Language Technology for Language Communities: An Overview based on Our Experience

Ixa group (Iñaki Alegria, Kepa Sarasola)

CinC 2017, Alcanena-Portugal

http://ixa.eus

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Ixa group

- **Almost 30 year** working on Language Technology
- **Basque**-centred research group but also other languages
- **Multidisciplinary**: computing, linguistics...
- **Text**-based resources and apps (speech in collaboration)
- **3 levels: r**esources, basic tools, applications
- **Local ↔ Global**
- Basque **community** ↔ International research community
- **Collaboration**: Basque academy, lexicography centre, publication
- **Alternative forums**: Basque Summer University, NGOs...

People Members

Created products

Education Masters

Master Ofiziala
Official Master'sDegree

Hizkuntzaren Azterketa eta Prozesamendua (HAP)

Language Analysis and Processing (LAP)

http://ixa.si.ehu.es/master

Language Technology Applications

Information Retrieval, Information Extraction and Question Answering

Papers; Projects: Kyoto, paths, Lcloud, opener, skater and Know2; Demo: Ihardetsi (QA system)

Machine Translation

Papers; Project: OpenMT-2, Takardi, qtleap; Demo: Opentrad-Matxin (Spanish to Basque MT system)

Language learning

Papers; Project: Irakazi

Linguistic processors

Morphology

Papers; Project: BER2TEK; Demos: Morfeus, Eustagger

Syntax-Morphosyntax

Papers; Project: BER2TEK; Demos: Zatiak (chunker), Maltixa (statistical parser)

Lexicography-Semantics

Papers; Project: Kyoto and Know2; Demos: Know2's demos, Eihera (name entities)

Linguistic Resources

Corpus

Papers; Project: Lexikoaren behatokia ; Demos: ZT, Ancora-EPEC , EuSemcor

Dictionaries

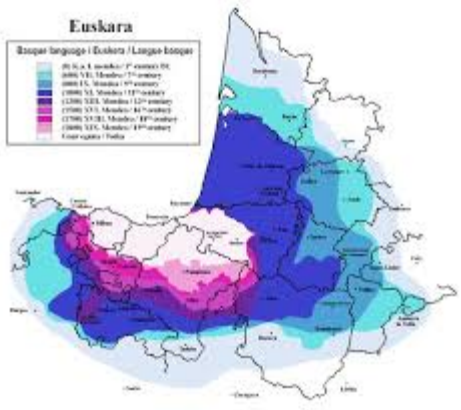
Papers; Project: BER2TEK; Demos: EDBL (lexical database), Xuxen (spelling checker)

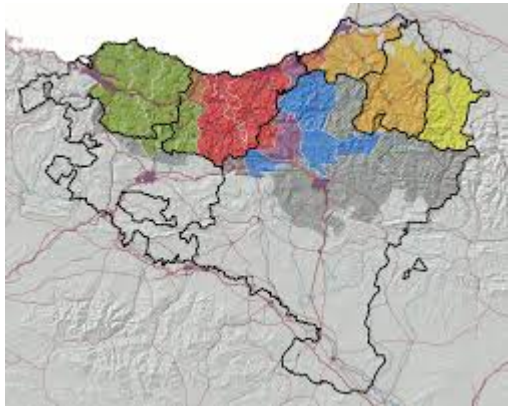
Ontologies

Papers; Project: Kyoto, Know2 and WNTERM; Demo: Basque Wordnet

Basque

- Old language (pre-Indo-European), rich dialects
- Near to 800.000 speakers (25-30% of people)
- Standardization in 1968 but rich dialects
- New political rule a bit later
- Basque schools: Ikastolak
- Declining on France side (non official)
- No monolingual speakers





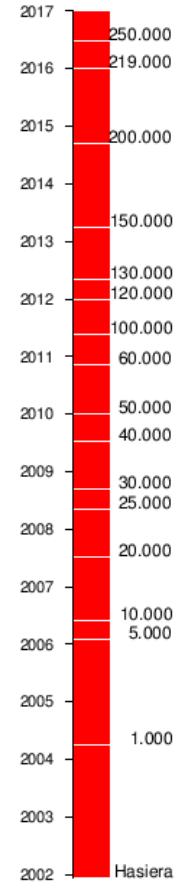
ixa

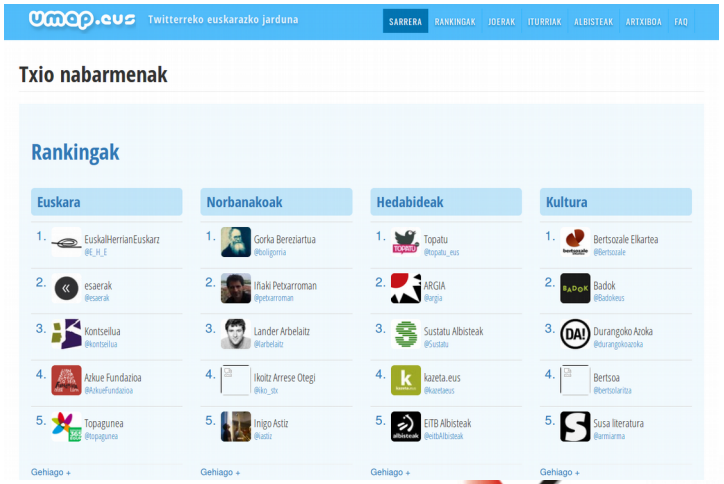
Universidad del País Vasco
Euskal Herriko Unibertsitatea

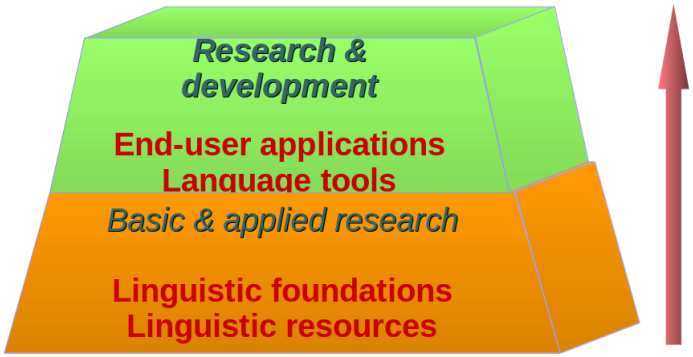
Basic matters

Looking to the texts...

- Standardization (1968)
- (Digital) Contents (school books, small dictionaries…)
  - Readers: School (Ikastolak) → University
- Open/Free software / open contents
- Wikimedia / Wikipedia
- Digital community
- Need of incremental design and development of language foundations, tools, and applications

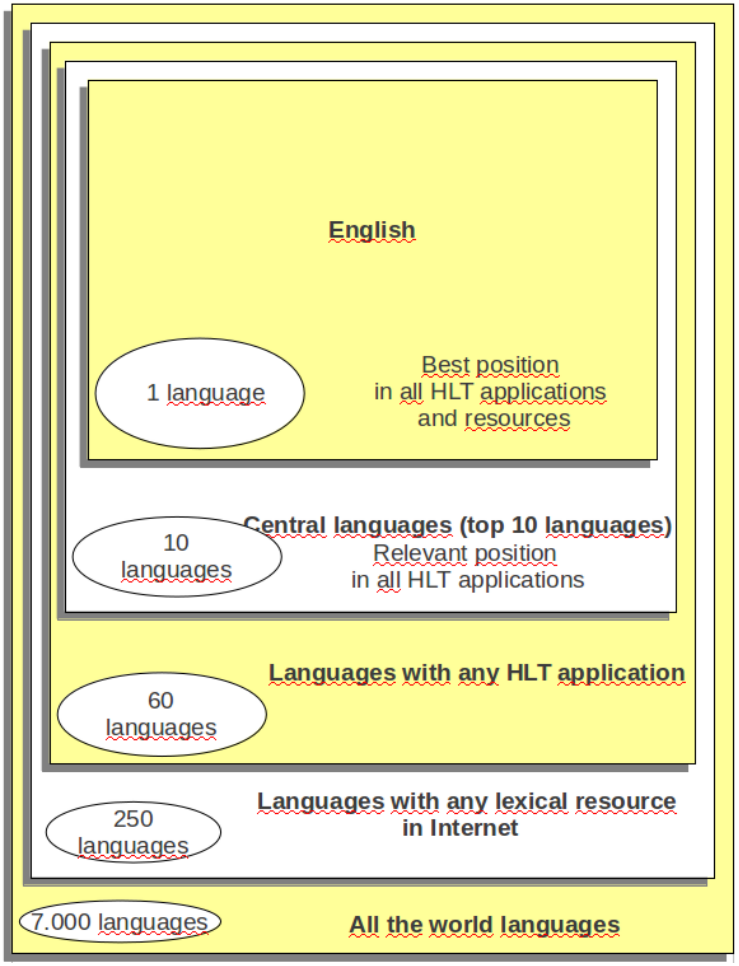






Resources and BLARK

- *Basic LAnguage Resource Kit* (BLARK) *Krauwer (2003)*
- Typology based on (digital) resources
- Associations:



mercator
European Research Centre on Multilingualism and Language Learning

SOAS University of London
100 Years
World Languages Institute

Foundation for Endangered Languages

SaLT
Speech And Language Technology for Minority Languages

ELRA
ASSOCIATION EUROPEAN LANGUAGE RESOURCES
elda

ixa

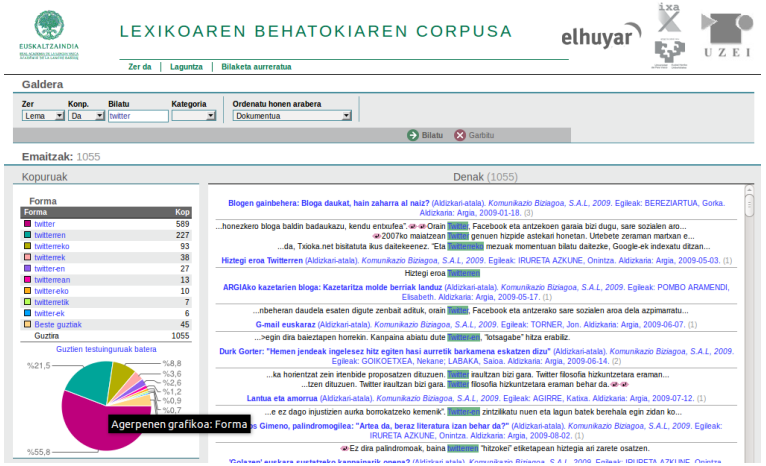
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Basic resources

- Corpora (digital texts)
- Dictionary (better a digital one)
- Normative grammar (even in paper)

Corpora (digital texts)

- Collecting corpora is not easy (“digital only for paper” → **error**)
- Sources: publishers, schools and **Wikipedia**
  - Alternative way: “*web as a corpus*” techniques or OCR
- Problems: copyrights and difficult formats (pdf, word...)
- Use: data for text mining and for **evaluation**
- An initial (small) digital corpus is a key start point
- Processes: enriching the dictionary, creating the spelling checker, learning language models...



CinC 2017

Universidad del País Vasco
Euskal Herriko Unibertsitatea

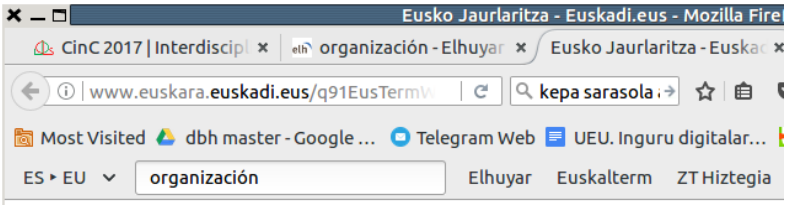
Basic tools and applications

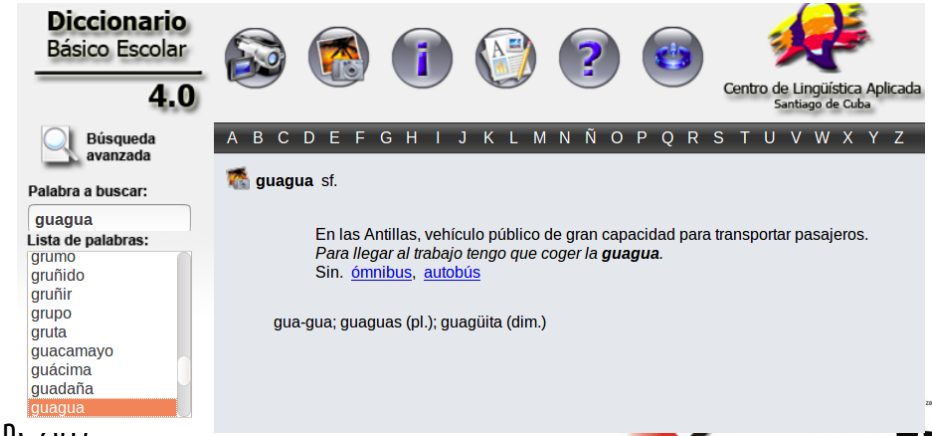
- On-line dictionary → Games
- Morphology → Spelling corrector
- Lemmatizer/POS_tagger → Search engine
- Machine translator or Normalizator
  (easier for close languages)

Dictionaries (mono- or bilingual)

- Basic tool for students, journalists and writers
- Historical evolution: Paper cards → Word Proc. → XML/TEI
- XML/TEI → **Multimedia**: DVD, Web/phone, paper
  - Unique maintenance → 3 products
- Integration: *Euskalbar* (browser)
- Some projects:

  Garabide NGO (Nahuatl) and Cuba (*DBE*).
  Scrable in Basque





Universidad del País Vasco
Euskal Herriko Unibertsitatea

Morphology / Spelling corrector

- Computational morphology is compulsory for most of the languages:
  - Dictionary + word-grammar
- The spelling corrector is a key application (only with big soft companies??)
  - Basic tool for students, journalists and writers
  - Key for standardization
- Integration/online: Microsoft, LbreOffice, Mozilla, Android…
- Basic tools:
  - *foma* and *hunspell* (free software)
- Projects: unified Basque, dialectal Basque, Quechua (Univ. Zurich and Cusco)

ixa

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Lemmatizer / Search engine

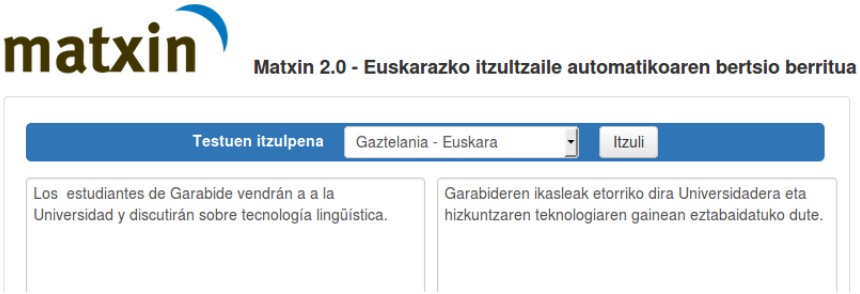
- Stemming → Lemmatizer (morphology) → POS tagging (learning)
  word → stem/lemma(root) → lemma in context
  juego → *jugar*(V)/*juego*(N) → *jugar*(V)
- used for information extraction
  + language identifier → Search engine
- A manually annotated corpus is needed to create a POS tagger
- Powerful tool for Information Retrieval and Information Extraction
- Some projects for Basque:

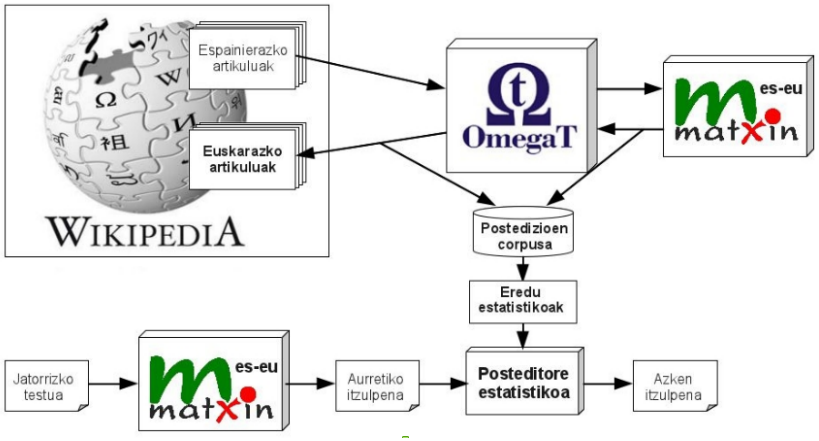
Cin

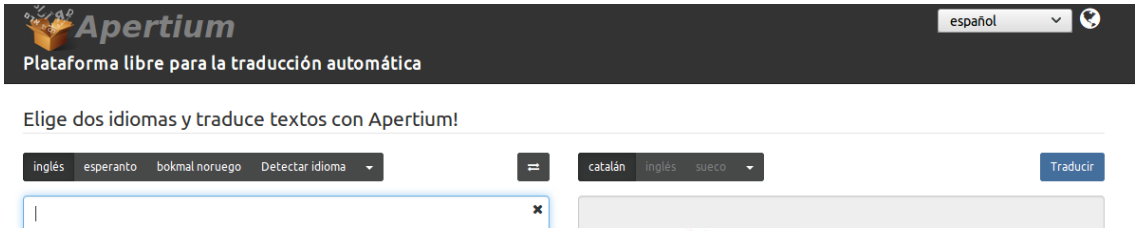
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Machine Translation

- It is not an easy-to-build tool (corpus, morphology...)
- Rule-based / data-driven (translation memories)
  - Nice performance for close languages
- Normalizer: translating/mapping dialects or variants (word-by-word)
- Possible improvements based on community crowdsourcing







ixa

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Discussion

- Sustainability / Cost:
  - Fast development could be expensive
  - Planning is a need for sustainability
- Integration on commercial software / open software

  Some popular applications are proprietary and

  the decision to add new languages depends on the company

  → interesting experience with Basque spelling checker.
  Microsoft become more interested ...
  after the localization and integration on LibreOffice
- Standardization/dialect/language/alphabet...

  Each community has to decide how to do it ...

  but standardization is crucial for text processing!

References

http://ixa.eus/argitalpenak (IXA publications, mainly in Basque and in English)

- Alegria, I., Artola, X., De Ilarraza, A. D., & Sarasola, K. (2011). Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque. Proceedings of 5th Language & Technology Conference: HLT as a Challenge for Computer Science and Linguistics. pp: 42-46, November 24-27, 2011, Poznan.
- Borin, L.(2009). Linguistic diversity in the information society. SALTMIL2009 Workshop: IR-IE-LRL. Information Retrieval and Information Extraction for Less Resourced Languages. University of theBasque Country.
- Forcada, M. (2006). Open source machine translation: an opportunity for minor languages. In Proc. of the Workshop Strategies for developing machine trans-lation for minority languages, LREC (Vol. 6, pp. 1-6).
- Krauwer, Steven. "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap." Proceedings of SPECOM 2003 (2003): 8-15.
- Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (Vol. 4, pp. 5-15).