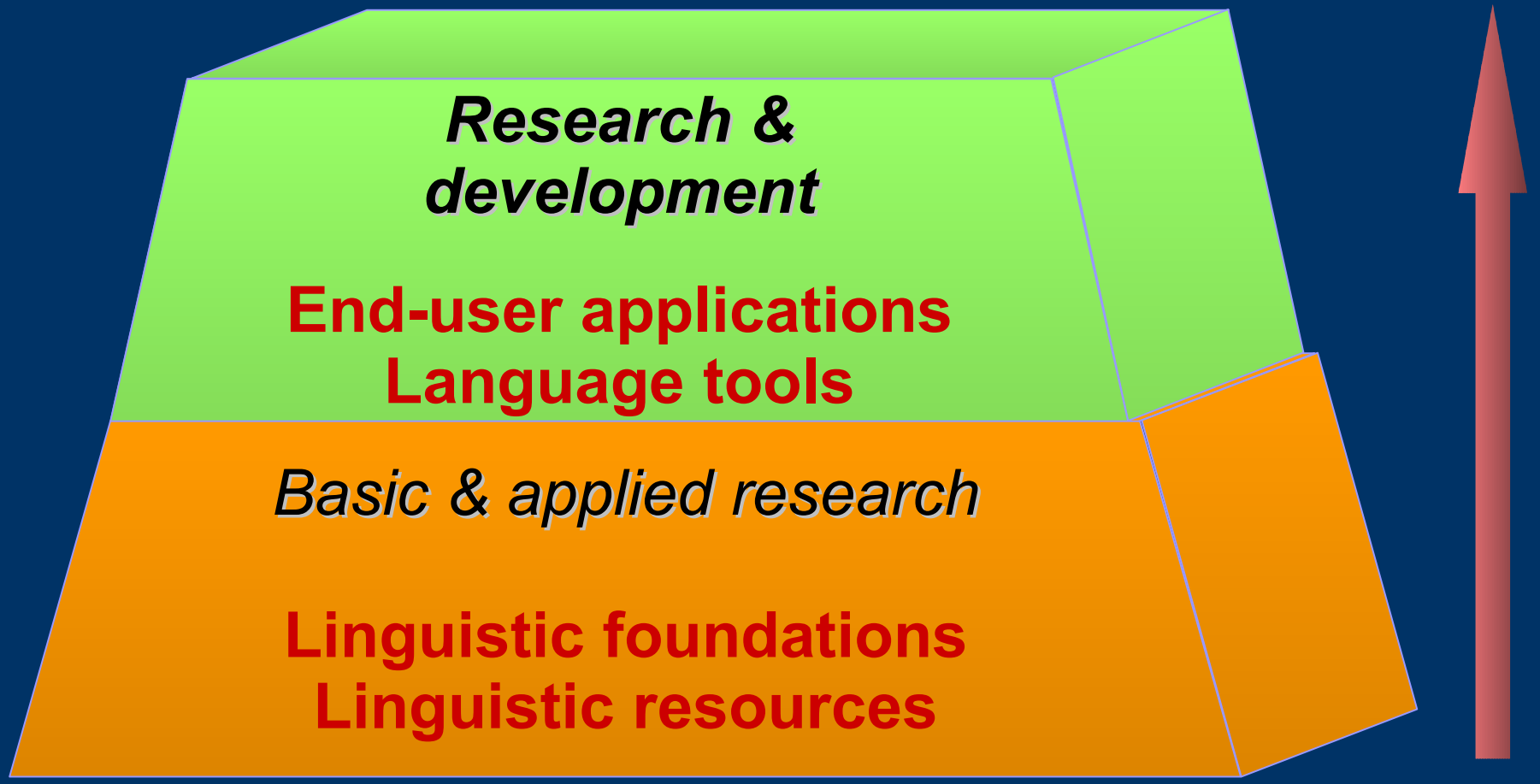
Underlying strategy

- 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

Example: 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.

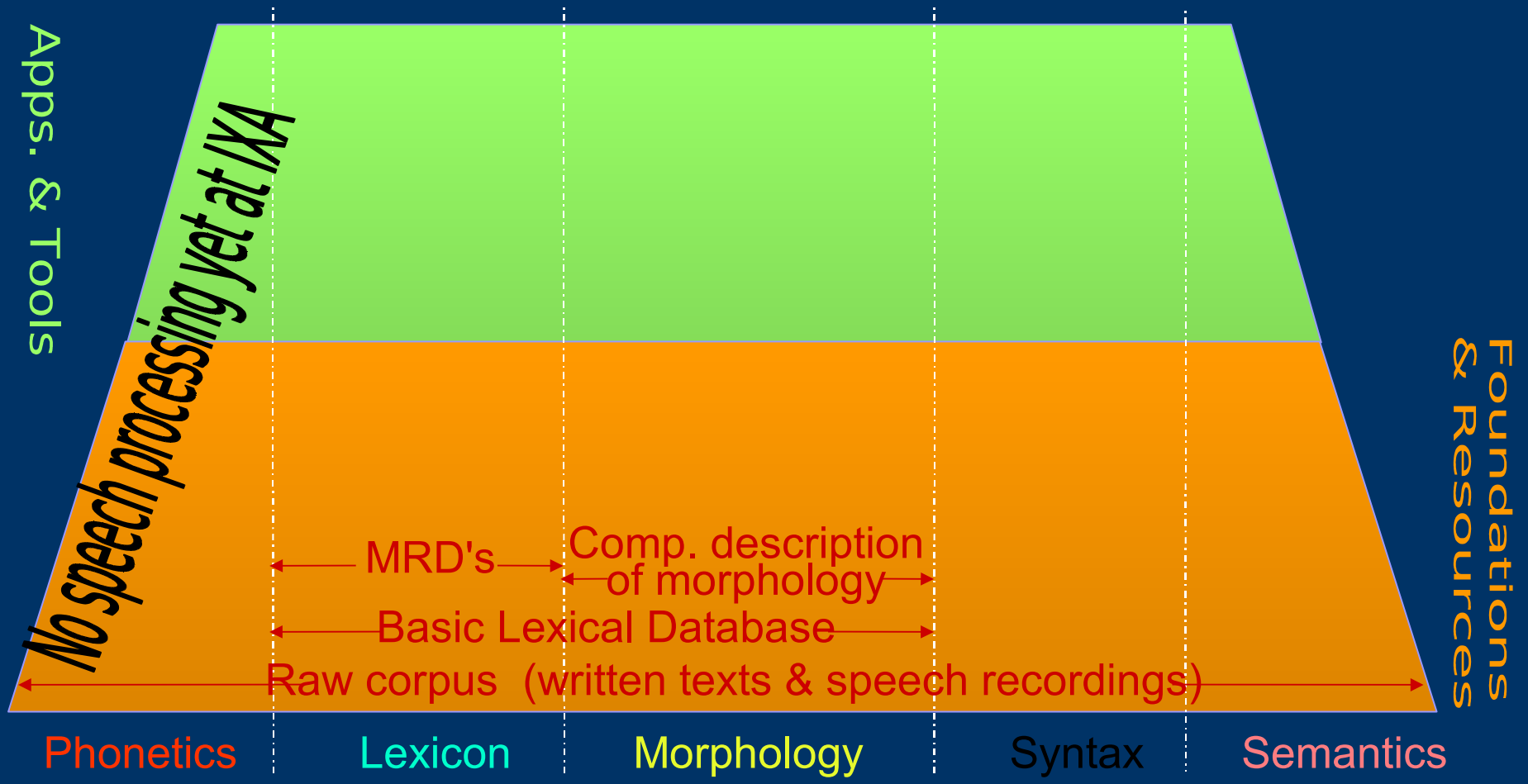
Strategic priorities: from basic research to application development



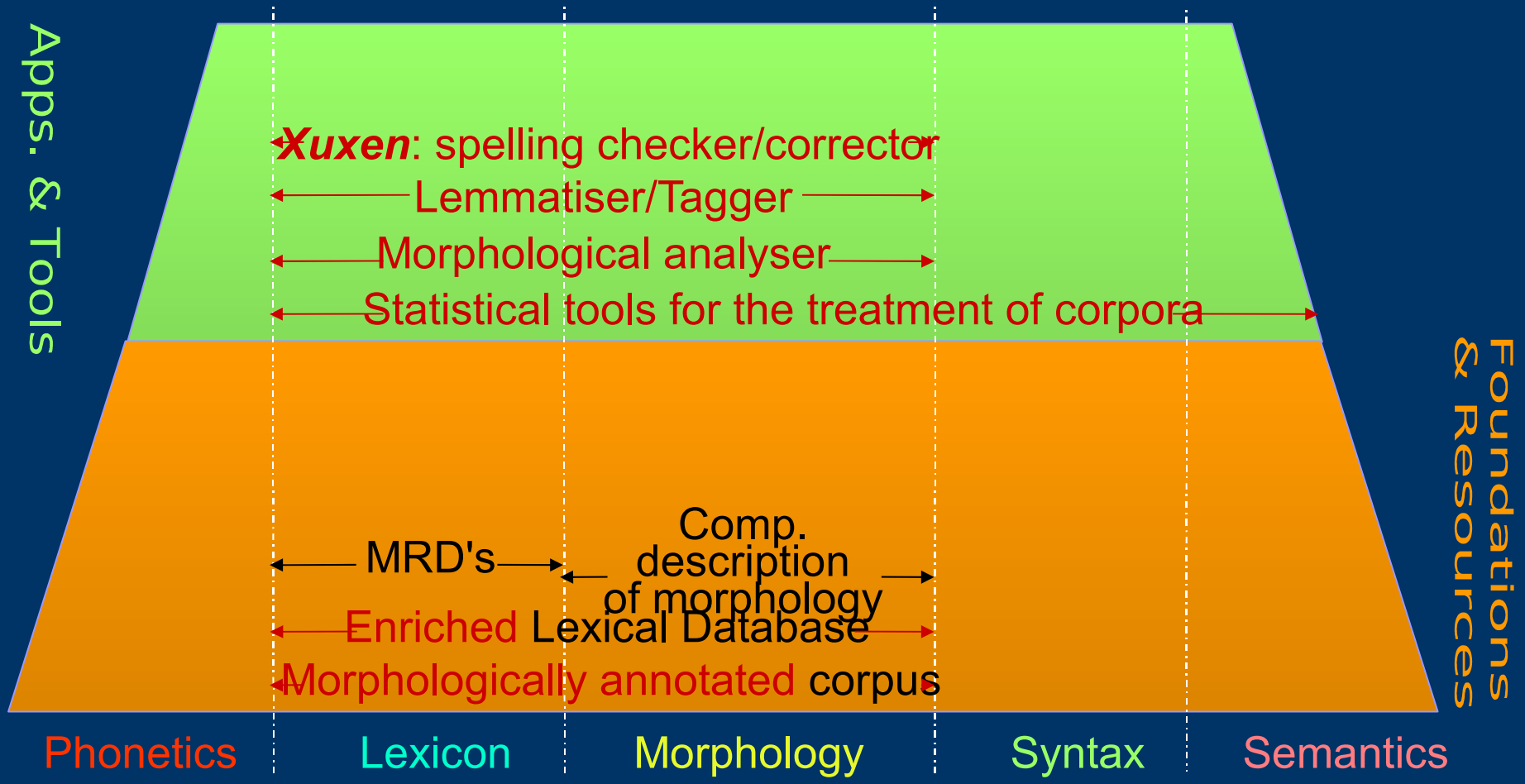
Linguistic foundations & resources, tools and applications

- 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.

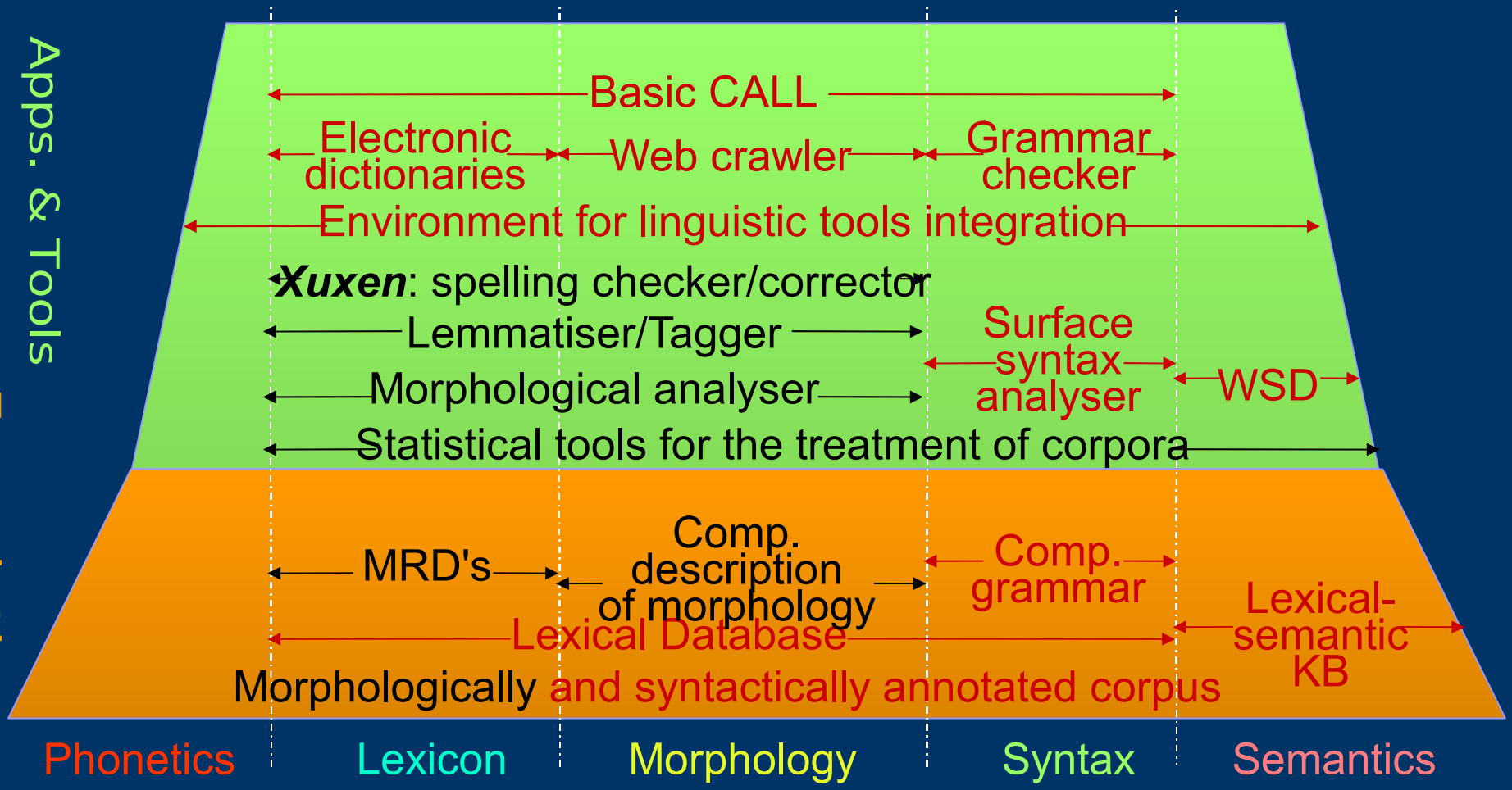
Phase I: laying foundations



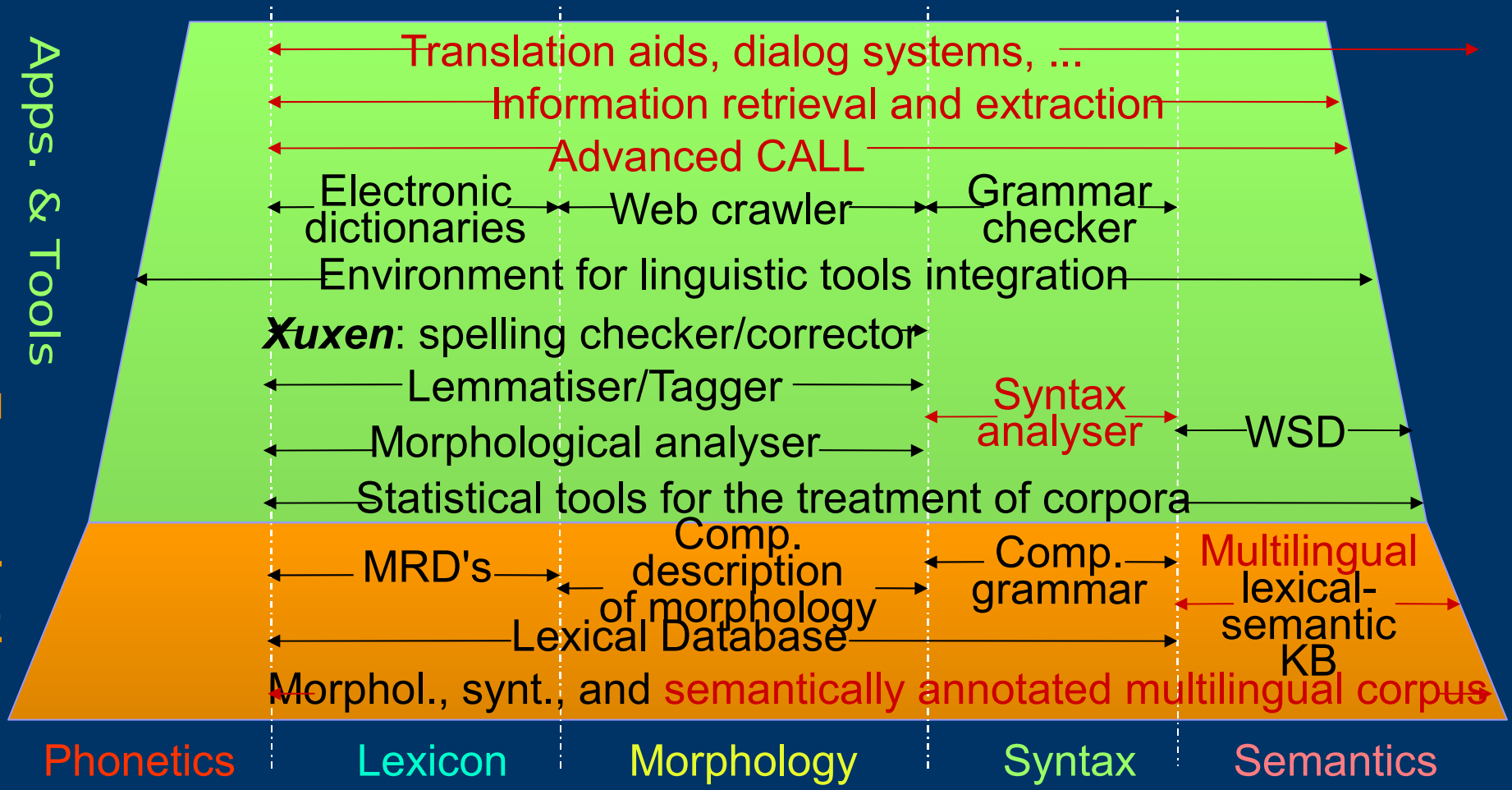
Phase II:
first basic tools and applications



Phase III: more advanced tools and applications



Phase IV: multilinguality and general applications



Outline

- Basque: a Less Resourced Language
- Strategy to develop HLT for Basque
- Useful applications and resources
- Conclusions

Useful applications and resources

Applications

- Spelling checker/corrector
- Spanish-Basque transfer based MT system
- Lemmatization based on-line dictionaries
- Lemmatization based search machine

Resources

- EDBL: Lexical Database for Basque
- BasWN: Basque Wordnet
- EPEC: syntactically annotated Text Corpus
- ZTC: Morphosyntactically Annotated Text Corpus

Basque spelling checker/corrector

Basque speakers have many doubts when trying writing in Basque:

- Out of educational systems until 1980
- Late standardization (1966)

- How to write the word for “tree” ?
  zuhaitza? zuhaitz? zugaitz? zuhaitx? zuhaitsa? sugatza?

Basque spelling checker/corrector

- The spelling-checker (Aduriz et al., 1997) has proven to be a very effective tool to resolve those doubts.
- It gives people more confidence in the text they are writing
- In fact, this program is one of the most powerful tools in the ongoing standardization of Basque.

Basque spelling checker/corrector

- Technically, the spelling checker is more complex than equivalent software for other languages.

  Rich morphology =>
  Difficulty to define the list of possible words in the language.

- Three different solutions:
  - Including the complete morphological analyser
    But we need propietary software to do it efficiently
    (Experimenting with new alternatives: *Foma* & *hfst*)
  - Using *hunspell* when possible in open software
  - Using *myspell* if *hunspell* not integrated *(Firefox...)*

Basque spelling checker/corrector

- Using *hunspell*
  - Adaptation of the two-level description to *hunspell* in a (semi)automatic way.
    - Stems
    - 2 sets of suffixes:
      the paradigms at first and second level

- Using *myspell*

  For Firefox and open tools that don't integrate *hunspell*
  - Adaptation combining the main paradigms (with less generation power for each one)
  - And adding the word forms appearing in a big corpus, after eliminating forms rejected by the complete spelling checker.

Basque spelling checker/corrector

- It is publicly available:
  - www.euskara.euskadi.net (>20,000 downloads)
    Versions for Office, OpenOffice, Mozilla, PC, Mac
  - addons.mozilla.org/en-US/firefox/addon/4020
    Version for Firefox (>100.000 downloads)
  - www.xuxen.com
    Online web service

Transfer based Machine Translation system

- 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.

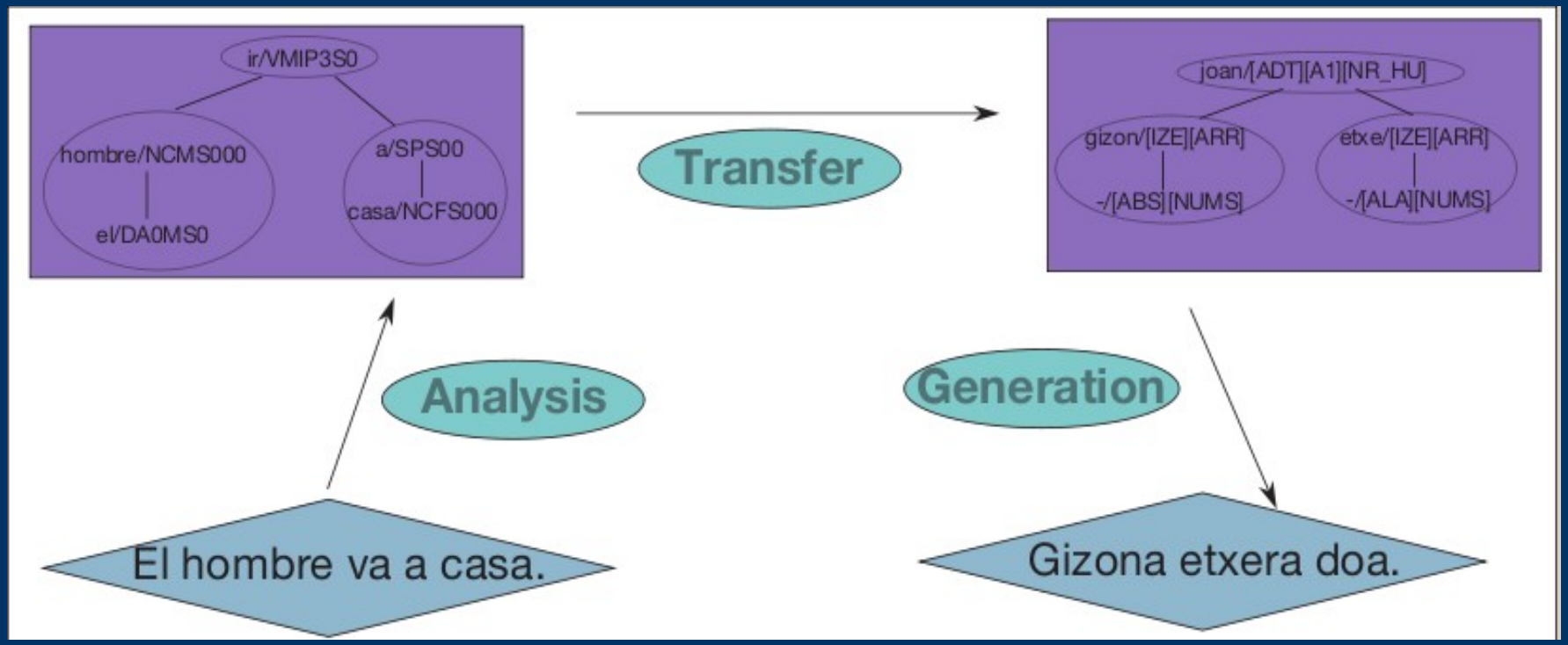
Transfer based Machine Translation system

Transfer based Machine Translation system

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

Machine Translation system Design



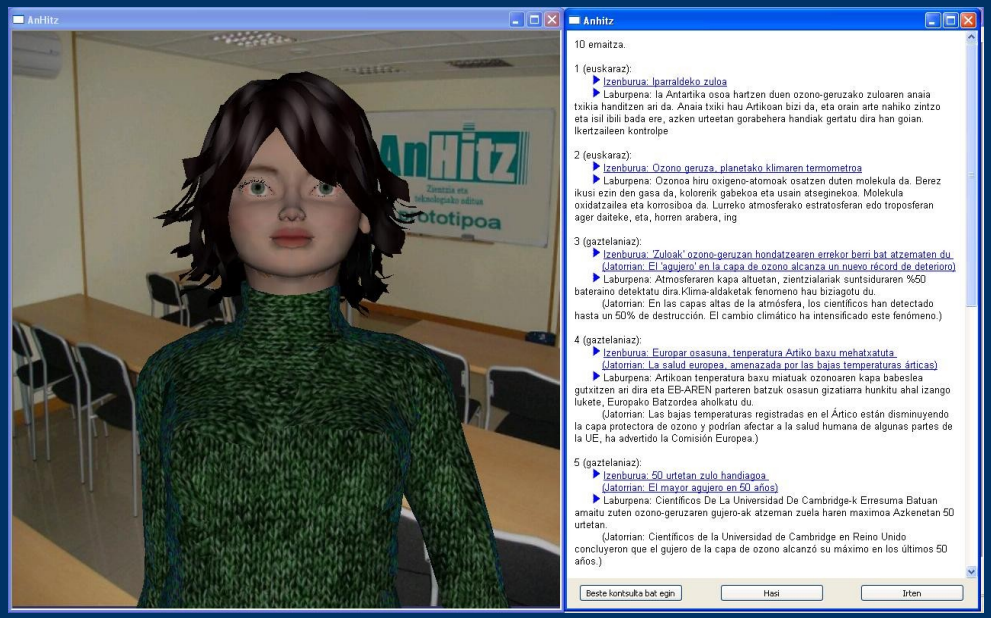
Machine Translation system Design

- **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/generator is reused.

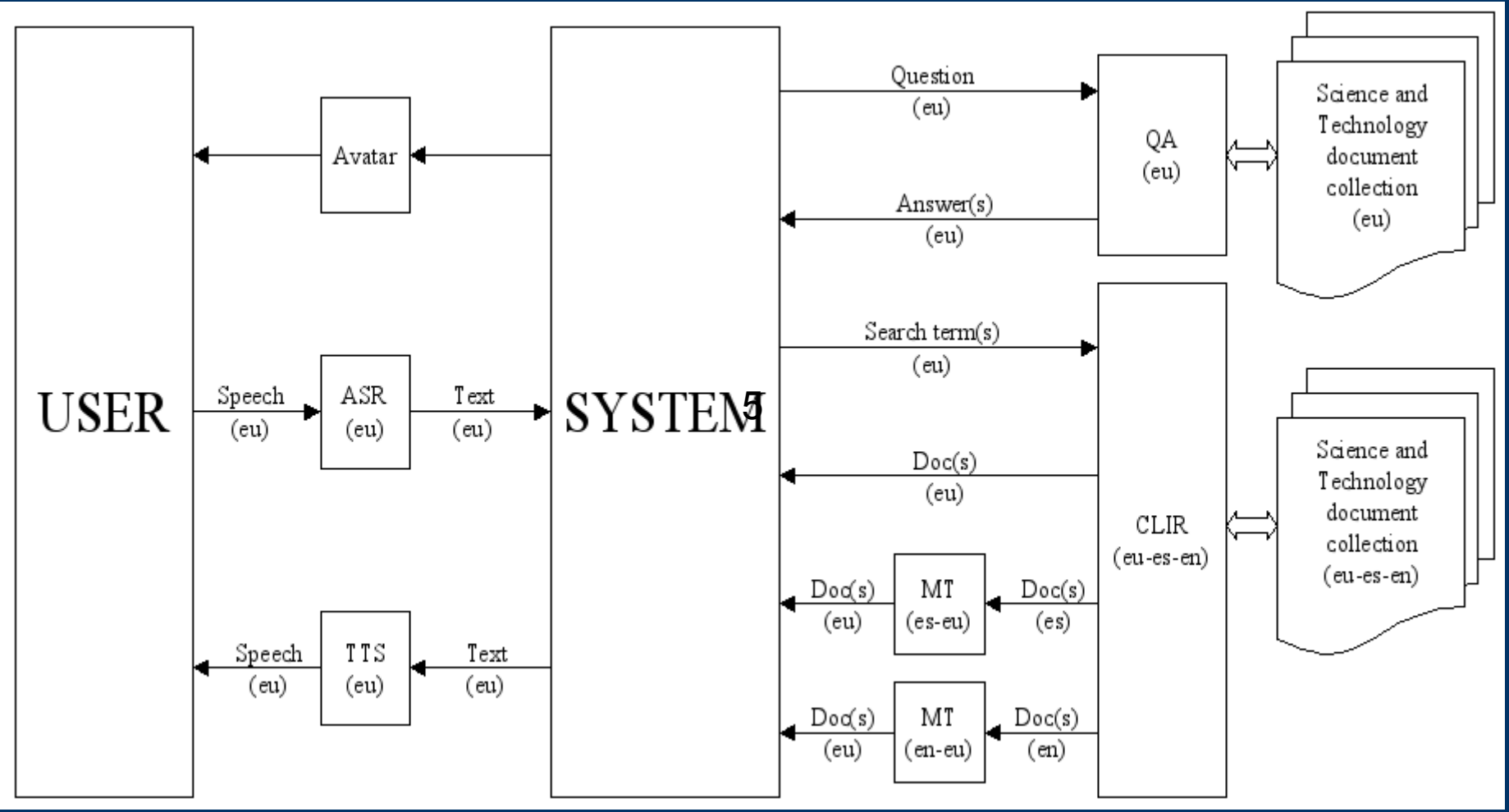
Matxin MT system
Evaluation in context *(IE-IR, MT, ASR, TTS)*

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



Matxin MT system
Evaluation in context (IE-IR, MT, ASR, TTS)



Matxin MT system
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

Matxin MT system

- 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

Combining Matxin and Corpus-based MT

Now we are building systems based on the other two MT approaches:

- EBMT
- SMT

And two hybrid MT systems:

- MEMT : Multi-Engine MT
  - EBMT + SMT + RBMT
- SPE: Statistical Post Edition
  - Statistical postediting of RBMT output

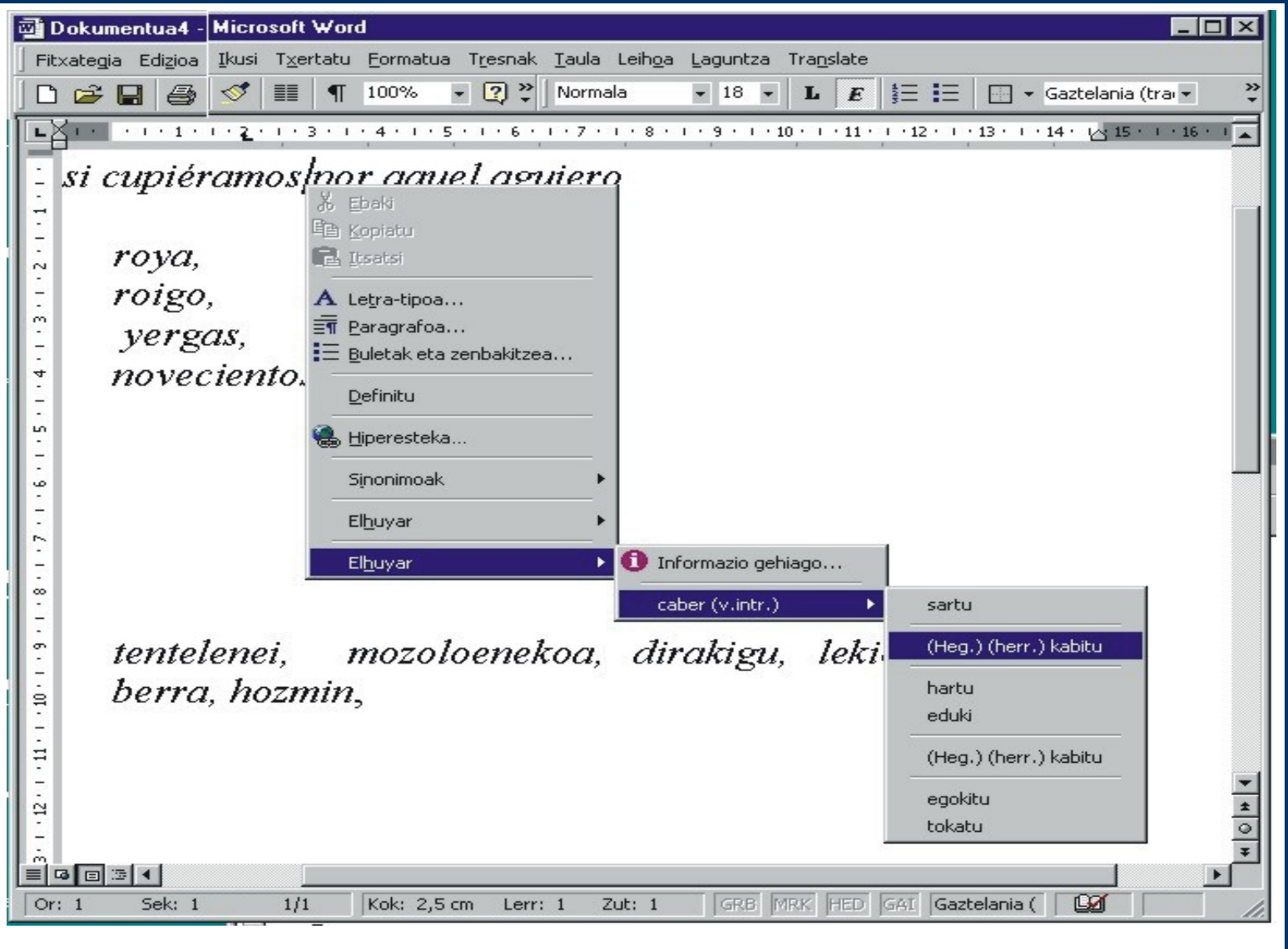
Collecting corpus for MT

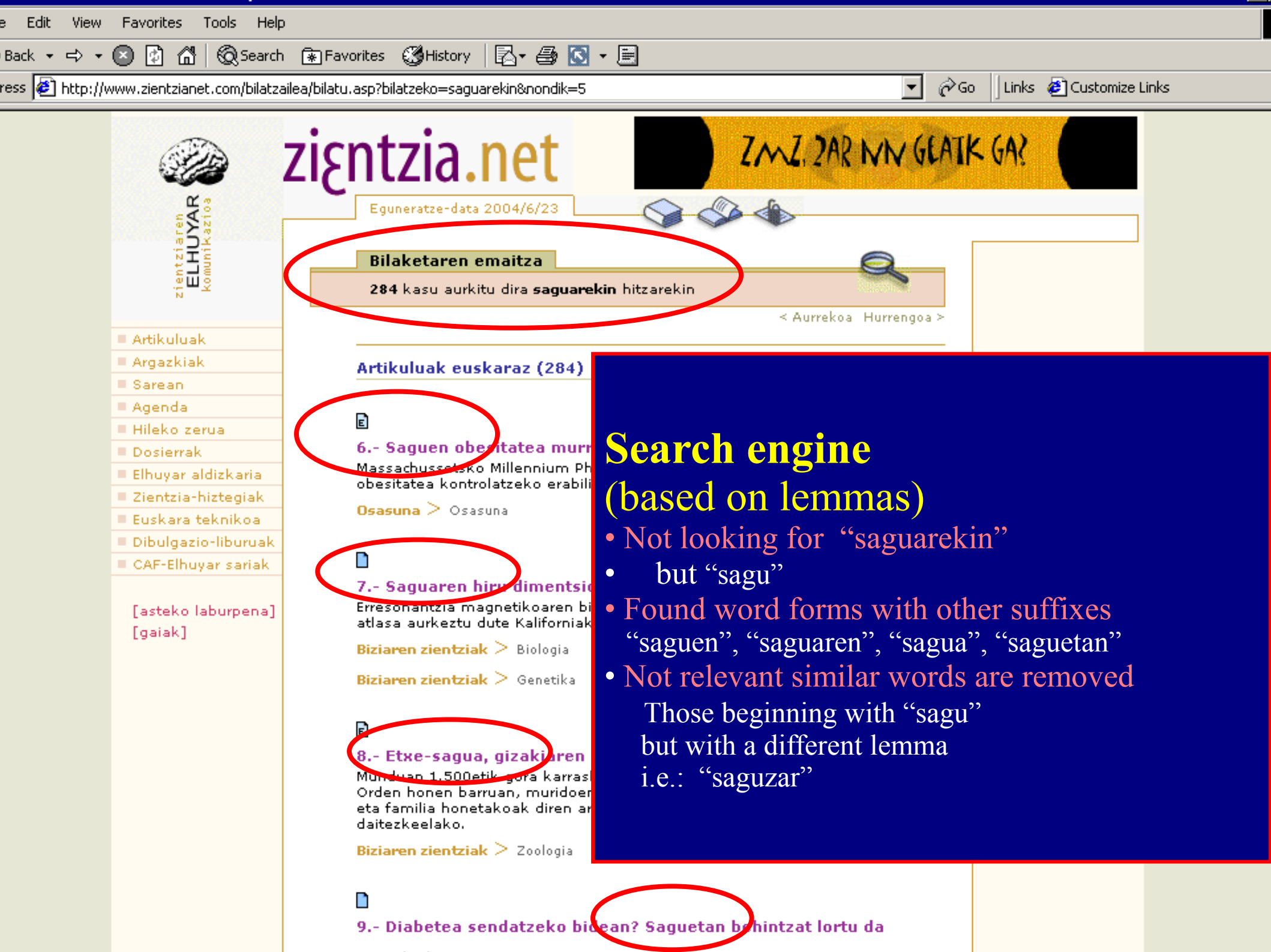
- Being Basque a less-resourced language, one of our main difficulties is getting a larger enough bilingual corpus.

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

- Labaka (2009)
  - 7 million Basque words bilingual corpus
  - Compare with Europarl (>30 million words)

Lemmatization based on-line bilingual dictionaries Basque-Spanish and Basque-French




Edit View Favorites Tools Help
Back Search Favorites History
http://www.zientzianet.com/bilatzailea/bilatu.asp?bilatzeko=saguarekin&nondik=5
Go Links Customize Links
zientzia.net
ZMZ, ZAR NN GLATK GA?
Eguneratze-data 2004/6/23
Bilaketaren emaitza
284 kasu aurkitu dira saguarekin hitzarekin
< Aurrekoa Hurrengoa >
Artikuluak
Argazkiak
Sarean
Agenda
Hileko zerua
Dosierrak
Elhuyar aldizkaria
Zientzia-hiztegiak
Euskara teknikoa
Dibulgazio-liburuak
CAF-Elhuyar sariak
[asteko laburpena]
[gaiak]
Artikuluak euskaraz (284)
6.- Saguen obesitatea mur
Massachusetsko Millennium Ph
obesitatea kontrolatzeko erabili
Osasuna > Osasuna
7.- Saguaren hiru dimentsio
Erresonantzia magnetikoaren bi
atlasa aurkeztu dute Kaliforniak
Biziaren zientziak > Biologia
Biziaren zientziak > Genetika
8.- Etxe-sagua, gizakiaren
Orden honen barruan, muridoen
eta familia honetakoak diren ar
daitezkeelako.
Biziaren zientziak > Zoologia
9.- Diabetea sendatzeko bidean? Saguetan behintzat lortu da
Search engine
(based on lemmas)
• Not looking for "saguarekin"
• but "sagu"
• Found word forms with other suffixes
"saguen", "saguaren", "sagua", "saguetan"
• Not relevant similar words are removed
Those beginning with "sagu"
but with a different lemma
i.e.: "saguzar"

Useful applications and resources

Applications

- Spelling checker/corrector
- Spanish-Basque transfer based MT system
- Lemmatization based on-line dictionaries
- Lemmatization based search machine

Resources

- EDBL: Lexical Database for Basque
- BasWN: Basque Wordnet
- EPEC: syntactically annotated Text Corpus
- ZTC: Morphosyntactically Annotated Text Corpus

EDBL: Lexical Database for Basque

- It has proved to be a multipurpose resource.
  - First developed as a lexical support for the spelling checker.
  - But nowadays, it supports :
    - speller checker,
    - morphological analyzer
    - the lemmatizer...
- It aims to reflect the general lexicon of standard Basque.
  - 100.000 entries, with morphological information.
- Development:
  - ORACLE V7 manager
  - UNIX operating system.
- Online consult: ixa2.si.ehu.es/edbl.

BasWN

- Lexical knowledge base that structures word meanings around lexical-semantic relations.
- It follows the specifications of EuroWordNet, a multilingual lexical knowledge base.
- It comprises 93.353 word senses and 59.948 words.
- Online interface which directly accesses the Basque, Spanish, Catalan and English WordNets
  ixa2.si.ehu.es/mcr/wei.html

EPEC: Syntactically Annotated Text Corpus

- 300,000 word corpus of standard written Basque
- Manually tagged at different levels:
  - morphosyntax,
  - syntactic phrases,
  - syntactic dependencies (BDT Basque Dependency Treebank)
  - BasWN word senses.
- Training corpus for the development and improvement of several NLP tools (Artola et al., 2009)

EPEC:
Syntactically Annotated Text Corpus

- First version (50,000 words) used for construction of:
  - a morphological analyzer,
  - a lemmatizer,
  - a shallow syntactic analyzer.
- This first version is publicly available
  - Ancora project (http://clic.ub.edu/ancora).
    Can be downloaded
    Can be consulted with a friendly graphic interface.
  - Natural Language Toolkit (http://www.nltk.org).

ZTC:
Morphosyntactically Annotated Text Corpus

- 10 millions words of standard written Basque texts on the subject of “Science and Technology”
  - All those words were automatically annotated
  - 1.8 million were manually revised and disambiguated.

- Still far away from the size of corpora for other languages
  - BNC corpus has 100 million words.

  However, ZTC is a very useful resource for Basque.

- On-line consult: www.ZTcorpusa.net.

ZTC:
Morphosyntactically Annotated Text Corpus

- The massive use of the lemmatizer was necessary.

  The creation of this resource would have been impossible without reusing the lemmatizer.

- We built new tools to help building ZTC:
  - corpus compilation
  - corpus annotation.
  - specific interface for advanced queries

Outline

- Basque: a Less Resourced Language
- Strategy to develop HLT for Basque
- Useful applications and resources
- Conclusions

Conclusions

- Less resourced languages have to do a great effort to face language technology.
- Ixa group has developed several applications that are effective tools to promote the use of Basque.
- From our experience R&D for less resourced languages should to be faced following this points:
  - High standardization
  - Open source
  - Reusing language foundations, tools, and applications
  - Incremental design and development of them

Conclusions

- Those guidelines seem to be trivial,
  but from our experience we know that they are not followed in many HLT projects related with these languages
- We think that if Basque is now in a quite good position in HLT is because those guidelines have been applied
  even though when it was easier to define
  "toy" resources and tools
    - useful to get good short term academic results,
    - but not reusable in future developments.

Conclusions

- This strategy has been completely useful to create HLT resources, tools and applications for Basque
- Obtaining satisfactory results in a short time
- Those products are valuable to support and to promote the use of Basque

Conclusions

- Other works related to general policies to develop resources and applications for less resourced languages:
  - (Streiter et al., 2006)
  - (Borin, 2009)
- We are planning to participate for Basque in:
  - BLARK (Krauwer, 2003)
  - CLARIN
  - LREC 2010 map of language resources Technologies and evaluation

Thank you very much!

ixa.si.ehu.es

Valuable Language Resources and Applications Supporting the Use of Basque

Iñaki Alegria, Maxux Aranzabe, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Aingeru Mayor and Kepa Sarasola

Ixa taldea.

University of the Basque Country

LTC 2009, Poznan