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

Ixa Group (Iñaki Alegria, Kepa Sarasola)

Haskòlinn Í Reykjavik

www.ixa.eus

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

ixa

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Scheme

- To explain **our experience** on Language Technology for Basque

  Perhaps some ideas **can be useful for Icelandic,**
  as Icelandic is a reference for us in the Basque Country

- **Sharing and discussing** these ideas with you

- **Technology** may be useful for language development but…

  A **core work** should be implemented before (or in parallel):

  - **Standardization**
  - **Digital Open Contents** and
  - **Open Source Software**

Scheme

- Introduction
  - "Ixa" research goup
  - Basque language
  - Basic concepts, basic strategies
  - Languages with scarce resources
- Basic tools and their applications

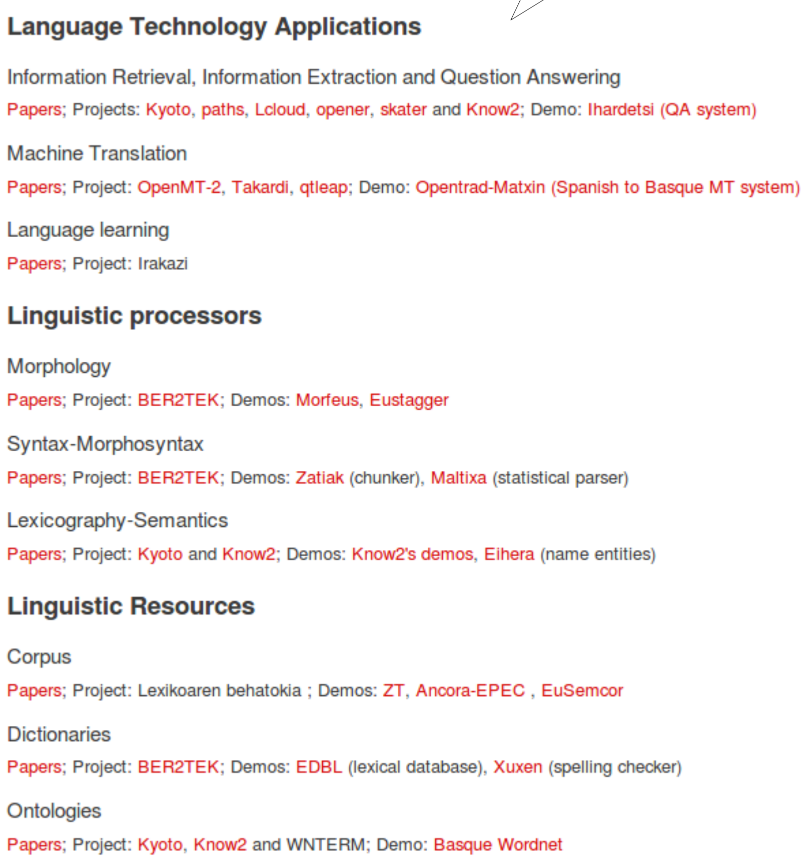
  Corpus, dictionaries, morphology, lemmatizer
- Language models.

  Good news for less-resourced languages (2020)
- Spanish and Basque plans for Language Technology
- Other associations working to promote the use of Basque
- Discussion

People

ixa

Results



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

Master

Master Ofiziala
Official Master'sDegree

Hizkuntzaren Azterketa eta Prozesamendua (HAP)

Language Analysis and Processing (LAP)

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

Ixa group (1988) - HiTZ Center (2020)

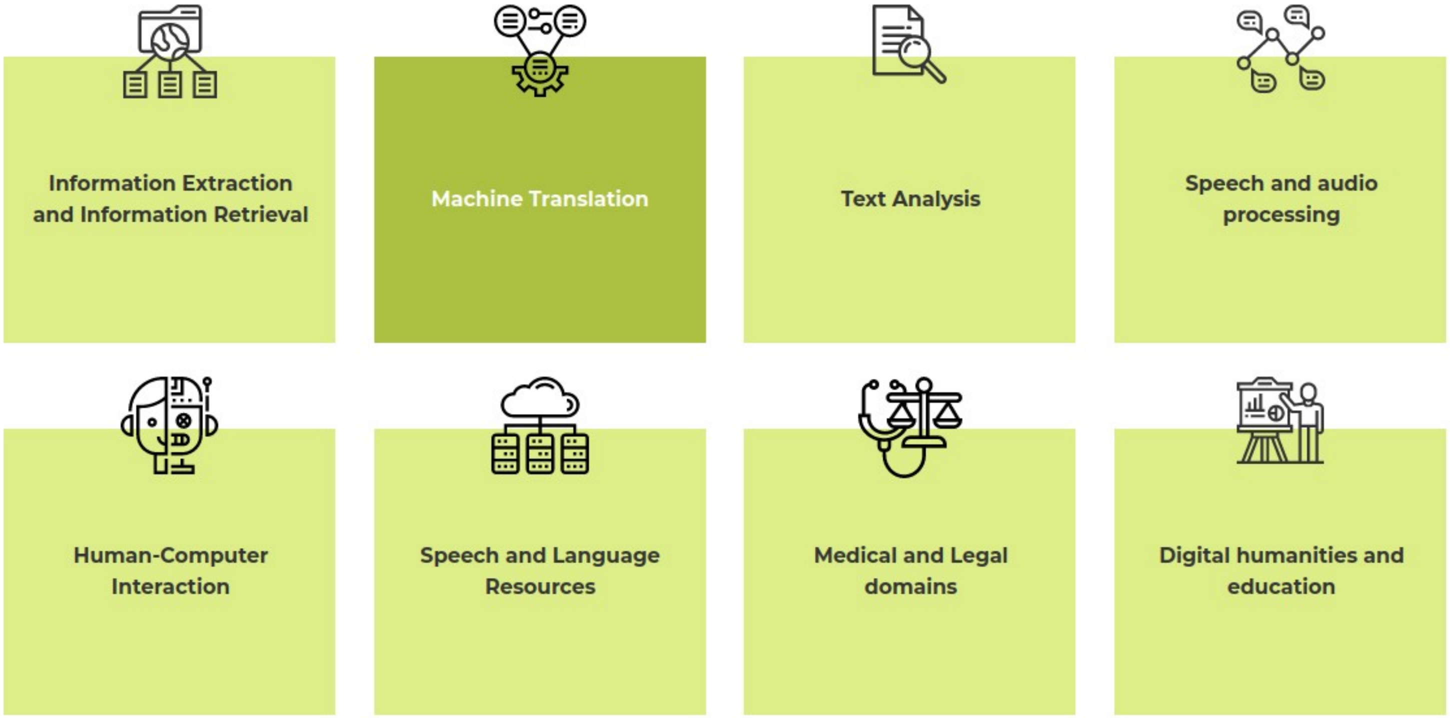
- **ixa.eus**
- **35 year** working on Language Technology
- **Basque**-centred research group… but also other languages
- **Multidisciplinary**: computing, linguistics...
- **Text**-based resources and tools (speech in collaboration)
- **3 levels:** resources, basic tools, applications
- **Local** and **Global**

  Basque **community** and International research **community**
- **Collaboration**: Basque academy, Governments, lexicography centres, publishers, schools...
- **Alternative forums**:

  Basque Summer University, Wikipedia, NGOs...

Hitz research center (hitz.eus)

- *Ixa* (text) and *Aholab* (speech) groups works together facing to new challenges (AI) collaboration with government, companies and associations



Information Extraction and Information Retrieval

Machine Translation

Text Analysis

Speech and audio processing

Human-Computer Interaction

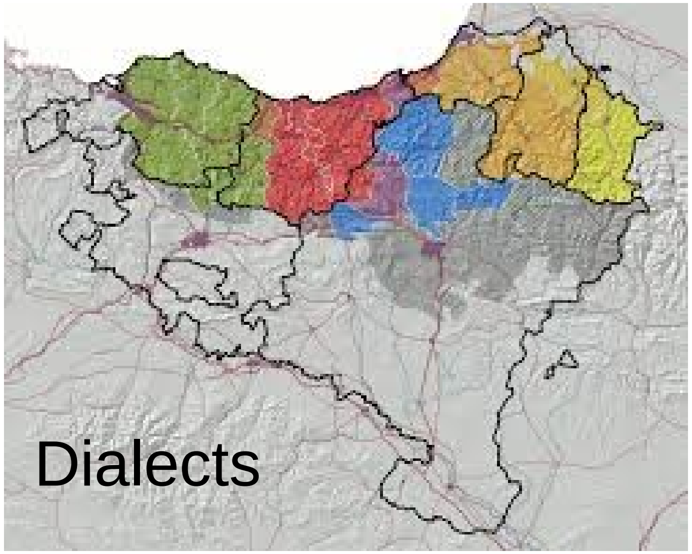
Speech and Language Resources

Medical and Legal domains

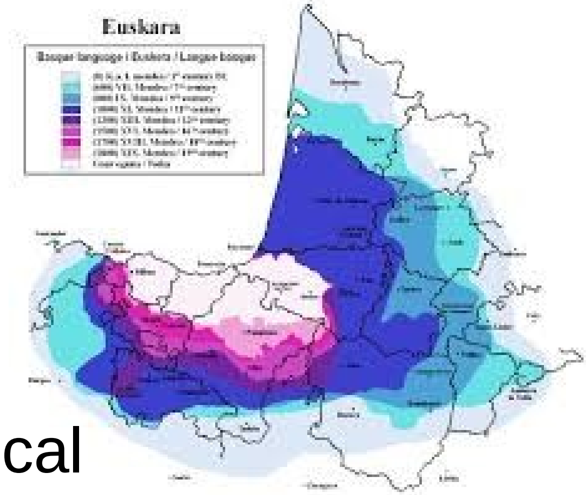
Digital humanities and education

Basque language

Basque Country



Dialects



Historical regression

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

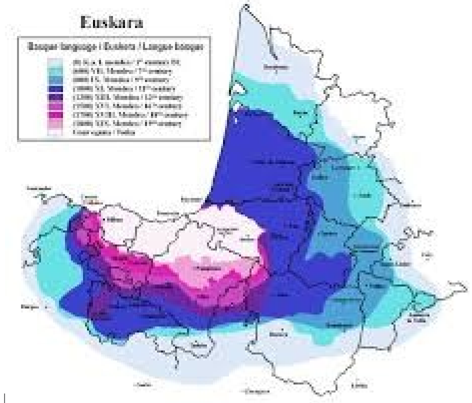
ixa

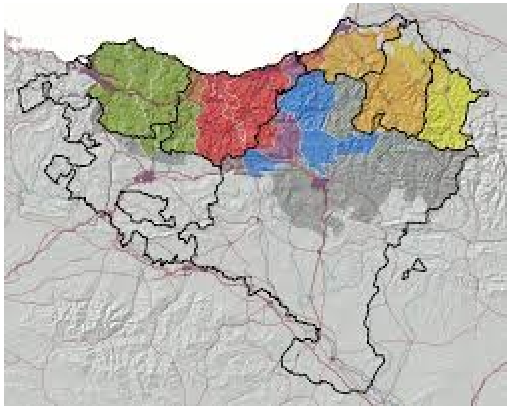
eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Basque language

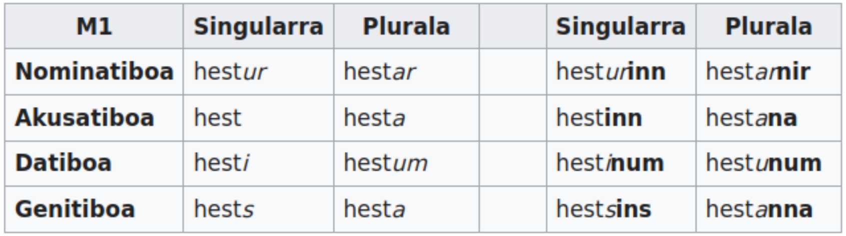
- Old language (pre-Indo-European)
- Near to 1 million speakers (~35%)
- Revitalization (1960-2023):
  - 1960: Basque schools (immersion)
  - 1968: Standard Basque
  - 1980: New political rule (before it was forbidden)
- Still declining on France side (non official status)
- Newspaper, TV channel, some radio channels and magazines
- No monolingual speakers





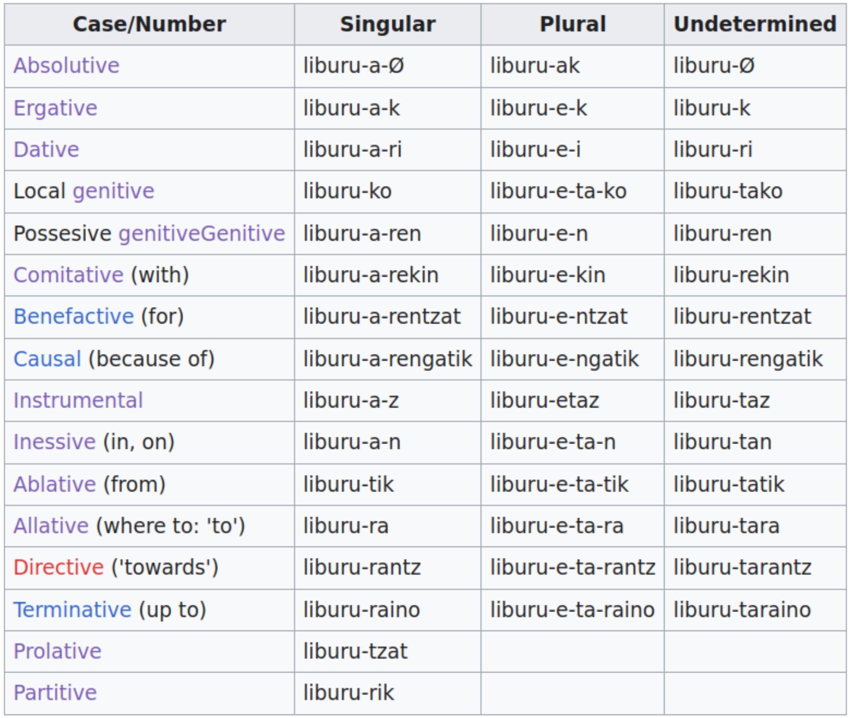
Basque language

- Four grammar-cases in Icelandic



M1	Singularra	Plurala		Singularra	Plurala
Nominatiboa	hest*ur*	hest*ar*		hest*ur***inn**	hest*ar***nir**
Akusatiboa	hest	hest*a*		hest**inn**	hest*a***na**
Datiboa	hest*i*	hest*um*		hest*i***num**	hest*u***num**
Genitiboa	hest*s*	hest*a*		hest*s***ins**	hest*a***nna**

- 16 cases in Basque (en.wikipedia)

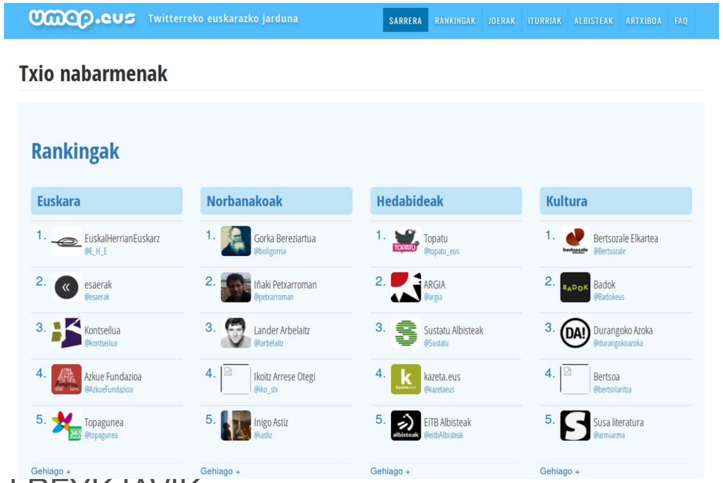


Case/Number	Singular	Plural	Undetermined
Absolutive	liburu-a-Ø	liburu-ak	liburu-Ø
Ergative	liburu-a-k	liburu-e-k	liburu-k
Dative	liburu-a-ri	liburu-e-i	liburu-ri
Local genitive	liburu-ko	liburu-e-ta-ko	liburu-tako
Possesive genitiveGenitive	liburu-a-ren	liburu-e-n	liburu-ren
Comitative (with)	liburu-a-rekin	liburu-e-kin	liburu-rekin
Benefactive (for)	liburu-a-rentzat	liburu-e-ntzat	liburu-rentzat
Causal (because of)	liburu-a-rengatik	liburu-e-ngatik	liburu-rengatik
Instrumental	liburu-a-z	liburu-etaz	liburu-taz
Inessive (in, on)	liburu-a-n	liburu-e-ta-n	liburu-tan
Ablative (from)	liburu-tik	liburu-e-ta-tik	liburu-tatik
Allative (where to: 'to')	liburu-ra	liburu-e-ta-ra	liburu-tara
Directive ('towards')	liburu-rantz	liburu-e-ta-rantz	liburu-tarantz
Terminative (up to)	liburu-raino	liburu-e-ta-raino	liburu-taraino
Prolative	liburu-tzat		
Partitive	liburu-rik		

Basic subjects in Language Technology

Looking for **texts in Basque**...

- Standardization (1968)
- (Digital) Contents (school books, small dictionaries, magazines…)
    - **Schools** (*Ikastolak*) → **University**
- Open/Free software / open contents
- Wikimedia / Wikipedia
- **Digital community**: websites, blogs, social networking, media...
- Need of **incremental design** and development of technology: **foundations** (resources), **tools**, and **applications**



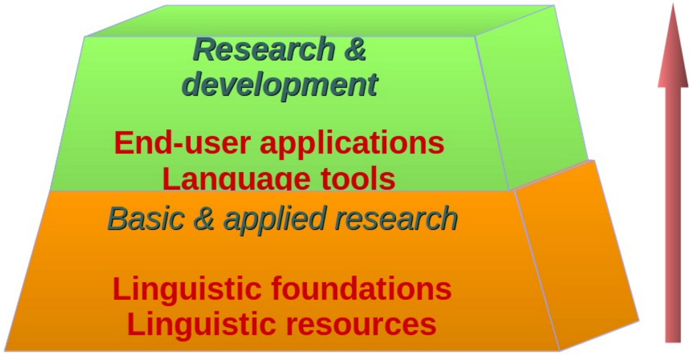
umap.eus Twitterreko euskarazko jarduna

SARRERA RANKINGAK JOERAK ITURRIAK ALBISTEAK ARTXIBOA FAQ

Txio nabarmenak

Rankingak

	Euskara	Norbanakoak	Hedabideak	Kultura
1.	EuskalHerrianEuskarz @E_H_E	Gorka Bereziartua @boligorria	Topatu @topatu_eus	Bertsozale Elkartea @Bertsozale
2.	esaerak @esaerak	Iñaki Petxarroman @petxarroman	ARGIA @argia	Badok @Badokeus
3.	Kontseilua @kontseilua	Lander Arbelaitz @larbelaitz	Sustatu Albisteak @Sustatu	Durangoko Azoka @durangokoazoka
4.	Azkue Fundazioa @AzkueFundazioa	Ikoitz Arrese Otegi @iko_stx	kazeta.eus @kazetaeus	Bertsoa @bertsolaritza
5.	Topagunea @topagunea	Inigo Astiz @iastiz	EiTB Albisteak @eitbAlbisteak	Susa literatura @armiarma
	Gehiago +	Gehiago +	Gehiago +	Gehiago +



Research & development

End-user applications

Language tools

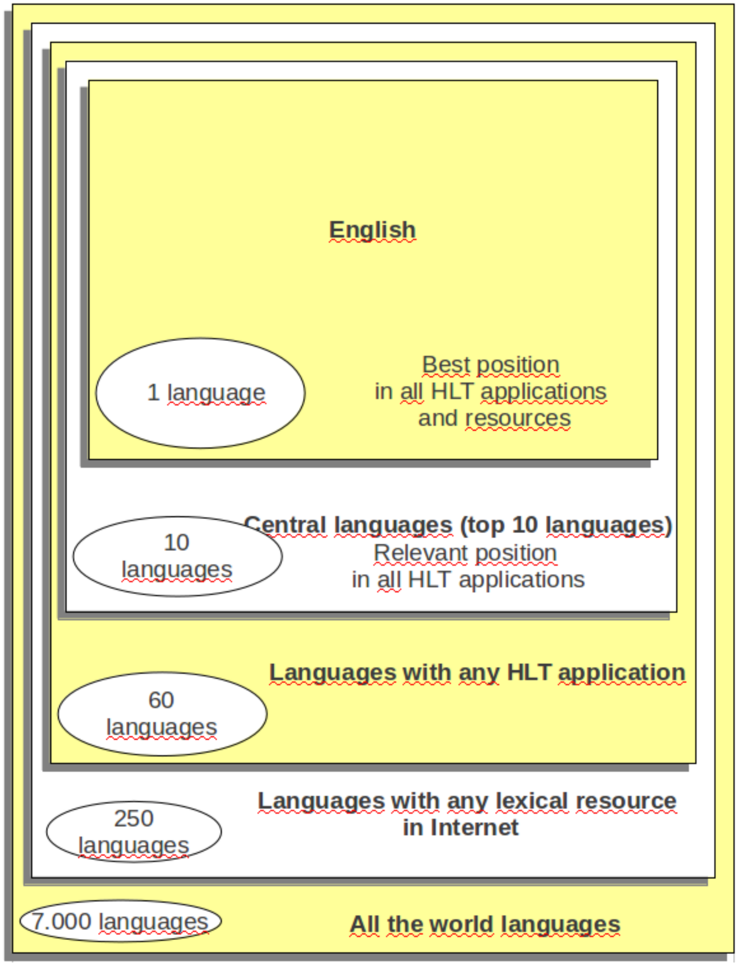
Basic & applied research

Linguistic foundations

Linguistic resources

Less resourced languages

- Typology based on (digital) resources



Languages with scare resources

- Typology based on (digital) resources
- Associations:

mercator
European Research Centre on Multilingualism and Language Learning

SOAS University of London
100 Years

World Languages Institute

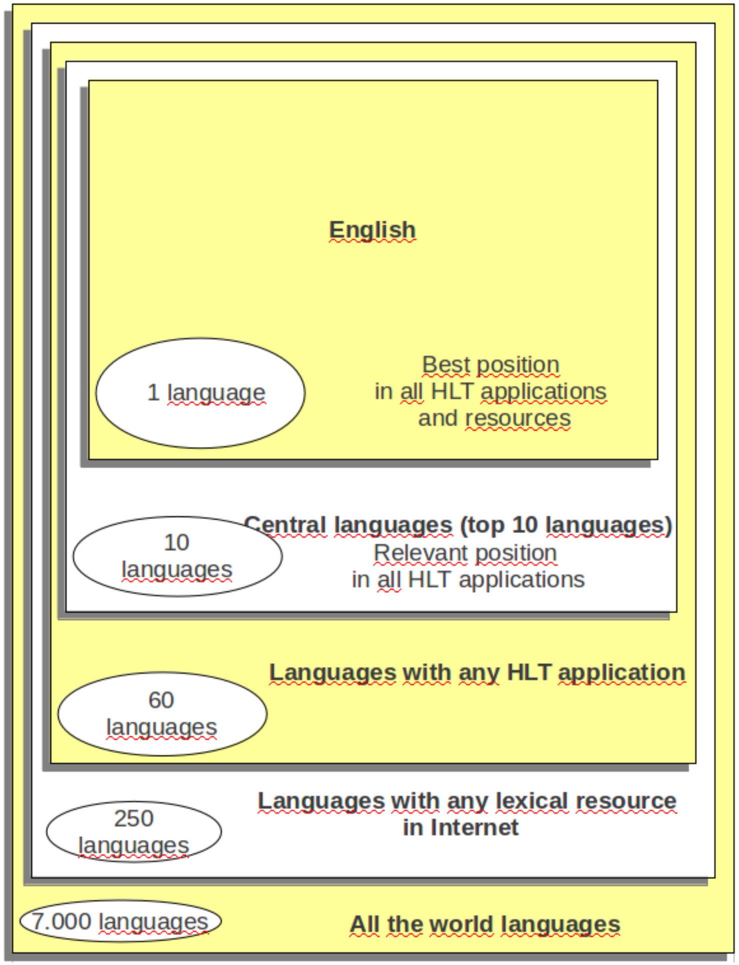
Foundation for Endangered Languages

Speech And Language Technology for Minority Languages

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

ELDA

SIGUL:
ISCA Special Interest Group.
Under-resourced Languages



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

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

ixa

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Basic resources

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

Basic tools and their applications

- On-line dictionary → Games
- Morphology → Spelling corrector
  → Language Learning
- Lemmatizer/POS_tagger → Search engine
- Text normalization
  and Machine translation

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 and OCR
- Problems: copyrights and 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/testing the spelling checker, learning (small) language models...

10 VENTAJAS DEL LIBRO DE TEXTO DIGITAL VS LIBRO DE TEXTO TRADICIONAL

MOTIVA A LOS ALUMNOS AL ESTUDIO Y LES AYUDA A APRENDER gracias a la diversidad de formatos digitales que adoptan sus contenidos.

PERMITE LA AUTOEVALUACIÓN. Los ejercicios autocorregibles favorecen que el alumno repase y obtenga resultados inmediatos sobre lo que ha aprendido, y suponen un importante ahorro de tiempo para el profesorado.

PROPORCIONA UN ENTORNO PERSONALIZABLE y adaptable a las necesidades y características de los alumnos.

AUMENTA LA INTERACCIÓN entre alumno, el contenido, y su aplicación al entorno.

AYUDA A LA COMPRENSIÓN DE CONCEPTOS a través de distintos lenguajes con gran potencial didáctico, como el audiovisual, o el interactivo.

FAVORECE LA ADQUISICIÓN DE LA COMPETENCIA DIGITAL, vital para manejarse en el mundo actual, en un entorno seguro.

ALIGERA LA MOCHILA frente a los libros convencionales, que cargan de peso la espalda de los alumnos.

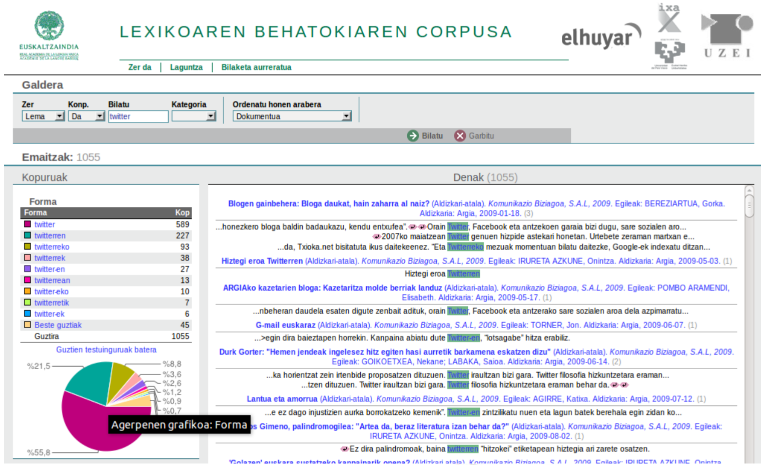
REPRESENTA UN IMPORTANTE AHORRO PARA LAS FAMILIAS frente al alto coste de los libros de texto tradicionales.

ES ACTUALIZABLE y permite relacionar los conocimientos con la actualidad y, por ello, acercarlo al contexto real del alumno

PUEDE CONSULTARSE EN CUALQUIER MOMENTO Y LUGAR, con o sin conexión a Internet.

www.aulaplaneta.com

aulaPlaneta



LEXIKOAREN BEHATOKIAREN CORPUSA

Galdera

Zer: Lema · Konp.: Da · Bilatu: twitter · Kategoria: · Ordenatu honen arabera: Dokumentua

Emaitzak: 1055

Kopuruak

Forma

Forma	Kop
twitter	589
twitterren	227
twitterreko	93
twitterrek	38
twitter-en	27
twitterrean	13
twitter-eko	10
twitterretik	7
twitter-ek	6
Beste guztiak	45
Guztira	1055

Guztien testuinguruak batera

Denak (1055)

Blogen gainbehera: Bloga daukat, hain zaharra al naiz? (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: BEREZIARTUA, Gorka. Aldizkaria: Argia, 2009-01-18. (3)

...honezkero bloga baldin badaukazu, kendu entxufea”.↵↵Orain Twitter, Facebook eta antzekoen garaia bizi dugu, sare sozialen aro...
↵2007ko maiatzean Twitter genuen hizpide astekari honetan. Urtebete zeraman martxan e...
...da, Txioka.net bisitatuta ikus daitekeenez. “Eta Twitterreko mezuak momentuan bilatu daitezke, Google-ek indexatu ditzan...

Hiztegi eroa Twitterren (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: IRURETA AZKUNE, Onintza. Aldizkaria: Argia, 2009-05-03. (1)

Hiztegi eroa Twitterren

ARGIAko kazetarien bloga: Kazetaritza molde berriak landuz (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: POMBO ARAMENDI, Elisabeth. Aldizkaria: Argia, 2009-05-17. (1)

...nbeheran daudela esaten digute zenbait adituk, orain Twitter, Facebook eta antzerako sare sozialen aroa dela azpimarratu...

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

...>egin dira baieztapen horrekin. Kanpaina abiatu dute Twitter-en, “lotsagabe” hitza erabiliz.

Durk Gorter: "Hemen jendeak ingelesez hitz egiten hasi aurretik barkamena eskatzen dizu" (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: GOIKOETXEA, Nekane; LABAKA, Saioa. Aldizkaria: Argia, 2009-06-14. (2)

...ka horientzat zein irtenbide proposatzen dituzuen. Twitter iraultzan bizi gara. Twitter filosofia hizkuntzetara eraman...
...tzen dituzuen. Twitter iraultzan bizi gara. Twitter filosofia hizkuntzetara eraman behar da.↵↵

Lantua eta amorrua (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: AGIRRE, Katixa. Aldizkaria: Argia, 2009-07-12. (1)

...e ez dago injustizien aurka borrokatzeko kemenik”. Twitter-en zintzilikatu nuen eta lagun batek berehala egin zidan ko...

...s Gimeno, palindromogilea: "Artea da, beraz literatura izan behar da?" (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: IRURETA AZKUNE, Onintza. Aldizkaria: Argia, 2009-08-02. (1)

↵Ez dira palindromoak, baina twitterren “hitzokei” etiketapean hiztegia ari zarete osatzen.

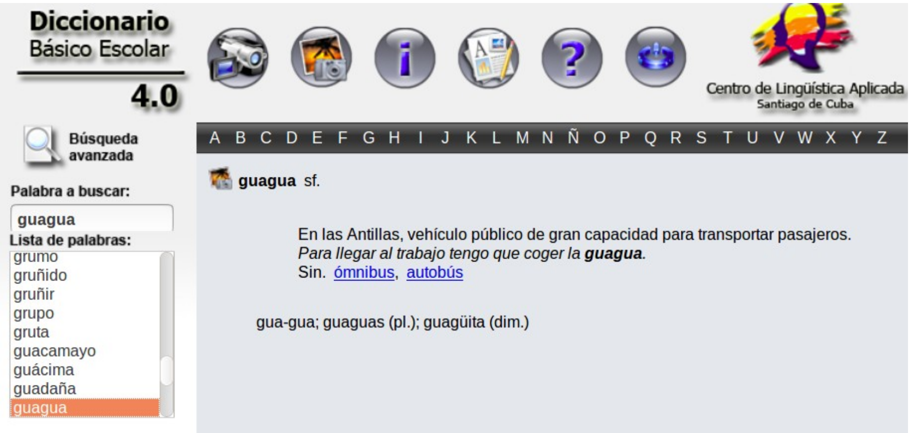
'Golazen' euskara sustatzeko kanpainarik onena? (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: IRURETA AZKUNE, Onintza

Dictionaries (mono- or bi-lingual)

- Basic tool for students, journalists, translators and writers
- Historical evolution: Paper cards → MSWord → XML/TEI
- XML/TEI → **Multimedia**: CD/DVD, Web/phone, Paper
  - Unique maintenance → 3 products
- Integration: *Euskalbar* (browser)
- Some projects in collaboration:
  - Small dictionary for Nahuatl, Garabide NGO

    Scholar dictionary in Cuba (*DBE*), CLA institute in Santiago de Cuba

    Scrable and other games in Basque



Diccionario
Básico Escolar
4.0
Centro de Lingüística Aplicada
Santiago de Cuba
A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z
Búsqueda avanzada

Palabra a buscar:

guagua

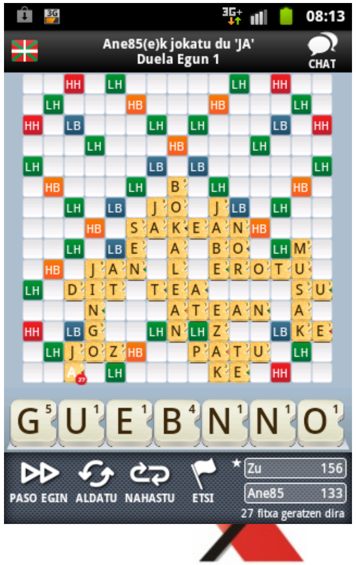
Lista de palabras:

grumo
gruñido
gruñir
grupo
gruta
guacamayo
guácima
guadaña
guagua

guagua sf.

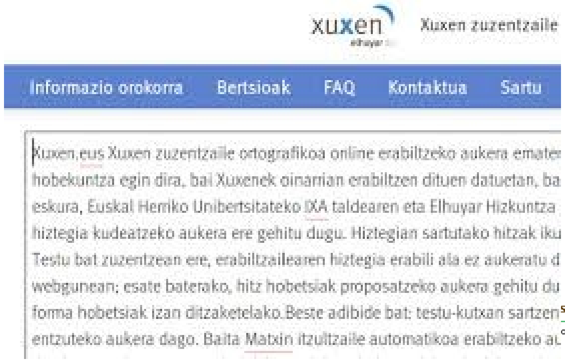
En las Antillas, vehículo público de gran capacidad para transportar pasajeros.
Para llegar al trabajo tengo que coger la ***guagua****.*
Sin. ómnibus, autobús

gua-gua; guaguas (pl.); guagüita (dim.)

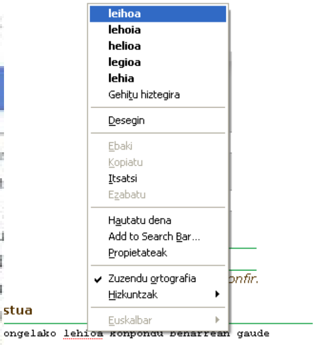


Morphology / Spelling corrector

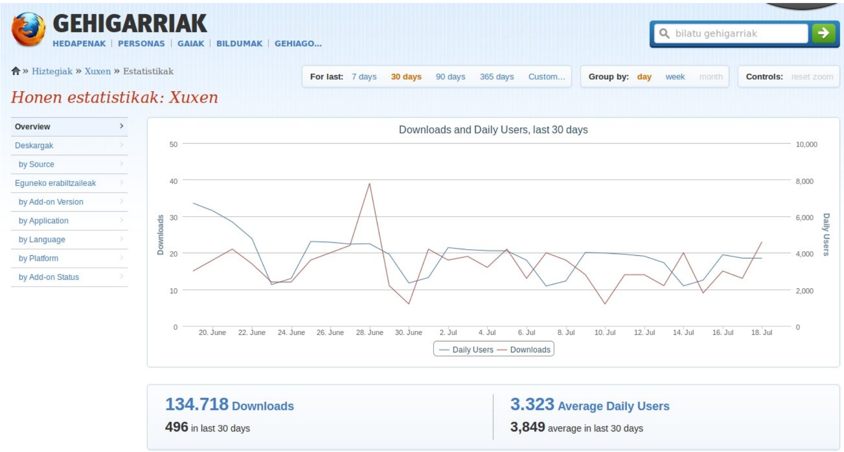
- Computational morphology is compulsory for most of the languages:
  - Dictionary + word-grammar
- The spelling corrector is a key application
  - Basic tool for students, journalists, translators and writers
  - **Key for basque standardization**
- Integration/online: MSoffice, LibreOffice, Mozilla, Android…
- Basic tools:
  - *foma* and *hunspell* (free software)
- Projects: unified Basque, dialectal Basque, Quechua (Univ. Zurich and Cusco), Mapudungun (Chile)



Xuxen.eus Xuxen zuzentzaile ortografikoa online erabiltzeko aukera emate
hobekuntza egin dira, bai Xuxenek oinarrian erabiltzen dituen datuetan, ba
eskura, Euskal Herriko Unibertsitateko IXA taldearen eta Elhuyar Hizkuntza
hiztegia kudeatzeko aukera ere gehitu dugu. Hiztegian sartutako hitzak iku
Testu bat zuzentzean ere, erabiltzailearen hiztegia erabili ala ez aukeratu d
webgunean; esate baterako, hitz hobetsiak proposatzeko aukera gehitu du
forma hobetsiak izan ditzaketelako.Beste adibide bat: testu-kutxan sartzen
entzuteko aukera dago. Baita Matxin itzultzaile automatikoa erabiltzeko au



ongelako lehioa konpondu beharrean gaude



134.718 Downloads
496 in last 30 days

3.323 Average Daily Users
3,849 average in last 30 days

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:

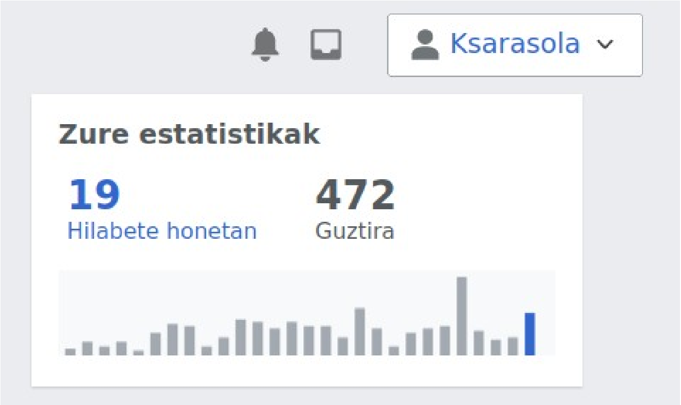
Successful applications since 2018

- Machine translation
- Use of the local language in Health services
- Digital humanities
- Speech synthesis
- Speech recognition
- Conversational interfaces, chatbots
- ...

Machine Translation

- Precondition: (Parallel) Corpora
- 2006 RBMT (Knowledge based)
  2009 SMT
  2018 NMT

  Collaboration with Elhuyar.eus



Ksarasola

Zure estatistikak

19 Hilabete honetan

472 Guztira

I translated to Basque 472 Wikipedia articles (2021-2023) :-)

- Neural revolution
  - Not perfect, it needs postedition but the quality is very high
  - Something incredible five years ago
  - We have 4 free good translation-services via web for Basque
  - This opens new ways

    where the use of translation could be extended to many situations
  - New horizons for under-resourced languages
- In most of the cases cross-lingual learning is used, but good results are also obtained even only using monolingual corpora (Artetxe et al., 2020)

  → Nice for languages with few parallel resources

Spanish and Basque plans for Language technology

- Plan for the Advancement of Language Technology

  2015-2020 90 M€

Spanish Plan for Artificial Intelligence

- 2021-2023
- Spanish Government.
- https://www.lamoncloa.gob.es/lang/en/presidente/news/Paginas/2020/20201202_enia.aspx

Basque Plan for Language technology

- 2021-2025
- Also the Basque Government.
- https://www.euskadi.eus/gaitu-plan-de-accion-de-las-tecnologias-de-la-lengua-2021-2024/web01-ejeduki/es/

Good news for less-resourced languages

There have been significant advances, even for less-resourced languages, in several areas:

- lexicon extraction (Artetxe et al., 2019),
- morphology induction (Anastasopoulos&Neubig, 2019)
- POS tagging (Kim et al., 2017),
- machine translation (Artetxe et al., 2017)
- chatbots

(Artetxe et al., 2020)

In most of the cases cross-lingual learning is used, but good results are also obtained even only using monolingual corpora,

→ Nice for languages with few parallel resources

Good news for less-resourced languages

(Agerri et al., 2020)

Word embeddings and pre-trained language models enabled improvements across most NLP tasks.

Unfortunately they are very expensive to train,

--> small companies and research groups tend to use big models provided by the big companies.

But our mono- & multilingual language BERT models have proven to be very useful in NLP tasks for Basque.

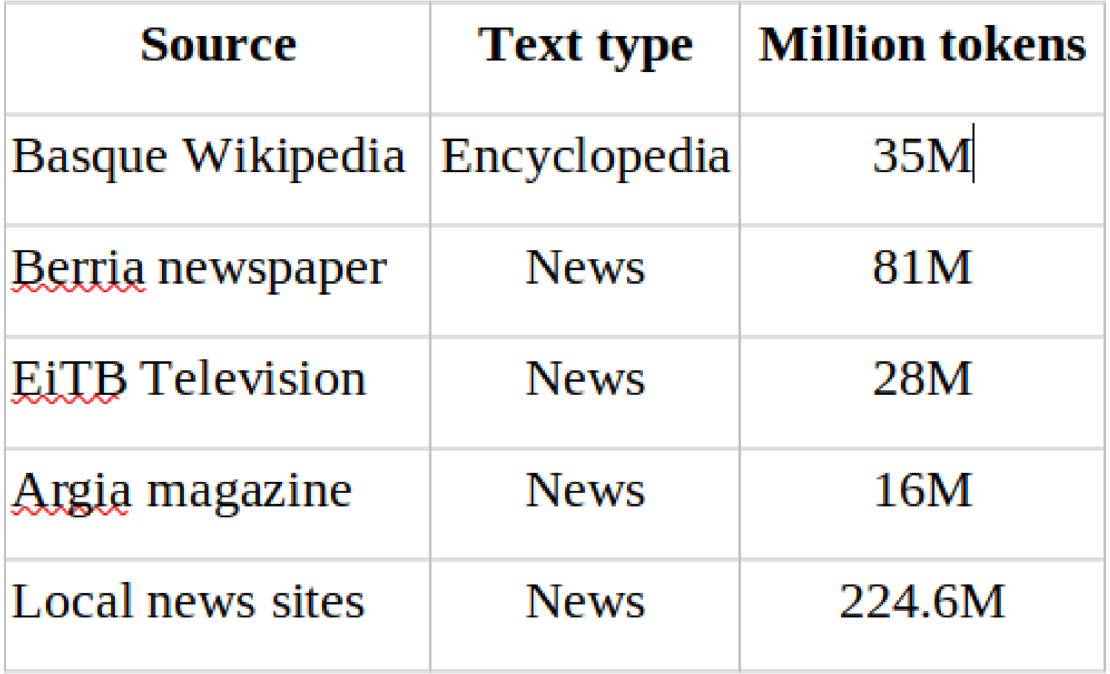
Eventhough they have been created:

- with a 500 times smaller corpus than the English one
- with a 80 times smaller wikipedia.

Good news for less-resourced languages

The original BERT language model for English was trained in 2018 using Google books corpus with 189 billion words.

Almost 500 times bigger than the Basque one (384 millions).



Source	Text type	Million tokens
Basque Wikipedia	Encyclopedia	35M
Berria newspaper	News	81M
EiTB Television	News	28M
Argia magazine	News	16M
Local news sites	News	224.6M

Good news for less-resourced languages

(Otegi et al., 2020)

A multilingual language model
pretrained only for English, Spanish and Basque, using:

- The monolingual Basque model
- English Wikipedia (2.5 Gword)
- Spanish Wikipedia (650 Mword)
  (80 and 20 times bigger than the Basque Wikipedia)

Successful to transfer knowledge from English to Basque in a conversational Question/Answering system

Better than the general Google's official mBERT model
(it covers too many languages, Basque is not well represented).

Other personal projects

- Basque Summer University

  https://en.wikipedia.org/wiki/Basque_Summer_University
- Association around the Basque Wikipedia

  https://eu.wikipedia.org
- Garabide NGO

  https://www.garabide.eus/english/
- Others:

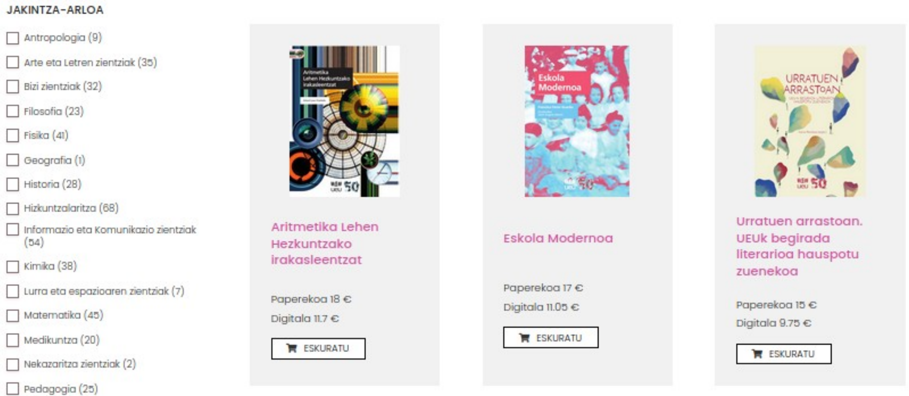
  Language schools and communities for practice (online also), association for localising open software (*Librezale*) …

  Repositories of talks (old people), bibliography in Basque, open literature and critics…

  .eus Internet domain

UEU: Basque Summer University

- The main public university (UPV/EHU) is bilingual
- UEU works only in Basque: community
  - Association of professors, lecturers and students
  - Promoting Basque at the University: teaching, research, dissemination
  - Some official titles in collaboration with official universities
  - Courses, conferences, books, databases…
- Two new interesting projects:
  - **Ikergazte**: pre- and post-doctoral Basque **researchers** together
  - **Open university in Basque**: today is not possible in Basque



JAKINTZA-ARLOA

- □ Antropologia (9)
- □ Arte eta Letren zientziak (35)
- □ Bizi zientziak (32)
- □ Filosofia (23)
- □ Fisika (41)
- □ Geografia (1)
- □ Historia (28)
- □ Hizkuntzalaritza (68)
- □ Informazio eta Komunikazio zientziak (54)
- □ Kimika (38)
- □ Lurra eta espazioaren zientziak (7)
- □ Matematika (45)
- □ Medikuntza (20)
- □ Nekazaritza zientziak (2)
- □ Pedagogia (25)

Aritmetika Lehen Hezkuntzako irakasleentzat

Paperekoa 18 €

Digitala 11.7 €

ESKURATU

Eskola Modernoa

Paperekoa 17 €

Digitala 11.05 €

ESKURATU

Urratuen arrastoan. UEUk begirada literarioa hauspotu zuenekoa

Paperekoa 15 €

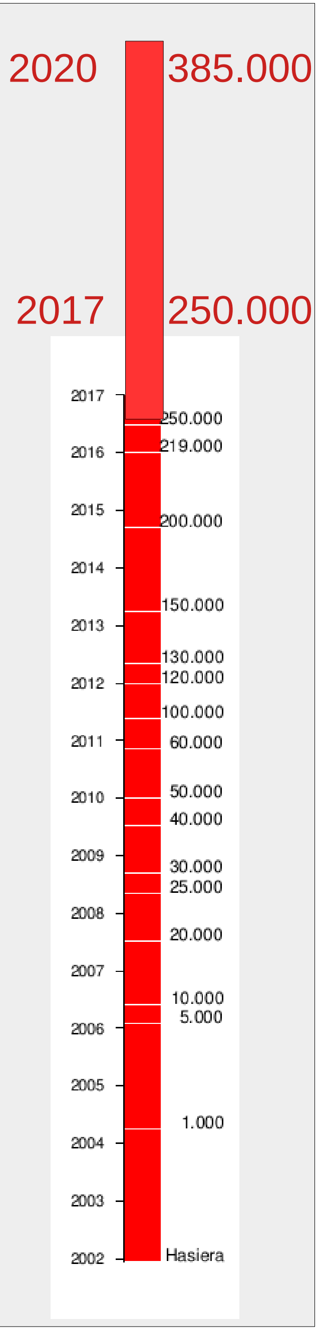
Digitala 9.75 €

ESKURATU

Wikipedia (open source text)

- No problems with copyrights and formats
- Applications:
  - **language models**
  - **text mining**
  - **Language Technology evaluation**
- Growing and growing
- 2017-2019
  Basque Government Education program
  for creating 1.000 basic articles
  in Basque Wikipedia
  for 12-16 years students
  created by students at the university
  :-))

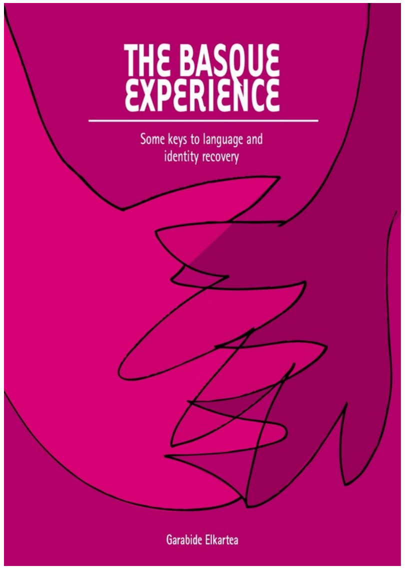
Articles
in Basque
Wikipedia
year
by
year



Garabide: NGO for language cooperation

- Cooperation with language communities,
  specially Latin America and north of Africa

  The Revitalization of the Basque Language as a Model for Other Minority Languages in the World
- https://issuu.com/di-da/docs/garabide-english
- Cooperation in pedagogy, sociolingusitics, technology...
- Master on language management (in Spanish) for leaders



THE BASQUE EXPERIENCE

Some keys to language and identity recovery

Garabide Elkartea

Basque in the digital era

- The Basque has achieved a good position

  “a successful case”
- Basque (data from a research in 2014 [1])
  - Translated books: 54th among 247
  - Wikipedia: 47th among 134 (34th in 2022)
  - Twitter: 36th among 70
- Globalization and Internet (young people) →

  main languages are more and more prevalent (Netflix...)

  young people learn Basque, have tools, but many do not use them

[1] https://www.pnas.org/doi/pdf/10.1073/pnas.1410931111

Discussion

- Sustainability / Cost:
- Planning is a need for sustainability
- Fast development could be *expensive*

- Integration on commercial software / open software

  Many popular applications are proprietary and

  the decision to add new languages depends on the company

  → interesting experience in Basque.

  Microsoft become more interested
  after the localization and integration on free software

  → spin-off company for apps

- Standardization/dialect/language/alphabet...

  Each community has to decide...

  But standardization is crucial for text processing!

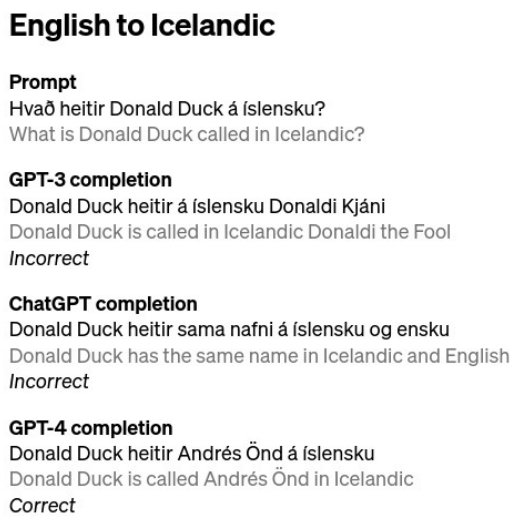
  A spelling corrector can be a “de facto” proposal

Discussion

How to face the development of “ChatGPT” for Basque?

- Using commercial software?

    https://openai.com/customer-stories/government-of-iceland



English to Icelandic

Prompt
Hvað heitir Donald Duck á íslensku?
What is Donald Duck called in Icelandic?

GPT-3 completion
Donald Duck heitir á íslensku Donaldi Kjáni
Donald Duck is called in Icelandic Donaldi the Fool
Incorrect

ChatGPT completion
Donald Duck heitir sama nafni á íslensku og ensku
Donald Duck has the same name in Icelandic and English
Incorrect

GPT-4 completion
Donald Duck heitir Andrés Önd á íslensku
Donald Duck is called Andrés Önd in Icelandic
Correct

- Or creating our open software by our own? Finland

Takk fyrir

Eskerrik asko

Thank you

It is always easier to open way
when you have close references

Technology makes it easier :-))

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

ixa

eman ta zabal zazu

Universidad del País Vasco Euskal Herriko Unibertsitatea

Spurningar, skoðanir...

Galderak, iritziak...

Questions, opinions...

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

Ixa Group (Iñaki Alegria, Kepa Sarasola)

Haskòlinn Í Reykjavik

www.ixa.eus

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

ixa

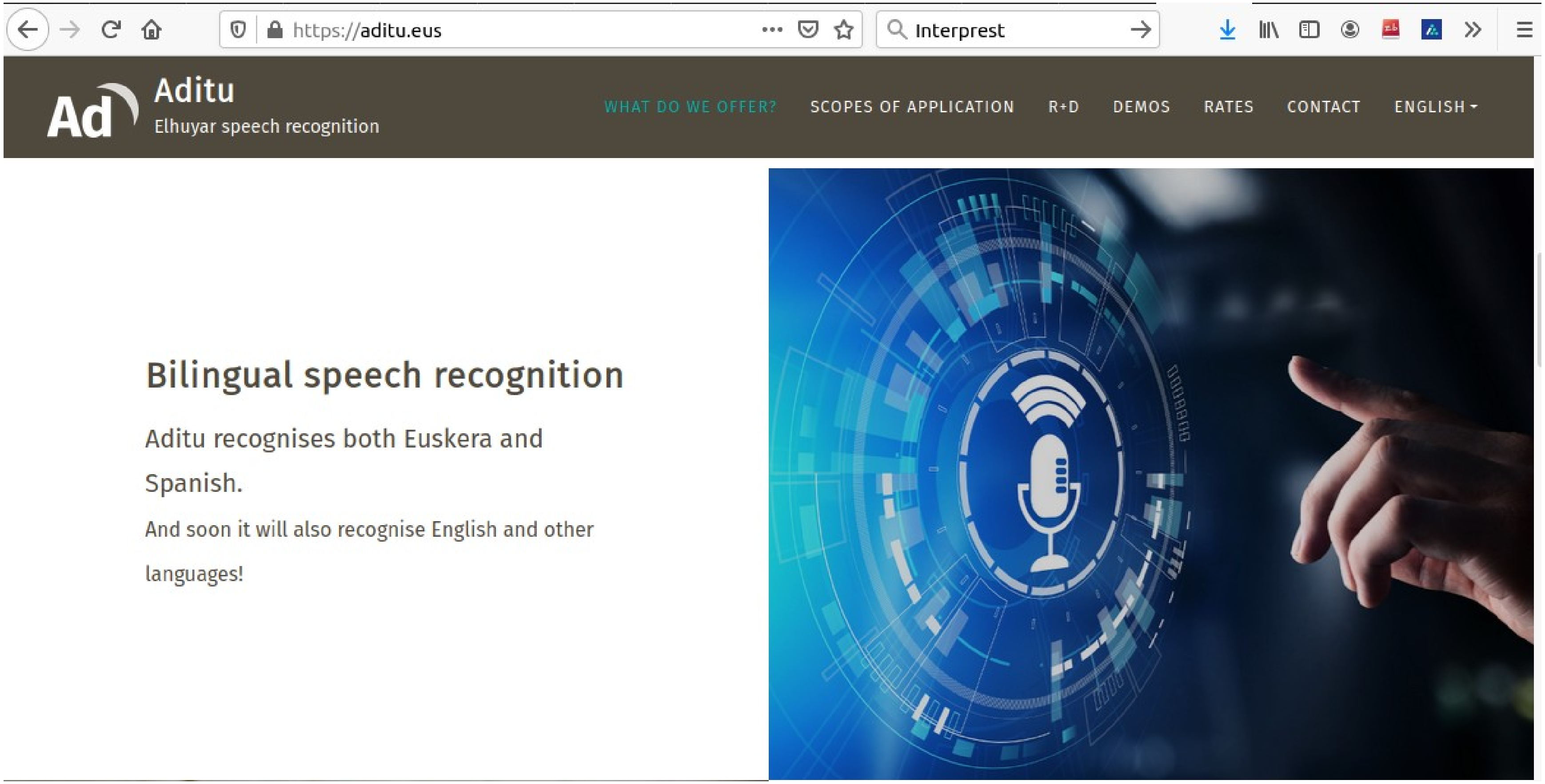
eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Extra slides (complementary)

- Some Useful applications
  - Aditu, bilingual speech recognition
  - Interprest,
    low cost and portable interpretation services
  - Bidaide. Web service to create and provide multingual contents.

Aditu, bilingual speech recognition

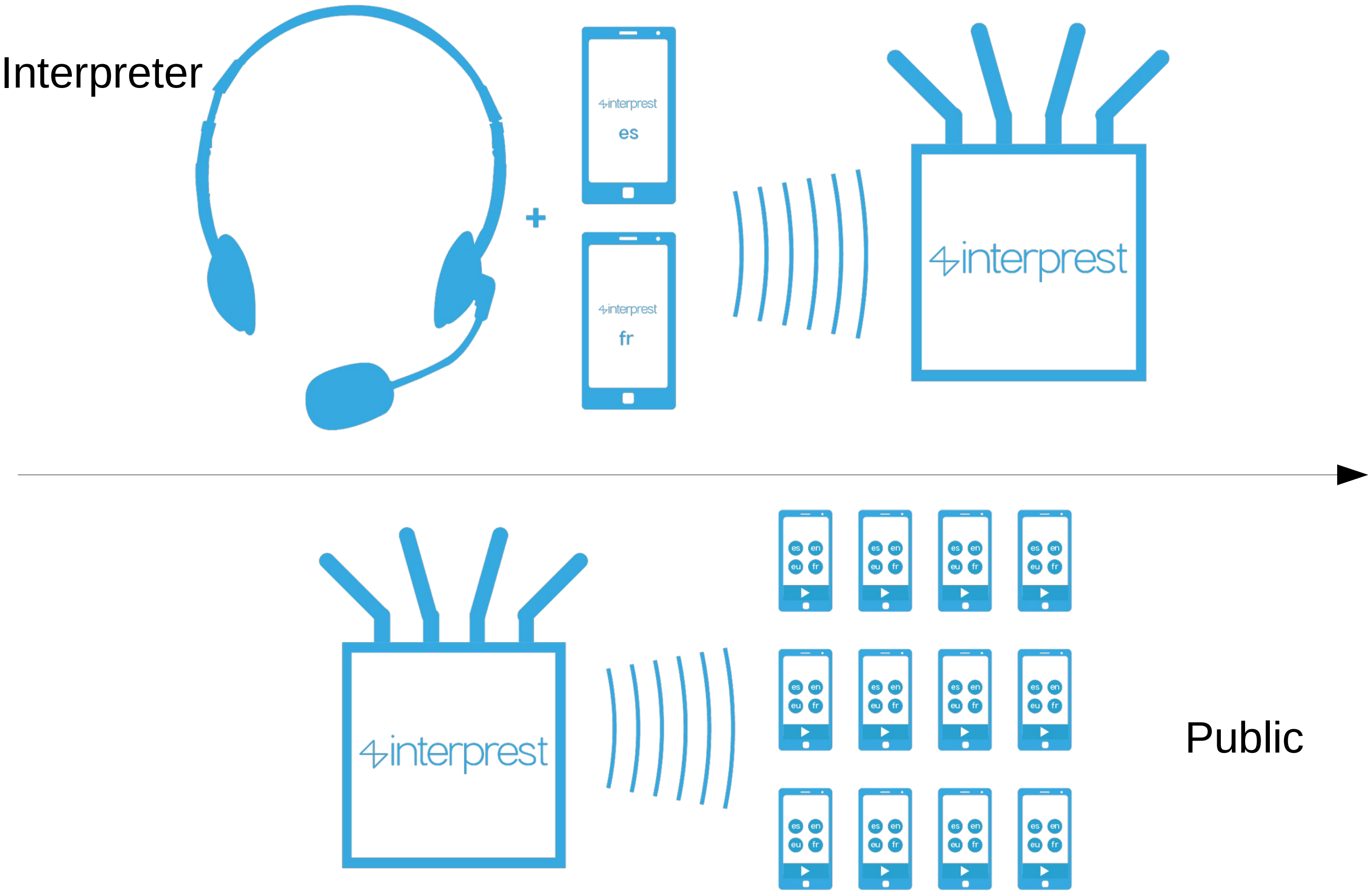


Bilingual speech recognition

Aditu recognises both Euskera and Spanish.

And soon it will also recognise English and other languages!

Interprest,
low cost and portable interpretation services



Interpreter

Public

Turin 2021

ixa

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

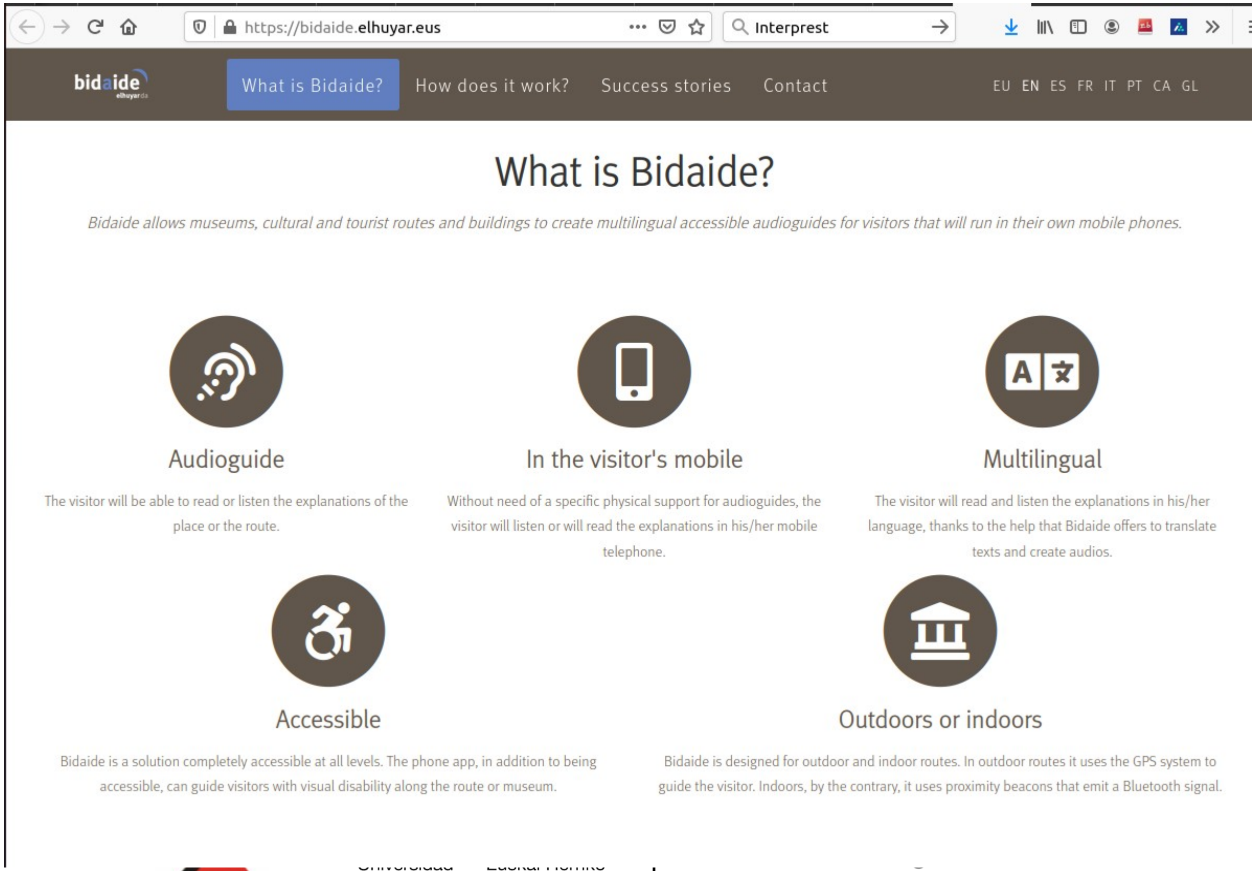
HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Bidaide

Web service to create and provide multingual contents.

In museums, routes or buildings.

Visitors read or listen to explanations on their own mobile



What is Bidaide?

Bidaide allows museums, cultural and tourist routes and buildings to create multilingual accessible audioguides for visitors that will run in their own mobile phones.

Audioguide

The visitor will be able to read or listen the explanations of the place or the route.

In the visitor's mobile

Without need of a specific physical support for audioguides, the visitor will listen or will read the explanations in his/her mobile telephone.

Multilingual

The visitor will read and listen the explanations in his/her language, thanks to the help that Bidaide offers to translate texts and create audios.

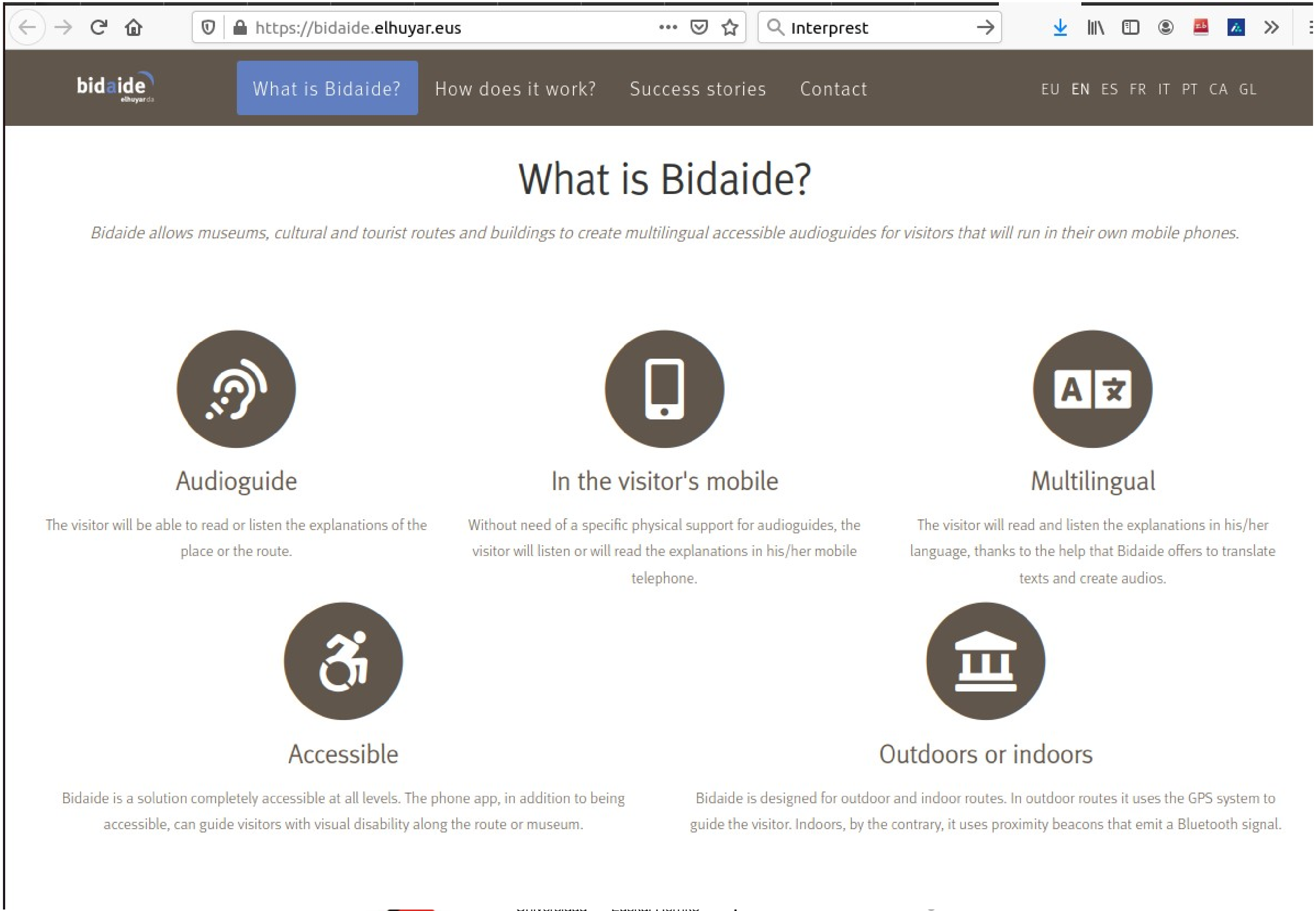
Accessible

Bidaide is a solution completely accessible at all levels. The phone app, in addition to being accessible, can guide visitors with visual disability along the route or museum.

Outdoors or indoors

Bidaide is designed for outdoor and indoor routes. In outdoor routes it uses the GPS system to guide the visitor. Indoors, by the contrary, it uses proximity beacons that emit a Bluetooth signal.

Bidaide



What is Bidaide?

Bidaide allows museums, cultural and tourist routes and buildings to create multilingual accessible audioguides for visitors that will run in their own mobile phones.

Audioguide

The visitor will be able to read or listen the explanations of the place or the route.

In the visitor's mobile

Without need of a specific physical support for audioguides, the visitor will listen or will read the explanations in his/her mobile telephone.

Multilingual

The visitor will read and listen the explanations in his/her language, thanks to the help that Bidaide offers to translate texts and create audios.

Accessible

Bidaide is a solution completely accessible at all levels. The phone app, in addition to being accessible, can guide visitors with visual disability along the route or museum.

Outdoors or indoors

Bidaide is designed for outdoor and indoor routes. In outdoor routes it uses the GPS system to guide the visitor. Indoors, by the contrary, it uses proximity beacons that emit a Bluetooth signal.