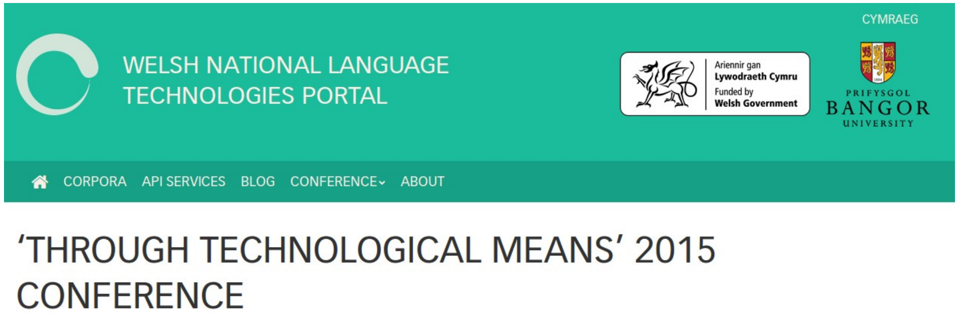
Promoting the Use of Basque via Language Technology

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


CYMRAEG
Ariennir gan Lywodraeth Cymru
Funded by Welsh Government
PRIFYSGOL BANGOR UNIVERSITY
CORPORA
API SERVICES
BLOG
CONFERENCE
ABOUT

WELSH NATIONAL LANGUAGE TECHNOLOGIES PORTAL

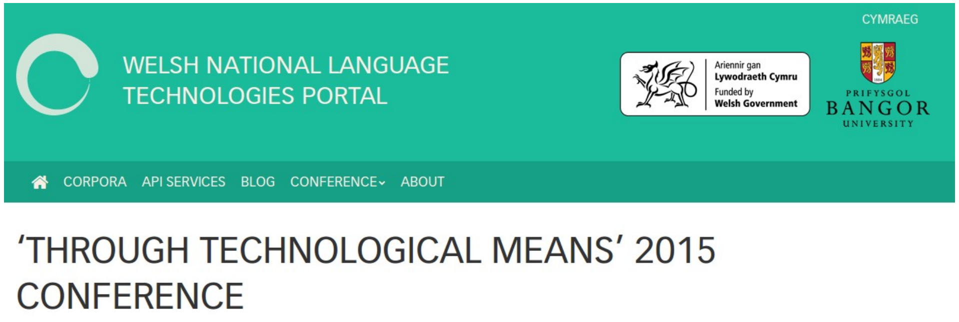
‘THROUGH TECHNOLOGICAL MEANS’ 2015 CONFERENCE

Trwy Ddulliau Technoleg /
Through Technological Means
The language technology conference
Bangor University
6th March 2015

What can we do?

What are we creating?

(Motivation)


CYMRAEG
Ariennir gan Lywodraeth Cymru
Funded by Welsh Government
PRIFYSGOL BANGOR UNIVERSITY
CORPORA
API SERVICES
BLOG
CONFERENCE
ABOUT

WELSH NATIONAL LANGUAGE TECHNOLOGIES PORTAL

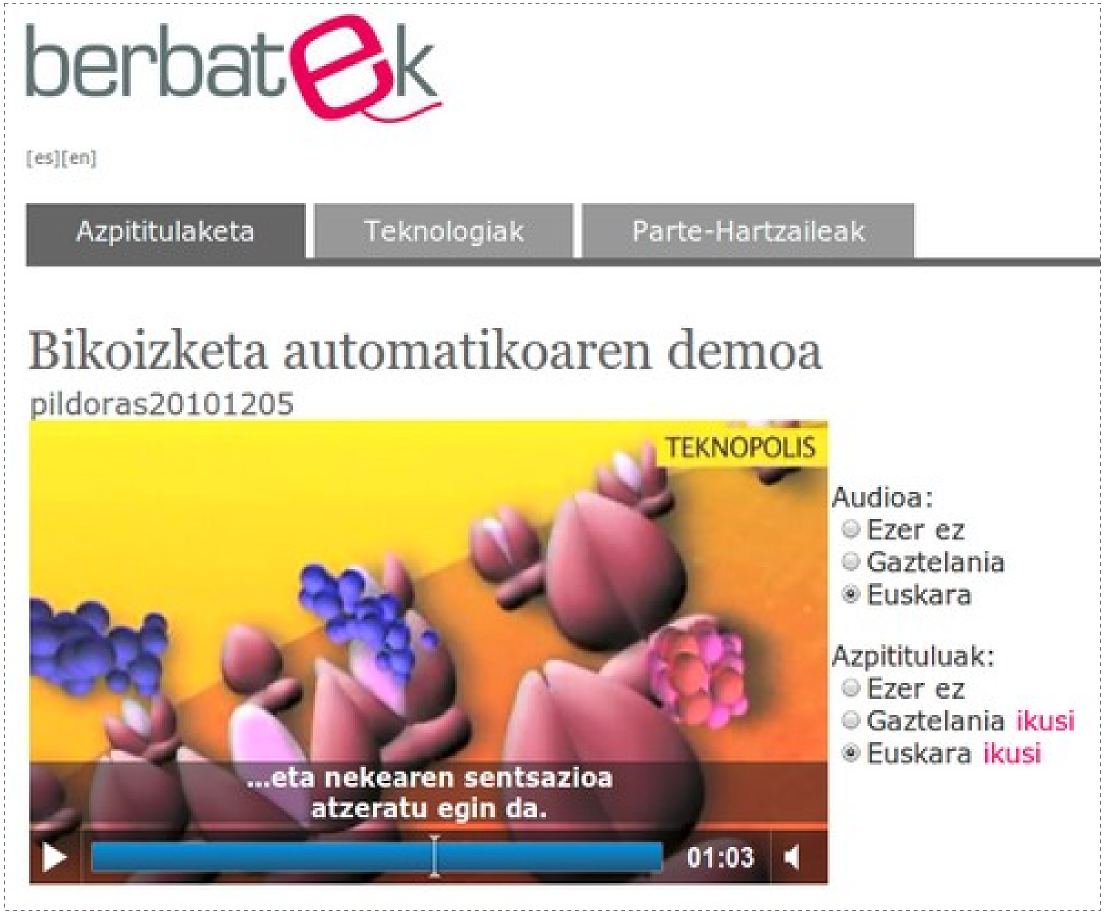
‘THROUGH TECHNOLOGICAL MEANS’ 2015 CONFERENCE

Trwy Ddulliau Technoleg /
Through Technological Means
The language technology
conference
Bangor University
6th March 2015

Translation, content management

(Leturia et al., 2013) TC3 Journal

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



Bikoizketa automatikoaren demoa

pildoras20101205

...eta nekearen sentsazioa
atzeratu egin da.

Audioa:
○ Ezer ez
○ Gaztelania
● Euskara

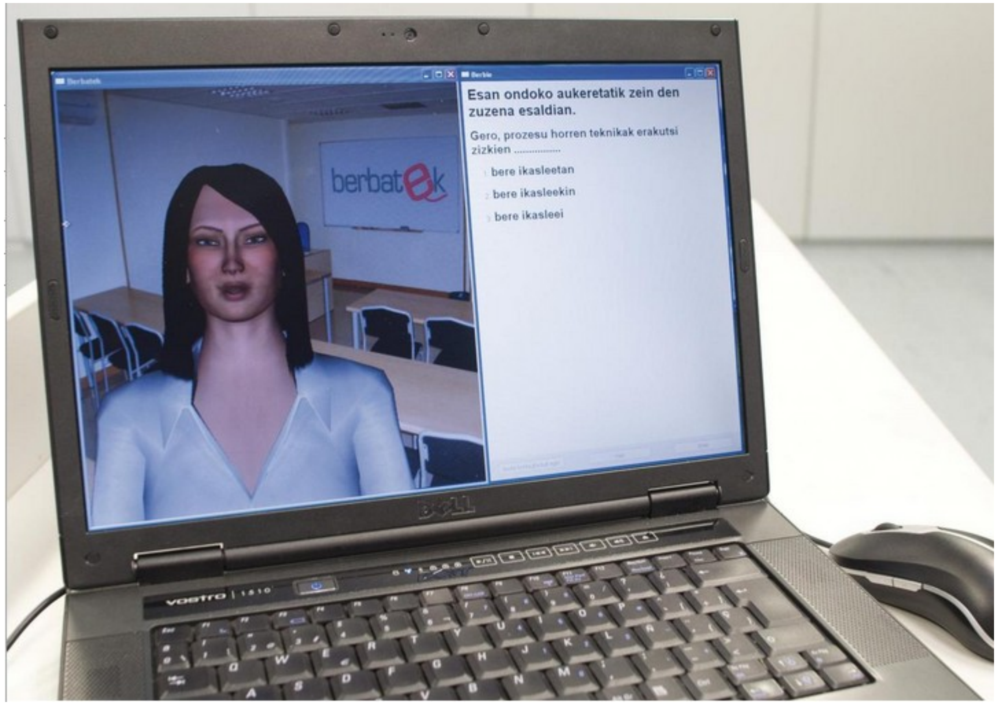
Azpitituluak:
○ Ezer ez
○ Gaztelania ikusi
● Euskara ikusi

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



Esan ondoko aukeretatik zein den zuzena esaldian.

Gero, prozesu horren teknikak erakutsi zizkien ....

1 bere ikasleetan
2 bere ikasleekin
3 bere ikasleei

Ixa Group: Some active projects

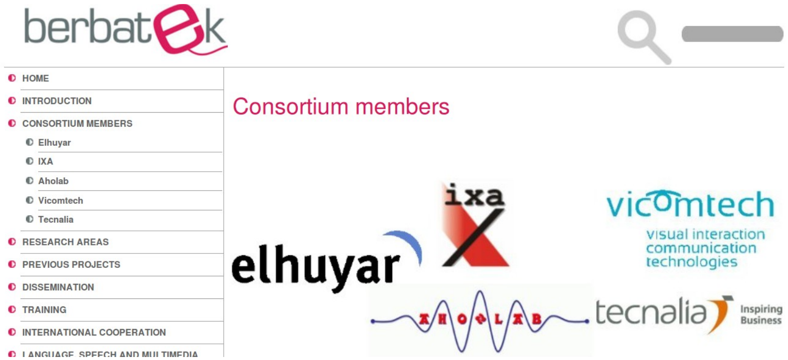
Euroean Commission:
Quality Translation by Deep Language Engineering Approaches.
(First european project developing LT for Basque)

qtleap
quality translation by deep language engineering approaches

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

Tacardi
Traduccion automática en contexto y aumentada con recursos dinámicos de internet

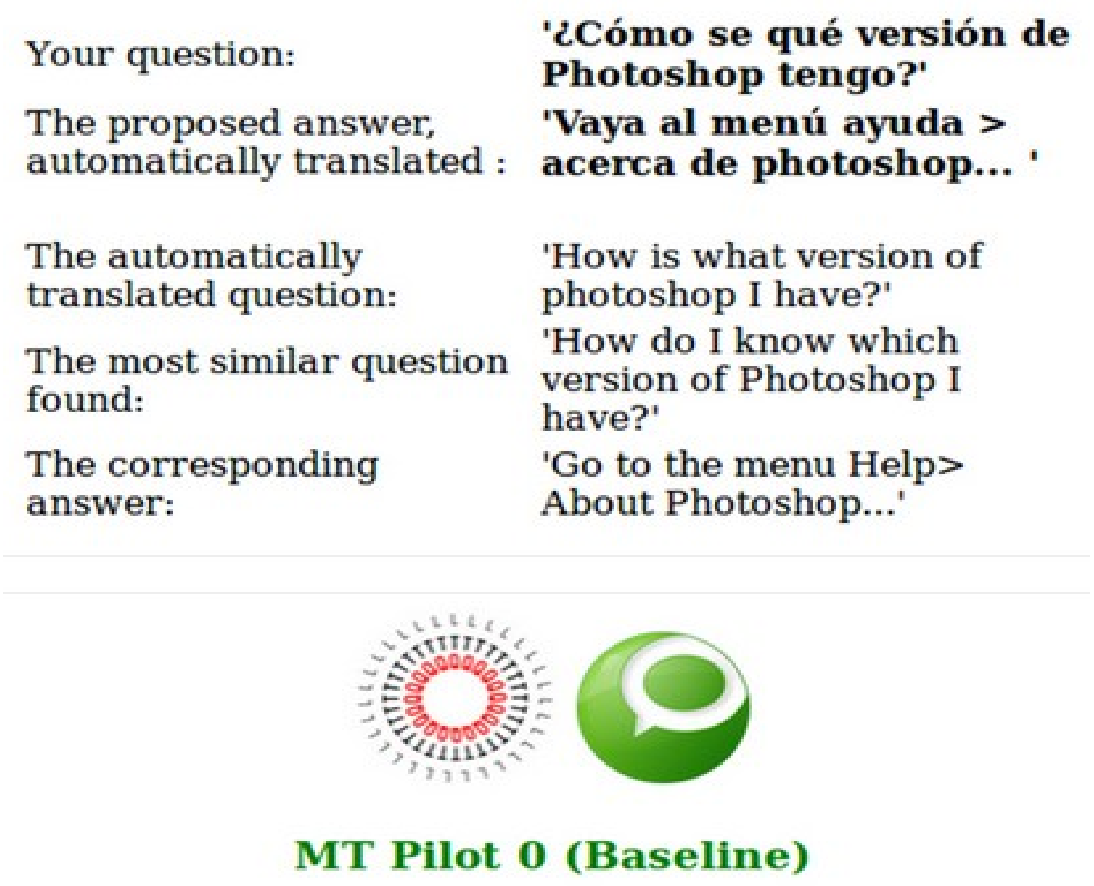
Basque Government:
LT Strategic Researche in the Basque Country. (Consortium)


berbatek

- HOME
- INTRODUCTION
- CONSORTIUM MEMBERS
  - Elhuyar
  - IXA
  - Aholab
  - Vicomtech
  - Tecnalia
- RESEARCH AREAS
- PREVIOUS PROJECTS
- DISSEMINATION
- TRAINING
- INTERNATIONAL COOPERATION
- LANGUAGE, SPEECH AND MULTIMEDIA

Consortium members

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

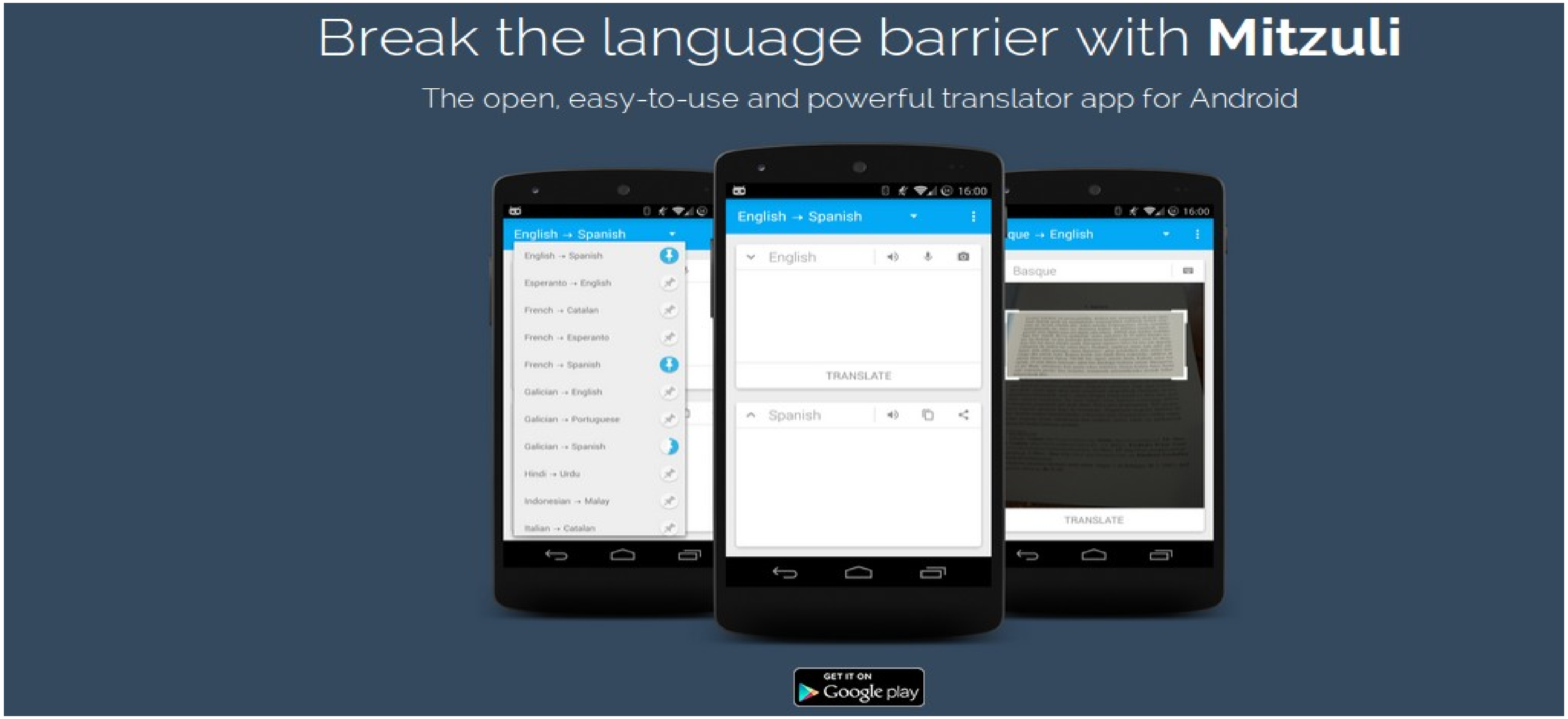
quality translation by deep language engineering approaches

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!

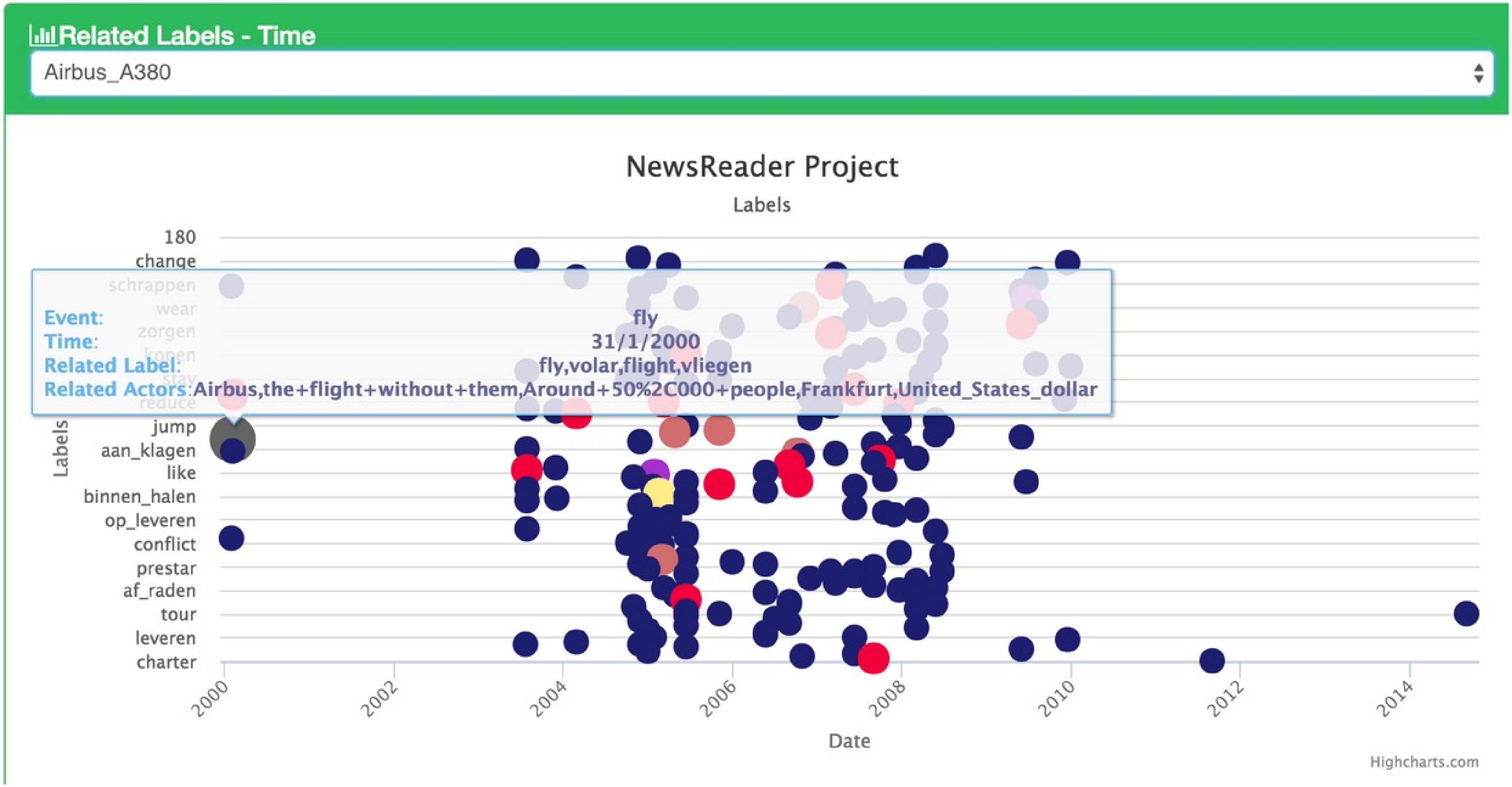


Break the language barrier with **Mitzuli**

The open, easy-to-use and powerful translator app for Android

Ixa Group: Some active projects

NewsReader: Building structured event indexes
of large volumes of financial and economic data for decision making



Related Labels - Time

Airbus_A380

NewsReader Project

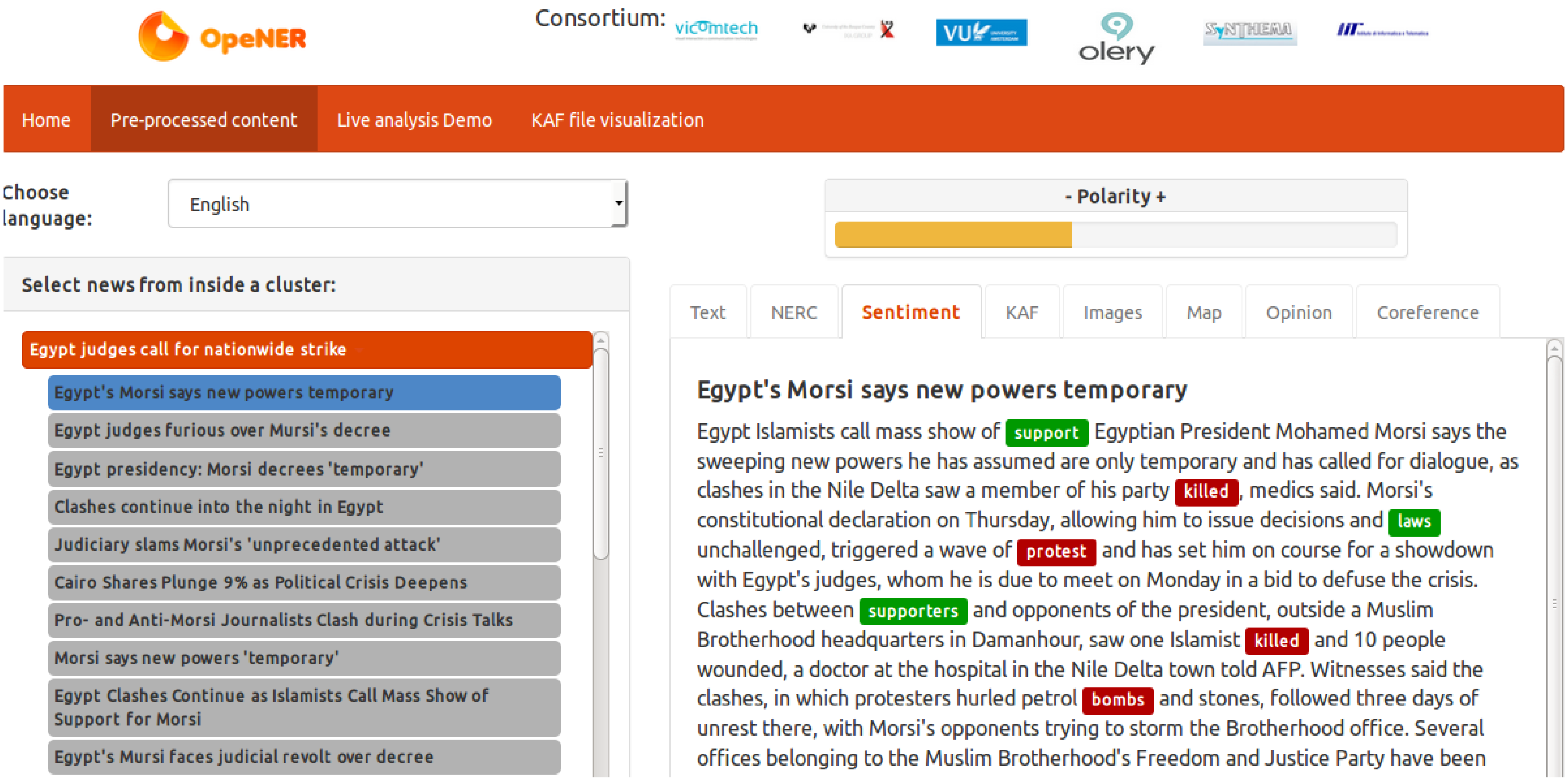
Labels

Event: fly
Time: 31/1/2000
Related Label: fly,volar,flight,vliegen
Related Actors: Airbus,the+flight+without+them,Around+50%2C000+people,Frankfurt,United_States_dollar

Ixa Group: Some active projects

Open Polarity Enhanced Named Entity Recognition

OpeNER:



Select news from inside a cluster:

- Egypt judges call for nationwide strike
  - Egypt's Morsi says new powers temporary
  - Egypt judges furious over Mursi's decree
  - Egypt presidency: Morsi decrees 'temporary'
  - Clashes continue into the night in Egypt
  - Judiciary slams Morsi's 'unprecedented attack'
  - Cairo Shares Plunge 9% as Political Crisis Deepens
  - Pro- and Anti-Morsi Journalists Clash during Crisis Talks
  - Morsi says new powers 'temporary'
  - Egypt Clashes Continue as Islamists Call Mass Show of Support for Morsi
  - Egypt's Mursi faces judicial revolt over decree

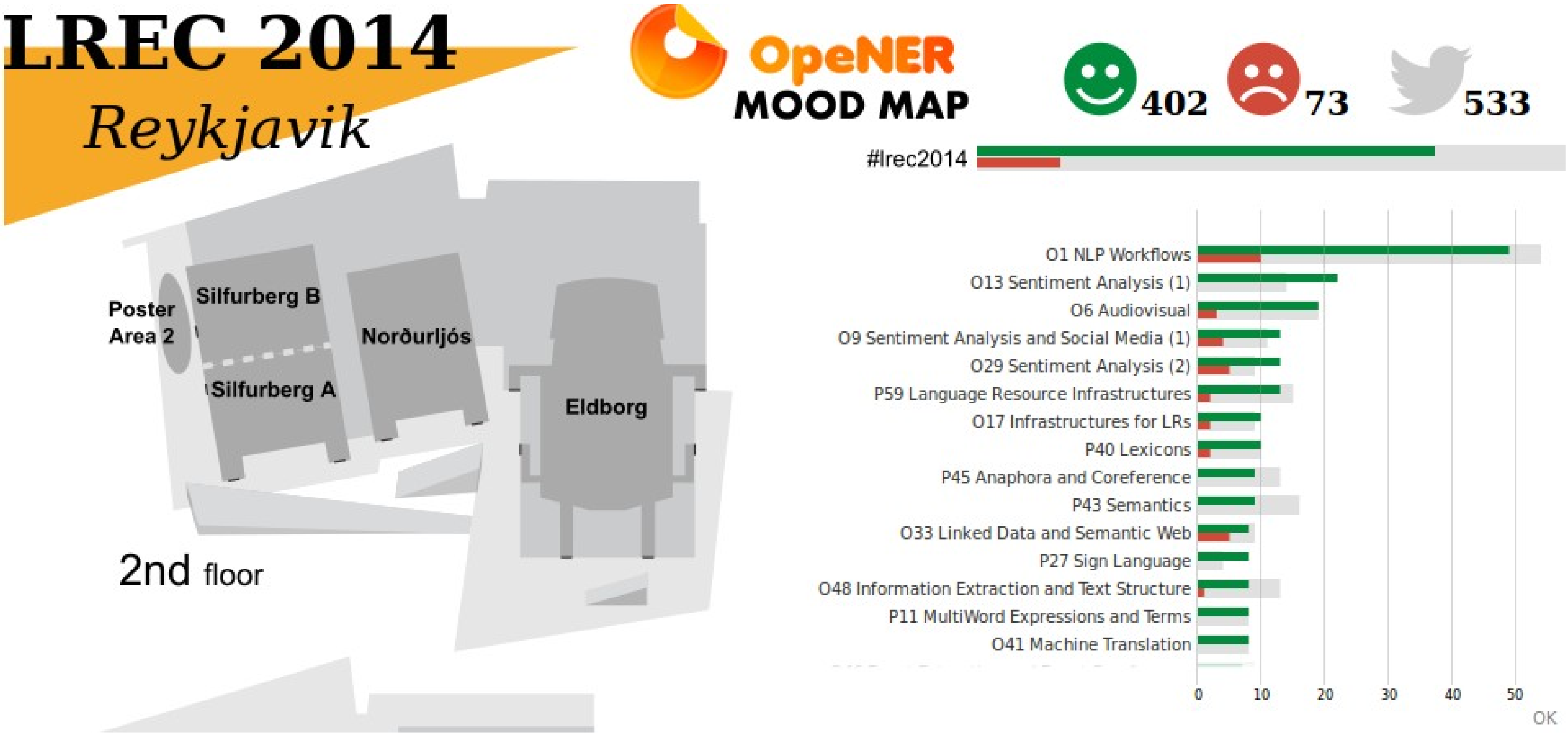
Egypt's Morsi says new powers temporary

Egypt Islamists call mass show of support Egyptian President Mohamed Morsi says the sweeping new powers he has assumed are only temporary and has called for dialogue, as clashes in the Nile Delta saw a member of his party killed, medics said. Morsi's constitutional declaration on Thursday, allowing him to issue decisions and laws unchallenged, triggered a wave of protest and has set him on course for a showdown with Egypt's judges, whom he is due to meet on Monday in a bid to defuse the crisis. Clashes between supporters and opponents of the president, outside a Muslim Brotherhood headquarters in Damanhour, saw one Islamist killed and 10 people wounded, a doctor at the hospital in the Nile Delta town told AFP. Witnesses said the clashes, in which protesters hurled petrol bombs and stones, followed three days of unrest there, with Morsi's opponents trying to storm the Brotherhood office. Several offices belonging to the Muslim Brotherhood's Freedom and Justice Party have been

Ixa Group: Some active projects

Open Polarity Enhanced Named Entity Recognition

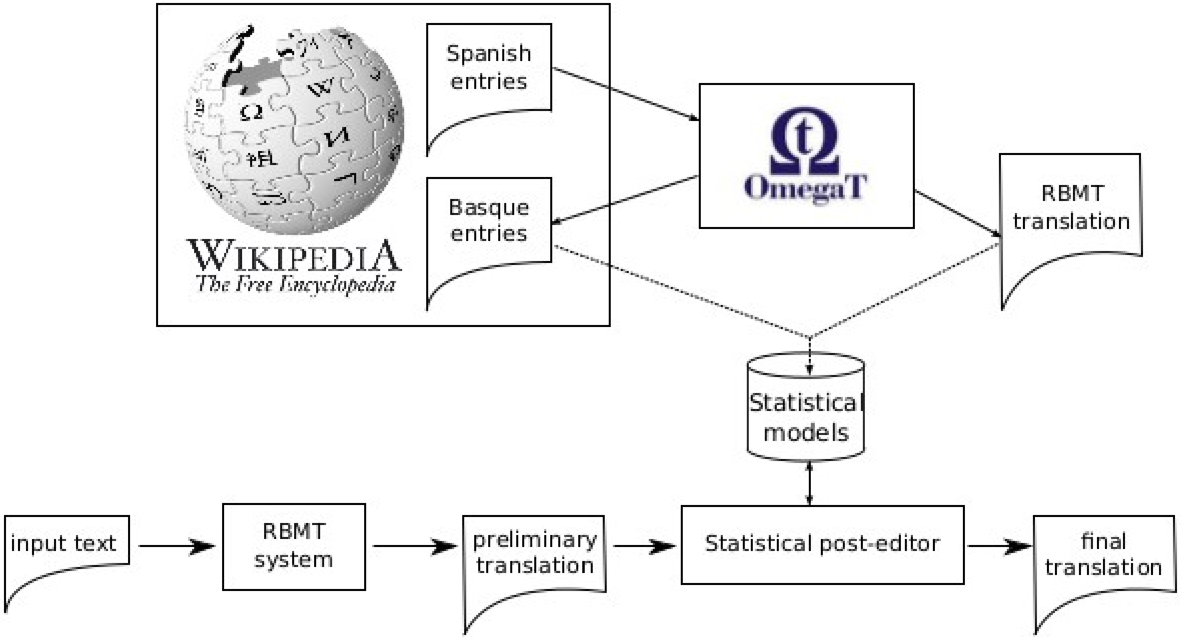
- Demo: Polarity analysis of tweets
  about the presentations in a conference:


LREC 2014
Reykjavik
Poster Area 2
Silfurberg B
Silfurberg A
Norðurljós
Eldborg
2nd floor
OpeNER
MOOD MAP
402
73
533
#lrec2014
O1 NLP Workflows
O13 Sentiment Analysis (1)
O6 Audiovisual
O9 Sentiment Analysis and Social Media (1)
O29 Sentiment Analysis (2)
P59 Language Resource Infrastructures
O17 Infrastructures for LRs
P40 Lexicons
P45 Anaphora and Coreference
P43 Semantics
O33 Linked Data and Semantic Web
P27 Sign Language
O48 Information Extraction and Text Structure
P11 MultiWord Expressions and Terms
O41 Machine Translation
0
10
20
30
40
50
OK

Crowd-sourcing LR creation

(Alegria eta 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.

Education

- **Grade on 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 thesis since 2010

Hizkuntzaren Azterketa eta Prozesamendua (HAP)
Language Analysis and Processing (LAP)

Lct

Erasmus Mundus

Boosting cooperation among the agents related to Language Industries

LANGUNE

THE BASQUE ASSOCIATION
OF 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

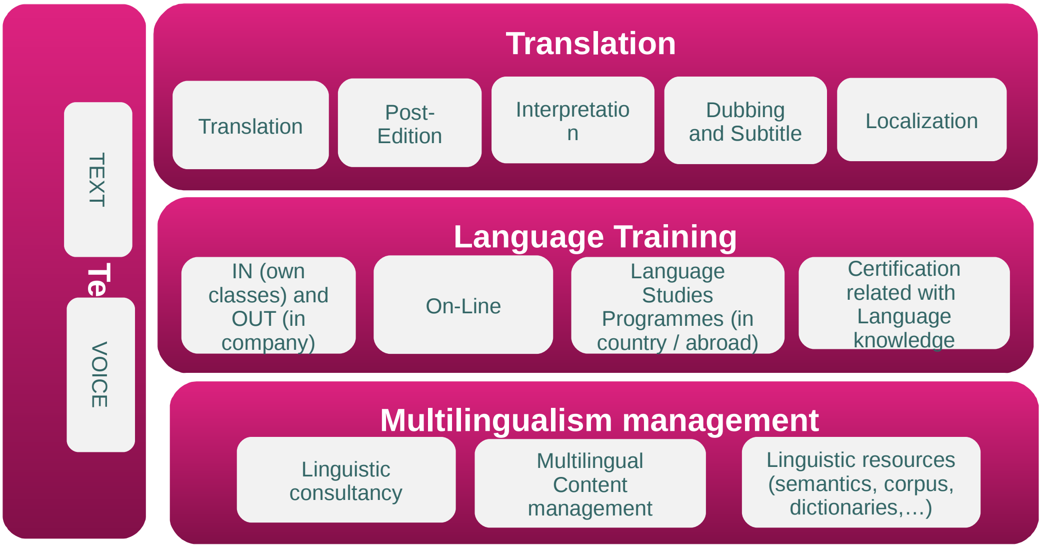
1. What does Langune work for?

- 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.

CORE BUSINESS OF LANGUNE



TEXT

Te

VOICE

Translation

Translation

Post-Edition

Interpretation

Dubbing and Subtitle

Localization

Language Training

IN (own classes) and OUT (in company)

On-Line

Language Studies Programmes (in country / abroad)

Certification related with Language knowledge

Multilingualism management

Linguistic consultancy

Multilingual Content management

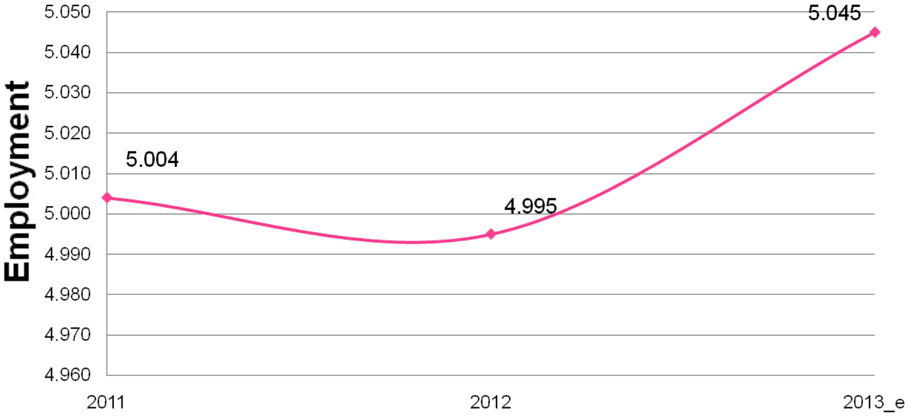
Linguistic resources (semantics, corpus, dictionaries,…)

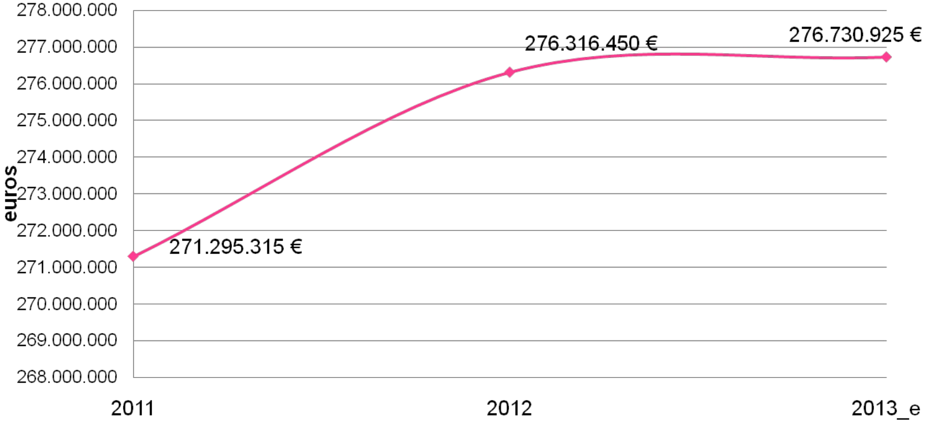
LANGUNE
THE BASQUE ASSOCIATION
OF LANGUAGE INDUSTRIES

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





Can help NLP less resouced 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 resouced languages to promote their use?
- Conclusions

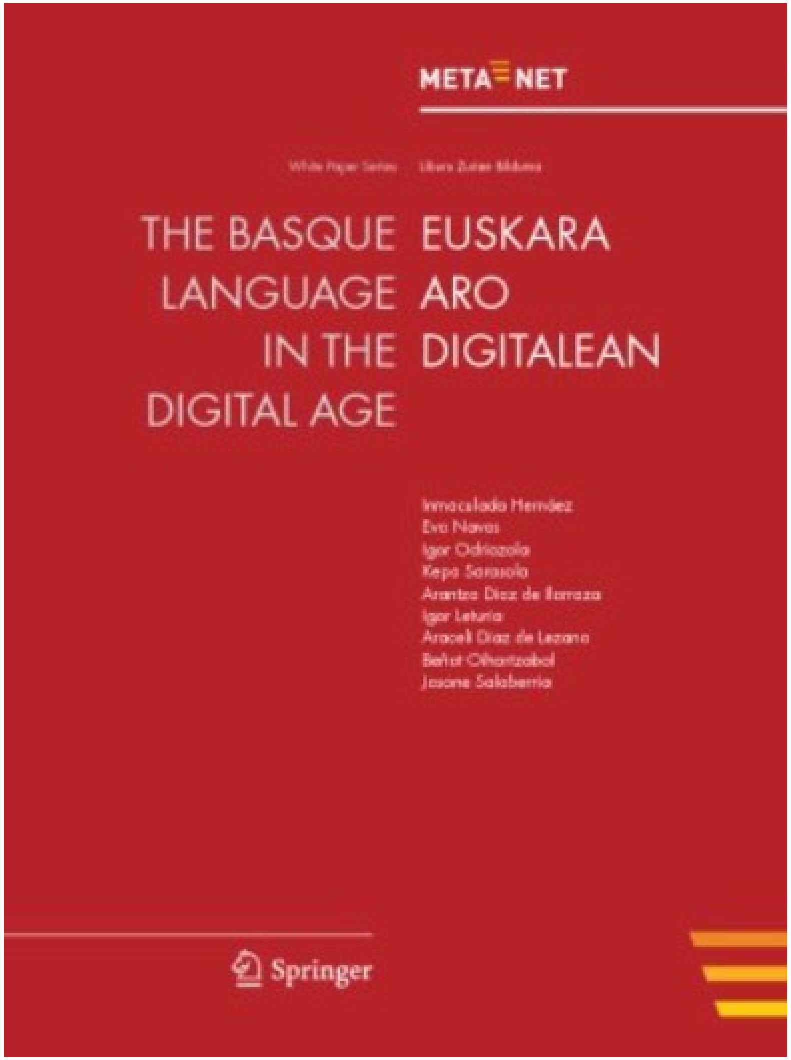
How are languages facing the ICT and HLT challenges?

- 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

White Paper Series Liburu Zurien Bilduma

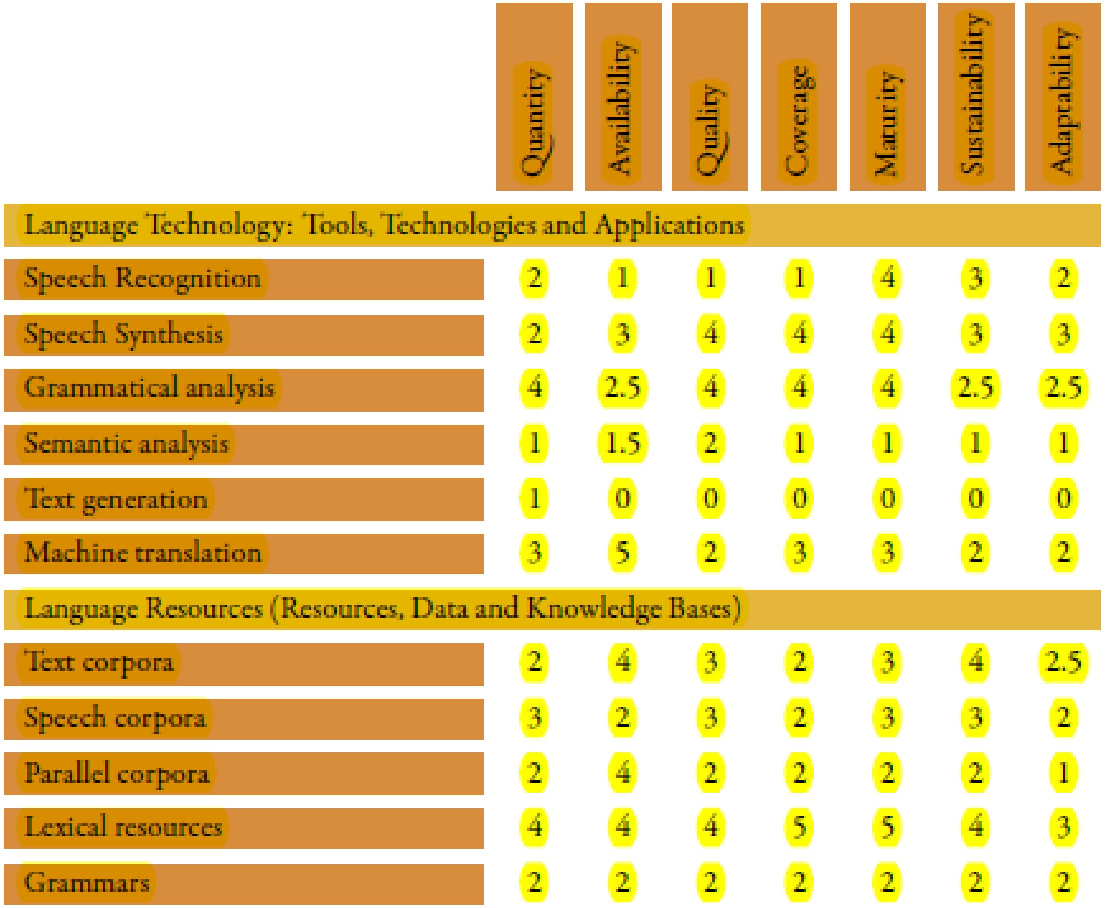
THE BASQUE LANGUAGE IN THE DIGITAL AGE

EUSKARA ARO DIGITALEAN

Inmaculada Hernáez
Eva Navas
Igor Odriozola
Kepa Sarasola
Arantza Díaz de Ilarraza
Igor Leturia
Araceli Díaz de Lezana
Beñat Oihartzabal
Josune Salaberria

Springer

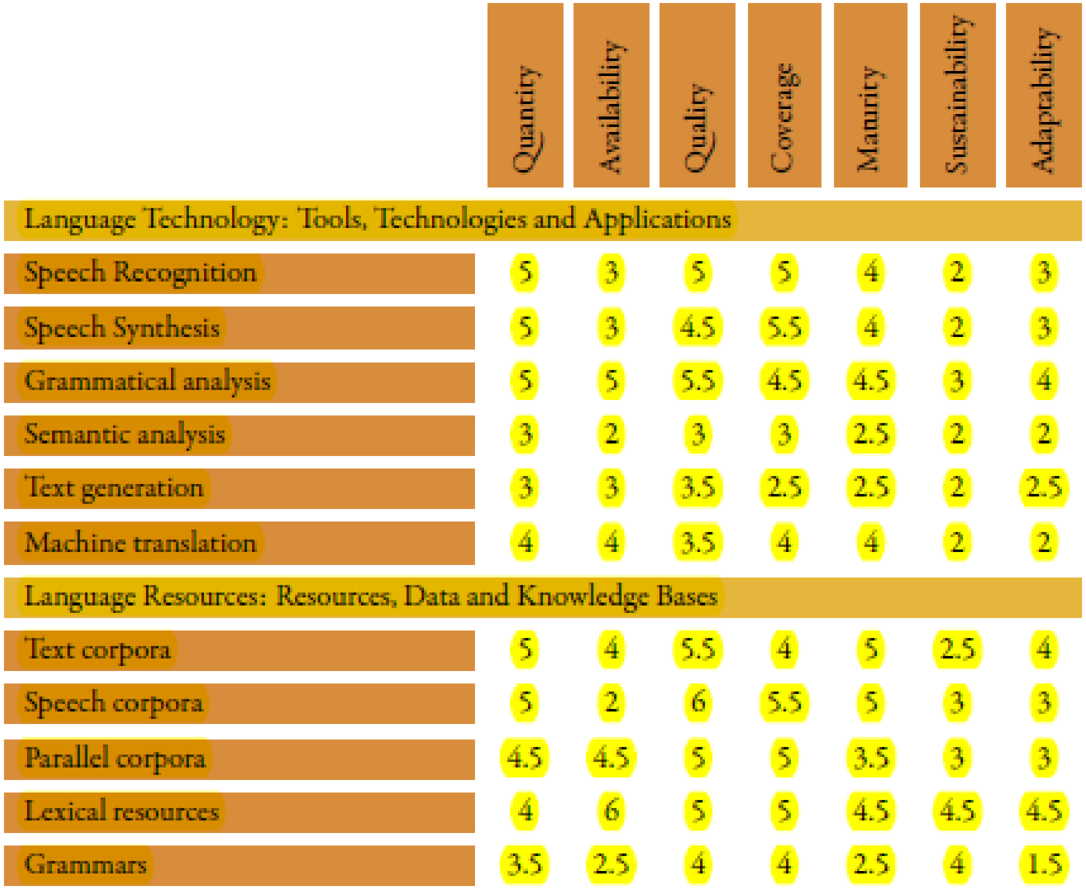
META-NET:
results for Basque



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

7: State of language technology support for Basque

META-NET: results for English

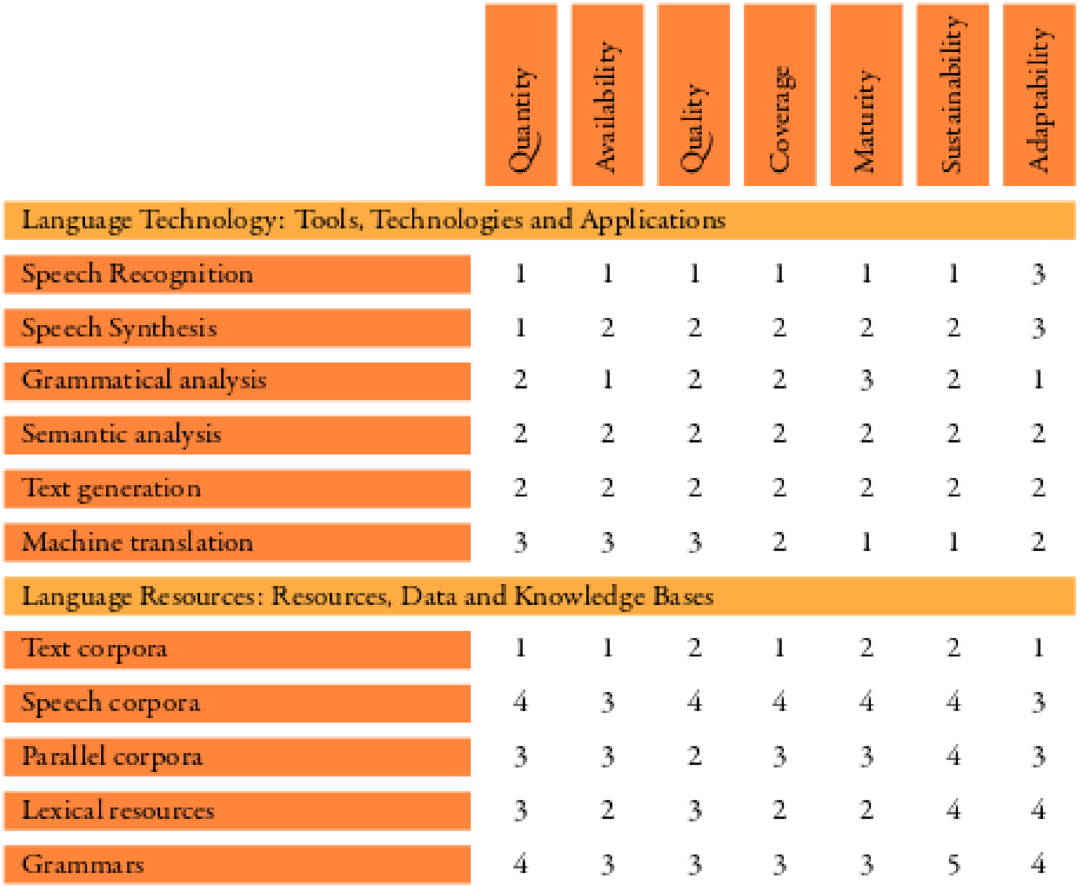


	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	5	3	5	5	4	2	3
Speech Synthesis	5	3	4.5	5.5	4	2	3
Grammatical analysis	5	5	5.5	4.5	4.5	3	4
Semantic analysis	3	2	3	3	2.5	2	2
Text generation	3	3	3.5	2.5	2.5	2	2.5
Machine translation	4	4	3.5	4	4	2	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	5	4	5.5	4	5	2.5	4
Speech corpora	5	2	6	5.5	5	3	3
Parallel corpora	4.5	4.5	5	5	3.5	3	3
Lexical resources	4	6	5	5	4.5	4.5	4.5
Grammars	3.5	2.5	4	4	2.5	4	1.5

8: State of language technology support for English

META-NET:
results for Welsh

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



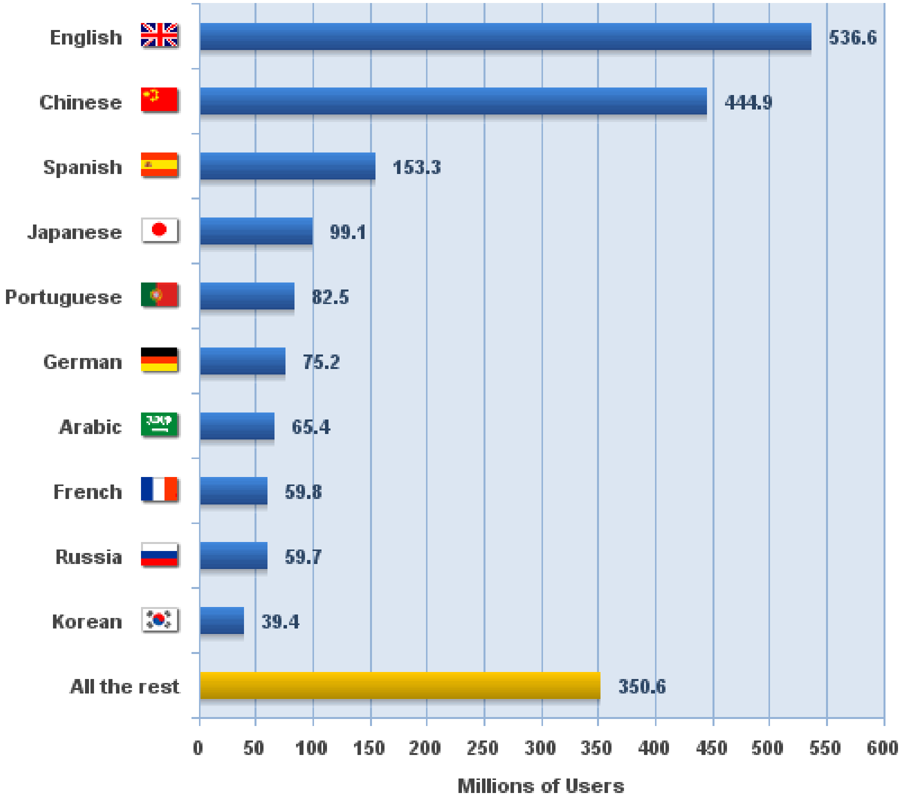
	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1	1	1	3
Speech Synthesis	1	2	2	2	2	2	3
Grammatical analysis	2	1	2	2	3	2	1
Semantic analysis	2	2	2	2	2	2	2
Text generation	2	2	2	2	2	2	2
Machine translation	3	3	3	2	1	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1	1	2	1	2	2	1
Speech corpora	4	3	4	4	4	4	3
Parallel corpora	3	3	2	3	3	4	3
Lexical resources	3	2	3	2	2	4	4
Grammars	4	3	3	3	3	5	4

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.
- 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.

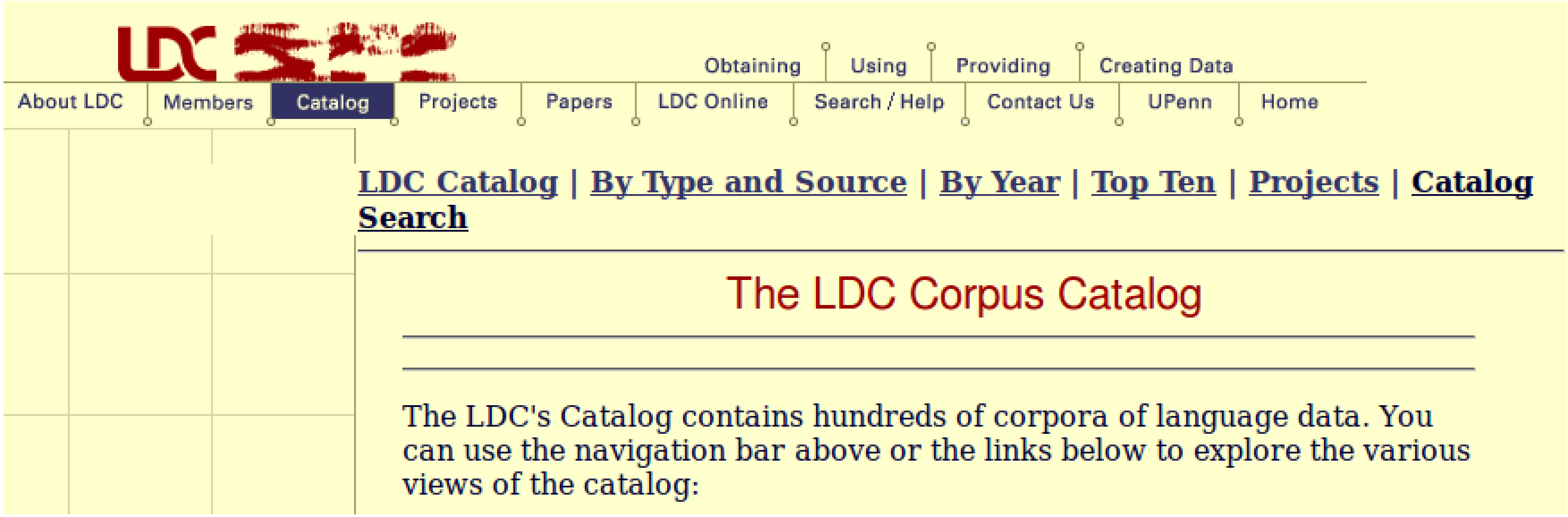
- > 1000 resources **for 60 languages**
- Resources distributed by ELRA agency

  (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



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

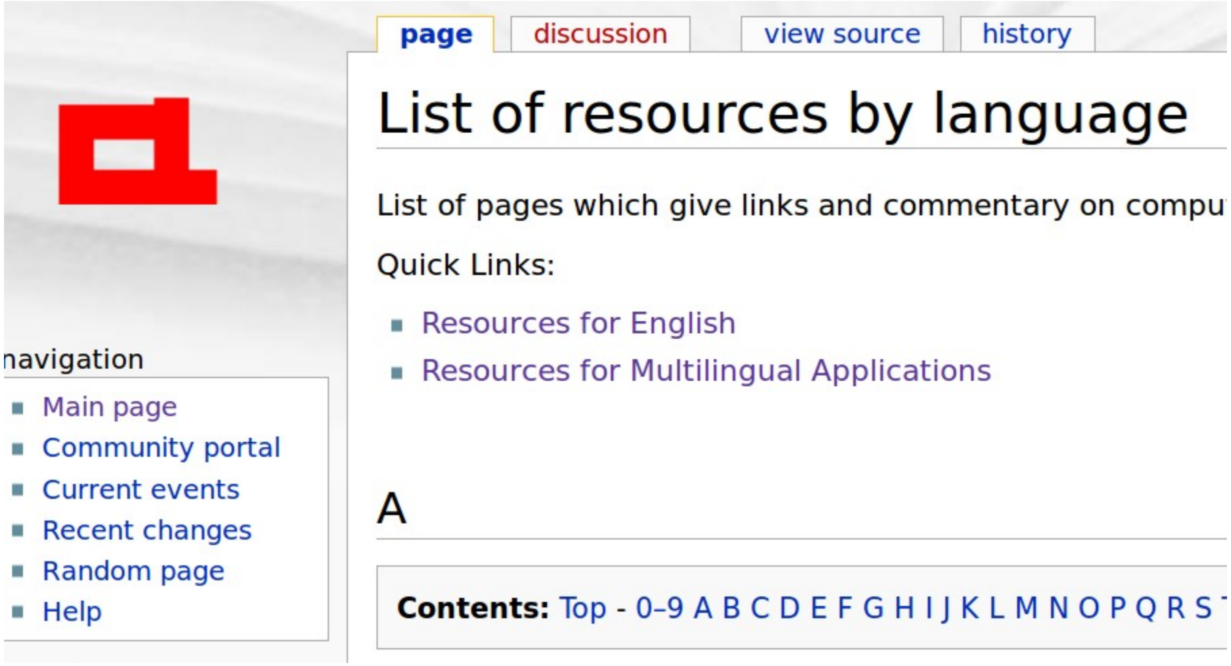
The LDC Corpus Catalog

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:

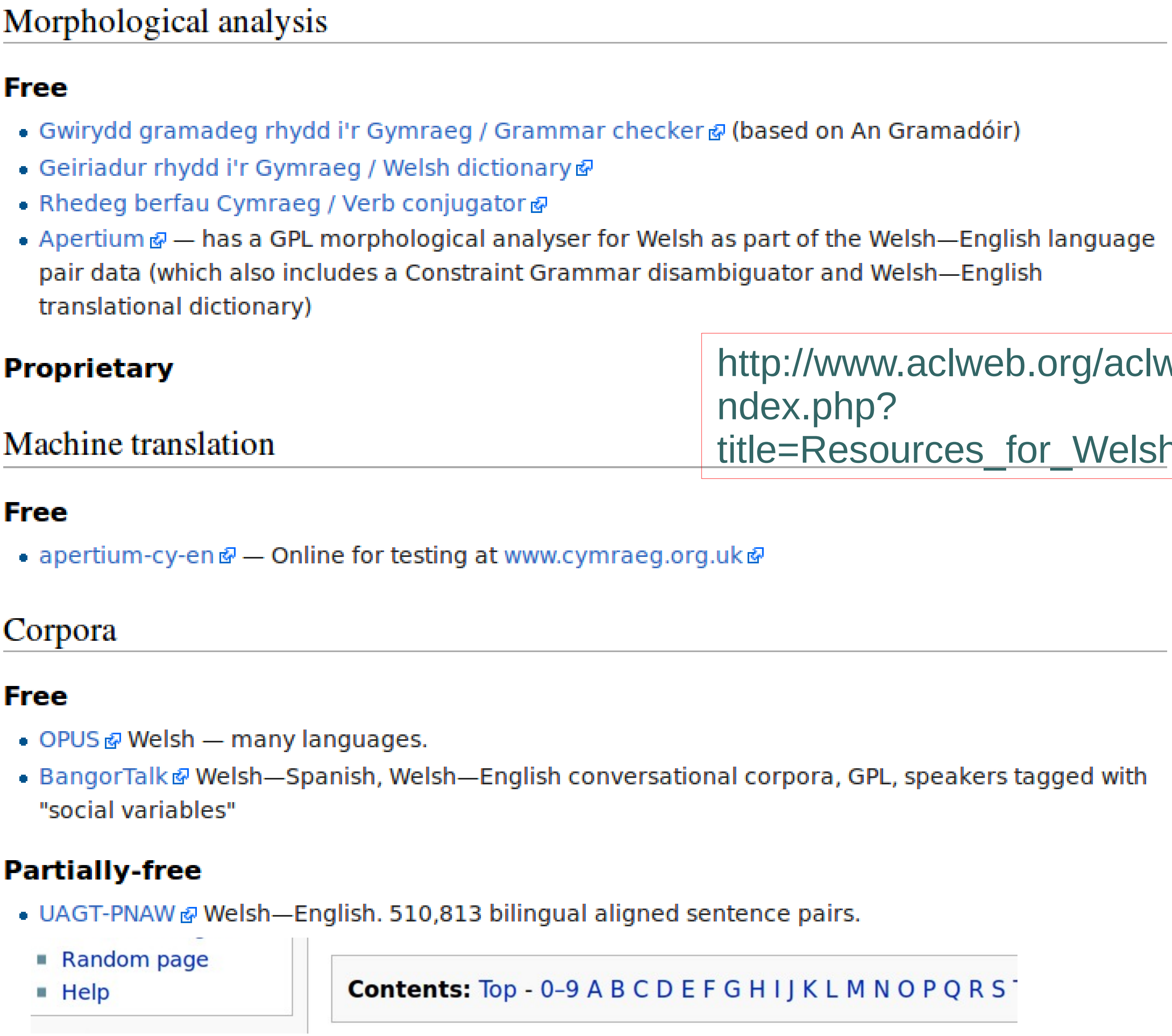
- Resources for English
- Resources for Multilingual Applications

A

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

I

tics



Morphological analysis

Free

- Gwirydd 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.

- Random page
- Help

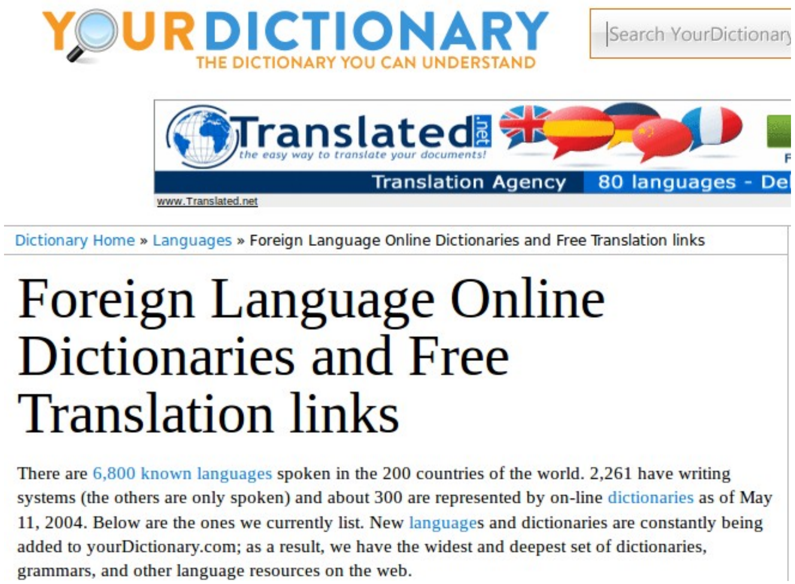
Contents: Top - 0–9 A B C D E F G H I J K L M N O P Q R S

http://www.aclweb.org/aclwiki/index.php?title=Resources_for_Welsh

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**

      Basque? No ; Welsh? No

- MT systems
  - Google-Translate: **100 languages** Basque? Yes ; Welsh? Yes
  - Babelfish: **13 languages**

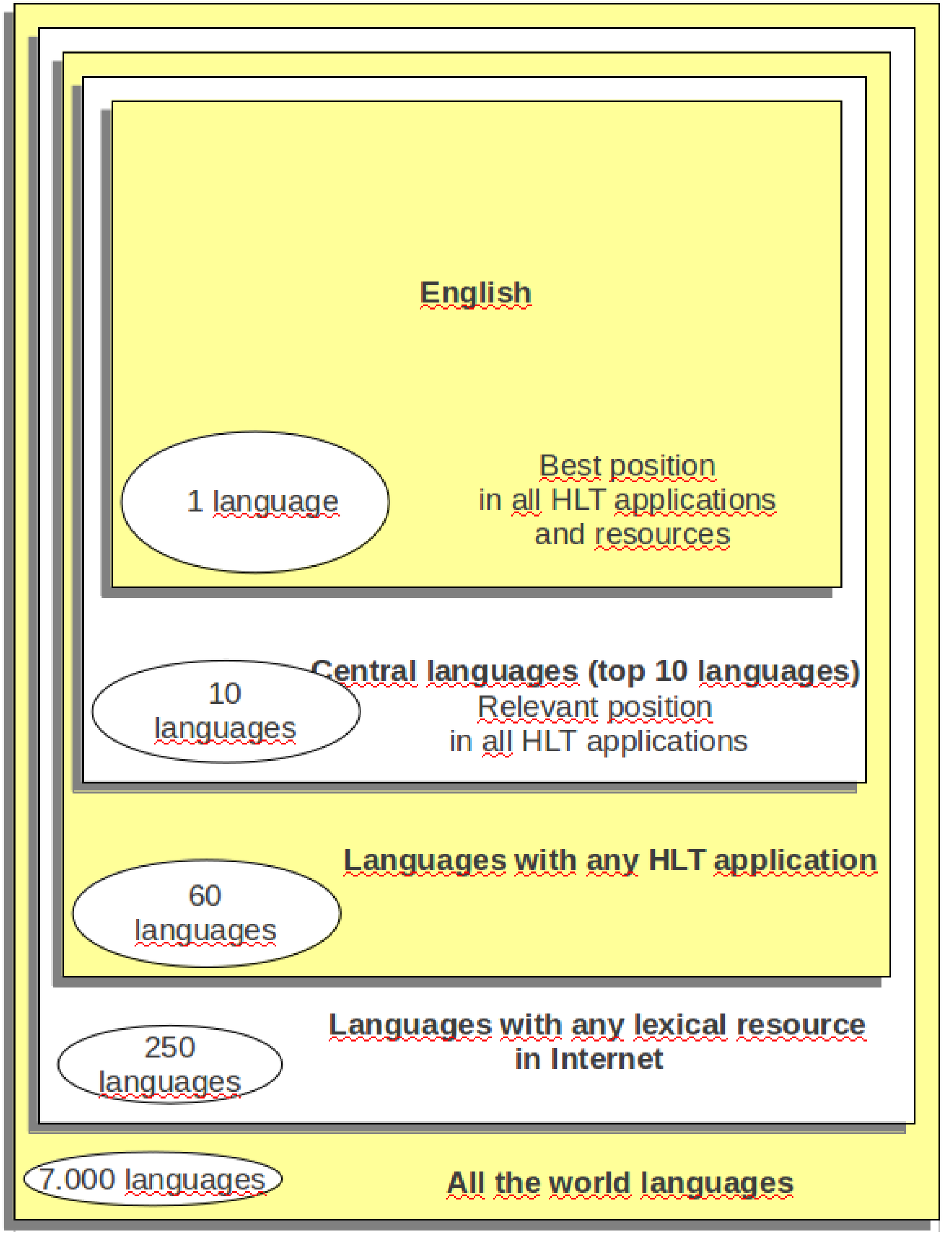
Outline

- 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



English

1 language

Best position
in all HLT applications
and resources

10
languages

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

60
languages

Languages with any HLT application

250
languages

**Languages with any lexical resource
in Internet**

7.000 languages

All the world languages

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

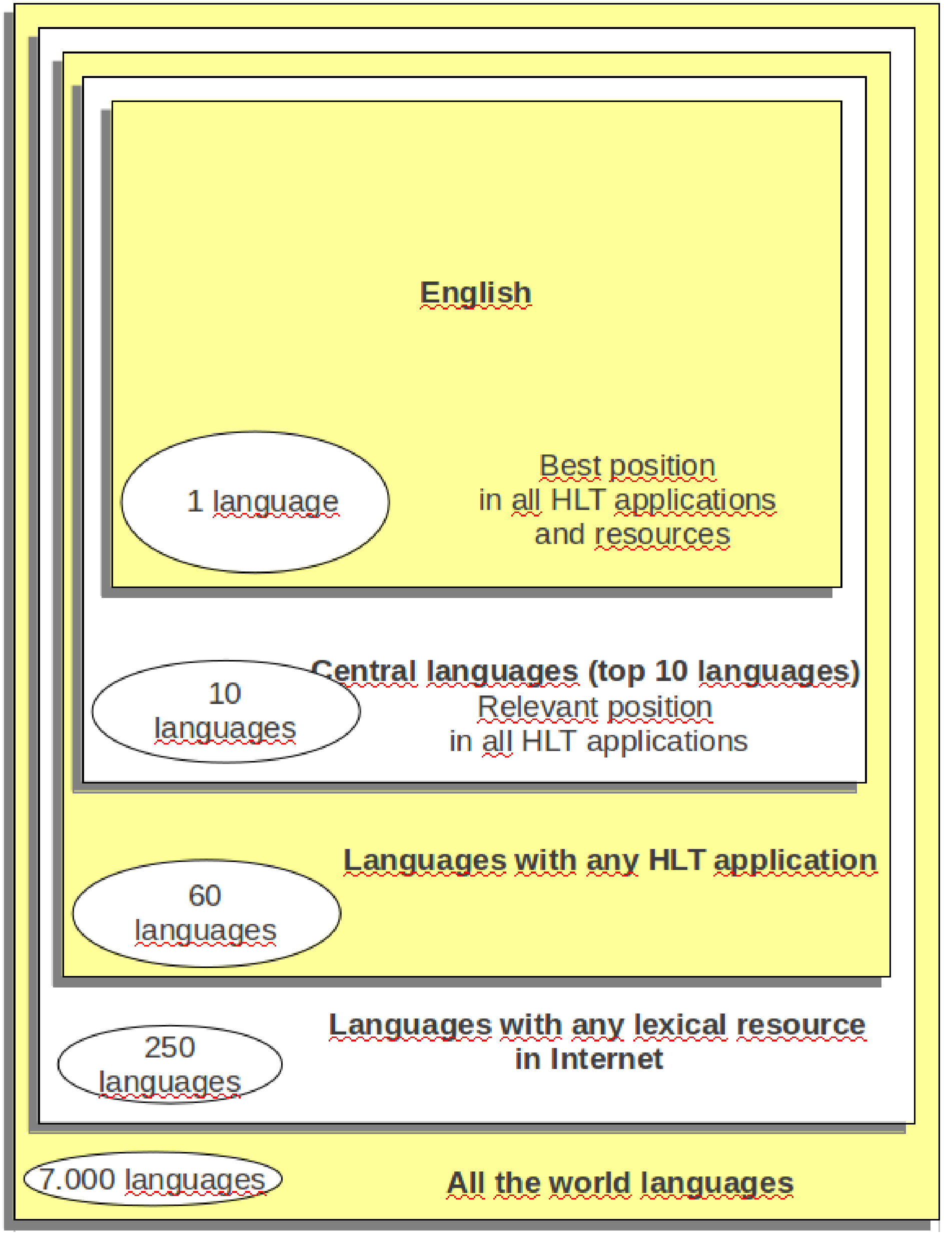
- Fifth level:
  Languages that have writing systems
  (Borin, 2009)
  - Here are included **other 2,014 languages**

- Sixth level:
  Only-spoken languages in the world
  - Here are included at least **other 4,500 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

60
languages

Languages with any HLT application

250
languages

**Languages with any lexical resource
in Internet**

7.000 languages

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
- 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 resouced 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

ixa

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

Underlying strategy

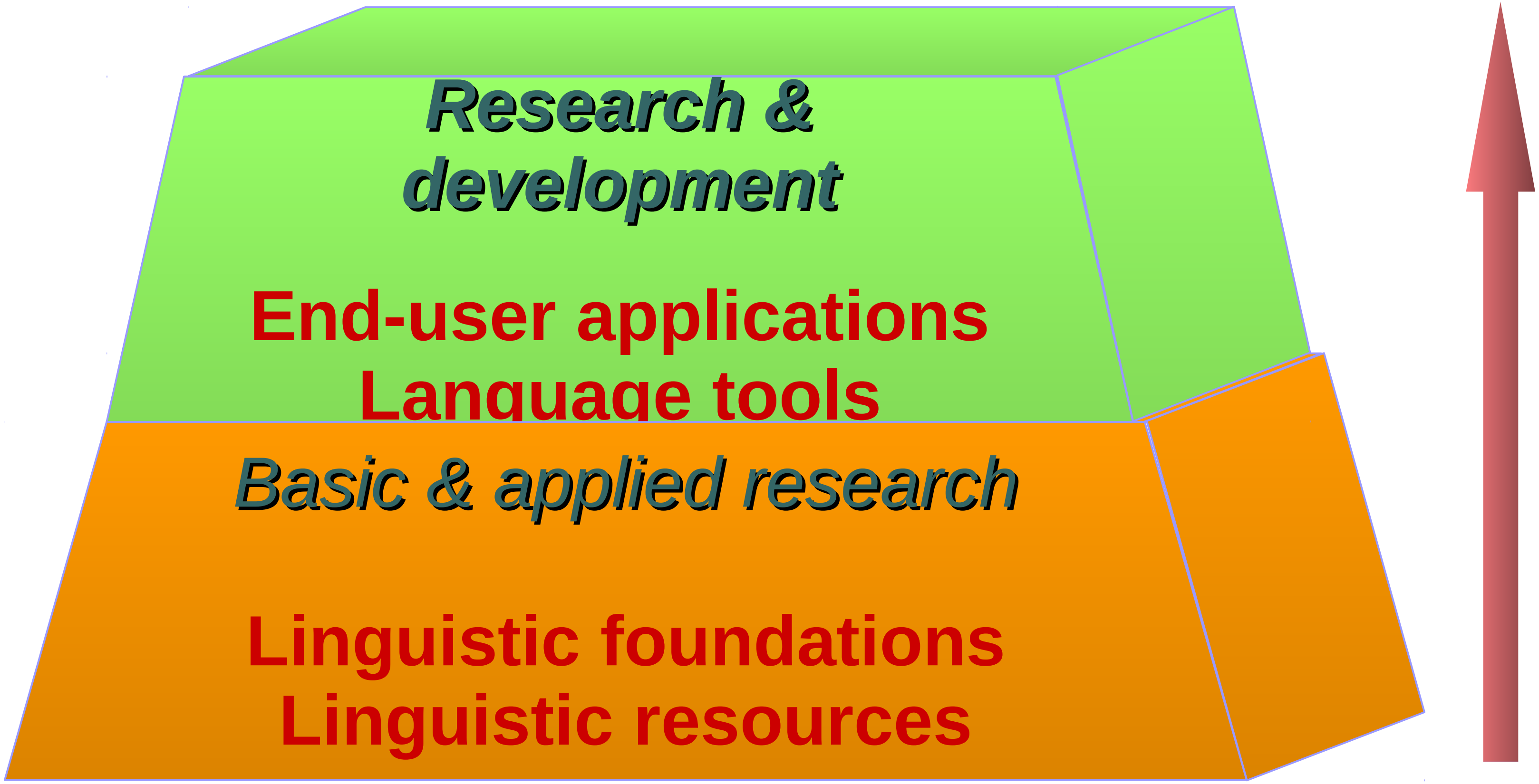
- Need of standardization of resources to be useful:
  - in different researches
  - in different tools
  - in different applications
- Need of incremental design and development of language foundations, tools, and applications
  - in a parallel and coordinated way
  - in order to get the best benefit from them

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 annotati on 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



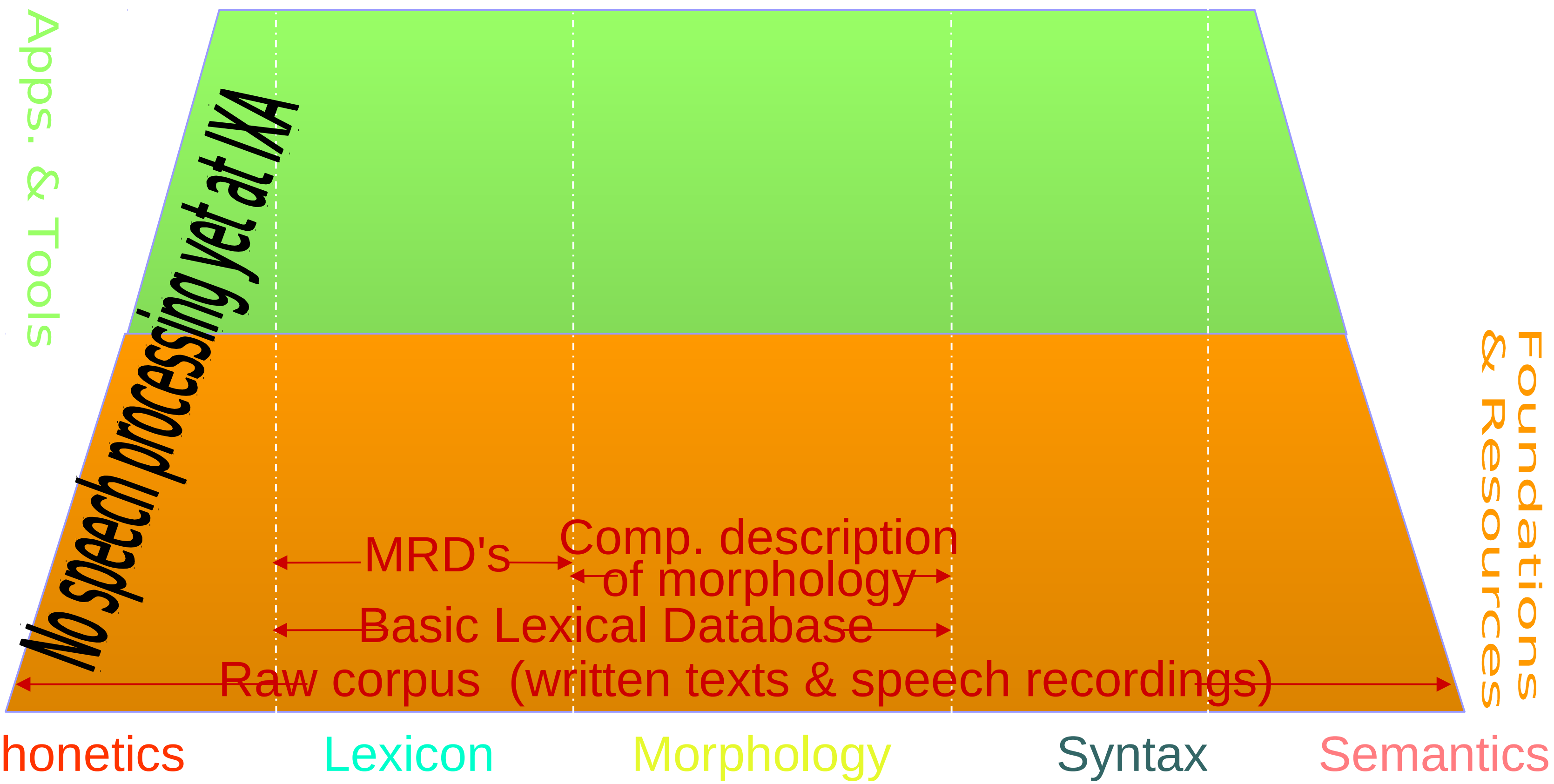
Research & development

End-user applications
Language tools

Basic & applied research

Linguistic foundations
Linguistic resources

Phase I: laying foundations



Apps. & Tools

No speech processing yet at IXA

MRD's

Comp. description of morphology

Basic Lexical Database

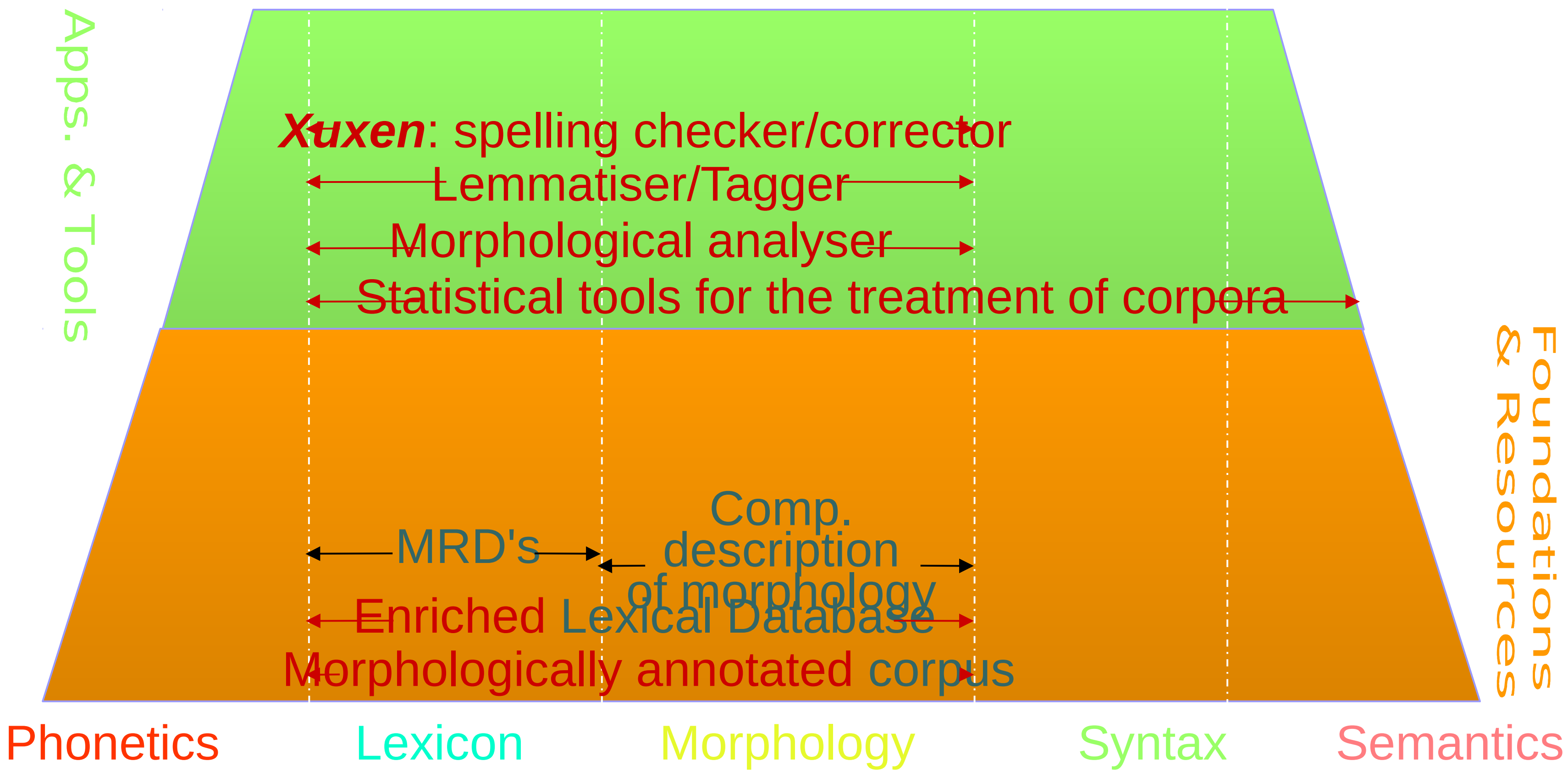
Raw corpus (written texts & speech recordings)

Foundations & Resources

Phonetics Lexicon Morphology Syntax Semantics

51

Phase II:
first basic tools and applications



Apps. & Tools

Xuxen: spelling checker/corrector

Lemmatiser/Tagger

Morphological analyser

Statistical tools for the treatment of corpora

MRD's

Comp. description of morphology

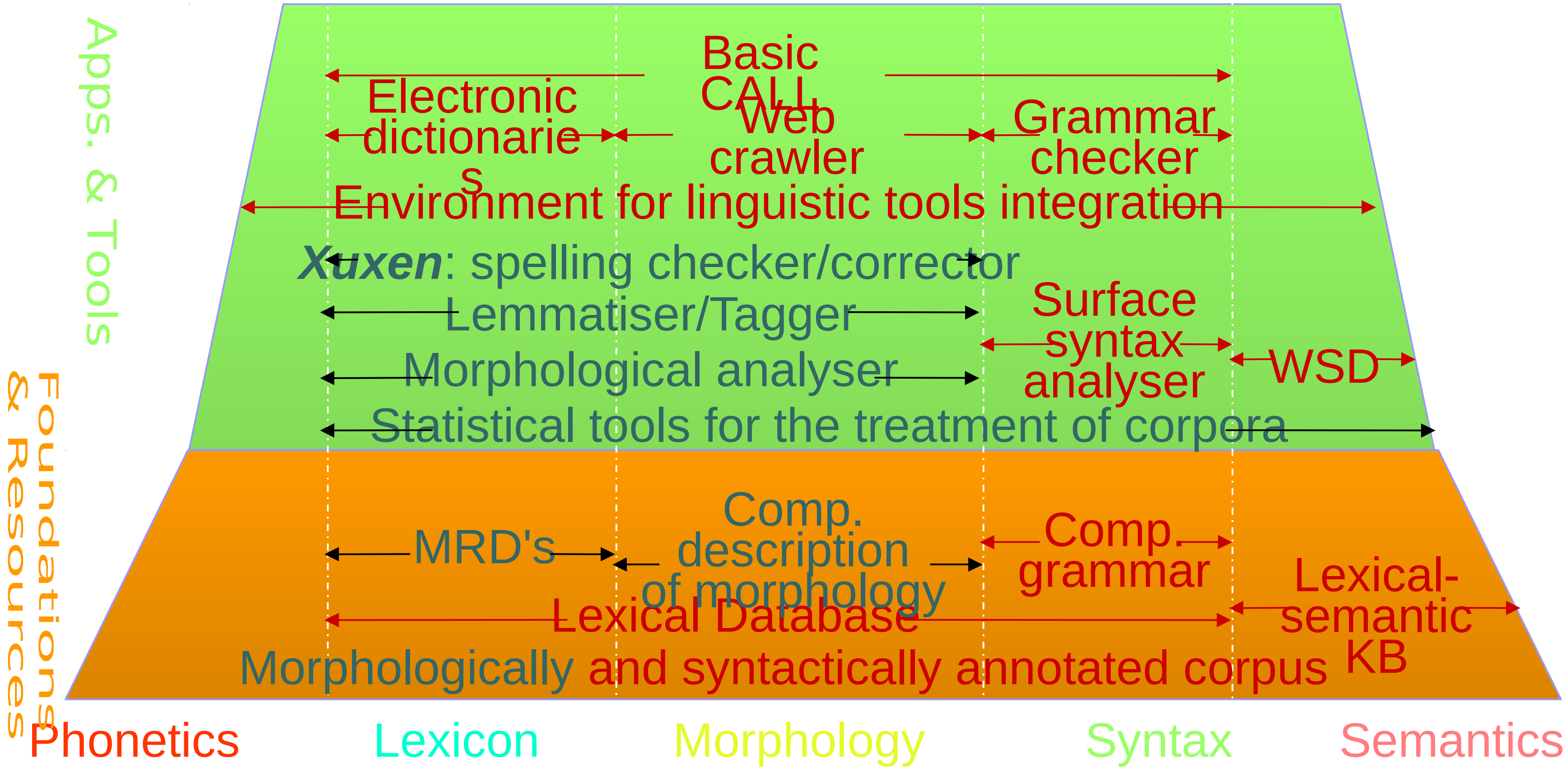
Enriched Lexical Database

Morphologically annotated corpus

Foundations & Resources

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

Morphological analyser

WSD

Statistical tools for the treatment of corpora

Foundations & Resources

Comp. description of morphology

MRD's

Comp. grammar

Lexical-semantic KB

Lexical Database

Morphologically and syntactically annotated corpus

Phonetics

Lexicon

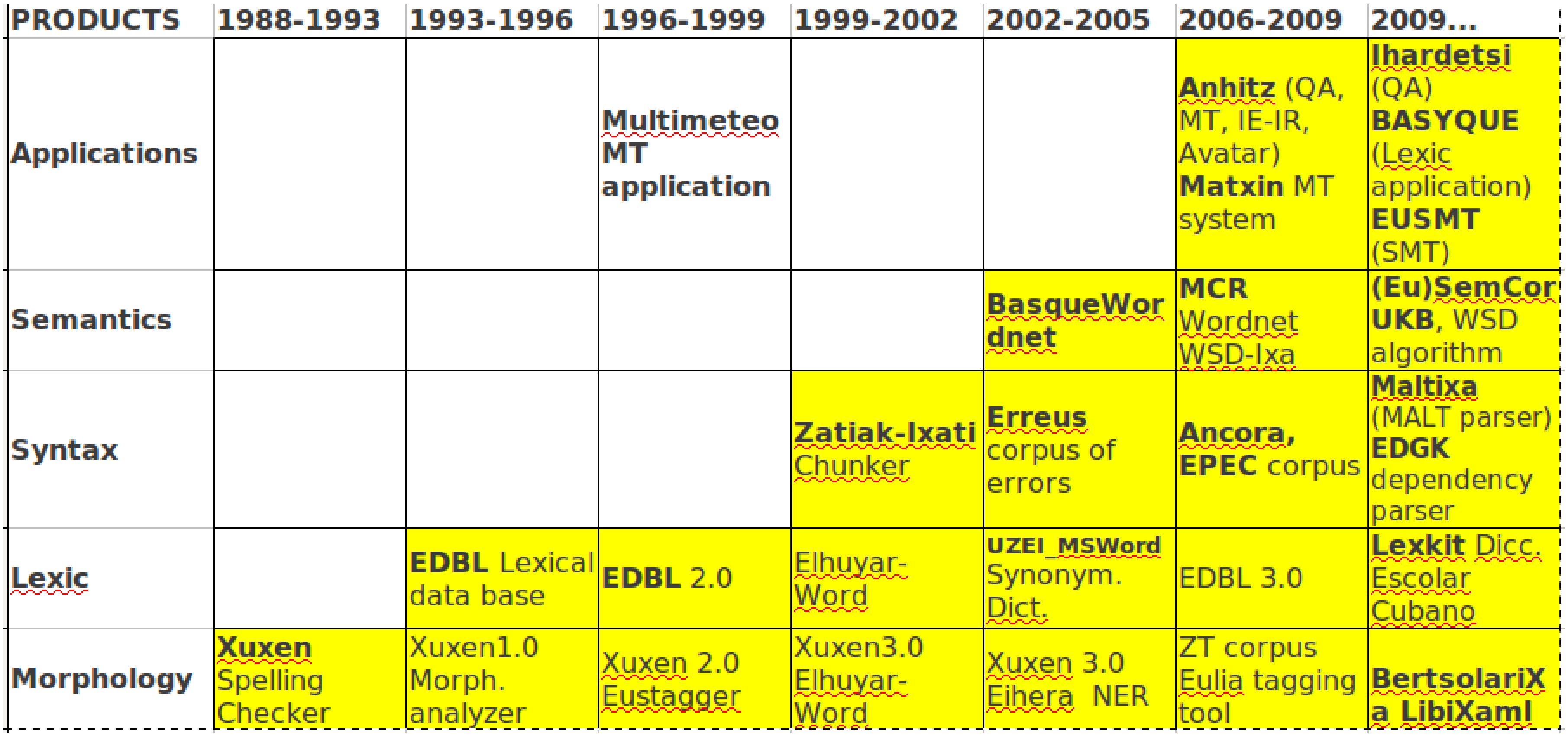
Morphology

Syntax

Semantics

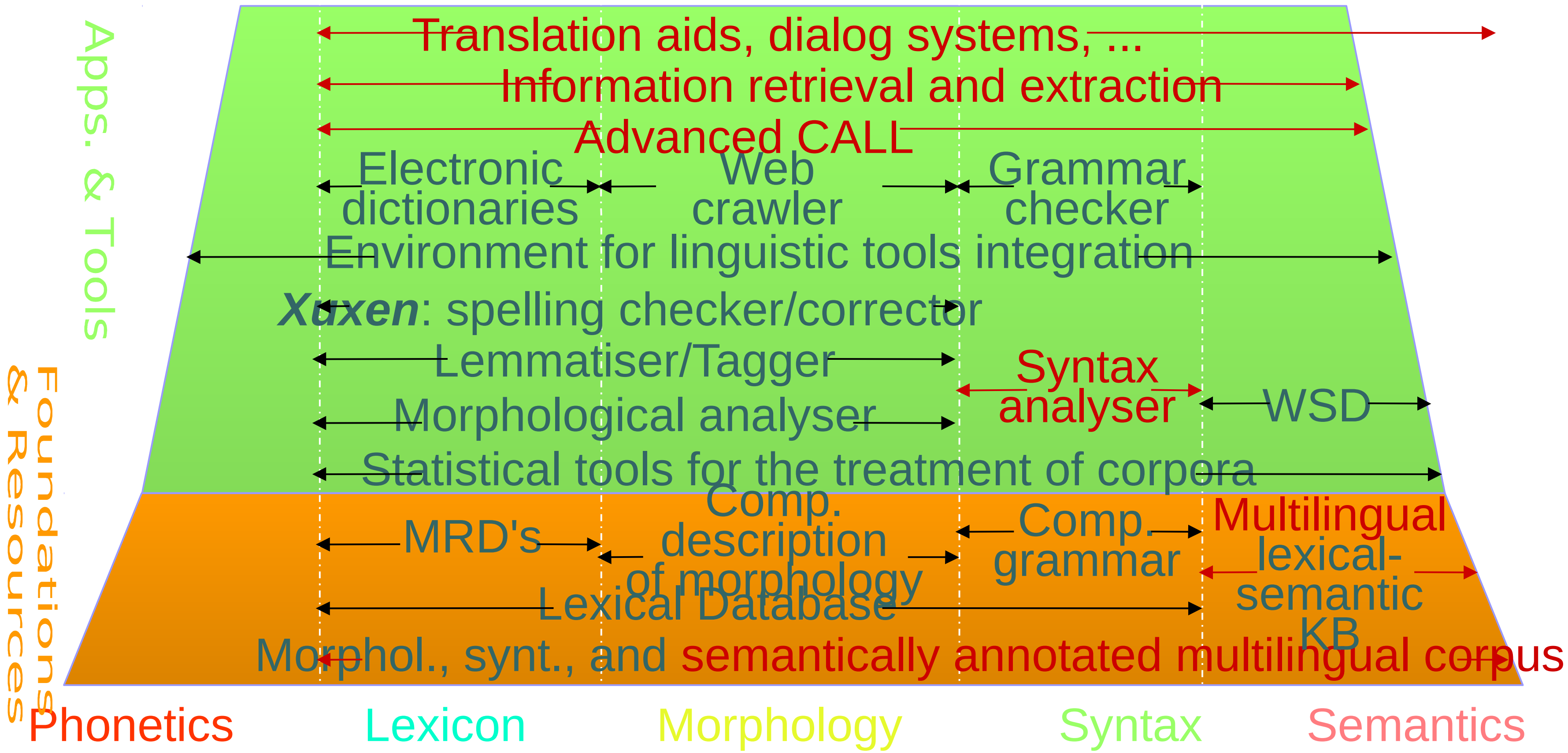
Created LRs and tools (1988-2010)

http://ixa.si.ehu.es/Ixa/Produktuak



PRODUCTS	1988-1993	1993-1996	1996-1999	1999-2002	2002-2005	2006-2009	2009...
Applications			**Multimeteo MT application**			**Anhitz** (QA, MT, IE-IR, Avatar) **Matxin** MT system	**Ihardetsi** (QA) **BASYQUE** (Lexic application) **EUSMT** (SMT)
Semantics					**BasqueWordnet**	**MCR** Wordnet WSD-Ixa	**(Eu)SemCor** **UKB**, WSD algorithm
Syntax				**Zatiak-Ixati** Chunker	**Erreus** corpus of errors	**Ancora,** **EPEC** corpus	**Maltixa** (MALT parser) **EDGK** dependency parser
Lexic		**EDBL** Lexical data base	**EDBL** 2.0	Elhuyar-Word	**UZEI_MSWord** Synonym. Dict.	EDBL 3.0	**Lexkit** Dicc. Escolar Cubano
Morphology	**Xuxen** Spelling Checker	Xuxen1.0 Morph. analyzer	Xuxen 2.0 Eustagger	Xuxen3.0 Elhuyar-Word	Xuxen 3.0 Eihera NER	ZT corpus Eulia tagging tool	**BertsolariXa LibiXaml**

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

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.

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