Promoting the Use of Basque via Language Technology

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

Ixa Taldea.
University of the Basque Country

http://ixa.si.ehu.es
What can we do?

What are we creating?

(Motivation)
Automatic dubbing of documentaries into Basque using subtitles in Spanish.
Aplications (2012)
Personal tutor in language learning
http://www.ehu.eus/ehusfera/ixa/2012/02/10/berbatek-projects-results-and-demos/

- Through a speech-driven avatar
- Automatically created grammar and comprehension exercises
- Writing aids (dictionaries, writing numbers, spelling...)
- Automatic evaluation of pronunciation
Ixa Group: Some active projects

European Commission: Quality Translation by Deep Language Engineering Approaches. (First European project developing LT for Basque)

Spanish Government: TACARDI: Traducción automática en contexto y aumentada con recursos dinámicos de internet.

Basque Government: LT Strategic Research in the Basque Country. (Consortium)
Ixa Group. Some active projects

QTLeap: Quality Translation...

Your question: ¿Cómo se qué versión de Photoshop tengo?
The proposed answer, automatically translated: Vaya al menú ayuda > acerca de Photoshop...
The automatically translated question: How is what version of Photoshop I have?
The most similar question found: How do I know which version of Photoshop I have?
The corresponding answer: Go to the menu Help> About Photoshop...

MT Pilot 0 (Baseline)

qtLeap
quality translation by deep language engineering approaches
Ixa Group. Some active projects

Mitzuli  http://mitzuli.com/?lang=en

The open easy-to-use and powerful translator app for Android
Created by one of our master students!
Ixa Group: Some active projects

NewsReader: Building structured event indexes of large volumes of financial and economic data for decision making.
Ixa Group: Some active projects
Open Polarity Enhanced Named Entity Recognition

OpeNER:
Ixa Group: Some active projects
Open Polarity Enhanced Named Entity Recognition

- Demo: Polarity analysis of tweets about the presentations in a conference:

![LREC 2014 Map](image)

- Creation of 100 new wikipedia entries
- 10% improvement in the MT output
- But ... huge work to engage volunteers.
Education

- **Graduate in Informatics**: Natural Language Processing (optional subject, since 1994)

- **Masters** (http://ixa.si.ehu.es/master)
  - Hiztek (in Basque, since 2001 to 2005)
  - HAP (in Basque, since 2005 ...)
  - Erasmus Mundus master on Language and Communication Technologies (in English, since 2014)
  - Language Analysis and Processing (in English, 2014...)

- **PhD Programme**: 11 PhD theses since 2010
Boosting cooperation among the agents related to Language Industries

- **Langune** association created in 2010: The Association of Language Industries of the Basque language
  1. What does Langune work for?
  2. Current reality of the LI in the Basque Country

See wider presentation: [www.langune.com](http://www.langune.com)
The Association of Language Industries of the Basque language – Langune, was officially set up in 2010, in order to promote the development and competitiveness of these industries, creating opportunities for collaboration and innovation in products / services, technologies and markets increasing the visibility and value added of this sector.

In 2012, the Department of Industry, Innovation, Trade and Tourism of the Basque Government conceded Langune the title of CLUSTER of Language Industry.

The comprehensive nature of the industry comes from having the entire value chain in a very reduced environment; from entities specialising in Translation to Language Training, Multilingualism management and Language Technology.
2. Current reality of the LI in the Basque Country

- The Basque language industry comprises **585 companies** with:
  - Turnover of around **276M€**.
  - Employment related to this sector **over 5,000 people**.
  - These figures represent **0.42% of Gross Domestic Product (GDP)**.
  - Tendency in 2013 around a **1% growth**.

**Growing tendency**

[Graph showing employment and turnover growth from 2011 to 2013]
Can help NLP less resourced languages to promote their use?

- Today **language technology** (LT) provides many powerful resources to make easier the use of human languages.
- But **all the languages are not able** to use this technology.
- Taking into account the **different levels in using LT**, we propose a classification for the 7000 languages in our world.
- **What language resources could be useful** to promote the use of less resourced languages?
- **Results achieved by IXA Group** in using LT to normalize and to promote the use of Basque.
Outline

How are languages facing the ICT and HLT challenges?

Which languages are "less resourced"? Six different levels

Can help NLP less resourced languages to promote their use?

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.
META-NET

META-NET, l’Alliance Technologique pour une Europe multilingue : réseau d’excellence soutenu par la Commission Européenne.

50+ laboratoires de recherche du domaine des sciences et technologies de la langue, dans une
META-NET: results for Basque

<table>
<thead>
<tr>
<th>Language Technology: Tools, Technologies and Applications</th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Maturity</th>
<th>Sustainability</th>
<th>Adaptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognition</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Speech Synthesis</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Grammatical analysis</td>
<td>4</td>
<td>2.5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Semantic analysis</td>
<td>1</td>
<td>1.5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Text generation</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Machine translation</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Language Resources (Resources, Data and Knowledge Bases)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text corpora</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2.5</td>
</tr>
<tr>
<td>Speech corpora</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Parallel corpora</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Lexical resources</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Grammars</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

7: State of language technology support for Basque
META-NET: results for English

<table>
<thead>
<tr>
<th>Language Technology: Tools, Technologies and Applications</th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Manufry</th>
<th>Sustainability</th>
<th>Adapability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognition</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Speech Synthesis</td>
<td>5</td>
<td>3</td>
<td>4.5</td>
<td>5.5</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Grammatical analysis</td>
<td>5</td>
<td>5</td>
<td>5.5</td>
<td>4.5</td>
<td>4.5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Semantic analysis</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Text generation</td>
<td>3</td>
<td>3</td>
<td>3.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Machine translation</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Resources: Resources, Data and Knowledge Bases</th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Manufry</th>
<th>Sustainability</th>
<th>Adapability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text corpora</td>
<td>5</td>
<td>4</td>
<td>5.5</td>
<td>4</td>
<td>5</td>
<td>2.5</td>
<td>4</td>
</tr>
<tr>
<td>Speech corpora</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>5.5</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Parallel corpora</td>
<td>4.5</td>
<td>4.5</td>
<td>5</td>
<td>5</td>
<td>3.5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Lexical resources</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Grammars</td>
<td>3.5</td>
<td>2.5</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>4</td>
<td>1.5</td>
</tr>
</tbody>
</table>

8: State of language technology support for English
META-NET: results for Welsh

http://www.meta-net.eu/whitepapers/volumes/welsh

<table>
<thead>
<tr>
<th>Language Technology: Tools, Technologies and Applications</th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Maturity</th>
<th>Sustainability</th>
<th>Adaptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognition</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Speech Synthesis</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Grammatical analysis</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Semantic analysis</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Text generation</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Machine translation</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Resources: Resources, Data and Knowledge Bases</th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Maturity</th>
<th>Sustainability</th>
<th>Adaptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text corpora</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Speech corpora</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Parallel corpora</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Lexical resources</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Grammars</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
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

Estimated internet users are 1,966,514,816 on June 30, 2010
Copyright © 2000 - 2010, Miriwalts 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)
  
  [Link](http://dtiil.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

- Articles in 282 languages (Mars 2015).
- Top 10 languages:
  - English (4.7 million articles),
  - German (1.8 M), French (1.6 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:
  - 17th: Catalan (454 K)
  - 34th: Basque (206 K)
  - 65th: Cymraeg Welsh (63K)
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. 1 for Welsh
- *The Universal Catalogue (5 products for Welsh)*
  - 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, neither for Welsh
How are languages facing HLT?

ACLwiki. Association for Computational Linguistics

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

**Free**
- Gwirdd gramadeg rhydd i'r Gymraeg / Grammar checker (based on An Gramadóir)
- Geiriadur rhydd i'r Gymraeg / Welsh dictionary
- Rhedeg berfau Cymraeg / Verb conjugator
- Apertium — has a GPL morphological analyser for Welsh as part of the Welsh—English language pair data (which also includes a Constraint Grammar disambiguator and Welsh—English translational dictionary)

**Proprietary**

**Machine translation**

**Free**
- apertium-cy-en — Online for testing at www.cymraeg.org.uk

**Corpora**

**Free**
- OPUS Welsh — many languages.
- BangorTalk Welsh—Spanish, Welsh—English conversational corpora, GPL, speakers tagged with "social variables"

**Partially-free**
- UAGT-PNAW Welsh—English. 510,813 bilingual aligned sentence pairs.

---

How are languages facing HLT?

On-line lexical resources for 300 languages
Search by language is allowed.
5 links to Basque resources (although they are >40)
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:**
    - Identifies **45 languages**
      - Basque? No; Welsh? No

- **MT systems**
  - **Google-Translate:** **100 languages**
    - Basque? Yes; Welsh? Yes
  - **Babelfish:** **13 languages**
How are languages facing the ICT and HLT challenges?

Which languages are "less resourced"?
Six different levels

Can help NLP less resouced languages to promote their use?

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
   - **Good** level (J. Mariani, regarding to LRs in LRE Map)
     - 37.9% of the users of Internet.
     - 45.00% of the web pages.
     - 62% of the HLT resources in LDC
     - 51% in ELRA.
     - With applications in almost all the types of HLT . .
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
  - Central languages (Streiter et al., 2006)
  - Relatively good level of support (J. Mariani)
Which languages are "less resourced"? Six different levels

- Third level: around 70 languages. Moderate and fragmentary support (J. Mariani)

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
  Weak support (J. Mariani)

Languages with any registered on-line lexical resource

- 307 languages in yourdictionary.com

- It is almost the same set of languages that is present in Wikipedia (286 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:**
  Only-spoken languages in the world
  - Here are included at least other 4,500 languages.
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.
- 3rd or the 4th are the levels of languages usually called as less-resourced in the HLT domain.
- We may consider that languages in the 5th and the 6th levels are really endangered,
Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- Can help NLP less resourced languages to promote their use?
- 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)
  - 2013: 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
We presented an open proposal for making progress in HLT (Aduriz et al., 1998)
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
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. *(ixa-pipes for English, Spanish and Basque)*

- 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

*Research & development*

End-user applications
Language tools

*Basic & applied research*

Linguistic foundations
Linguistic resources
Phase I: laying foundations

- Phonetics
- Lexicon
- Morphology
- Syntax
- Semantics

- MRD's
- Comp. description of morphology
- Basic Lexical Database
- Raw corpus (written texts & speech recordings)
Phase II: first basic tools and applications

- **Phonetics**
- **Lexicon**
- **Morphology**
- **Syntax**
- **Semantics**

- Lemmatiser/Tagger
- Morphological analyser
- Statistical tools for the treatment of corpora

- **Xuxen**: spelling checker/corrector

- **MRD's**
- Described morphology
- Enriched Lexical Database
- Morphologically annotated corpus

- Comp. description of morphology

- Foundations & Resources

Apps. & Tools
Phase III: more advanced tools and applications

Phonetics
Lexicon
Morphology
Syntax
Semantics

Apps. & Tools

Foundations & Resources

MRD's
Lexical Database
Morphologically and syntactically annotated corpus
Comp. description of morphology
Lexical-semtantic KB
Comp. grammar
WSD
Surface syntax analyser
Lemmatiser/Tagger
Morphological analyser
Xuxen: spelling checker/corrector
Statistical tools for the treatment of corpora
Grammar checker
Web crawler
Basic CALL
Electronic dictionarie
Environment for linguistic tools integration
Created LRs and tools (1988-2010)
http://ixa.si.ehu.es/Ixa/Produktuak

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications</td>
<td></td>
<td><strong>Multimeteo MT</strong></td>
<td><strong>application</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantics</td>
<td></td>
<td></td>
<td></td>
<td><strong>BasqueWordnet</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntax</td>
<td></td>
<td></td>
<td><strong>Zatiak-Ixati Chunker</strong></td>
<td><strong>Erreus corpus of errors</strong></td>
<td><strong>Ancora, EPEC corpus</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexic</td>
<td><strong>EDBL</strong> Lexical data base</td>
<td><strong>EDBL 2.0</strong></td>
<td><strong>Elhuyar-Word</strong></td>
<td><strong>UZEI MSWord Synonym. Dict.</strong></td>
<td><strong>EDBL 3.0</strong></td>
<td><strong>Lexkit</strong> Dicc. Escolar Cubano</td>
<td></td>
</tr>
<tr>
<td>Morphology</td>
<td><strong>Xuxen</strong> Spell. Checker</td>
<td><strong>Xuxen1.0 Morph. analyzer</strong></td>
<td><strong>Xuxen 2.0 Eustagger</strong></td>
<td><strong>Xuxen3.0 Elhuyar-Word</strong></td>
<td><strong>Xuxen 3.0 Eihera NER</strong></td>
<td><strong>ZT corpus Eulia tagging tool</strong></td>
<td><strong>BertsolariX a LibiXaml</strong></td>
</tr>
</tbody>
</table>
Phase IV: multilinguality and general applications

- Translation aids, dialog systems, ...
- Information retrieval and extraction
- Advanced CALL
- Electronic dictionaries
- Web crawler
- Grammar checker
- Environment for linguistic tools integration
- Lemmatiser/Tagger
- Xaxen: spelling checker/corrector
- Morphological analyser
- Syntax analyser
- Statistical tools for the treatment of corpora
- WSD
- MRD's
- Comp. description of morphology
- Comp. grammar
- Multilingual lexical-semantic KB
- Lexical Database
- Morphol., synt., and semantically annotated multilingual corpus
- Morphology
- Phonetics
- Lexicon
- Apps. & Tools
- Foundations & Resources
- Semantics
- Syntax
- 56
Conclusions

- From our experience we defend that research and development for less resourced languages should be faced to build a BLARK following these 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.
Conclusions

- We promoted the creation of Langune (The Association of Language Industries of the Basque language)
  - 578 companies,
  - 276M€,
  - 5,000 people,
  - 0.42% GDP
Diolch
Eskerrik asko
Thanks

kepa.sarasola@ehu.es
ixa.si.ehu.es