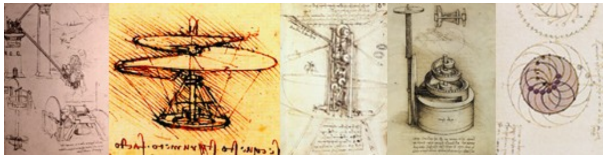
Can NLP help less resourced languages to promote their use?

Kepa Sarasola
Ixa Taldea.
University of the Basque Country
http://ixa.si.ehu.es



Joint Culture & Technology and CLARIN-D Summer School
"Digital Humanities & Language Resources"
22nd of July to 1st of August 2014

Can NLP help 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 NLP help less resourced languages to promote their use?
- Related work
- Conclusions

How are languages facing the ICT and HLT challenges?

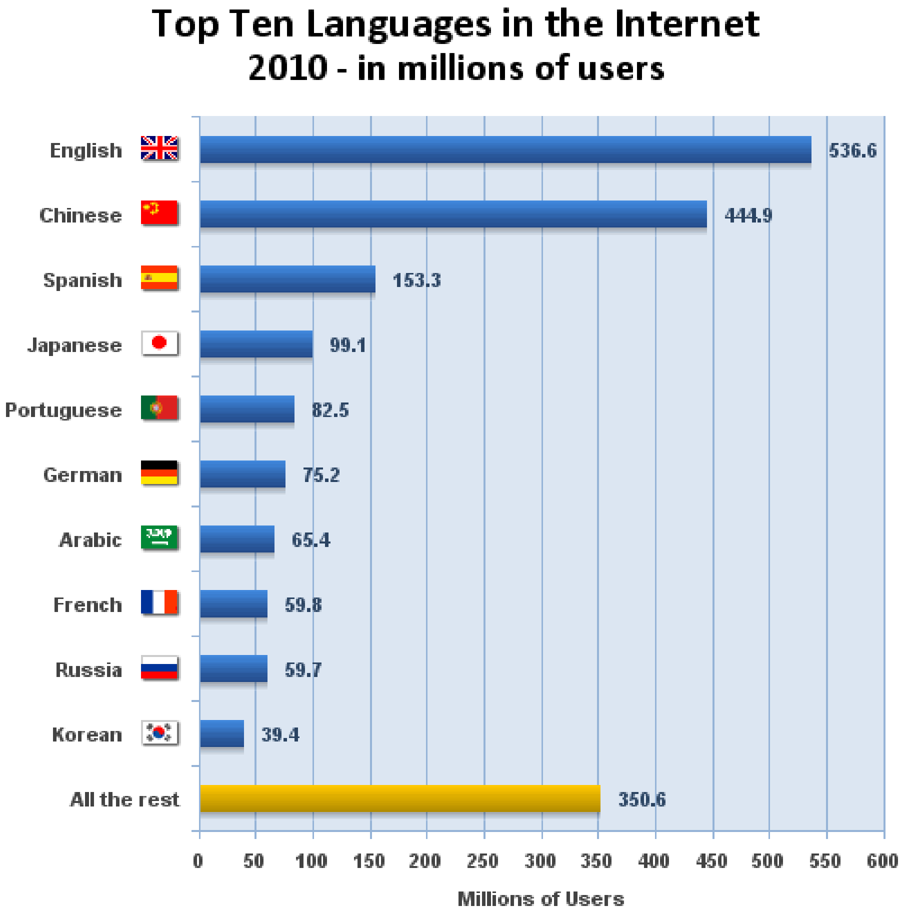
- Figures about amounts of resources on the Internet for different languages are not easy to obtain

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

How are languages facing ICT?

Number of users

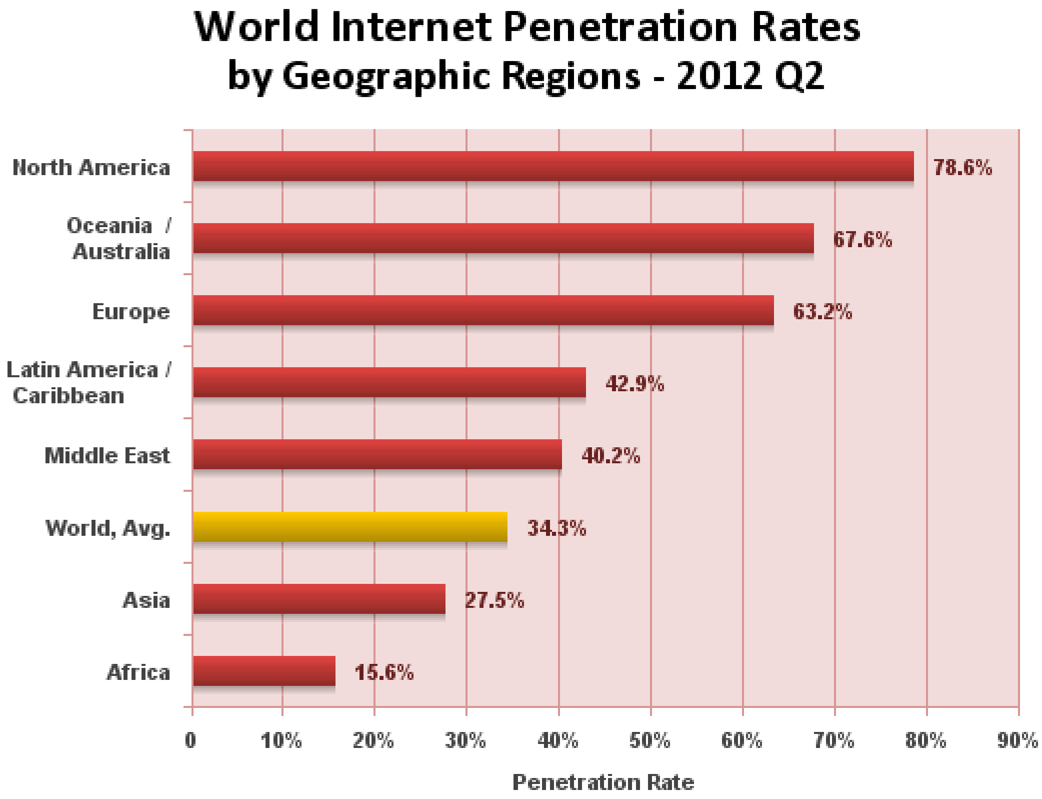
- Internet World Stats 2010
- English :
  - 636 million users
  - 30%
- Top ten languages
  - 1.600 million users
  - 82.2%
- Rest of the languages
  - 360 million users
  - 17,8% of users
  - 36% of world population



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

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

How are languages facing ICT?



World Internet Penetration Rates
by Geographic Regions - 2012 Q2

Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 7,017,846,922
and 2,405,518,376 estimated Internet users on June 30, 2012.
Copyright © 2012, 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 287 languages (July 2014).
- Top 10 languages:
  English (4,6 million articles),
  German (1.7 M), French (1.5 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 (431 K)
  - 33th: Esperanto (199 K)
  - 36th: Basque (189 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
  - They 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.

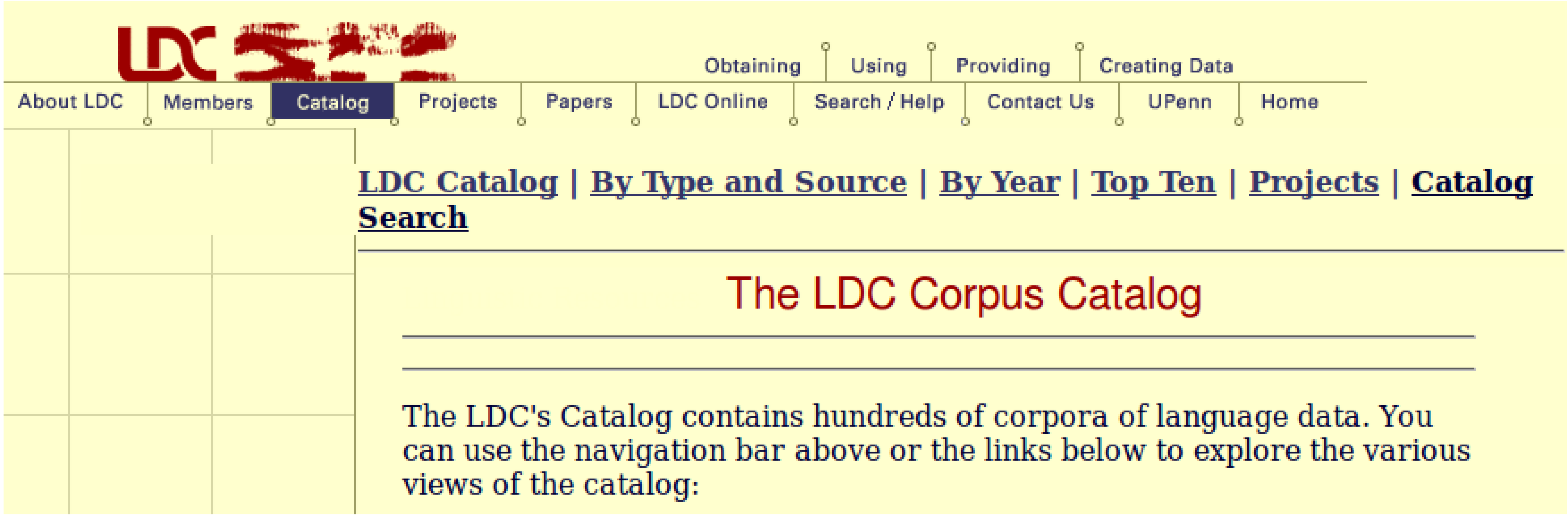
- > 1000 resources **for 60 languages**
- Resources distributed by ELRA agency

  (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.
  - 519 for English, 462 for German,
    16 for catalan, 6 for Basque

How are languages facing HLT?

LDC. Linguistic Data Consortium

- > 500 resources **for 82 languages**
- Search by language is allowed.
- 370 products for English, no products for Basque,



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

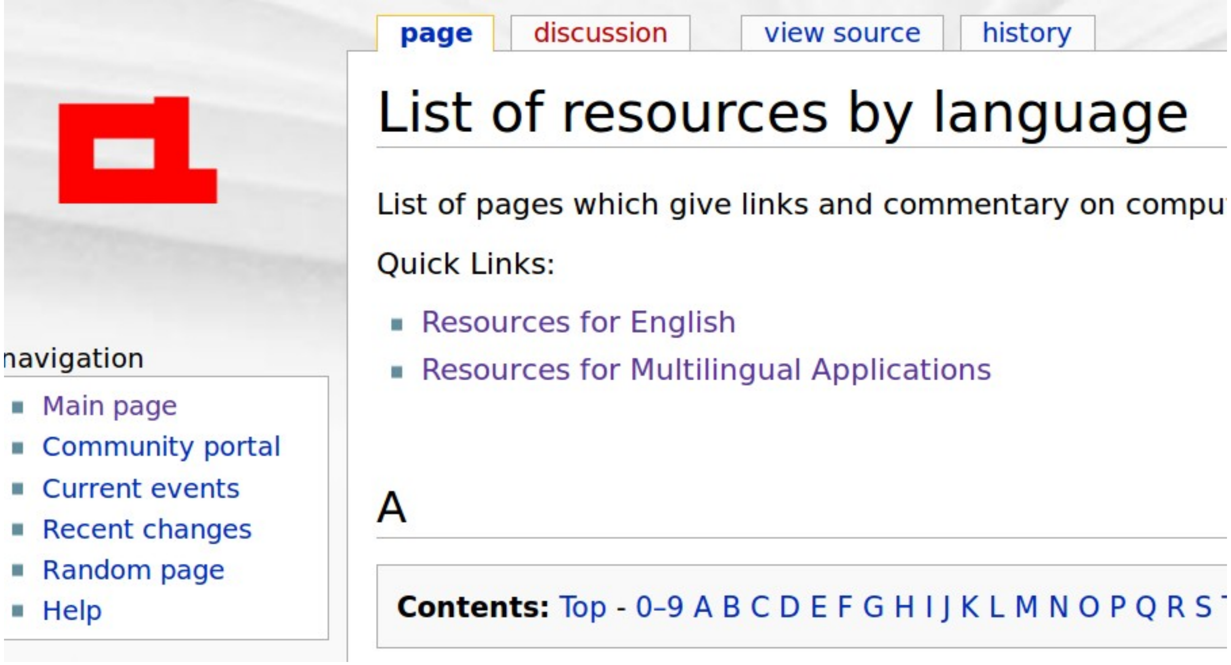
The LDC Corpus Catalog

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

How are languages facing HLT?

ACLwiki. Association for Computational Linguistics

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



List of resources by language

List of pages which give links and commentary on compu

Quick Links:

- Resources for English
- Resources for Multilingual Applications

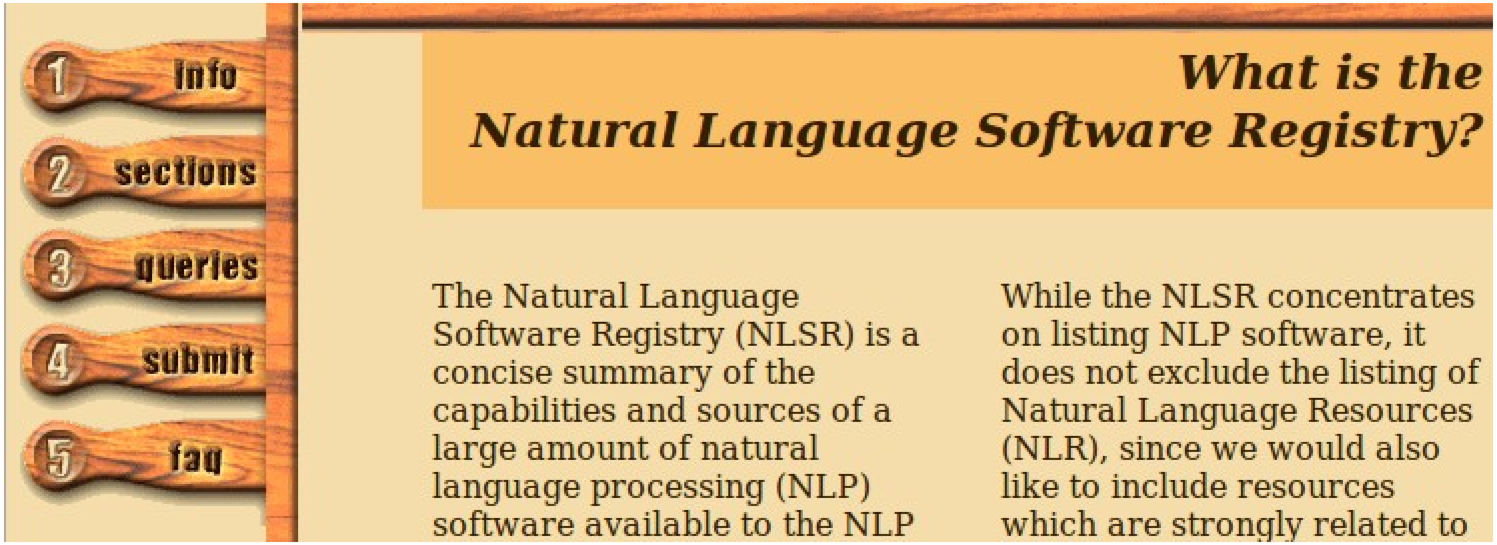
A

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

How are languages facing HLT?

NLSR. Natural Language Software Registry (DFKI)

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



What is the Natural Language Software Registry?

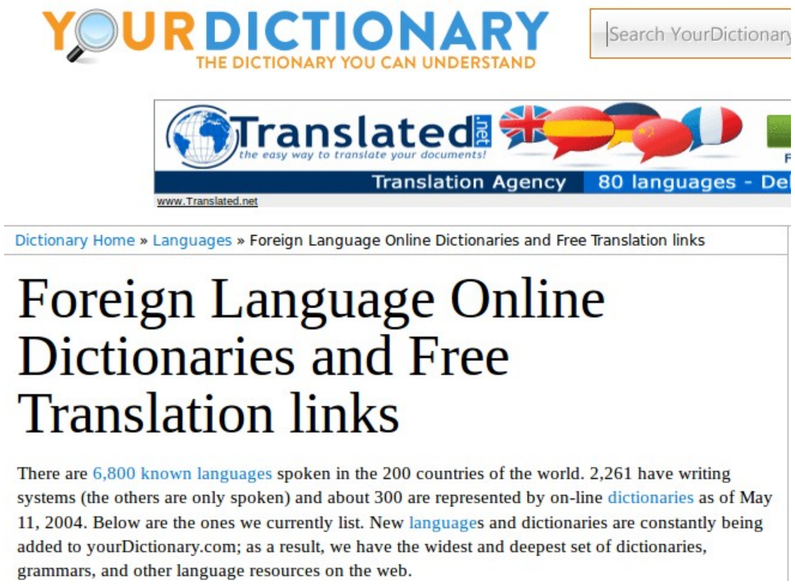
The Natural Language Software Registry (NLSR) is a concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP

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 more than 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**

          (54 languages free download local languages)

            - **Basque, Catalan, Quechua**
    - Libreoffice
        - **104 languages**

          Basque, Catalan, Quechua??

How are languages facing HLT?

Presence/absence in the most popular linguistic services

- Search engines
  - Google:
    - Interfaces in 152 languages
    - Identificates **50 languages**
- MT systems
  - Babelfish: **13 languages**
  - Google-Translate: **80 languages**

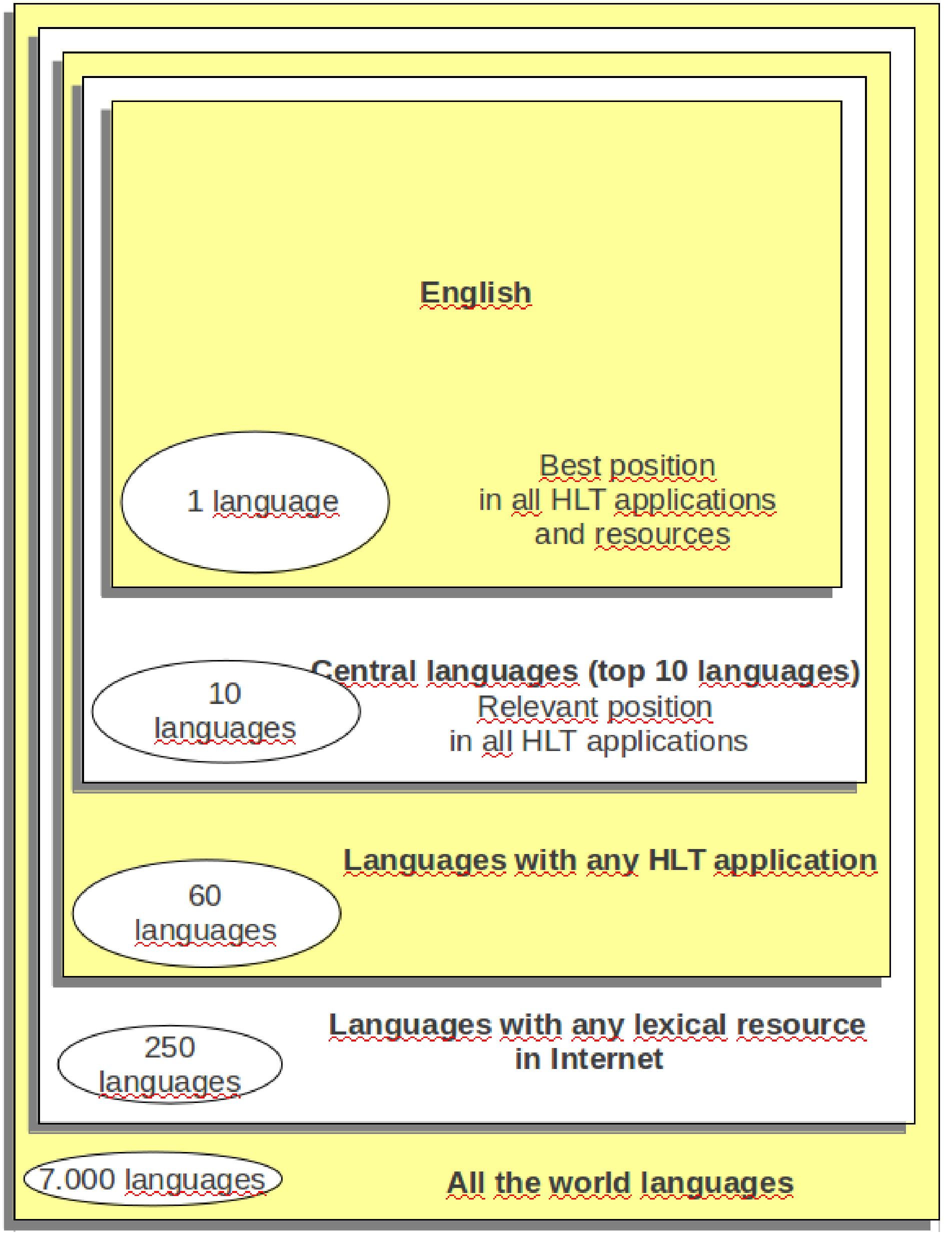
Outline

- How are languages facing the ICT and HLT challenges?
- **Which languages are "less resourced"? Six different levels**
- Can NLP help less resourced 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 of support (Mariani, 2013) 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 (Mariani, 2013)

Which languages are "less resourced"? Six different levels

- Third level: around 70 languages.
  Moderate and fragmentary support (Mariani, 2013)

  Languages with any HLT resource registered
  - 60 languages in ELRA,
  - 82 in LDC,
  - 73 in ACLWiki
  - 30 in NLSR.

  Google dentificates **50 languages**

  Google-Translate: **80 languages**

Which languages are "less resourced"? Six different levels

- Fourth level: Around 300 languages
  Weak support in (Mariani, 2013)

  Languages with any lexical resource on-line registered

  - 307 languages in *yourdictionary.com*
  - It is almost the same set of languages that are present in Wikipedia (287 languages).

Which languages are "less resourced"? Six different levels

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

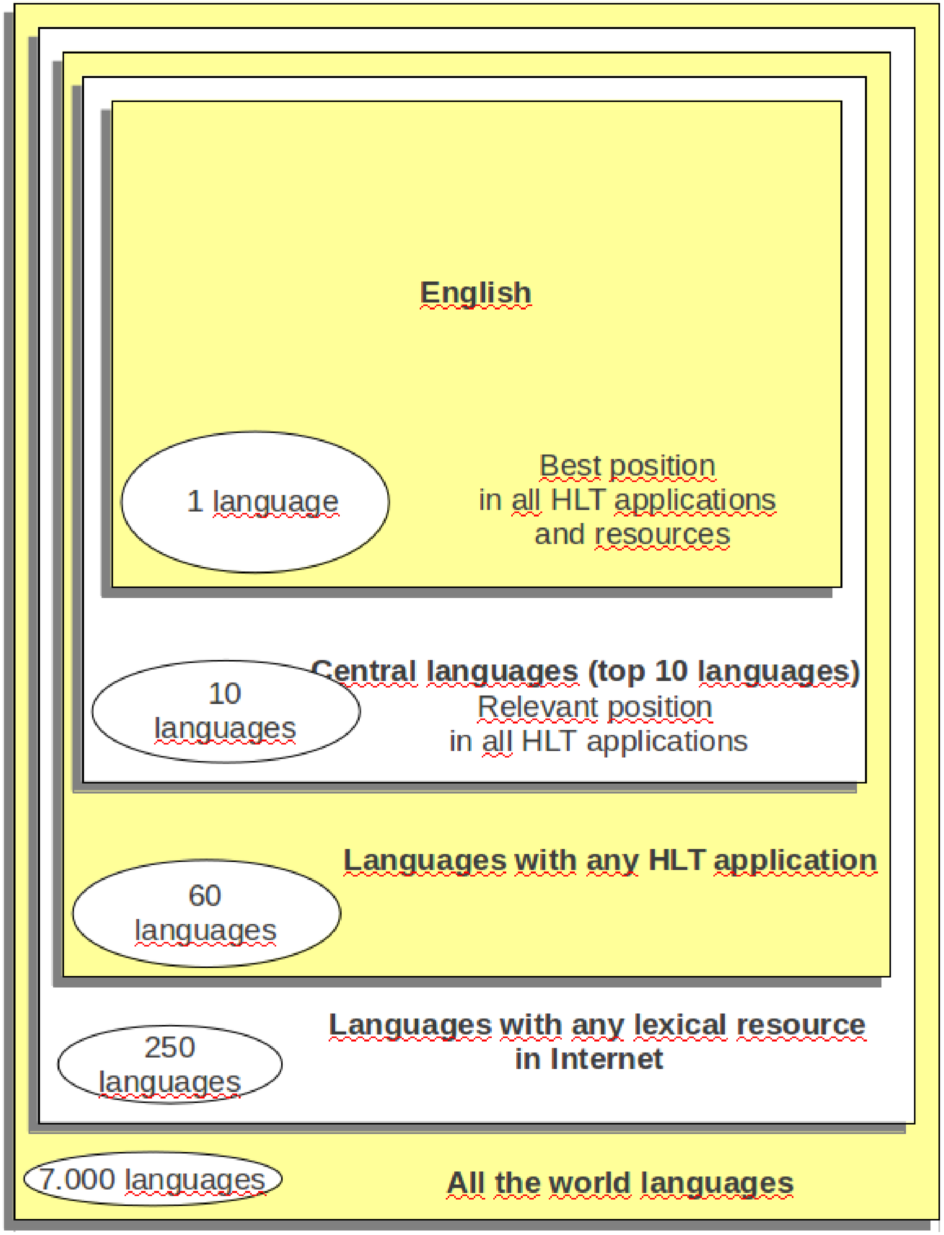
- Sixth level:
  the big bag including only-spoken languages in the world
  - At least **other 4,500 languages**

Both 5th and 6th correspond to Languages with No Application support (Mariani, 2013)

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 NLP help less resourced languages to promote their use?**
- Related work
- Conclusions

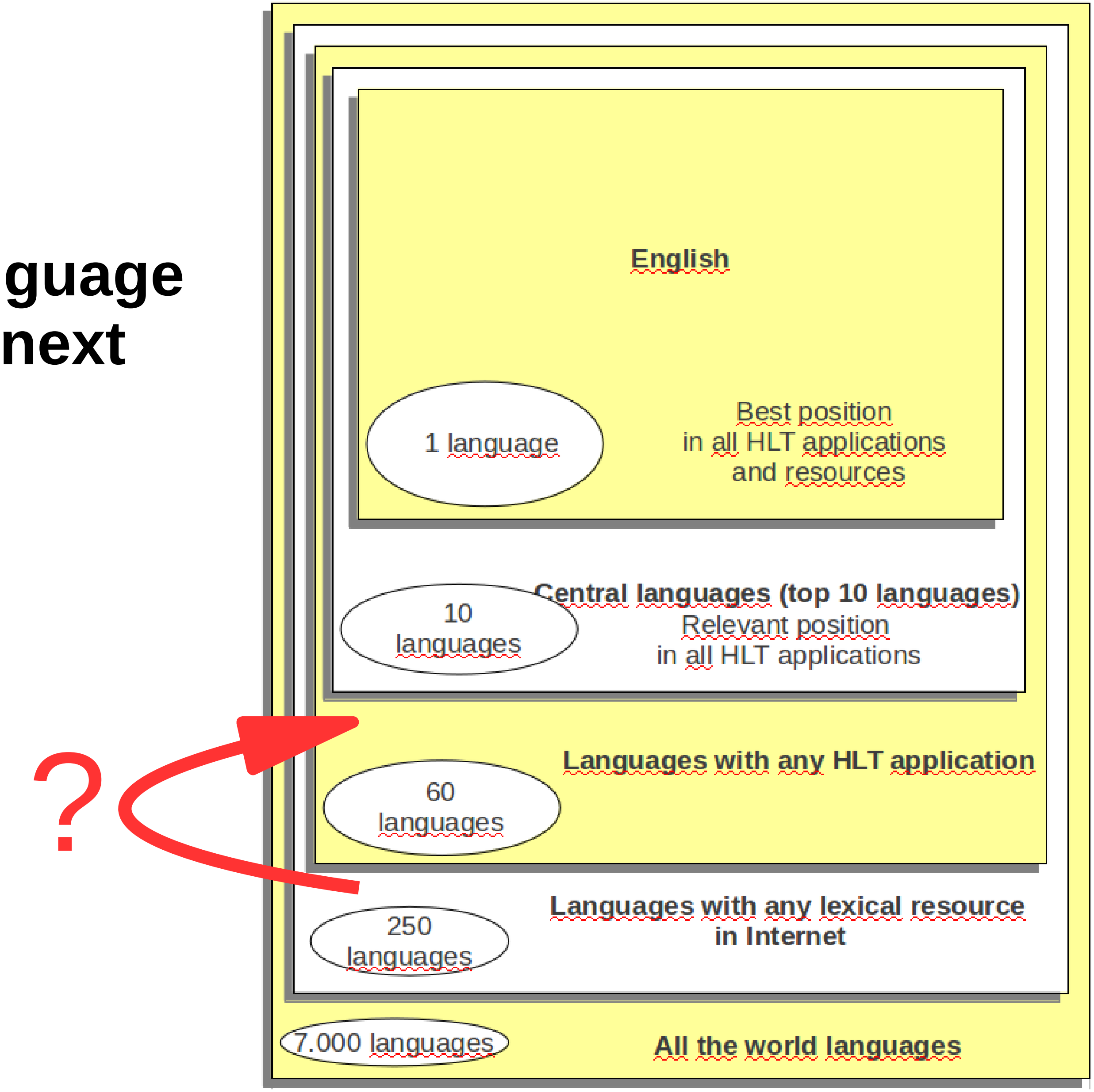
Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- **Can NLP help less resourced languages to promote their use?**
- Related work
- Conclusions

Can NLP help?

Helping the language to climb to the next level?

- **Basque** from 4th level to 5th level? (1988-2009)
- **Quechua** ? (2012 - ...)



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

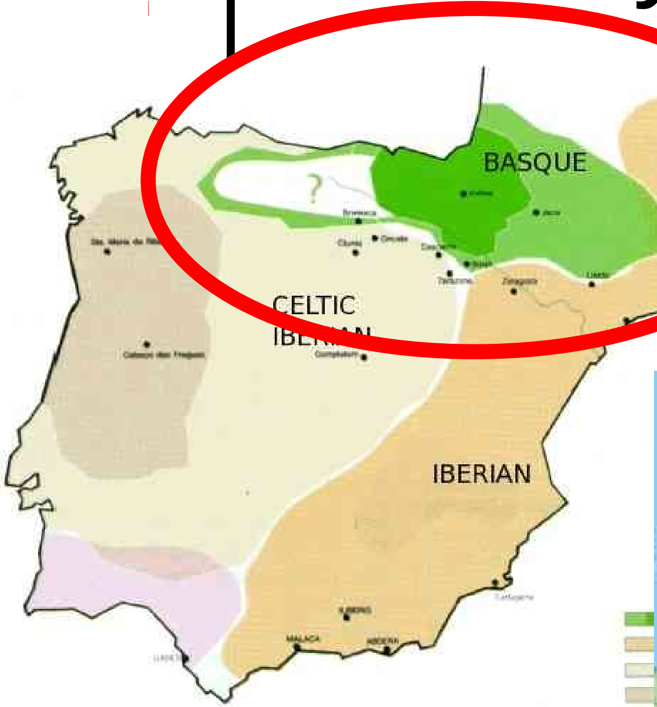
250 languages

Languages with any lexical resource
in Internet

7.000 languages

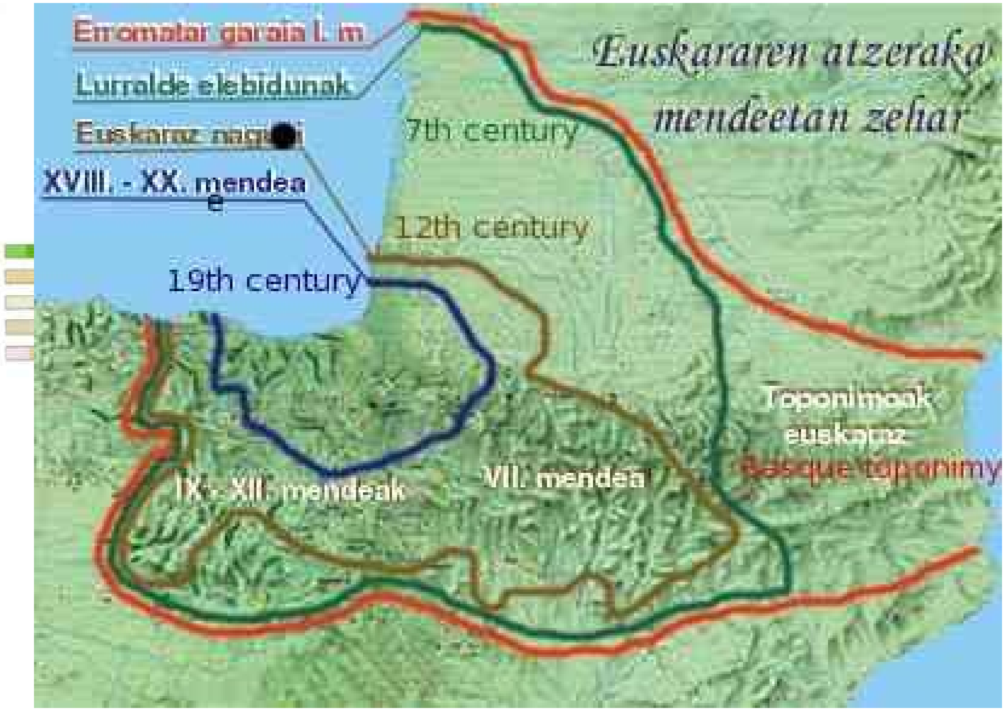
All the world languages

History of Basque

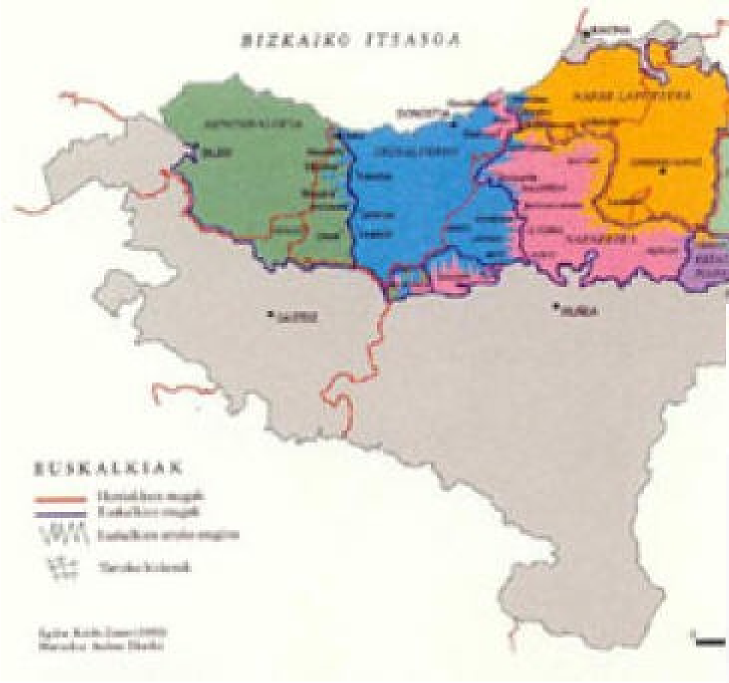


Prerromanic languages in Spain

Basque in 7th, 12th and 19th centuries



Basque nowadays

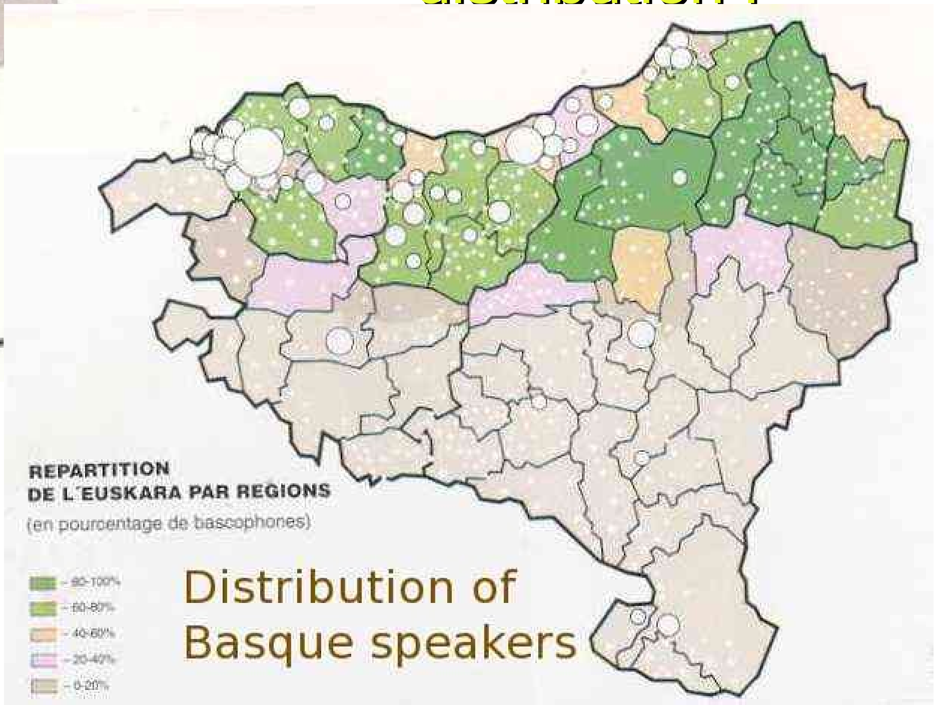


x different dialects !

1,033,900 Speakers

(First lang.: 700,000)

Non homogeneous distribution !



Distribution of Basque speakers

Main reasons of Basque regression for centuries

- No official language
- Out of the education system
- 6 dialects!
- Out of media
- Out of industry

Main reasons of Basque regression for centuries

	But since 1980...
• No official language	Coofficial language
• Out of the education system	Integrated in education (even at university)
• 6 dialects!	Unified Basque (1966)
• Out of media	TV, newspaper...
• Out of industry	Out of new ICTs ???

Basque. Linguistic features: Agglutinative language

Case	Undet.	Det.sing.	Det.Pl.	CloserPl.
Absolutive	*katu*	*katua*	*katuak*	*katuok*
Ergative	*katuk*	*katuak*	*katuek*	*katuok*
Dative	*katuri*	*katuari*	*katuei*	*katuoi*
Genitive1	*katuren*	*katuaren*	*katuen*	*katuon*
Associative	*katurekin*	*katuarekin*	*katuekin*	*katuokin*
...				
...				
...				

~with cat   ***with the cat***   ***with the cats***   ***~with these cats***

14 different cases b

**In fact, at least 360 possible word forms
for every noun or adjective**

**In theory, more than one million word forms
are possible for them**

Basque. Linguistic features:

Case suffixes and free order of components

The dog brought the newspaper in his mouth

Txakur-rak	egunkari-a	aho-an	zekarren.
The-dog	the-newspaper	in-his-mouth	brought
ergative-3-s	absolutive-3-s	inessive-3-s	
Subject	Object	Modifier	Verb

Alternative possible orders:

Txakur-rak	aho-an	egunkari-a	zekarren.
Txakur-rak	aho-an	zekarren	egunkari-a.
Egunkari-a	txakur-rak	zekarren	aho-an.

...

Basque. Linguistic features:

Ergative language & multiple agreement

- Ergative case. Subject of transitive verbs

I am	Ni naiz	(absolutive)
I saw the cat	Nik katua ikusi nuen	(ergative)

- Agreement in number and person between verb and (subject, object and indirect object)

I saw the cat	Nik katua ikusi nuen
I saw the cats	Nik katuak ikusi nituen
I saw you	Nik zu ikusi zintudan

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.

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

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

Underlying strategy

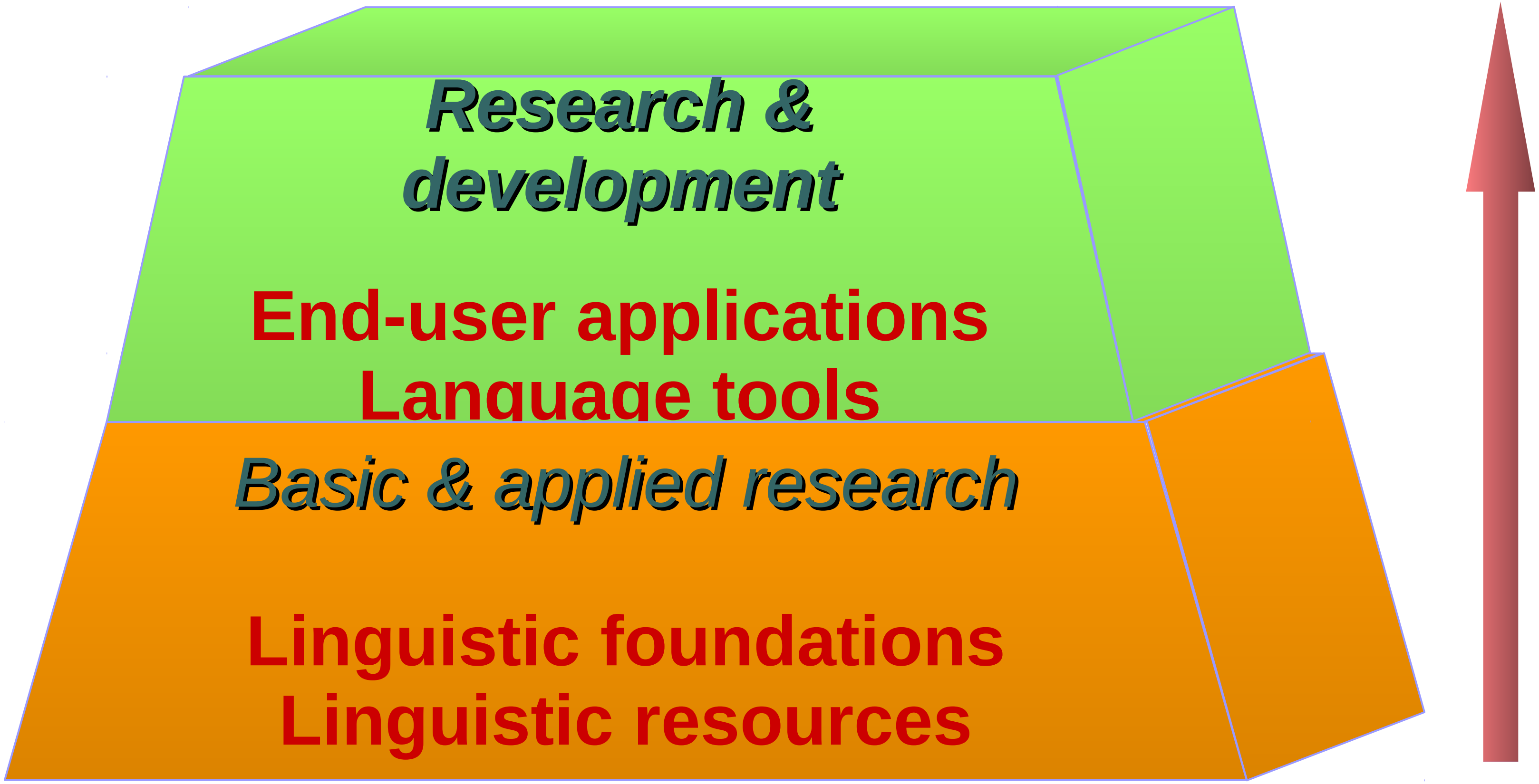
- Need of standardization of resources to be useful:
  - in different researches
  - in different tools
  - in different applications
- Need of incremental design and development of language foundations, tools, and applications
  - in a parallel and coordinated way
  - in order to get the best benefit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



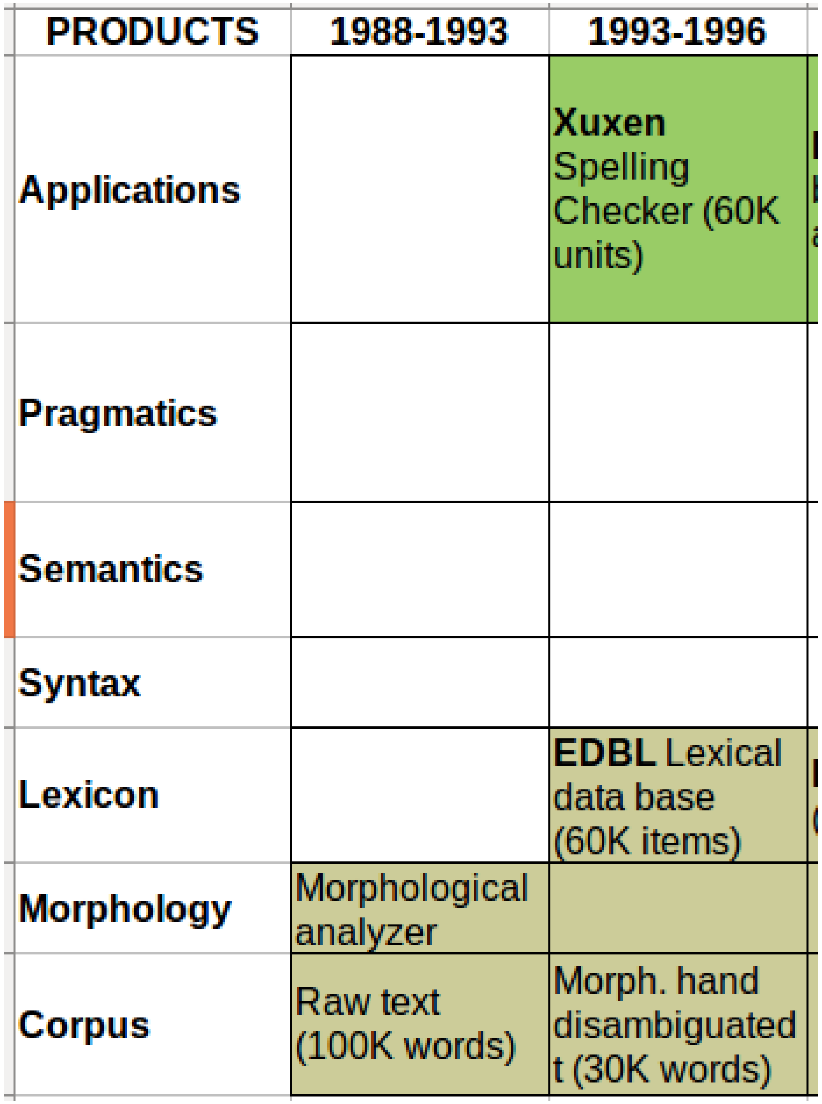
Research & development

End-user applications
Language tools

Basic & applied research

Linguistic foundations
Linguistic resources

Phase I: laying foundations

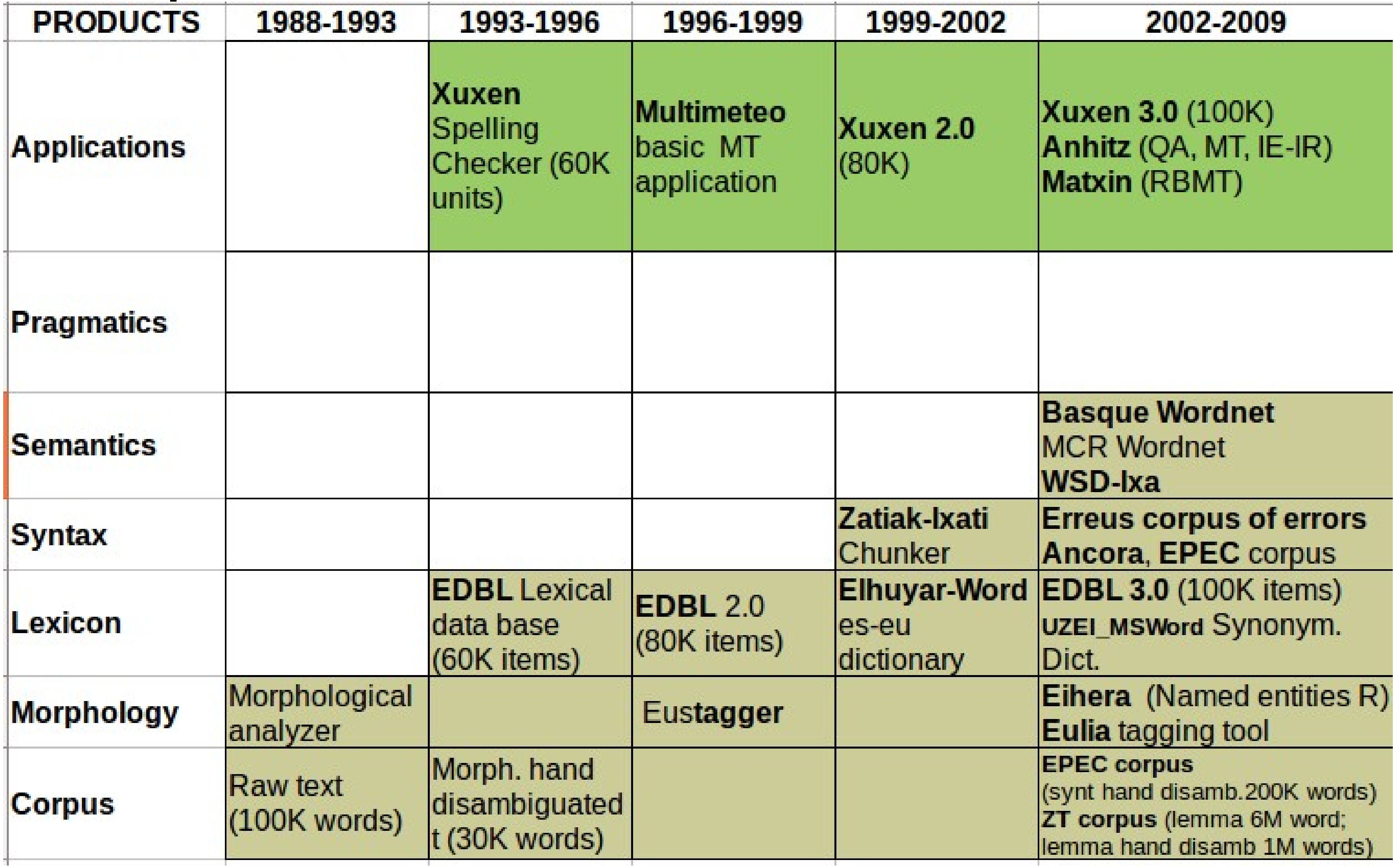


PRODUCTS	1988-1993	1993-1996
Applications		**Xuxen** Spelling Checker (60K units)
Pragmatics		
Semantics		
Syntax		
Lexicon		**EDBL** Lexical data base (60K items)
Morphology	Morphological analyzer	
Corpus	Raw text (100K words)	Morph. hand disambiguated t (30K words)

Phase II:
first basic tools and applications

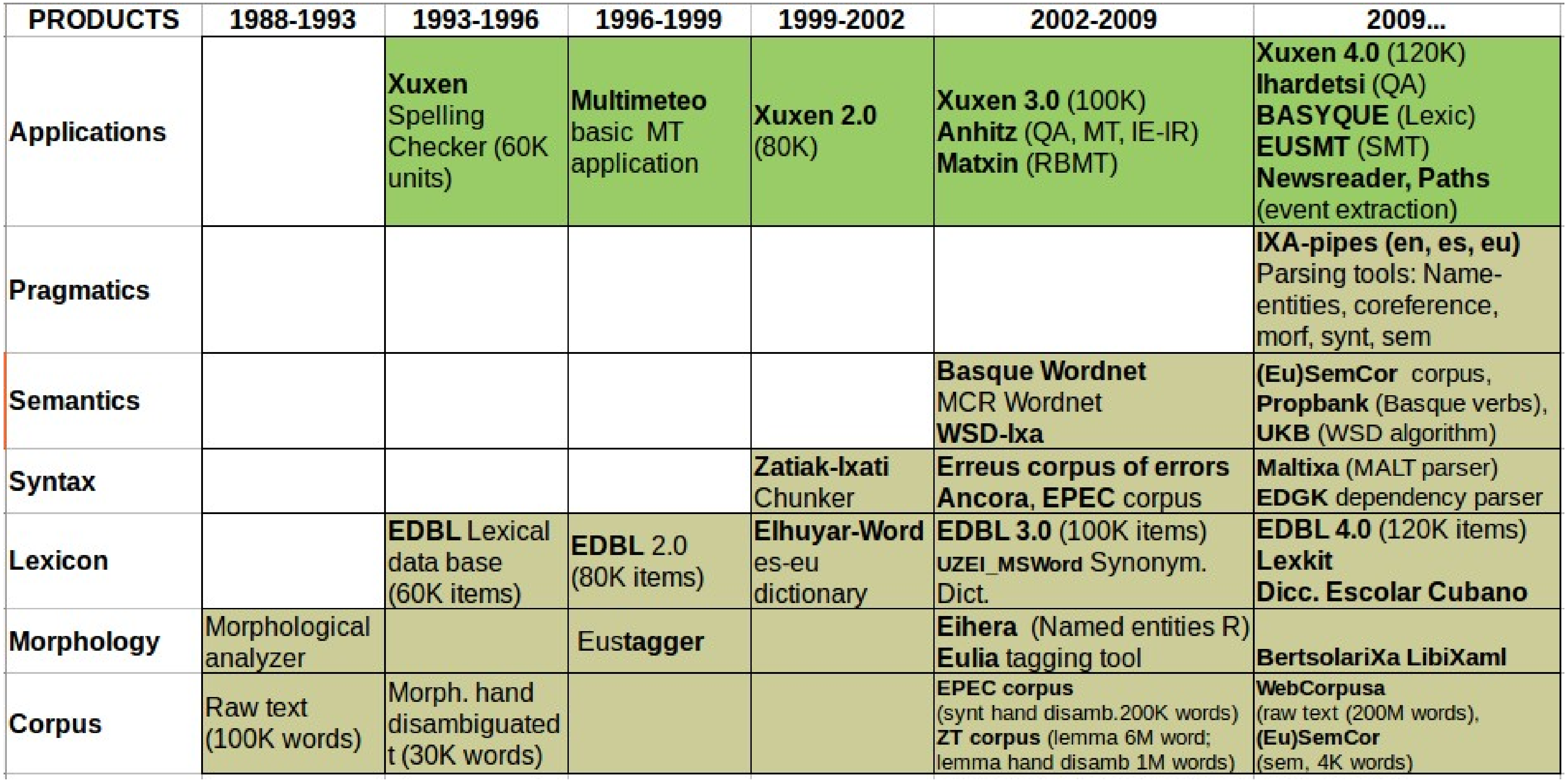
PRODUCTS	**1988-1993**	**1993-1996**	**1996-1999**	**1999-2002**
Applications		**Xuxen** Spelling Checker (60K units)	**Multimeteo** basic MT application	**Xuxen 2.0** (80K)
Pragmatics				
Semantics				
Syntax				**Zatiak-Ixati** Chunker
Lexicon		**EDBL** Lexical data base (60K items)	**EDBL** 2.0 (80K items)	**Elhuyar-Word** es-eu dictionary
Morphology	Morphological analyzer		Eus**tagger**	
Corpus	Raw text (100K words)	Morph. hand disambiguated t (30K words)		

Phase III: more advanced tools and applications



PRODUCTS	1988-1993	1993-1996	1996-1999	1999-2002	2002-2009
Applications		**Xuxen** Spelling Checker (60K units)	**Multimeteo** basic MT application	**Xuxen 2.0** (80K)	**Xuxen 3.0** (100K) **Anhitz** (QA, MT, IE-IR) **Matxin** (RBMT)
Pragmatics					
Semantics					**Basque Wordnet** MCR Wordnet **WSD-Ixa**
Syntax				**Zatiak-Ixati** Chunker	**Erreus corpus of errors** **Ancora**, **EPEC** corpus
Lexicon		**EDBL** Lexical data base (60K items)	**EDBL** 2.0 (80K items)	**Elhuyar-Word** es-eu dictionary	**EDBL 3.0** (100K items) **UZEI_MSWord** Synonym. Dict.
Morphology	Morphological analyzer		Eus**tagger**		**Eihera** (Named entities R) **Eulia** tagging tool
Corpus	Raw text (100K words)	Morph. hand disambiguated t (30K words)			**EPEC corpus** (synt hand disamb.200K words) **ZT corpus** (lemma 6M word; lemma hand disamb 1M words)

Phase IV: multilinguality and general applications

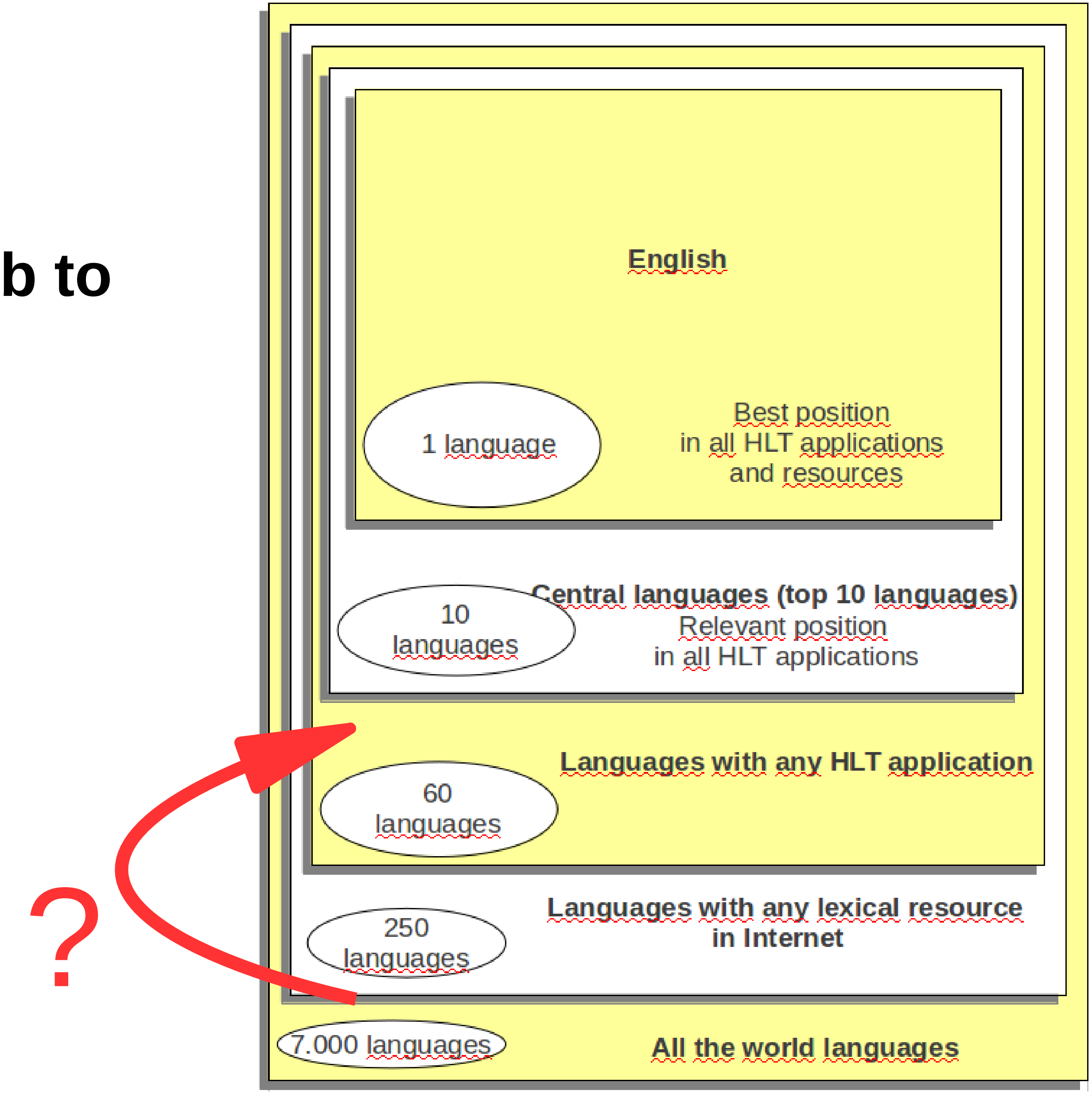


PRODUCTS	1988-1993	1993-1996	1996-1999	1999-2002	2002-2009	2009...
Applications		**Xuxen** Spelling Checker (60K units)	**Multimeteo** basic MT application	**Xuxen 2.0** (80K)	**Xuxen 3.0** (100K) **Anhitz** (QA, MT, IE-IR) **Matxin** (RBMT)	**Xuxen 4.0** (120K) **Ihardetsi** (QA) **BASYQUE** (Lexic) **EUSMT** (SMT) **Newsreader, Paths** (event extraction)
Pragmatics						**IXA-pipes (en, es, eu)** Parsing tools: Name-entities, coreference, morf, synt, sem
Semantics					**Basque Wordnet** MCR Wordnet **WSD-Ixa**	**(Eu)SemCor** corpus, **Propbank** (Basque verbs), **UKB** (WSD algorithm)
Syntax				**Zatiak-Ixati** Chunker	**Erreus corpus of errors** **Ancora, EPEC** corpus	**Maltixa** (MALT parser) **EDGK** dependency parser
Lexicon		**EDBL** Lexical data base (60K items)	**EDBL** 2.0 (80K items)	**Elhuyar-Word** es-eu dictionary	**EDBL 3.0** (100K items) **UZEI_MSWord** Synonym. Dict.	**EDBL 4.0** (120K items) **Lexkit** **Dicc. Escolar Cubano**
Morphology	Morphological analyzer		**Eustagger**		**Eihera** (Named entities R) **Eulia** tagging tool	**BertsolariXa LibiXaml**
Corpus	Raw text (100K words)	Morph. hand disambiguated t (30K words)			**EPEC corpus** (synt hand disamb.200K words) **ZT corpus** (lemma 6M word; lemma hand disamb 1M words)	**WebCorpusa** (raw text (200M words), **(Eu)SemCor** (sem, 4K words)

Can NLP help?

Helping to climb to the next level?

- Basque from 4th level to 5th level? (1988-2009)
- Quechua ? (2012 - ...)



Quechua. Linguistic features

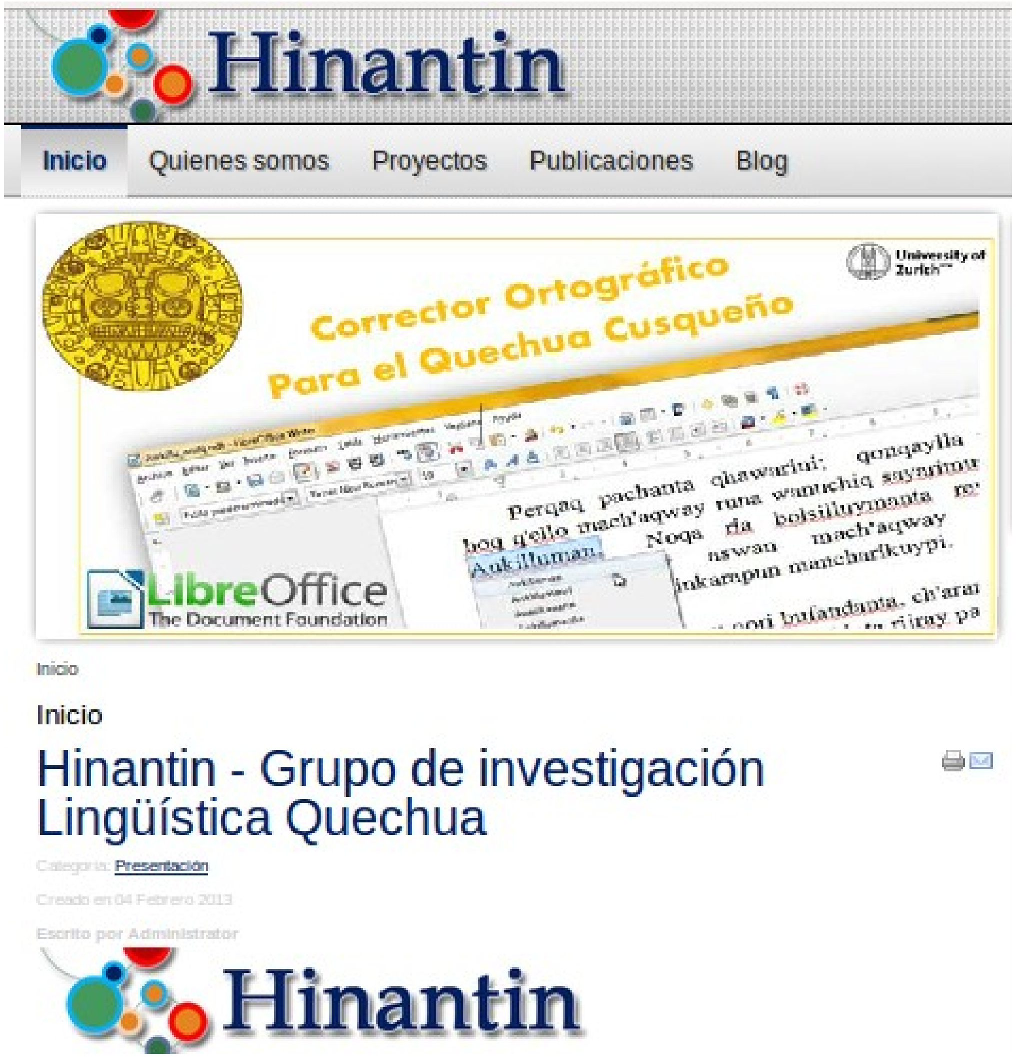
- Aglutinative language. 130 suffixes
- No ergative language no multiple multiple agreement

		But since 2000...
• No official language	→	~Coofficial language
• Out of the education system	→	Small integration in education
• Several dialects	→	Standard (1994) still in discussion
• Out of media	→	
• Out of industry	→	

Hinantin Group

Working for Quechua

Colaborating with:
- Univ. of Zurich
- Univ. Of Basque Country (Ixa)



Hinantin

Inicio | Quienes somos | Proyectos | Publicaciones | Blog

Inicio

Inicio

Hinantin - Grupo de investigación Lingüística Quechua

Categoría: Presentación

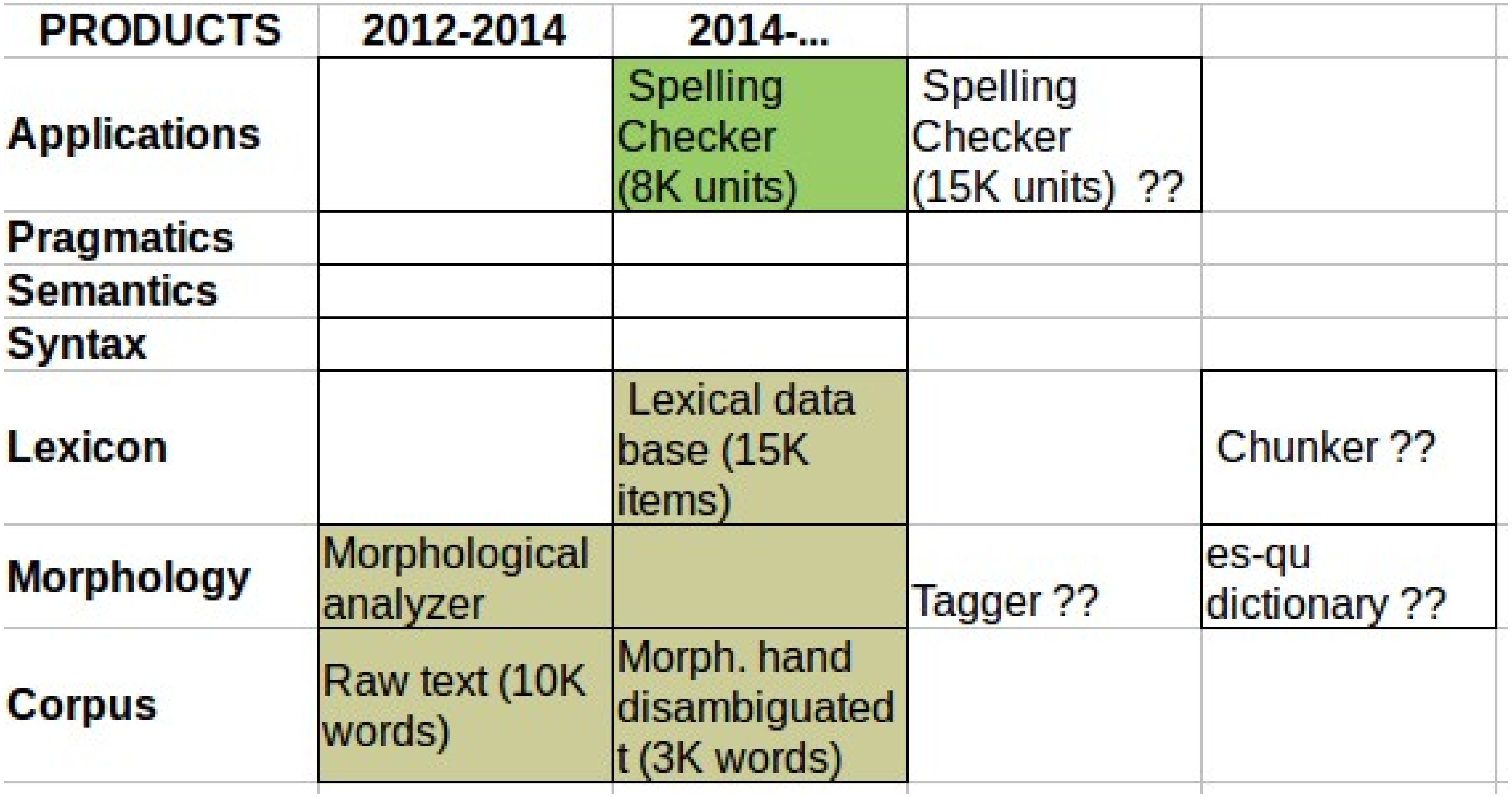
Creado en 04 Febrero 2013

Escrito por Administrator

Hinantin

Hinantin Group

Working for Quechua



PRODUCTS	2012-2014	2014-...		
Applications		Spelling Checker (8K units)	Spelling Checker (15K units) ??	
Pragmatics				
Semantics				
Syntax				
Lexicon		Lexical data base (15K items)		Chunker ??
Morphology	Morphological analyzer		Tagger ??	es-qu dictionary ??
Corpus	Raw text (10K words)	Morph. hand disambiguated t (3K words)		

Can NLP help to languages in the 5th and 6th levels?

- Fifth level:
  Languages that only have writing systems
  - **Other 2,014 languages** are included here

- Sixth level:
  Only-spoken languages
  - At least **other 4,500 languages**

Outline

- How are languages facing the ICT and HLT challenges?
- Which languages are "less resourced"? Six different levels
- Can NLP help less resourced languages to promote their use?
- **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)" (Krauwer, 2003)
    - Joint initiative between ELSNET and ELRA in1998.
    - 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 equiped with a parser and a multi-million-word corpus of text*".
  - He analyzes the relation among the sociology of language and HLT, and guises 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
- SALTMIL ("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 a 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.
- We colaborate with Hinantin group in creating LT for Quechuan