

# Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque



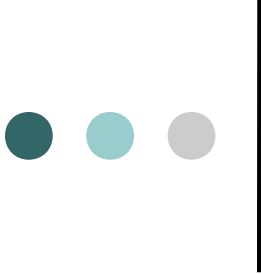
Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza  
and **Kepa Sarasola**

Ixa Taldea. University of the Basque Country

<http://ixa.si.ehu.es>



Language Technology Conference, LTC2011 Poznan



# LTC 2011 topics

- The idea is to discuss availability, quality, maturity, sustainability, and gaps of the **LR and LT for a number of languages and technologies.**
- **Recommendations on the way to address these gaps based on experience from well resourced languages**
- **Experience in the production, validation and distribution of LR for less-resourced 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 the ICT and HLT challenges?

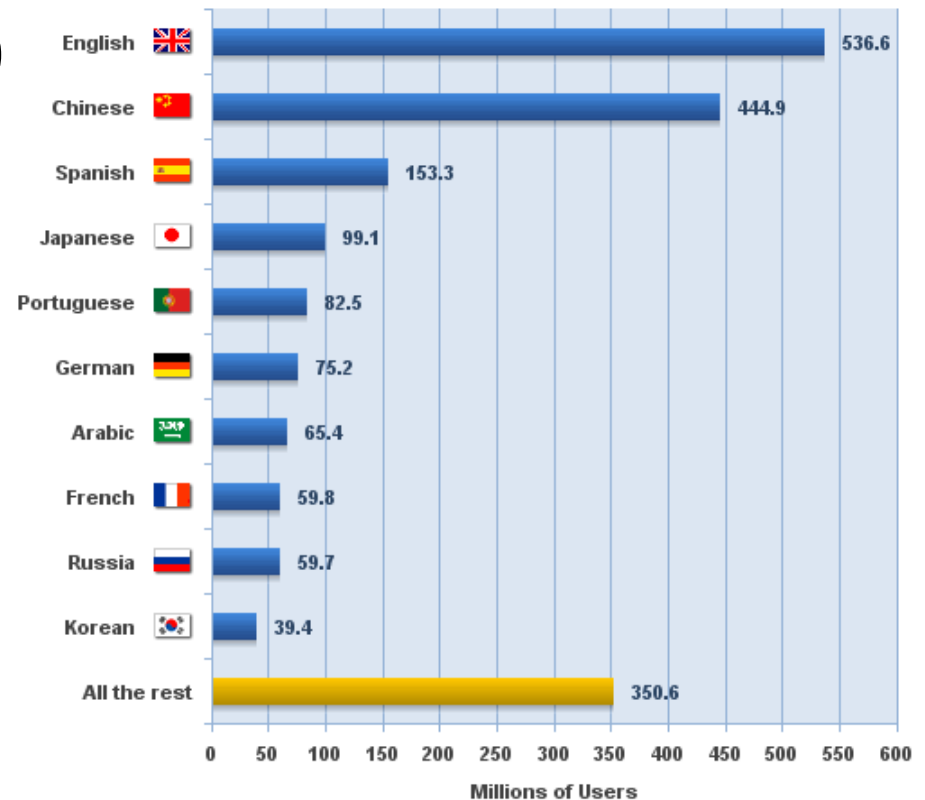
- Figures about amounts of resources on the Internet for different languages are not easy to obtain
- We should use more specific public rankings
  - Internet users,
  - Internet documents
  - Wikipedia's articles.

# 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](http://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://dttil.unilat.org/LI/2007/ro/resultados\\_ro.htm](http://dttil.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.
- Alternative way:
  - "Web as a Corpus" (Kilgarriff & Grefenstette, 2003)
  - Obtain figures for a language using APIs of search engines (if recognized by the engine)



# How are languages facing ICT?

## Number of articles in Wikipedia

[http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](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/absence in the most popular linguistic services

- word processing
- search engines
- machine-translation engines





# How are languages facing HLT?

Several public repositories:

- ELRA
- LDC
- ACLWiki
- NLSR

These information sources are not always complete

- Repositories refer to the products they offer
  - manage resources and sell some of them
- Wiki-like sites only to those entered by volunteers
  - just for consulting

# How are languages facing HLT?



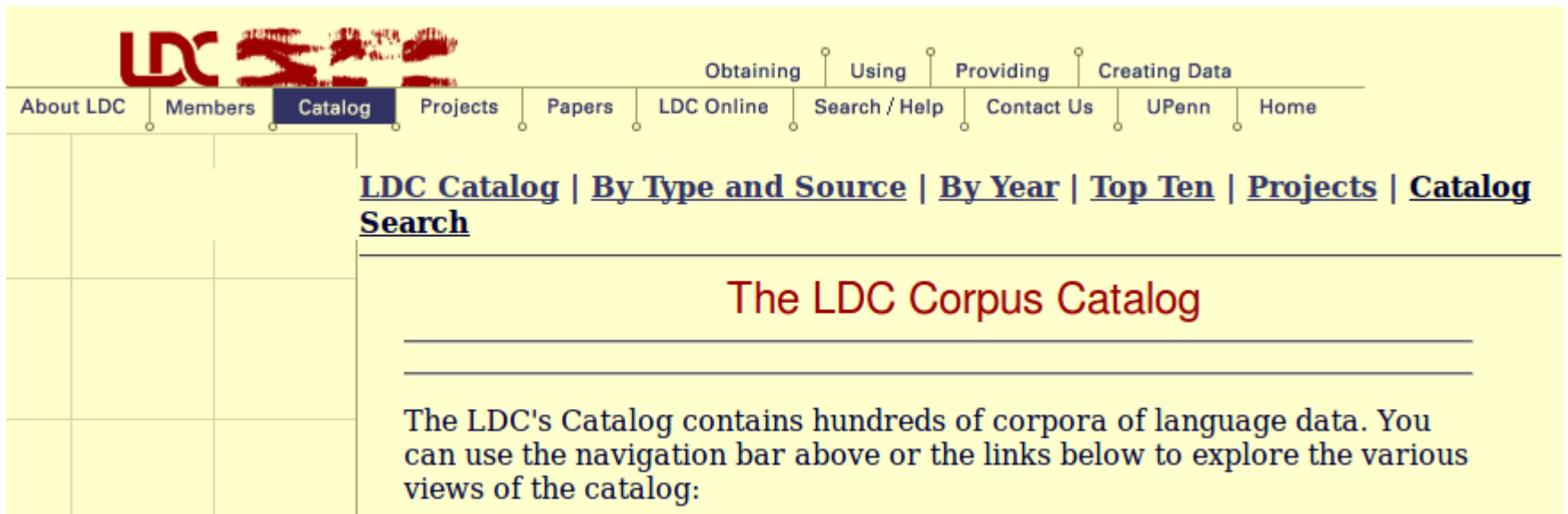
## **ELRA European Language Resources Association.**

- > 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

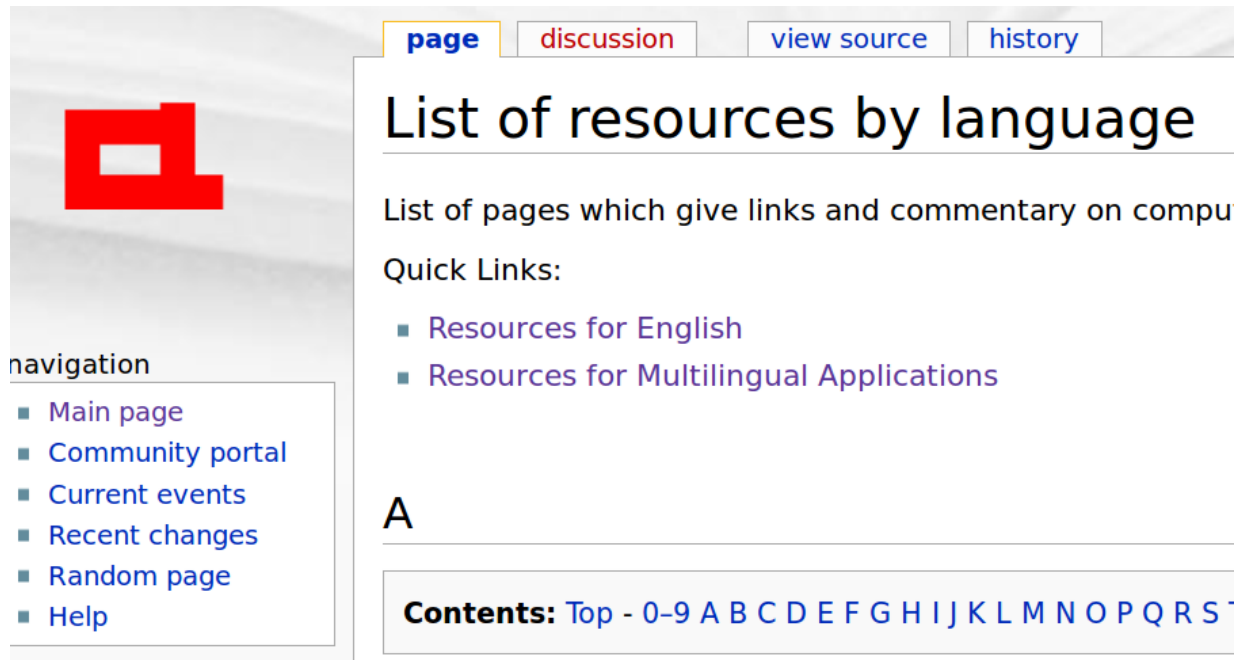


The screenshot shows the LDC Catalog website. At the top left is the LDC logo. A navigation bar contains links: About LDC, Members, Catalog (highlighted), Projects, Papers, LDC Online, Search / Help, Contact Us, UPenn, and Home. Above the main content area, there are links for Obtaining, Using, Providing, and Creating Data. Below the navigation bar, there are links for [LDC Catalog](#), [By Type and Source](#), [By Year](#), [Top Ten](#), [Projects](#), and [Catalog Search](#). The main heading is "The LDC Corpus Catalog". Below this, a paragraph states: "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

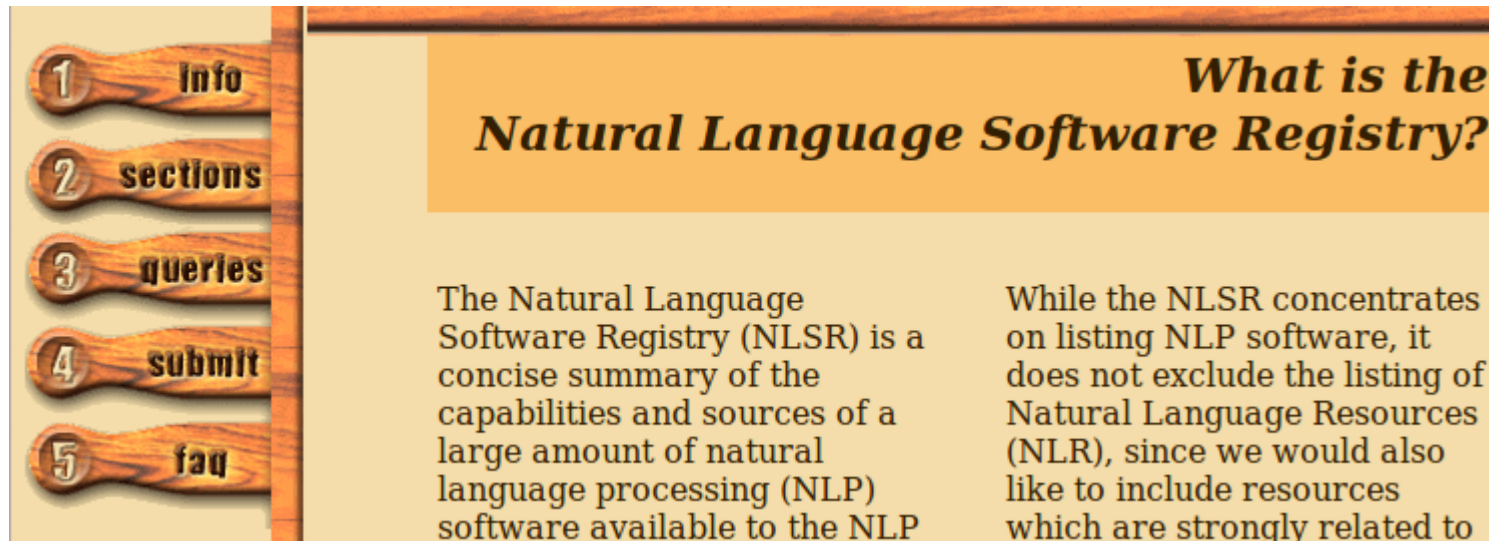


The screenshot shows the ACLwiki website interface. At the top, there are navigation tabs: [page](#), [discussion](#), [view source](#), and [history](#). The main heading is "List of resources by language". Below this, there is a description: "List of pages which give links and commentary on compu". Underneath, there is a "Quick Links:" section with two items: [Resources for English](#) and [Resources for Multilingual Applications](#). On the left side, there is a navigation menu with a red logo above it, containing links for [Main page](#), [Community portal](#), [Current events](#), [Recent changes](#), [Random page](#), and [Help](#). At the bottom, there is a "Contents:" section with a list of letters: [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?

## NLSR. Natural Language Software Registry (DFKI)

- Resources for **30 languages**
- Search by language is allowed.
- 3 products for Basque
- 59 products for “any language”



The image shows a screenshot of the NLSR website. On the left, there is a vertical navigation menu with five items: 1 info, 2 sections, 3 queries, 4 submit, and 5 faq. The main content area has a title "What is the Natural Language Software Registry?" in a bold, italicized font. Below the title, there are two columns of text. The left column starts with "The Natural Language Software Registry (NLSR) is a concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP". The right column starts with "While the NLSR concentrates on listing NLP software, it does not exclude the listing of Natural Language Resources (NLR), since we would also like to include resources which are strongly related to".

**1 info**

**2 sections**

**3 queries**

**4 submit**

**5 faq**

### *What is the Natural Language Software Registry?*

The Natural Language Software Registry (NLSR) is a concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP

While the NLSR concentrates on listing NLP software, it does not exclude the listing of Natural Language Resources (NLR), since we would also like to include resources which are strongly related to

# 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)



[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/absence in the most popular linguistic services

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

Basque is in both



# How are languages facing HLT?

Presence/absence in the most popular linguistic services

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





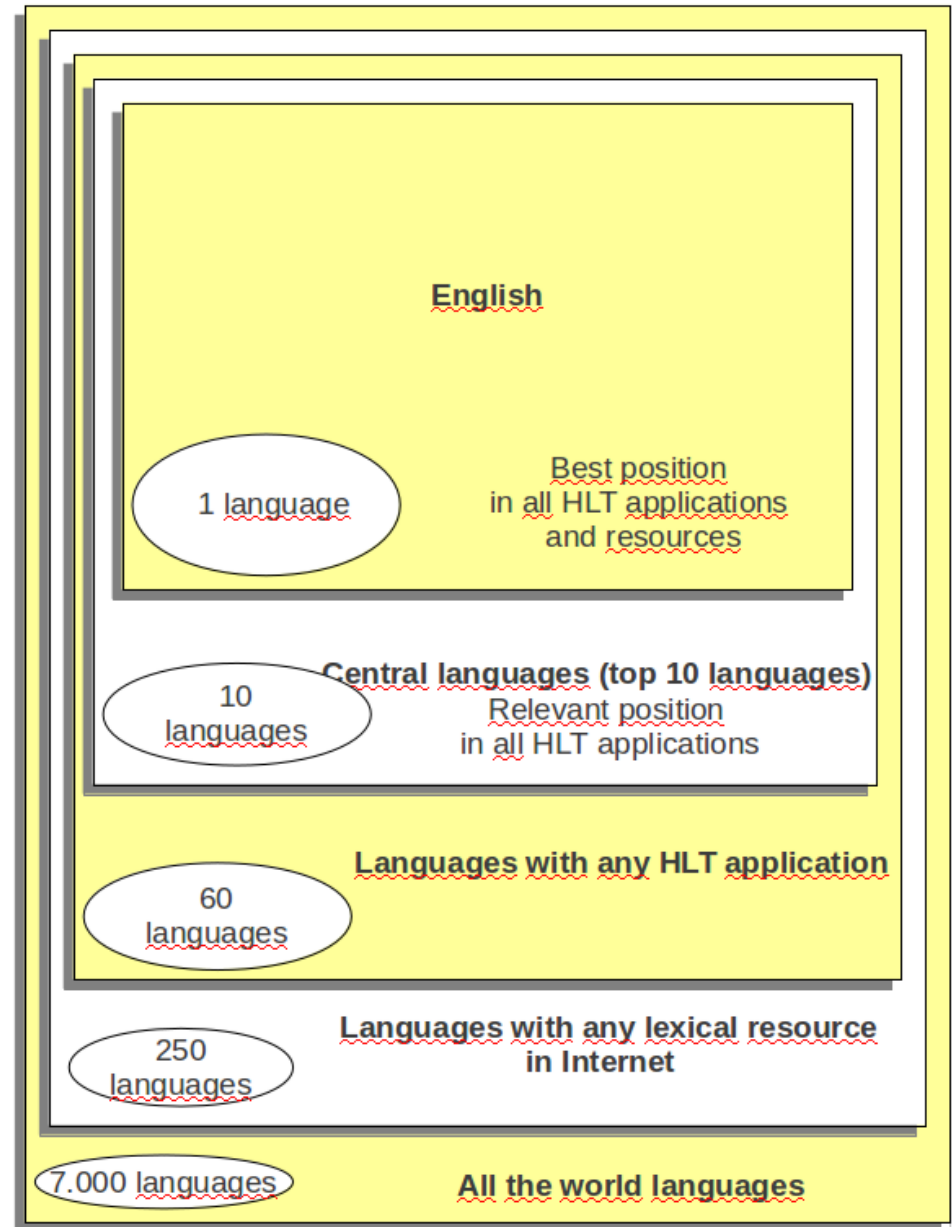
# Outline

- 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





## 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 continues
  - Most major categories of HLT are represented
  - Most of the HLT kind of 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 the term "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
- 30 in NLSR.



## 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"?

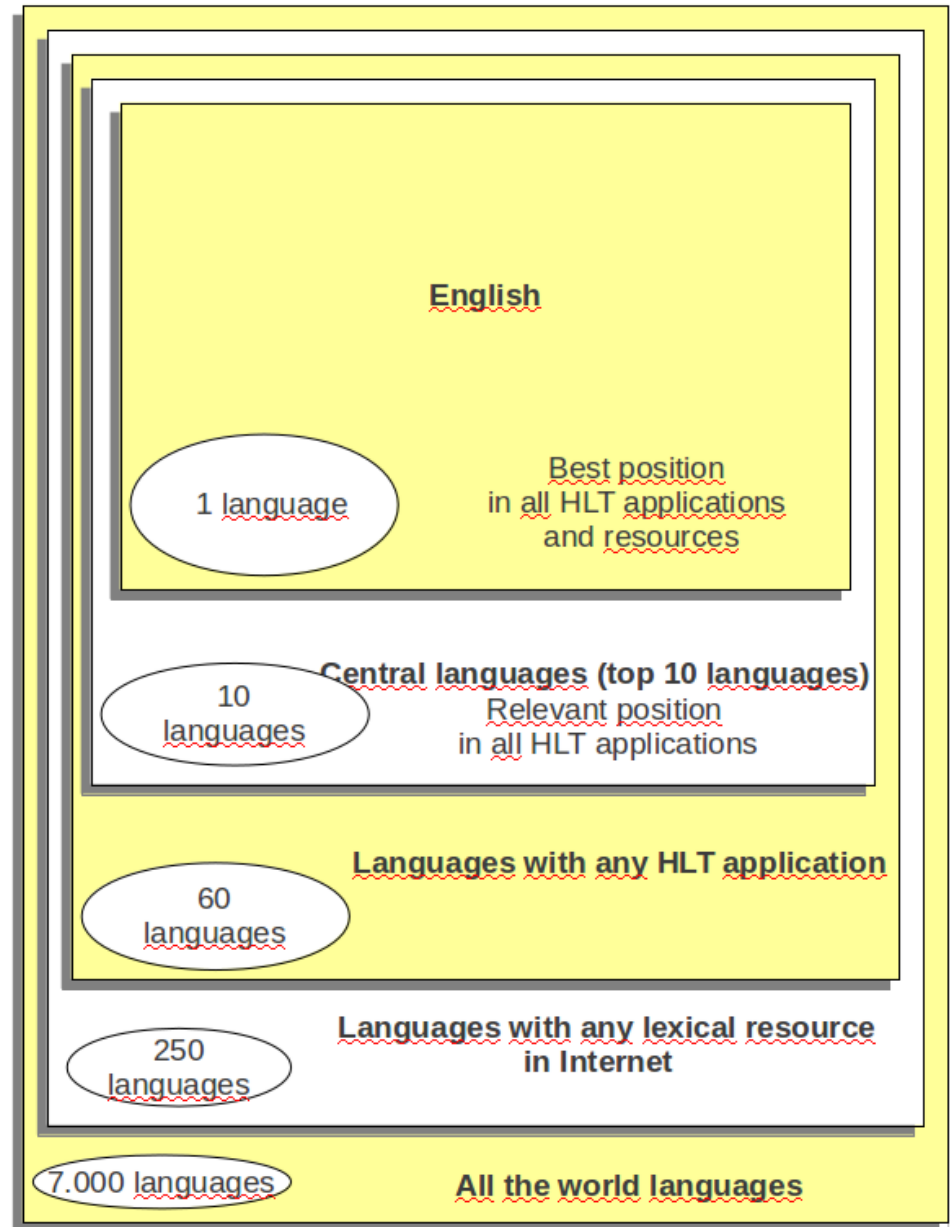
### Six different levels

- 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







## Which languages are "less resourced"?

### Six different levels

This 6 level typology gives a **relative definition of less-resourced languages**

- Comparing with English all the other languages could be considered less-resourced
- Or ...except the 10 top languages the rest can be considered less-resourced.
- The languages of the third level are lesser resourced than the languages of the second level, by definition
- 3<sup>rd</sup> or the 4<sup>th</sup> are the levels of languages usually called as less-resourced in the HLT domain.
- We may consider that languages in the 5<sup>th</sup> and the 6<sup>th</sup> levels are really endangered,



## 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.



# Which languages are "less resourced"?

## Six different levels

- However, there are some risks on the application of these indicators:
  - Languages with very active proponents may have a high visibility on Wikipedia which may not be significative of the presence of the language on the general Internet.
    - For example, Catalan appears in the 13<sup>th</sup> position in the ranking of the number of articles in Wikipedia, but Catalan is usually taken as a less resourced language
  - Nevertheless the Wikipedia indicator is:
    - highly accessible,
    - automatically updated for all the languages,
    - useful when used in conjunction with other indicators.



# 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



# Strategy to develop HLT in Basque IXA Research Group

- IXA group: research group created in 1988.
- Our aim was to face the challenge of adapting Basque to HLT.
  - 1986: 5 university lecturers (computer science)
  - 2011: Interdisciplinary team
    - *31 computer scientists and 10 linguists*
- *Collaborating with 7 companies from Basque Country and 5 from abroad*
- *Involved in the birth of two new spin-off companies*
- *10 HLT products valuable to promote use of Basque.*

<http://ixa.si.ehu.es>

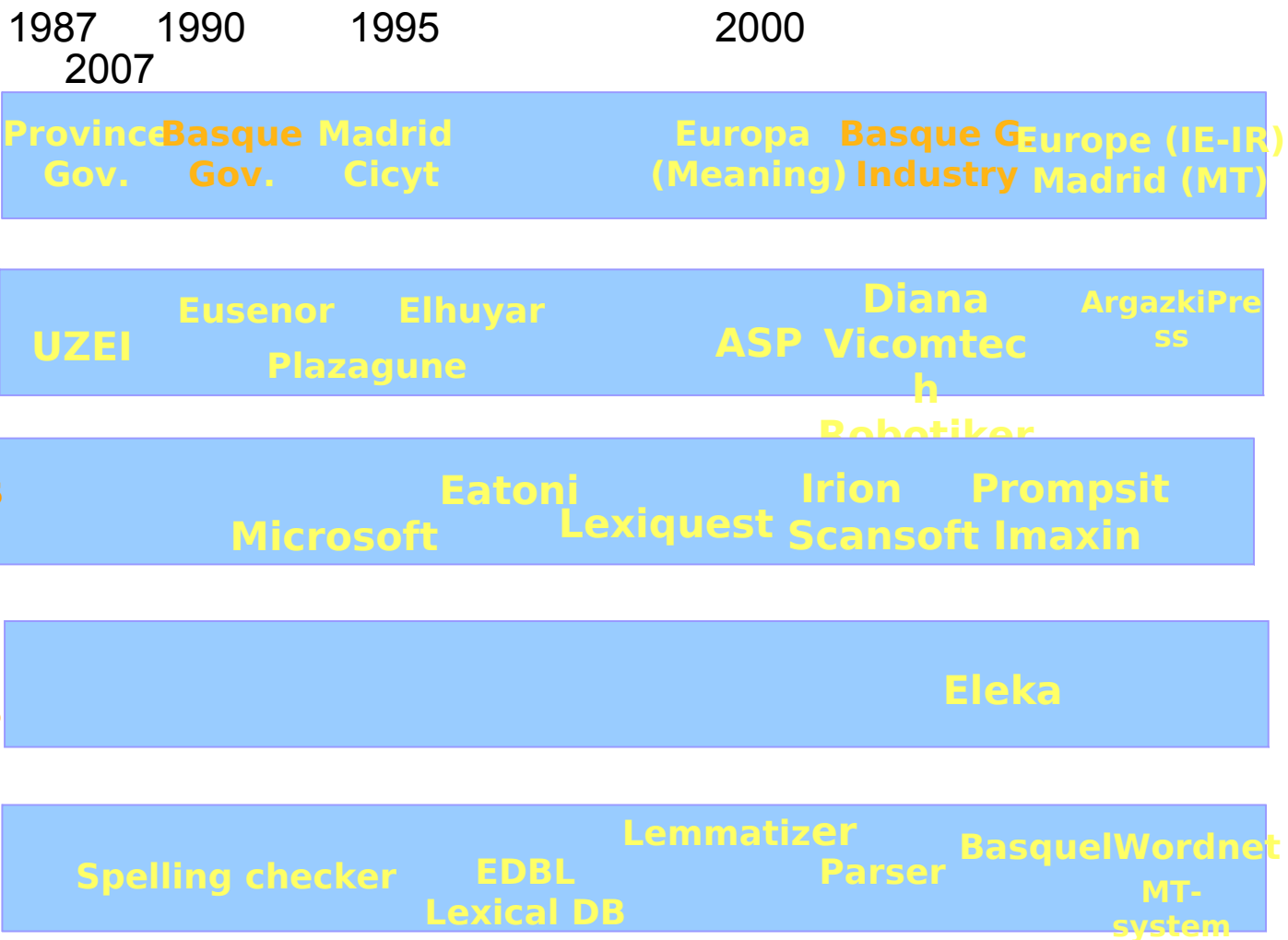


# Strategy to develop HLT in Basque IXA Research Group

We presented an open proposal for making progress in HLT (Aduriz et al., 1998).

- Anyway, the steps proposed did not correspond exactly with those observed in the history of the processing of English
  - Resources available for the treatment of Basque allowed facing problems in a different way
  - English LRs did not evolve as the result of a single coordinated plan.
  - Instead many independent efforts produced these English LRs to address specific project needs.

# 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



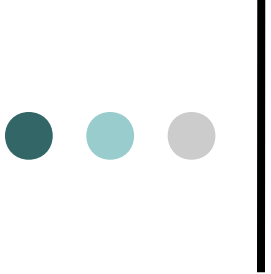


# Strategy to develop HLT in Basque IXA Research Group

- 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
  - definition of a general methodology for corpus annotation (Artola et al., 2009).
- Taking as reference our experience in incremental design and development of resources/tools,
  - We propose four phases as a general strategy for language processing (Alegria et al., 2011)

● ● ● Strategic priorities:  
from basic research to  
application development

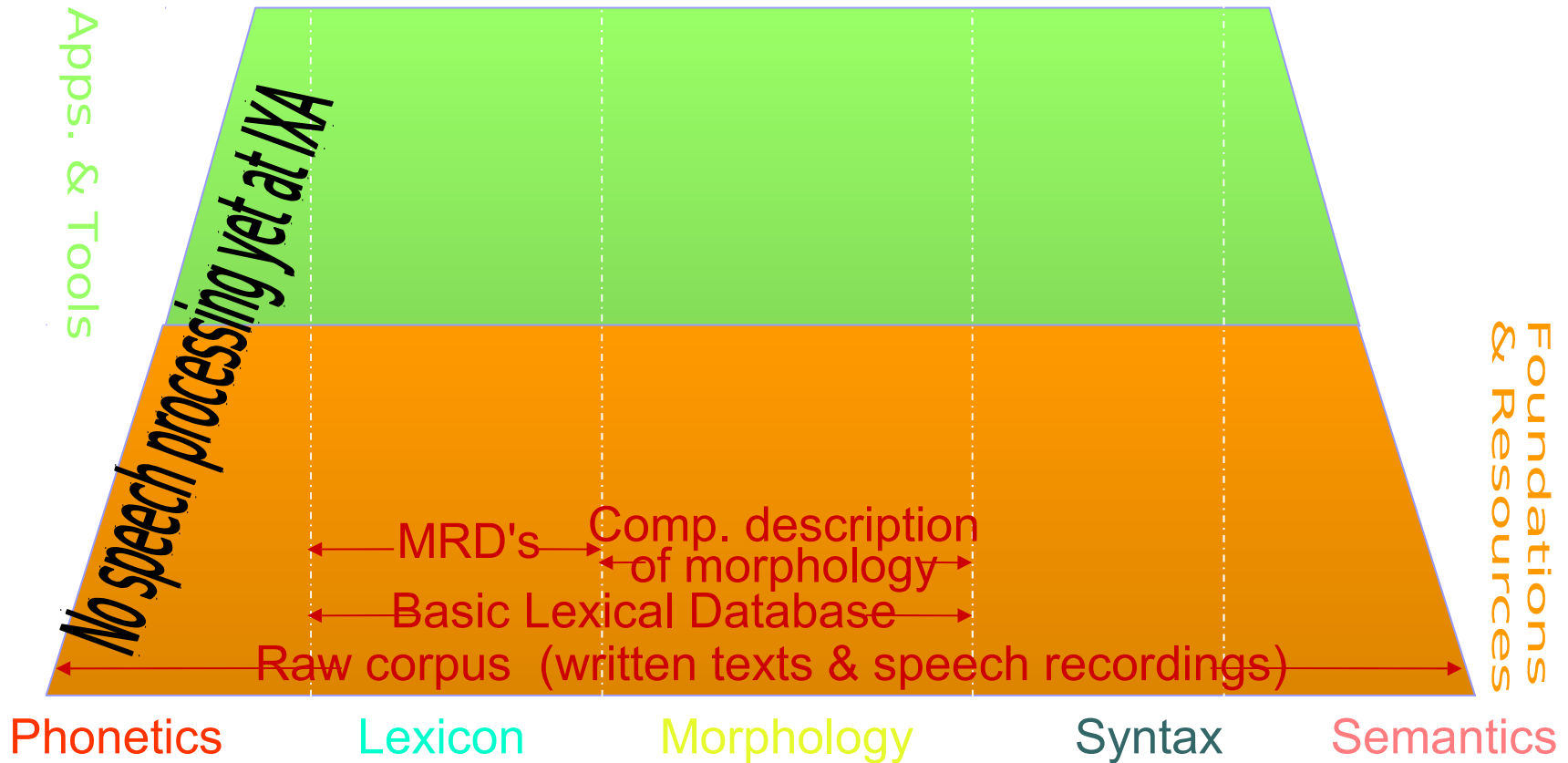




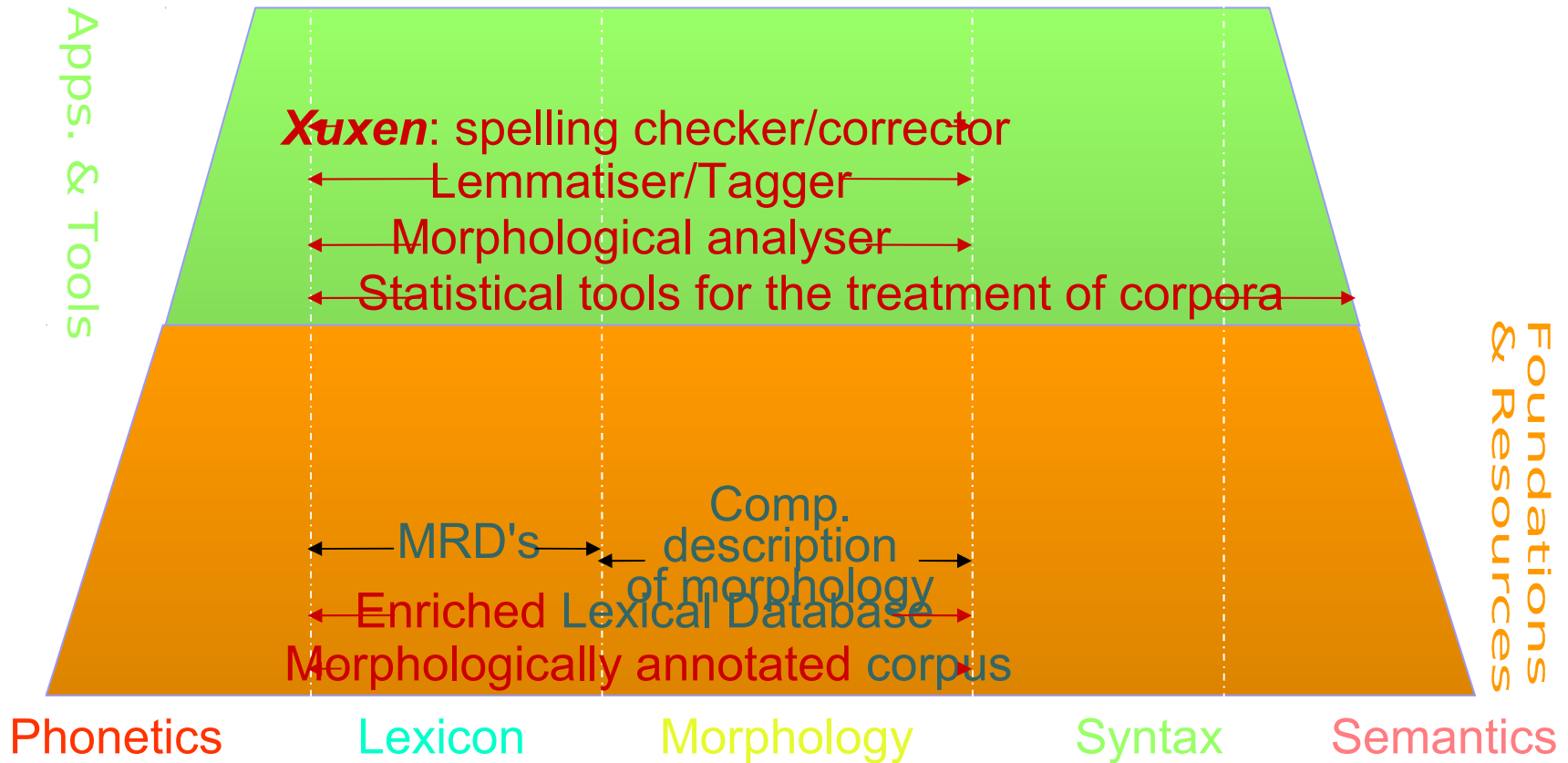
# Linguistic foundations & resources, tools and applications

- **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.

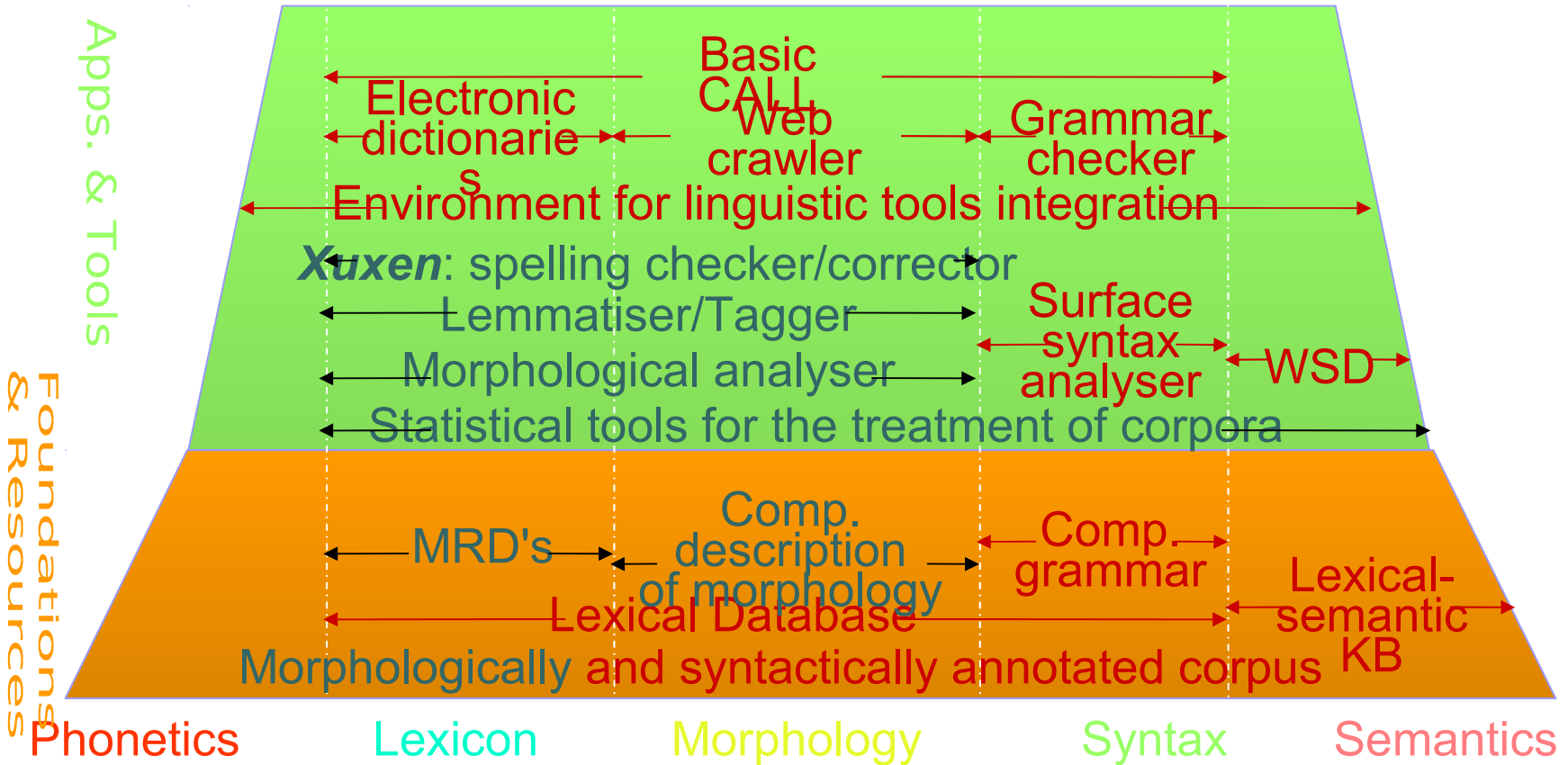
# Phase I: laying foundations



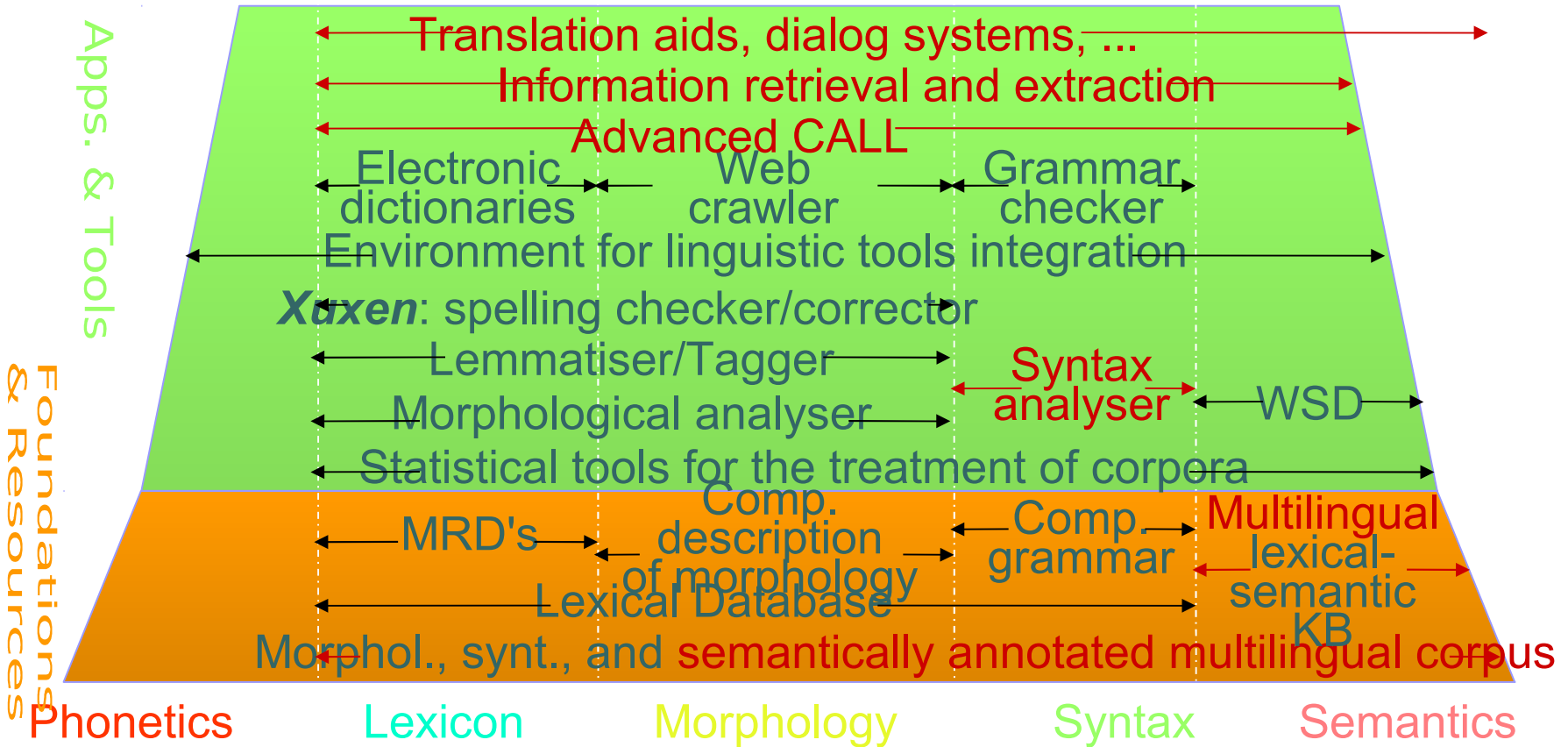
# Phase II: first basic tools and applications



# Phase III: more advanced tools and applications



# Phase IV: multilinguality and general applications





## Strategy to develop HLT in Basque

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

- BLARK, Basic Language Resources Kit (Krauwert, 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





# Strategy to develop HLT in Basque

## Open source

Success of the open source initiatives and the 2.0 communities

- Using open-source programs is a key factor of success, because **efforts are not repeated** and because 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** (in English).
  - But simultaneously this is a **key factor of quality** and so, **sustainability**.
- **Tool version control systems** as SVN, and **public repositories** 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**.



# 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

- *Corpus linguistics around the world* (Wilson et al., 2006) describes corpus resources on several languages.
- Roadmap of tools:
  - "Basic toolkit for HLT"(Agirre et al. 2002) (IXA group)
  - "Basic Language Resource Kit (BLARK)" (Krauwert, 2003)
    - Joint initiative between ELSNET and ELRA in 1998.
    - Maegaard et al. (2004) describe a BLARK for Arabic
    - Simov et al. (2004) for Bulgarian.
    - The term BLARK has been very successful and it is used in a large number of papers in the area.



# Related work

- Streiter et al. (2006) report on HLT projects for noncentral languages and proposes instructions for funding bodies and strategies for developers.
  - They use the *non-central* term and
  - Benefits and unsolved problems when using open source software for non-central languages is very interesting.
- Forcada (2006) remarks the opportunity of using open source machine translation for minor languages.



# Related work

- The ELSNET network of excellence prepared definitions for a language resources and evaluation roadmap, using for that the HLT Roadmap System, a framework for implementing technology roadmaps (Busemann & Uszkoreit, 2004).
  - Several different roadmaps have been published.
  - As in our first proposal in 2002 the elements in the diagram (HLT products) are classified into three equivalent subsets: (Language Resources / Language Processing / Language Usage) in their roadmap, and Language resources/ Language Tools / Language Applications) in our strategy.
  - 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



# Related work

- Borin (2006 and 2009)
  - points to the promise of the HLT for lesser-known languages and describes the linguistic diversity in the information society.
  - He cites the paper from Ostler "*a language will not get by in the world of today unless it is equipped with a parser and a multi-million-word corpus of text*".
  - He analyzes the relation among the sociology of language and HLT, and gives us some strategic considerations, i.e. "*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
- SALT MIL ("Speech And Language Technology for Minority Languages") has been organizing seven conferences related to HLT and less-resourced languages.



# Conclusions

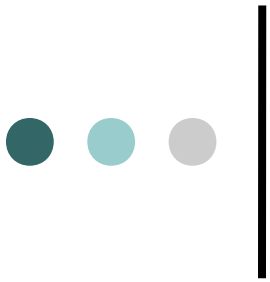
- From our experience we defend that research and development for less resourced languages should to be faced to build a BLARK following this points:
  - 1) high standardization
  - 2) open-source
  - 3) reusing language foundations, tools, and applications
  - 4) incremental design and development of 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.**





# Conclusions

- We know that any HLT project related with a less privileged language should follow those guidelines, but from our experience we know that in most cases they do not.
- We think that if Basque is now in an good position in HLT is because during the last twenty years 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 always reusable in future developments.
- Similar experiences with other languages:  
Czech is another exception to the correlation between language size and LR scarcity; the excessive rich body of LRs for Czech is due to the coordinated efforts of a few ambitious and productive researchers.



Thanks

Eskerrik asko

[Kepa.sarasola@ehu.es](mailto:Kepa.sarasola@ehu.es)

[ixa.si.ehu.es](http://ixa.si.ehu.es)