Application of Language Technologies for Less-Resourced Languages, the case of Basque

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
Outline

- Which languages are "less resourced"?
  Six different levels
- Strategy to develop Language Technologies for less-resourced languages

- Conclusions
- ELRA18 topics
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 of support in yesterday Joseph Mariani's presentation regarding to the number of 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.
   - 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.
- Relatively good level of support in Mariani's presentation
Which languages are "less resourced"?

Six different levels

- **Third level**: around 70 languages. Moderate and fragmentary support in Joseph's pres.

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 in Mariani's presentation

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.

Both 5th and 6th correspond to **NA (Not Appl.)** support in Mariani's presentation
How are languages facing HLT?

Which languages are "less resourced"?

- The answer is relative
- Six different levels

Diagram:
- 1 language
- 10 languages
- 60 languages
- 250 languages
- 7,000 languages

Legend:
- English
- Best position in all HLT applications and resources
- Central languages (top 10 languages) Relevant position in all HLT applications
- Languages with any HLT application
- Languages with any lexical resource in Internet
- All the world languages
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.
- $3^{rd}$ or the $4^{th}$ are the levels of languages usually called as less-resourced in the HLT domain.
- We may consider that languages in the $5^{th}$ and the $6^{th}$ levels are really endangered,
Outline

- Which languages are "less resourced"? Six different levels
- **Strategy to develop Language Technologies for less-resourced languages**
- Conclusions
strategy to develop hlt in basque
ixa research group

- ix a 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).

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.
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

**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)

Apps. & Tools

No speech processing yet at IXA

Foundations & Resources
Phase II: first basic tools and applications

- Lemmatiser/Tagger
- Morphological analyser
- Statistical tools for the treatment of corpora
- Xuxen: spelling checker/corrector
- MRD's
- Enriched Lexical Database
- Morphologically annotated corpus

Phonetics | Lexicon | Morphology | Syntax | Semantics
Phase III: more advanced tools and applications

- Phonetics
- Lexicon
- Morphology
- Syntax
- Semantics

**Applications & Tools**

- Electronic dictionaries
- Environment for linguistic tools integration
- Xuxen: spelling checker/corrector
- Lemmatiser/Tagger
- Morphological analyser
- Statistical tools for the treatment of corpora

**Foundations & Resources**

- MRD's
- Lexical Database
- Morphologically and syntactically annotated corpus

**Basic CALL**

- Web crawler
- Grammar checker
- Surface syntax analyser
- WSD

**Compilation description of morphology**

- Comp. grammar
- Lexical-semantic KB
Created LRs and tools (1988-2010)

<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>Multimeteo MT application</td>
<td></td>
<td></td>
<td>Anhitz (QA, MT, IE-IR, Avatar)</td>
<td></td>
<td>Ihardetsi (QA) BASYQUE (Lexic application)</td>
</tr>
<tr>
<td>Semantics</td>
<td></td>
<td></td>
<td></td>
<td>BasqueWordnet</td>
<td>MCR Wordnet WSD-Ixas</td>
<td></td>
<td>(Eu)SemCor UKB, WSD algorithm</td>
</tr>
<tr>
<td>Syntax</td>
<td></td>
<td>Zatiak-IxatI Chunker</td>
<td>Erreus corpus of errors</td>
<td>Ancora, EPEC corpus</td>
<td></td>
<td>Maltixa (MALT parser) EDGK dependency parser</td>
<td></td>
</tr>
<tr>
<td>Lexic</td>
<td>EDBL Lexical data base</td>
<td>EDBL 2.0</td>
<td>Elhuyar-Word</td>
<td>UZEI MSWord Synonym. Dict.</td>
<td>EDBL 3.0</td>
<td></td>
<td>Lexkit Dic. Escolar Cubano</td>
</tr>
<tr>
<td>Morphology</td>
<td>Xuxen Spelling Checker</td>
<td>Xuxen1.0 Morph. analyzer</td>
<td>Xuxen 2.0 Eustagger</td>
<td>Xuxen3.0 Elhuyar-Word</td>
<td>Xuxen 3.0 Eihera NER</td>
<td>ZT corpus Eulia tagging tool</td>
<td>BertsonariX a LibiXam!</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
  - Morphological analyser
  - Statistical tools for the treatment of corpora
  - Syntax analyser
  - WSD
  - Xuxen: spelling checker/corrector

- MRD's
- Comp. description of morphology
- Comp. grammar
- Multilingual lexical-semantic KB

- Morphol., synt., and semantically annotated multilingual corpus

Foundations & Resources

Apps. & Tools

Lexicon

Morphology

Syntax

Semantics

Phonetics
Aplications (2012)
Translation, content management and learning
(Leturia et al., 2013) TC3 Journal

- Automatic dubbing of documentaries into Basque using subtitles in Spanish.

- Personal tutor in language learning
  - through a speech-driven avatar
  - automatically created grammar and comprehension exercises
  - writing aids (dictionaries, writing numbers, spelling...)
  - automatic evaluation of pronunciation
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
1. What does Langune work for?

- 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**
Conclusions

- From our experience we defend that research and development for less resourced languages should 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.
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
ELRA18 topics

- ISLRN?

- Impact Factor?
  Fine. It will help us to present the real value of our products.
  - In general / per language
  - Number of downloads
  - Number of citations
ELRA18 topics

- Crowd-sourcing evaluation?
  - 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.
Crowd-sourcing LR creation
(Alegria et al, 2013).
'The People’s Web Meets NLP: Collaboratively Constructed LRs', Springer

Reciprocal Enrichment Between Basque Wikipedia and Machine Translation

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