

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

Ixa group (Iñaki Alegria, Kepa Sarasola)

CinC 2017, Alcanena-Portugal

<http://ixa.eus>

# 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**: resources, basic tools, applications
- **Local ↔ Global**
- Basque **community** ↔ International research community
- **Collaboration**: Basque academy, lexicography centre, publication
- **Alternative forums**: Basque Summer University, NGOs...

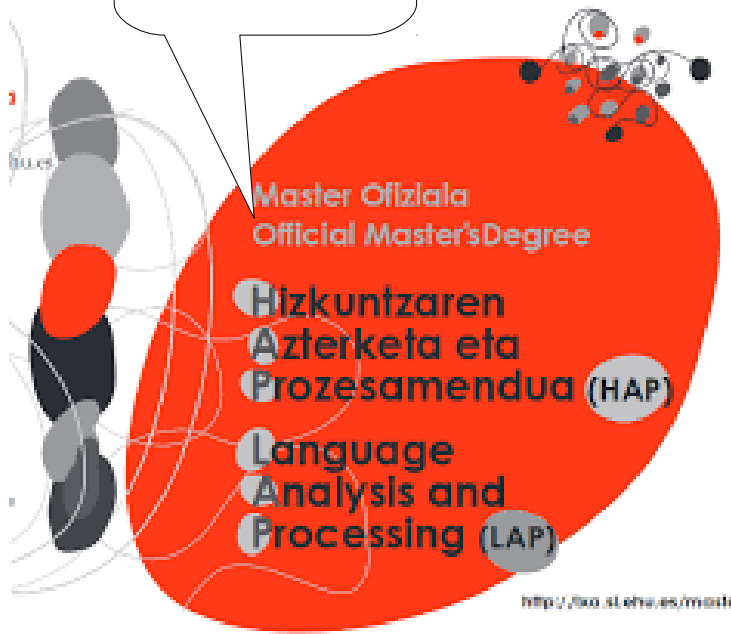


**People  
Members**



**Created  
products**

**Education  
Masters**



Master Ofiziala  
Official Master's Degree

Hizkuntzaren  
Azterketa eta  
Prozesamendua (HAP)

Language  
Analysis and  
Processing (LAP)

<http://ixa.sleha.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*

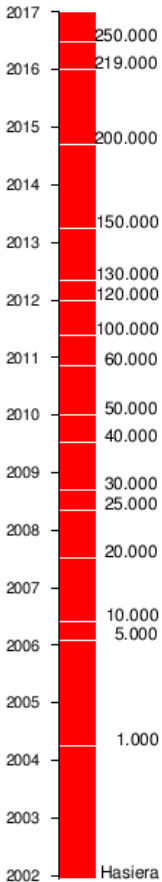




# Basic matters

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



**Rankingak**

Euskara	Norbanakoak	Hedabideak	Kultura
1. EuskaHerriaEuskarz @E_H_E	1. Sorio Bereziartua @Bergina	1. Topatu @topatu_eus	1. Bertsozale Elkarte @bertsozale
2. esarak @esarak	2. Itaki Petxarroman @petxaroman	2. REGIA @regia	2. Badok @badok
3. Kontseilua @kontseilua	3. Lander Arbelaitz @arbelaitz	3. Sustatu Albisteak @sustatu	3. DAL Durangoko Azoka @durangokoazoka
4. Azkue Fundazioa @azkuefundazioa	4. Holtz Arrese Olegi @hltz_olegi	4. kazeta.eus @kazetaeus	4. Bertsoa @bertsolanta
5. Topagunea @topagunea	5. Imigo Asti @imigo_asti	5. EITB Albisteak @eitbalbisteak	5. Susa literatura @susa



C 2017

# Resources and BLARK

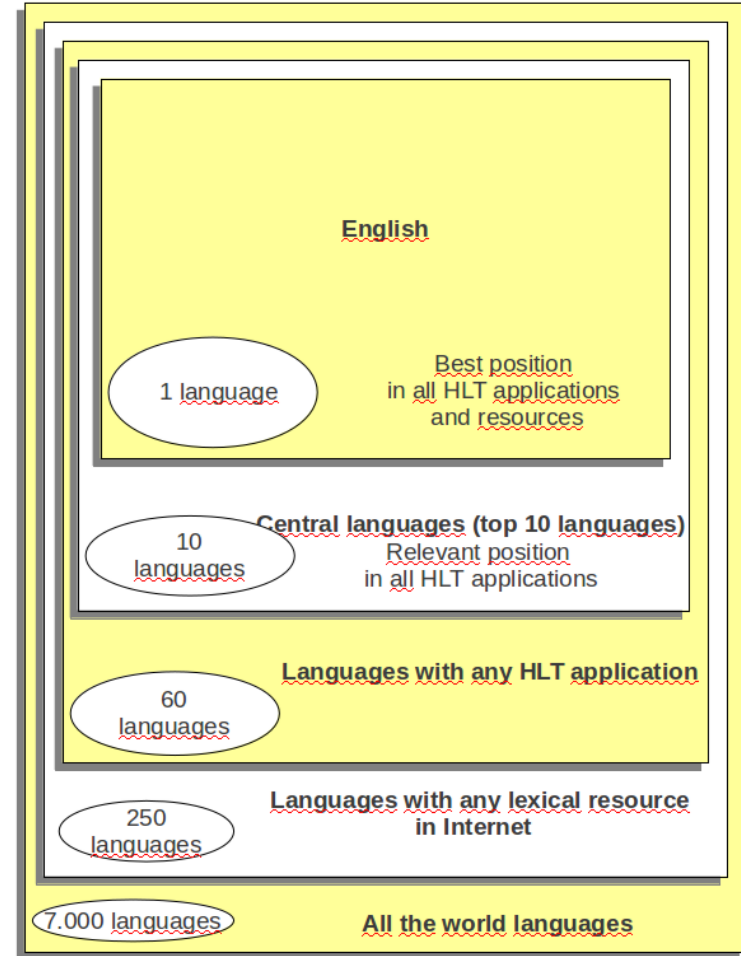
- *Basic LAnguage Resource Kit (BLARK) Krauwer (2003)*
- Typology based on (digital) resources
- Associations:



World Languages Institute



Foundation for Endangered Languages



# 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...



**LEXIKOAREN BEHATOKIAREN CORPUSA**

Zer da | Laguntza | Bilaketa aurreraba

Galdera

Zer | Komp. | Bilatu | Kategoría | Ordenatu honen arabera

Lema | Da | twitter | Dokumentua

Bilatu | Gabatu

Emailtzak: 1055

Kopuruak

Forma	Kop
twitter	589
littereris	227
tuttereko	93
tutterrek	38
tutter-en	27
tutterman	13
tuttereko	10
tuttererik	7
tutterek	6
Beste guztiak	45
Guztia	1055

Guztien testuinguruak batan

Forma	%
twitter	55,8
littereris	21,5
tuttereko	8,8
tutterrek	3,6
tutter-en	2,6
tutterman	1,2
tuttereko	0,9
tuttererik	0,7
tutterek	0,7
Beste guztiak	0,7
Guztia	0,7

Agerpenen grafikoa: Forma

Denak (1055)

Blogen gainbehera: Bloga daukat, hain zaharra al nait? (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: BEREZIARTUA, Gona, Aldizkaria: Argia, 2009-01-18. (3)

...honezkeri bloga baldin badaukazu, kendu entsefuz? (Aldizkari-atala) Facebook eta antzekoen garaia bizi duguz, sare sozialen aro... (2007ko maiatzan) gureen hizpide atsekari honetan. Unebete zientzian mautan e... ..da. Tuokla net bidaatuta kus daitelkeenez. 'Ela (Aldizkari-atala) mezuak momentan bilatu daitzake. Google-ek inderatu ditzan...

Hitzegi erroa Twitterren (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: IRURETA AZKUNE, Oiniza, Aldizkaria: Argia, 2009-05-03. (1)

Hitzegi erroa (Aldizkari-atala)

ARGIAko kazetarien blog: Kazetaritza modde berriak landuz (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: POMBO ARAMENDI, Elisabet, Aldizkaria: Argia, 2009-05-17. (1)

...niberhan daukela esaten digute zenbat aduk, orain (Aldizkari-atala) Facebook eta antzeko sare sozialen aroa dela aspirmatu...

G-mail euskaraz (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: TORNER, Jon, Aldizkaria: Argia, 2009-06-07. (1)

...egin dira baiztapan horekin. Kanpaina abiatu dute (Aldizkari-atala) 'lotsagabe' hitza erabiz.

Durk Gorter: "Hemen jendeak ingelesez hitz egiten hasi aurretik barkamena eskatzen dizut" (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: CORRETTEA, Nekane, LARAKA, Saira, Aldizkaria: Argia, 2009-05-14. (1)

...ka horietzat zen itenbide proposatzen dituzten. (Aldizkari-atala) iraultzan bizi gara. Twitter filosofia hizkuntzetara eraman... ..izen dituzten. Twitter iraultzan bizi gara. (Aldizkari-atala) filosofia hizkuntzetara eraman behar da. (Aldizkari-atala)

Lantua eta amorua (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: AGIRRE, Kaitia, Aldizkaria: Argia, 2009-07-12. (1)

... ez dago nigitzen auzia bonasteko herenak" (Aldizkari-atala) zirkulazio ruon eta lagun batek berehala egin zidan ko...

Gimeno, palindromogilea: "Artea, beraz literatura izan behar da?" (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: IRURETA AZKUNE, Oiniza, Aldizkaria: Argia, 2009-08-02. (1)

... Ez dira palindromak, baina (Aldizkari-atala) 'hizokeri' esketapean hitzega arri zante osatzen.

"Hizokeri" artea esateko berriz ere (Aldizkari-atala) Komunikazio Blogopoa, S.A.I., 2009. Egileak: IRURETA AZKUNE, Oiniza, Aldizkaria: Argia, 2009-08-02. (1)



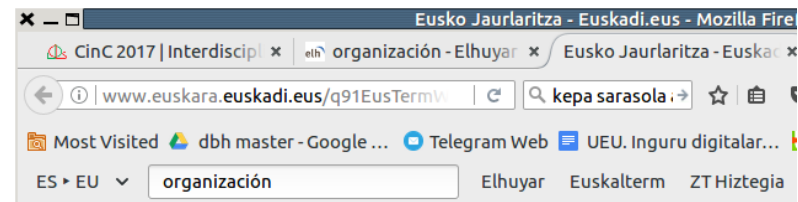
# Basic tools and applications

- 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*).  
Scrabble in Basque



# 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, LibreOffice, Mozilla, Android...
- Basic tools:
  - *foma* and *hunspell* (free software)
- Projects: unified Basque, dialectal Basque, Quechua (Univ. Zurich and Cusco)

The screenshot shows the Xuxen website interface. At the top, there is a logo for 'xuxen' and the text 'Xuxen zuzentzaile'. Below the logo is a navigation bar with links: 'Informazio orokorra', 'Bertsioak', 'FAQ', 'Kontaktua', and 'Sartu'. A dropdown menu is open, listing options: 'lelhoa', 'leholia', 'legioa', 'lehia', 'Gehitu hiztegia', 'Desein', 'Ebak', 'Kopiatu', 'Itzatsi', 'Egabatu', 'Hautatu dena', 'Add to Search Bar...', 'Eropietateak', 'Zuzendu ortografia Hizuntzak', and 'Euskarbar'. The main content area contains text about the Xuxen website's purpose as an online orthographic correction tool.

The screenshot shows the 'GEHIGARRIAK' website statistics page for Xuxen. The page title is 'Honen estatistikak: Xuxen'. It features a line chart titled 'Downloads and Daily Users, last 30 days'. The chart shows two data series: 'Daily Users' (blue line) and 'Downloads' (red line). Below the chart, there are summary statistics: '134.718 Downloads' and '496 in last 30 days' for Downloads; and '3.323 Average Daily Users' and '3,849 average in last 30 days' for Daily Users. The page also includes a search bar and navigation links like 'Hiztegiak', 'Xuxen', and 'Estatistikak'.

# Lemmatizer / Search engine

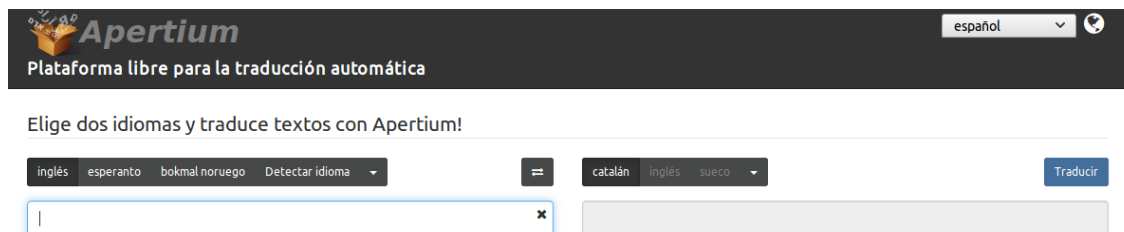
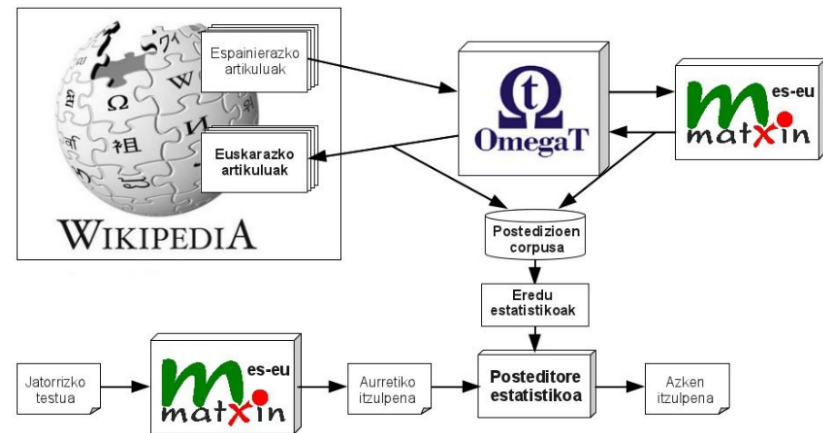
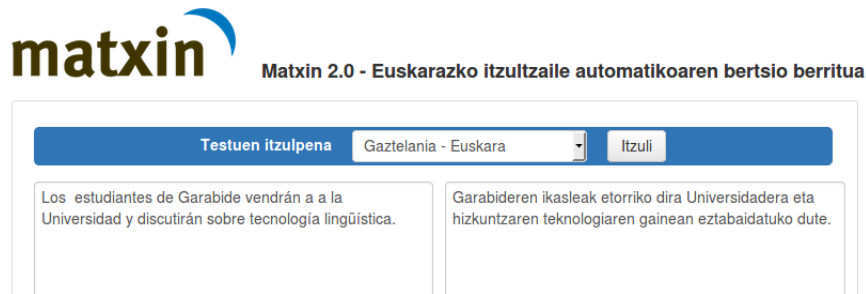
- Stemming → Lemmatizer (morphology) → POS tagging (learning)  
word → stem/lemma(root) → lemma in context  
*juego* → *jugar(V)/juego(N)* → *jugar(V)*
- used for information extraction  
+ language identifier → Search engine
- A manually annotated corpus is needed to create a POS tagger
- Powerful tool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



Cin

# 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



# 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. SALT MIL2009 Workshop: IR-IE-LRL. Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country.
- Forcada, M. (2006). Open source machine translation: an opportunity for minor languages. In Proc. of the Workshop Strategies for developing machine translation 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).