Putting a Less Resourced Language in the Forefront: the Case of Basque

Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza and **Kepa Sarasola**

Ixa Taldea. University of the Basque Country

http://ixa.si.ehu.es

CoCoFLaRE workshop: “Reinforcing International Collaboration in LRE”

Istanbul, May 26, 2012

Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- Strategy to develop Language Technologies for less-resourced languages
- Related work
- Conclusions

How are languages facing the ICT and HLT challenges?

- Figures about amounts of resources on the Internet for different languages are not easy to obtain

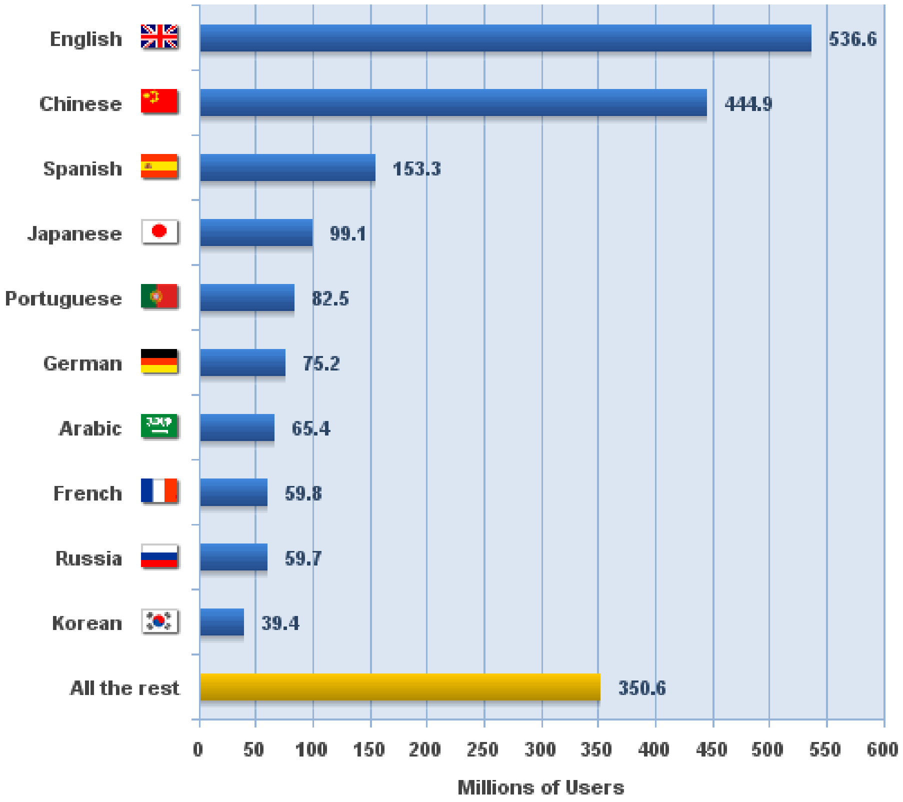
- We should use more specific public rankings
  - Internet users,
  - Internet documents
  - Wikipedia's articles.

How are languages facing ICT?

Number of users

- Internet World Stats 2010
- English :
  - 636 million users
  - 30%
- Top ten languages
  - 1.600 million users
  - 82.2%
- Rest of the languages
  - 360 million users
  - 17,8% of users
  - 36% of world population

Top Ten Languages in the Internet 2010 - in millions of users



Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,966,514,816 on June 30, 2010
Copyright © 2000 - 2010, Miniwatts Marketing Group

How are languages facing ICT?

Number of Internet documents

- Reliable statistics for different languages are scarce
- A study on the presence of Romance languages (2007)
  http://dtil.unilat.org/LI/2007/ro/resultados_ro.htm
  - 45% of the webpages were written in English,
  - 5.9% in German, 3.80% in Spanish, 4.41% in French, 2.66% in Italian, 1.39% in Portuguese,
    0.28% in Romanian, and 0.14% in Catalan.

How are languages facing ICT?

Number of articles in Wikipedia

http://meta.wikimedia.org/wiki/List_of_Wikipedias

- Articles in 282 languages (October 2011).
- Top 10 languages:
  English (3.8 million articles),
  German (1.3 M), French (1.2 M),
  Dutch, Italian, Polish, Spanish, Russian, Japanese, and Portuguese.
  - Chinese, Arabic and Korean are not in this second top list, instead of them Polish, Italian and Dutch are included.
- Surprisingly:
  - 13th: Catalan (357 K)
  - 27th: Esperanto (156 K)
  - 36th: Basque (106 K)

How are languages facing HLT?

Several public repositories:

- ELRA, LDC, ACLWiki, NLSR

Presence in the most popular linguistic services

- word processing
- search engines
- machine-translation engines

How are languages facing HLT?

ELRA European Language Resources Association.

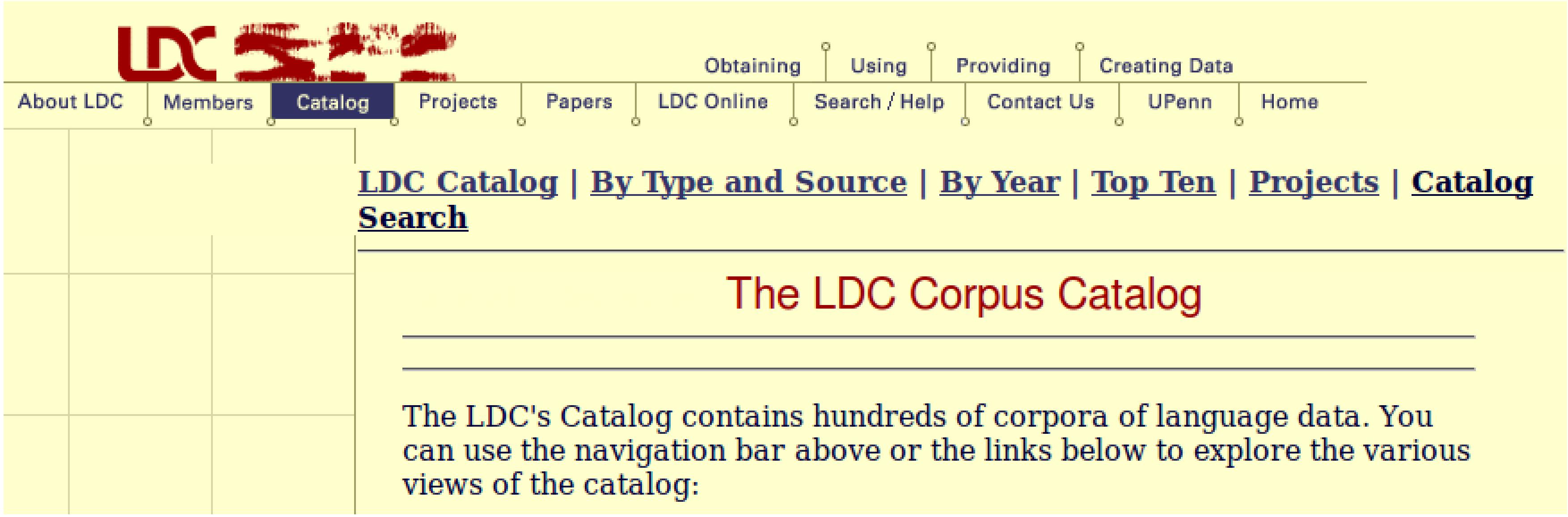
- > 1000 resources **for 60 languages**
- Resources distributed by ELRA agency

  (some products are free for research)
- 6 products for Basque.
- *The Universal Catalogue*
  - Collaborative enriching and comprising information
  - Recently added by ELRA
  - Other products not distributed by ELRA.
  - The catalog does not offer “Search by language” functionality.

How are languages facing HLT?

LDC. Linguistic Data Consortium

- > 500 resources **for 82 languages**
- Search by language is allowed.
- No products for Basque



LDC Catalog | By Type and Source | By Year | Top Ten | Projects | Catalog Search

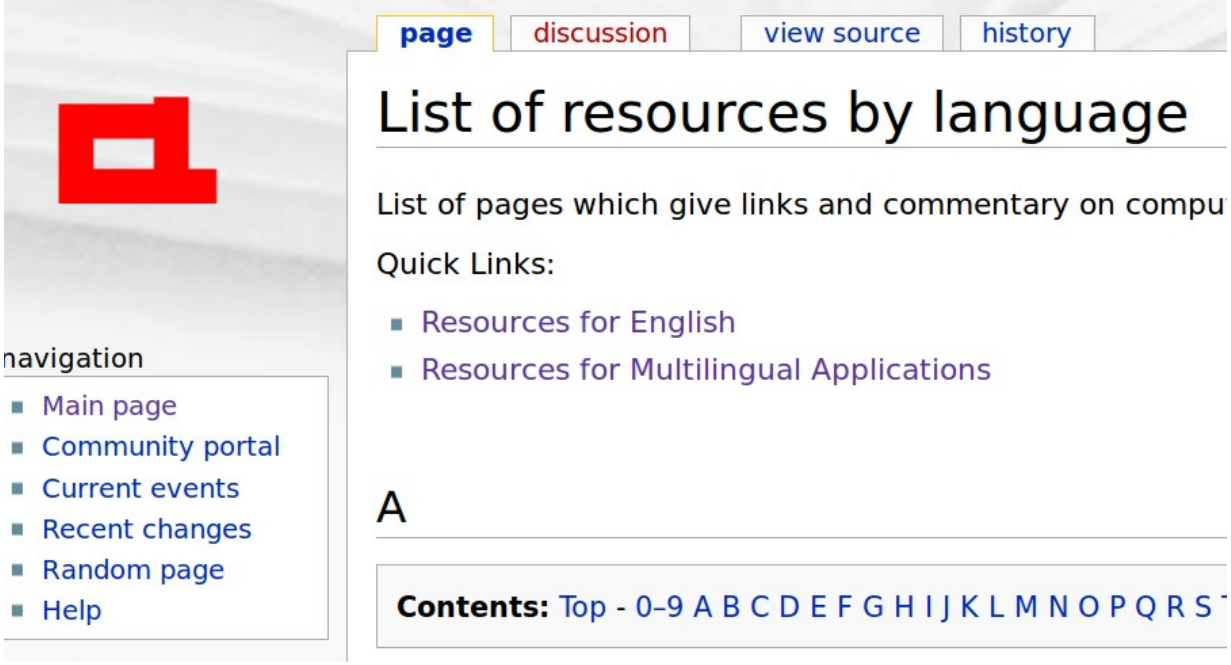
The LDC Corpus Catalog

The LDC's Catalog contains hundreds of corpora of language data. You can use the navigation bar above or the links below to explore the various views of the catalog:

How are languages facing HLT?

ACLwiki. Association for Computational Linguistics

- Resources **for 73 languages**
- Search by language is allowed.
- 15 products for Basque



List of resources by language

List of pages which give links and commentary on compu

Quick Links:

- Resources for English
- Resources for Multilingual Applications

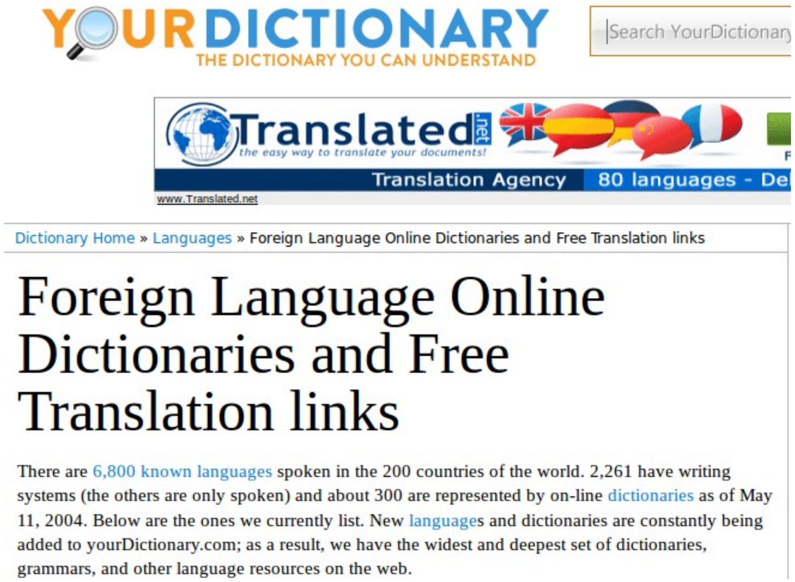
A

Contents: Top - 0–9 A B C D E F G H I J K L M N O P Q R S

How are languages facing HLT?

yourdictionary.com

- On-line lexical resources **for 300 languages**
- Search by language is allowed.
- 5 links to Basque resources
  (although they are >40)


YOURDICTIONARY
THE DICTIONARY YOU CAN UNDERSTAND
Search YourDictionary
Translated.net
the easy way to translate your documents!
Translation Agency
80 languages - De
www.Translated.net
Dictionary Home » Languages » Foreign Language Online Dictionaries and Free Translation links

Foreign Language Online Dictionaries and Free Translation links

There are 6,800 known languages spoken in the 200 countries of the world. 2,261 have writing systems (the others are only spoken) and about 300 are represented by on-line dictionaries as of May 11, 2004. Below are the ones we currently list. New languages and dictionaries are constantly being added to yourDictionary.com; as a result, we have the widest and deepest set of dictionaries, grammars, and other language resources on the web.

How are languages facing HLT?

Presence in the most popular linguistic services

- Word processing
    - MSWord
        - **91 languages**
    - Libreoffice
        - **104 languages**

How are languages facing HLT?

Presence in the most popular linguistic services

- Search engines
  - Google:
    - Identificates **45 languages**
- MT systems
  - Babelfish: **13 languages**
  - Google-Translate: **63 languages**

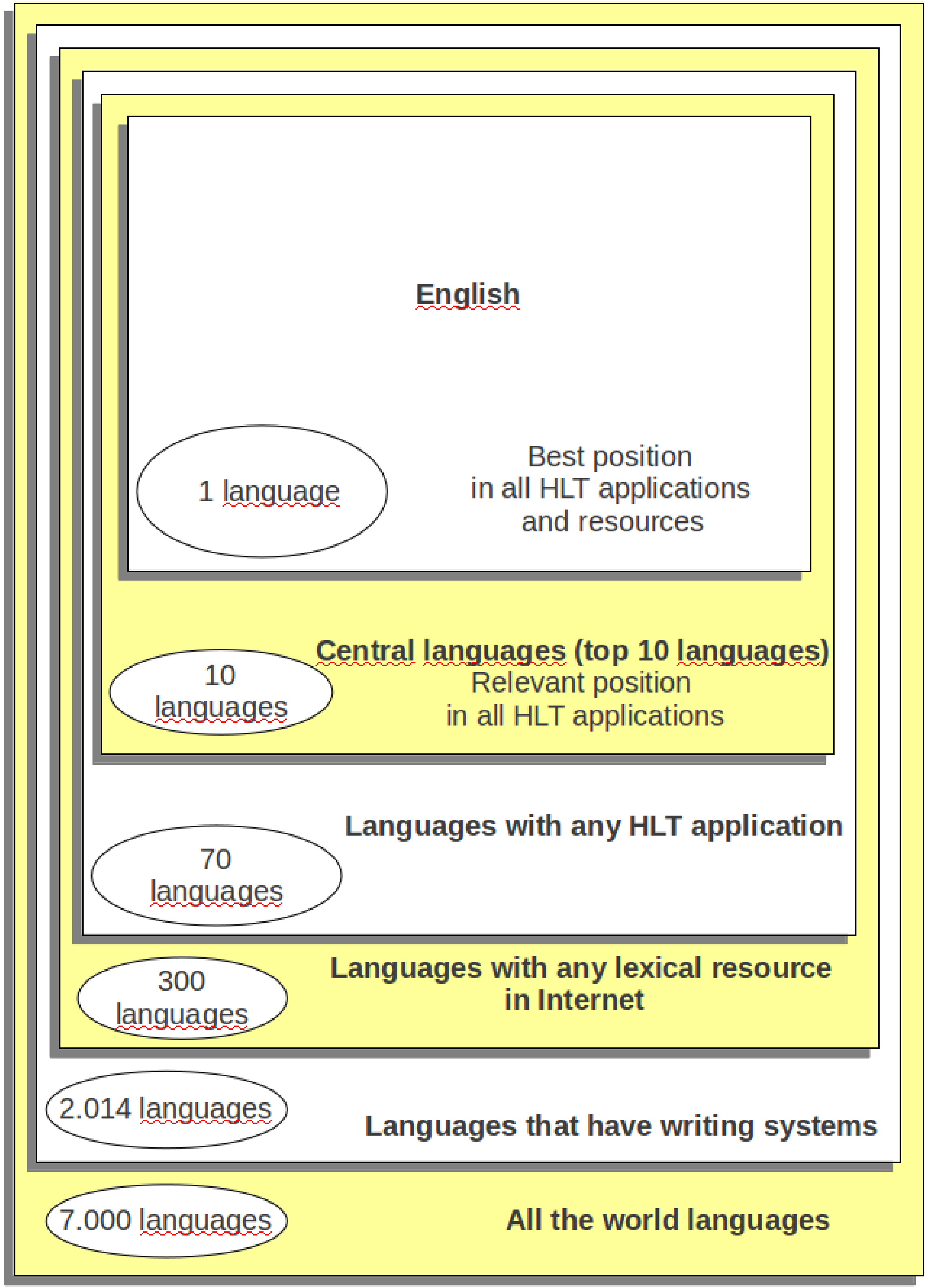
Outline

- How are languages facing the ICT and HLT challenges?
- **Which languages are "less resourced"? Six different levels**
- Strategy to develop Language Technologies for less-resourced languages
- Related work
- Conclusions

How are languages facing HLT?

Which languages are "less resourced"?

- The answer is relative
- Six different levels



English

1 language

Best position
in all HLT applications
and resources

10 languages

Central languages (top 10 languages)
Relevant position
in all HLT applications

70 languages

Languages with any HLT application

300 languages

Languages with any lexical resource in Internet

2.014 languages

Languages that have writing systems

7.000 languages

All the world languages

Which languages are "less resourced"? Six different levels

- 1. First level: English.
  - 37.9% of the users of Internet.
  - 45.00% of the web pages.
  - 62% of the HLT resources in LDC
  - 51% in ELRA.
  - Almost all the types of HLT applications.

Which languages are "less resourced"? Six different levels

- Second level: top 10 languages in the web
  - 82.2% of the Internet users (55.4% excluding English)
  - Active LR development
  - Most major categories of HLT are represented
  - Most of the resources described in LDC or ELRA are available for those languages
    - 45.79% for German, 41.27% for French,
      40.76% for Spanish; 36.24% for Italian,
    - 31.31% for Portuguese

  - Streiter et al. (2006) use "**central languages**" to refer to this set of languages.

Which languages are "less resourced"?
Six different levels

- Third level: around 70 languages.

  Languages with any HLT resource registered

  - 60 languages in ELRA,
  - 82 in LDC,
  - 73 in ACLWiki

Which languages are "less resourced"? Six different levels

- Fourth level: Around 300 languages

  Languages with any lexical resource on-line registered

  - 307 languages in *yourdictionary.com*
  - It is almost the same set of languages that is present in Wikipedia (282 languages).

Which languages are "less resourced"? Six different levels

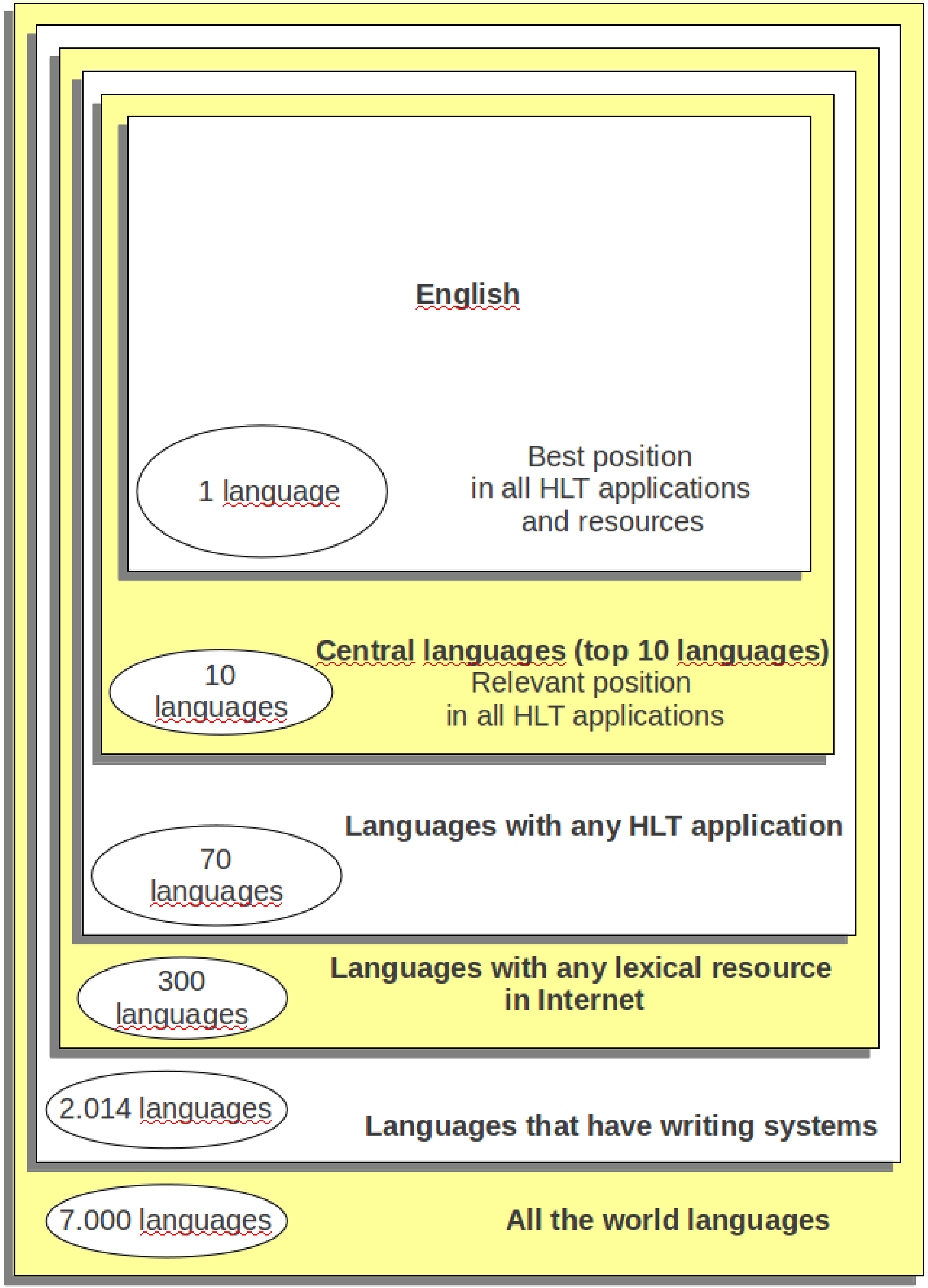
- Fifth level:
  Languages that have writing systems
  (Borin, 2009)
  - Here are included **other 2,014 languages**

- Sixth level:
  the big bag also including only-spoken languages in the world
  - Here are included at least **other 4,500 lang**.

How are languages facing HLT?

Which languages are "less resourced"?

- The answer is relative
- Six different levels



English

1 language

Best position
in all HLT applications
and resources

10 languages

Central languages (top 10 languages)
Relevant position
in all HLT applications

70 languages

Languages with any HLT application

300 languages

Languages with any lexical resource in Internet

2.014 languages

Languages that have writing systems

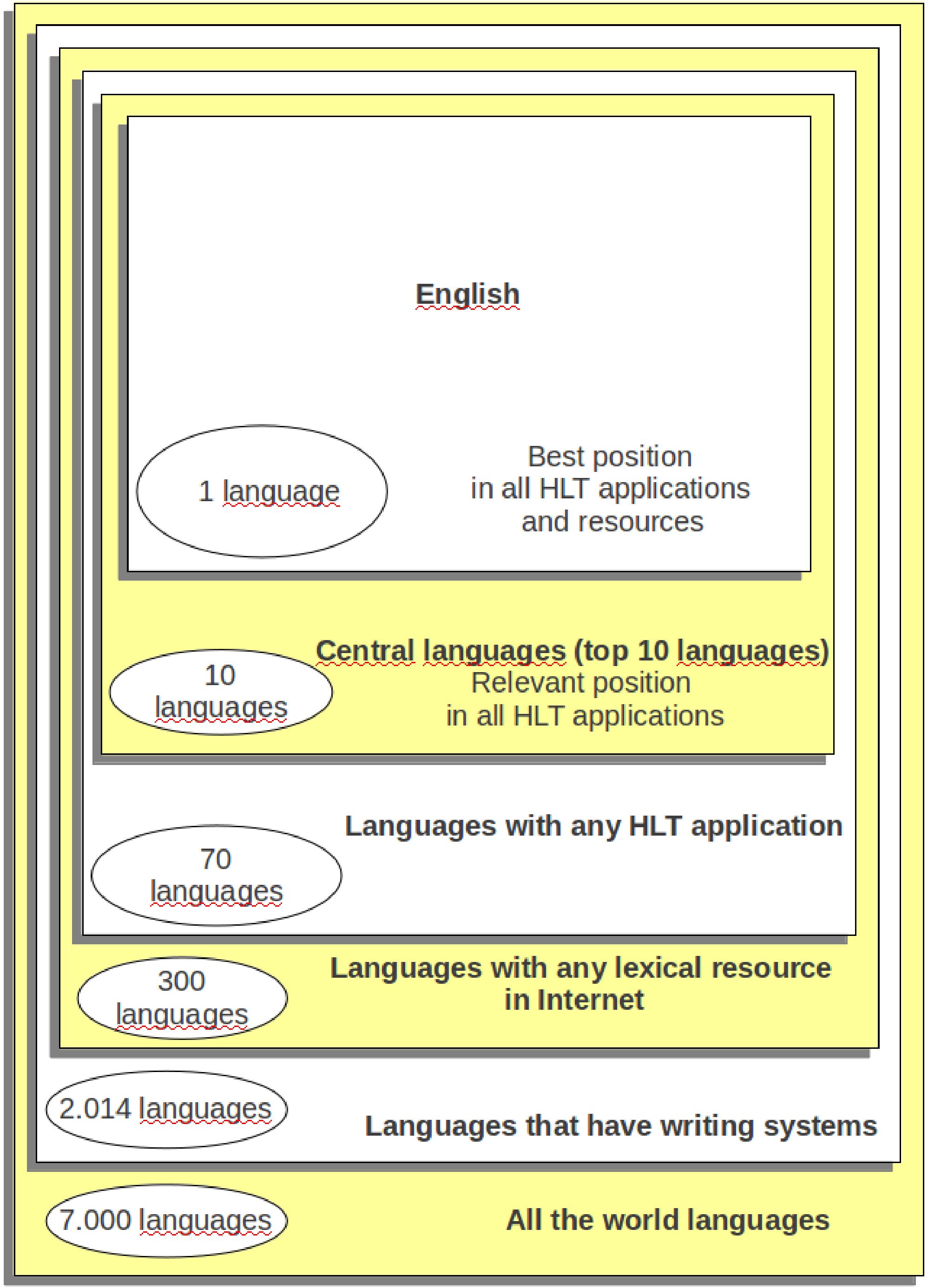
7.000 languages

All the world languages

How are languages facing HLT?

Which languages are "less resourced"?

- The 3rd or the 4th are the levels of languages usually called as **Less Resourced** in the HLT domain.
- Languages in the 5th and the 6th levels are really **endangered**



English

1 language

Best position
in all HLT applications
and resources

10 languages

Central languages (top 10 languages)
Relevant position
in all HLT applications

70 languages

Languages with any HLT application

300 languages

Languages with any lexical resource
in Internet

2.014 languages

Languages that have writing systems

7.000 languages

All the world languages

Which languages are "less resourced"? Six different levels

- This classification is not strict,
- but it may be useful to recognize application domains (sets of languages) for possible different strategies in the development of HLT resources.

Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- **Strategy to develop Language Technologies for less-resourced language**s
- Related work
- Conclusions

Strategy to develop HLT in Basque

IXA Research Group

Basque language

- < 1 million speakers.
- Very different linguistically. Alone in its language family.
- In regression for centuries. But revitalising process in the last 40 years.
- It is not an official language in Europe (partially in Basque Country).

IXA research group (created in 1988)

- Our aim was to face the challenge of adapting Basque to HLT.
- 1988: 5 members 2012: 31 computer scientists and 10 linguists
- 10 HLT products valuable to promote use of Basque.

http://ixa.si.ehu.es

Strategy to develop HLT in Basque

IXA Research Group

We presented an open proposal for making progress in HLT:

Aduriz et al., 1998

A framework for the automatic processing of Basque

First LREC Conference. Granada. 1998.

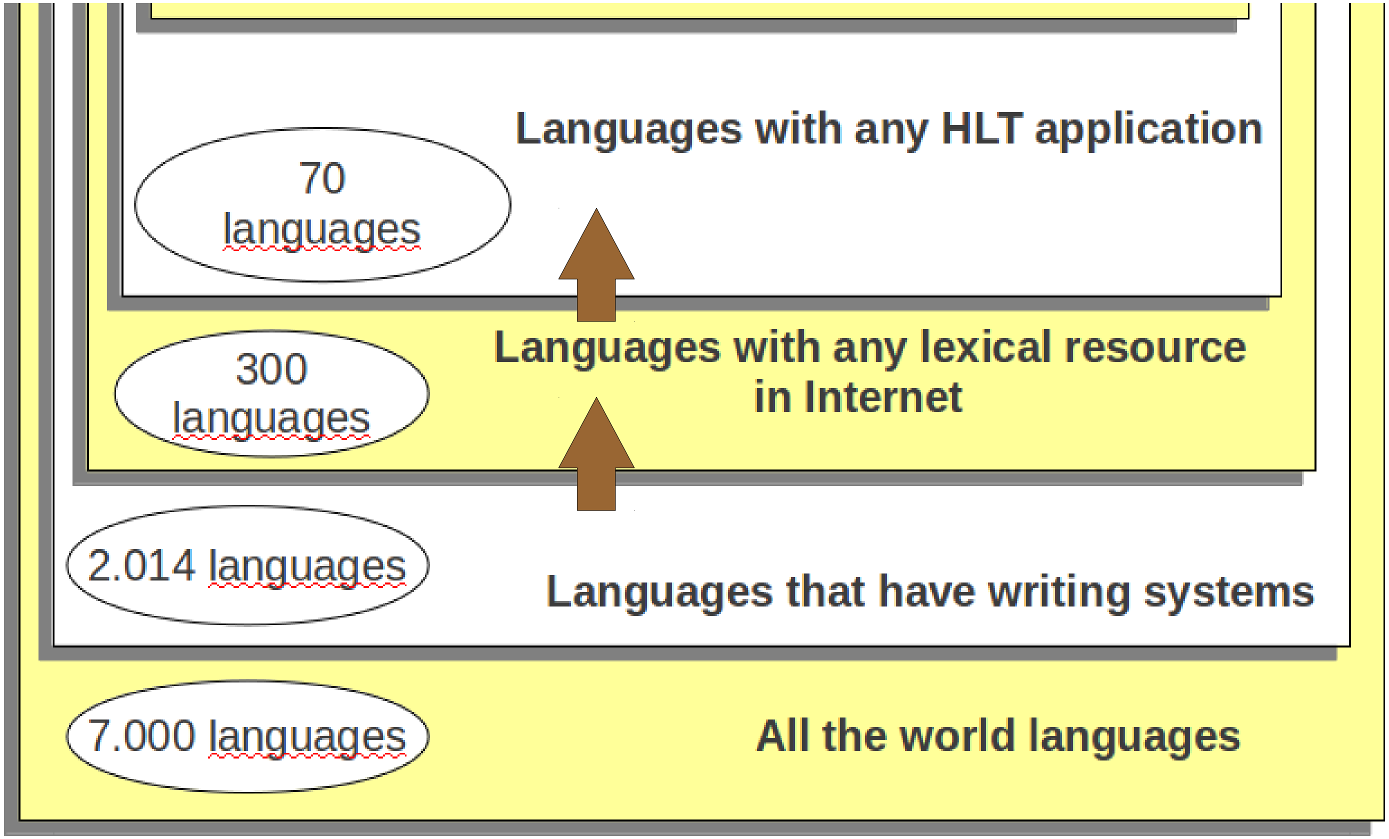
Anyway, the steps proposed did not correspond exactly with those observed in the history of the processing of English

- Different kinds of resources available
- a single coordinated plan <=> many independent efforts

Strategy to develop HLT in Basque

IXA Research Group

We consider it may be useful to promote languages from the 5th level to the 4th or from the 4th to the 3rd.


Languages with any HLT application
70 languages
Languages with any lexical resource in Internet
300 languages
2.014 languages
Languages that have writing systems
7.000 languages
All the world languages

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

- Based on open source

Strategy to develop HLT in Basque Standardization

Our steps on standardization of resources brought us

- to adopt TEI and XML standards as a basis for linguistic annotation at the different levels of processing
- to define a general methodology for corpus annotation (stand-off, representing multiple interpretations)

  - X. Artola, A. Diaz de Ilarraza, A. Soroa, A. Sologaistoa 2009
    Dealing with Complex Linguistic Annotations within a Language Processing Framework
    IEEE Transactions on Audio, Speech, and Language
  - Rigau G., Soroa A., W. Bosma, P. Vossen, M. Tesconi, A. Marchetti, M. Monachini, C. Aliprandi 2009
    KAF: a generic semantic annotation format
    Generative Lexicon 2009. pp 145-152

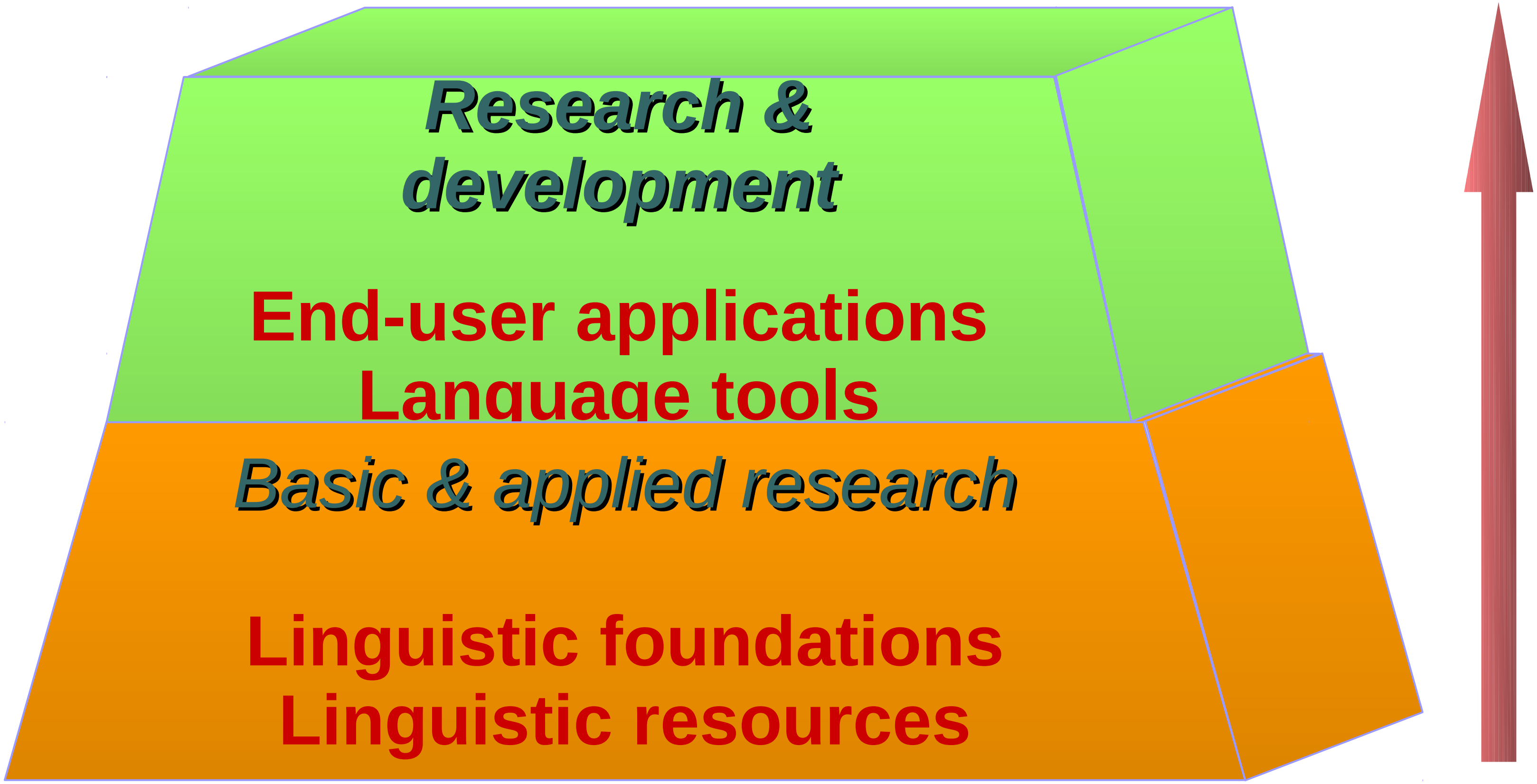
Strategy to develop HLT in Basque

Incremental design and development

We propose four phases as a general strategy for language processing

Alegria I., Aranzabe M., Arregi X., Artola X., Díaz de Ilarraza A., Mayor A., Sarasola K. 2011
Valuable Language Resources and Applications Supporting the Use of Basque
Z. Vetulani (Ed.): LTC 2009, Lecture Notes in Artifitial Intelligence LNAI 6562,

Strategic priorities: from basic research to application development



Research & development

End-user applications
Language tools

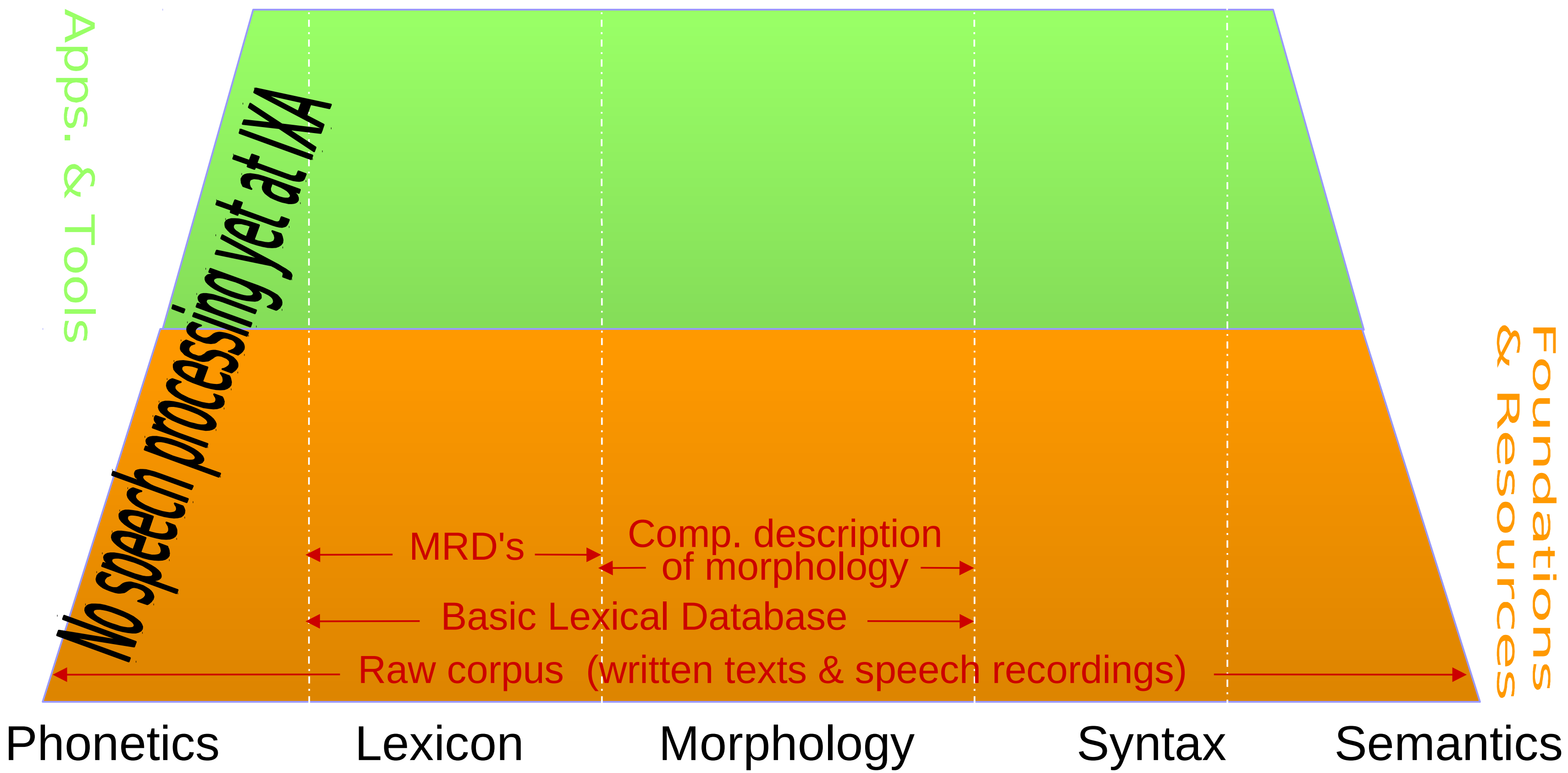
Basic & applied research

Linguistic foundations
Linguistic resources

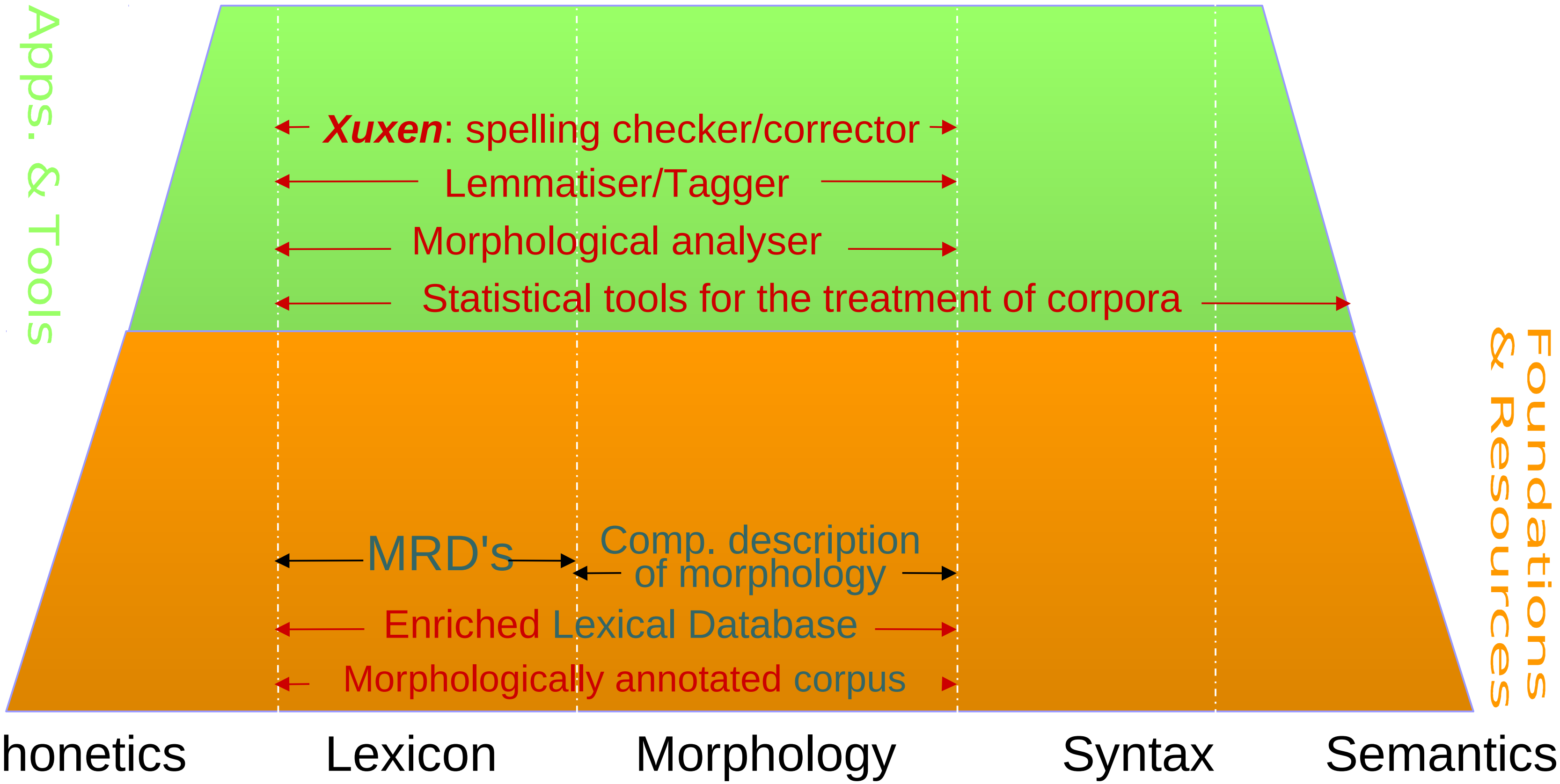
Strategy to develop HLT in Basque

Four phases

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

Foundations & Resources

MRD's

Comp. description of morphology

Enriched Lexical Database

Morphologically annotated corpus

Phonetics

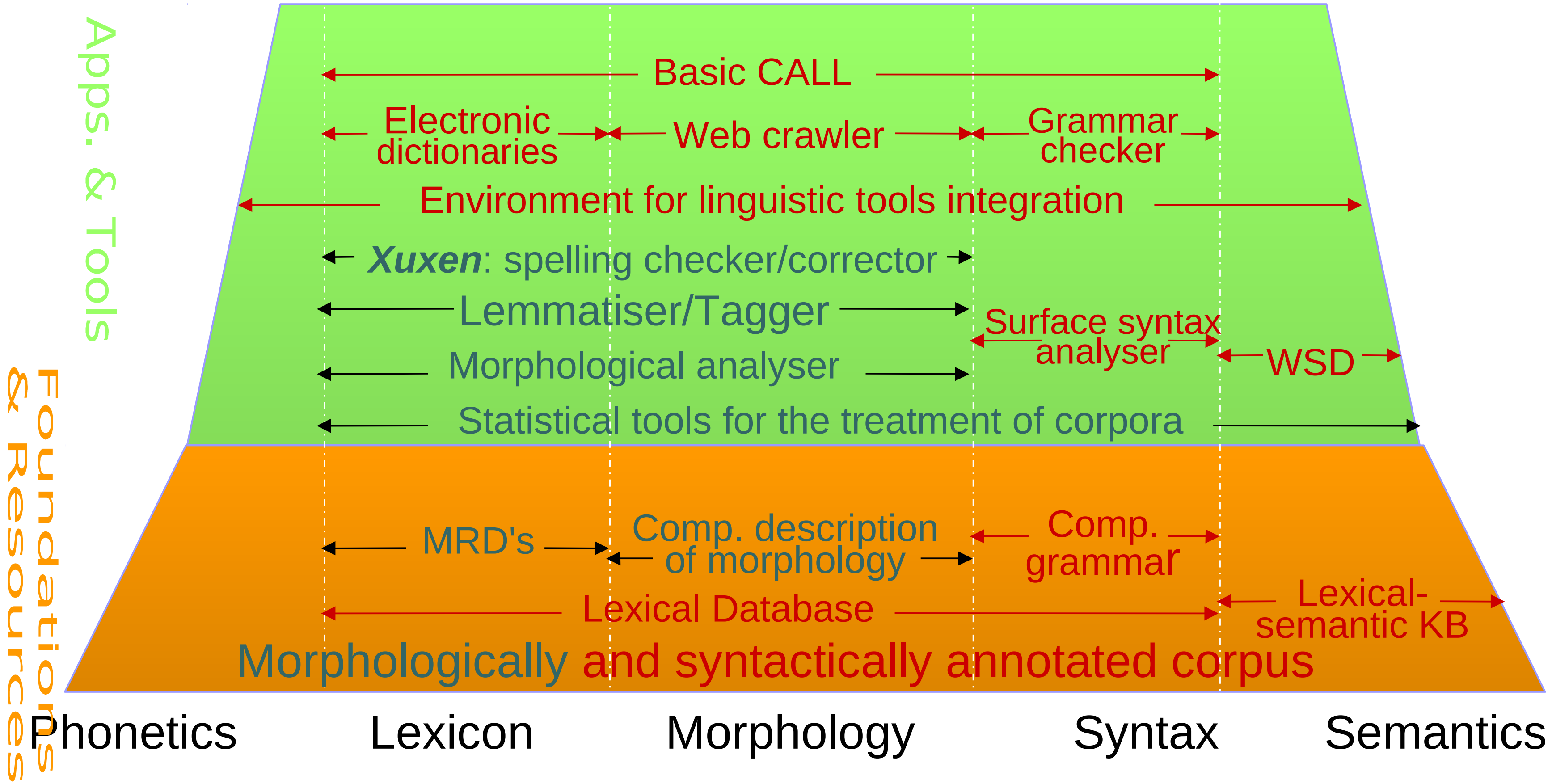
Lexicon

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

WSD

Morphological analyser

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

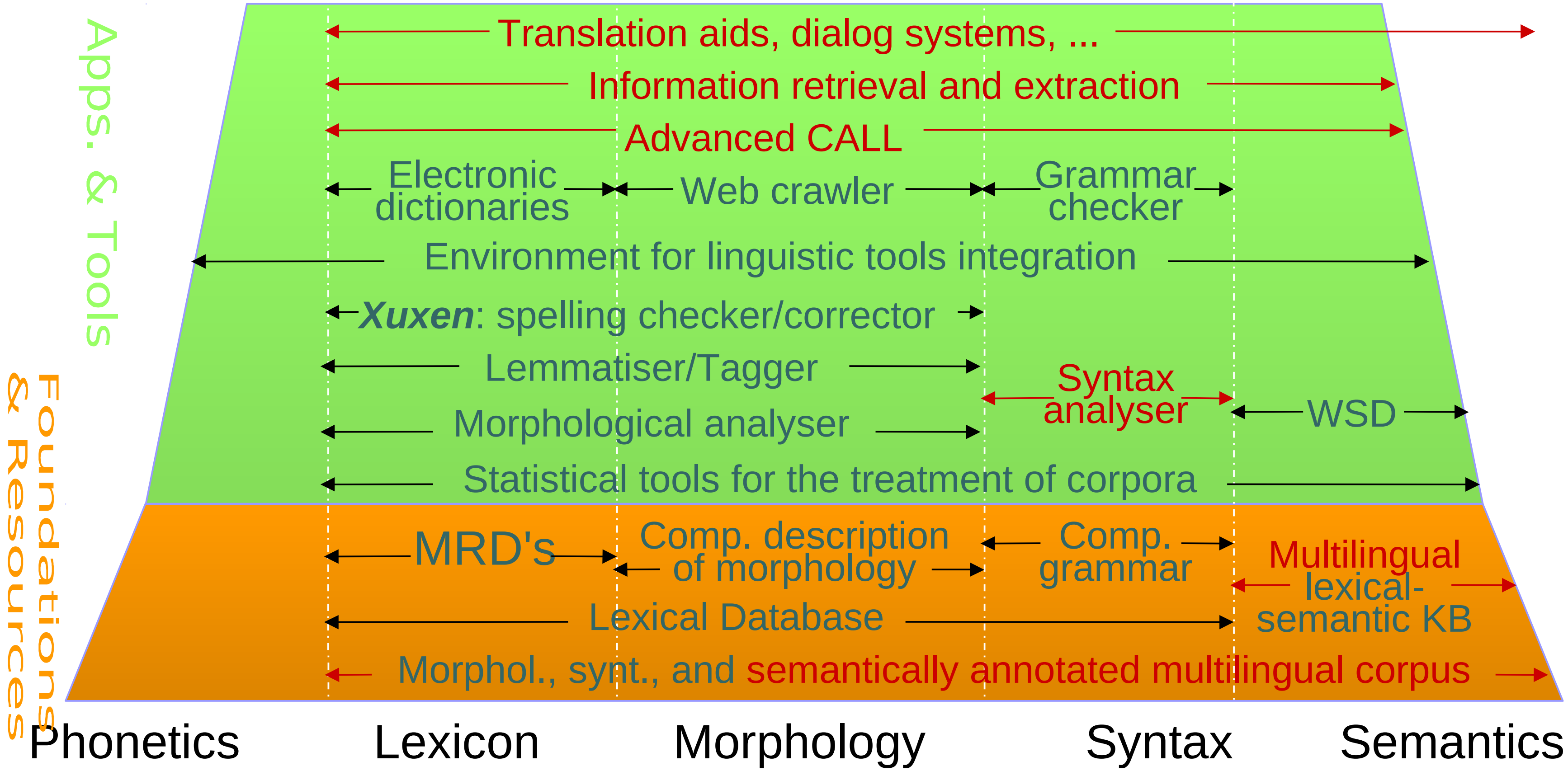
Lexicon

Morphology

Syntax

Semantics

Phase IV: multilinguality and general applications



Apps. & Tools

Foundations & Resources

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

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

Strategy to develop HLT in Basque
Open source

Using open-source programs is a key factor of success,

- **Efforts are not repeated**
  and there is a more or less **widespread making contribution**.
- Developing open-source code is more difficult and laborious, because it is necessary to **structure the programs** and **prepare good documentation**.
  - But simultaneously this is a **key factor of quality** and so, **sustainability**.
- **Tool version control systems** as SVN, and **public reposities** brings us to a better methodology and so, easier reuse.

However, arranging communities to help in enriching resources for less-resourced languages is not an easy task, **without a substantial critical mass of collaborators this kind of processes is inviable**.

Strategy to develop HLT in Basque

The strategy established a good position
to adopt those initiatives emerging during the last years:

- BLARK, Basic Language Resources Kit (Krauwer, 2003).
  Its aim was the definition of the minimal set of language resources necessary to do any precompetitive research and education,
- CLARIN (Váradi et al.2008),
  an interoperable research infrastructure of language resources and language technology that would allow to offer a stable, persistent, accessible and extendable infrastructure for the research in eHumanities;
- META-NET Network of Excellence
- Flarenet

Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- Strategy to develop Language Technologies for less-resourced languages
- **Related work**
- Conclusions

Related work

Roadmaps of tools (I)

- "Basic toolkit for HLT" (Agirre et al. 2002)   (IXA group)

- “Basic Language Resource Kit (BLARK)" (Krauwer, 2003)
  - Joint initiative between ELSNET and ELRA in 1998.
  - Maegaard et al. (2004) BLARK for Arabic
  - Simov et al. (2004)   BLARK for Bulgarian ...
  - The term BLARK has been very successful and it is used in a large number of papers in the area.

Related work

Roadmaps of tools (II)

- The ELSNET network of excellence prepared definitions for a language resources and evaluation roadmap (Busemann & Uszkoreit, 2004).
  - Several different roadmaps have been published.
  - Their level of granularity in the diagram elements is very much fine than ours
  - Definition of a roadmap for “central languages”, mainly for the main European official languages
- Meta-NET white papers (2012)

Related work

- Streiter et al. (2006)
  - They propose instructions for funding bodies and strategies for developers.
  - They use the *non-central* term and
  - Benefits and problems when using open source software for non-central languages.

- Forcada (2006)
  - He remarks the opportunity of using open source machine translation for minor languages.
  - Apertium initiative

Related work

- Ostler (1998):
  - "*a language will not get by in the world of today unless it is equiped with* ***a parser*** *and* ***a multi-million-word corpus of text***".
- Borin (2006 and 2009)
  - relation among the sociology of language and HLT
  - Some strategic considerations, "*those languages for which* ***information extraction resources and tools*** *will be available will probably exhibit a more secure and prominent presence on the Semantic Web than those lacking such resources, and as a consequence, acquire the status in the eyes of their speakers that such a presence confers*".

Related work

- Efforts to create, coordinate and make language resources and technology available and readily usable for a big number of languages
  - Clarin
  - Flarenet
  - MetaNet
  - ELRA

  - SALTMIL (http://ixa2.si.ehu.es/saltmil)
    Speech And Language Technology for Minority Languages
  - AfLaT (http://AfLaT.org)
    Language technology research for African languages

Conclusions

- From our experience we defend that research and development for less resourced languages should to be faced to build a BLARK following this points:
  - 1) high standardization
  - 2) open-source
  - 3) reusing language foundations, tools, and applications
  - 4) incremental design and development of them.
- We have defined six different sets of languages attending to their penetration on HLT technologies.
- We think that our strategy to develop language technologies could be **useful for several hundred languages:**
  those that have developed a **written standard**
  and perhaps also some **initial lexical resources**
  but that are **still very far from central languages.**

Thanks

Eskerrik asko

kepa.sarasola@ehu.es

ixa.si.ehu.es

Putting a Less Resourced Language in the Forefront: the Case of Basque

Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza and **Kepa Sarasola**

Ixa Taldea. University of the Basque Country

http://ixa.si.ehu.es

CoCoFLaRE workshop:

“Reinforcing International Collaboration in LRE”

Istanbul, May 26, 2012