PRIFYSGOL BANGOR UNIVERSITY

Symposiwm Academaidd Technolegau Iaith Cymru 2020
Wales Academic Symposium on Language Technologies 2020

Language Technology for Language Communities: An Overview based on Basque Experience 2020

Kepa Sarasola, Olatz Perez-de-Viñaspre. Iñaki Alegria
Ixa group. HiTZ Center
University of the Basque Country

Bangor 2020

ixa

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

PRIFYSGOL
BANGOR
UNIVERSITY

Symposiwm Academaidd Technolegau Iaith Cymru 2020
Wales Academic Symposium on Language Technologies 2020

Happily year after year we feel that
the recovery processes of Welsh and Basque
go hand in hand,

They are quite parallel processes.

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

Technology makes it easier   :-))

ixa

eman ta zabal zazu
Universidad
del País Vasco
Euskal Herriko
Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

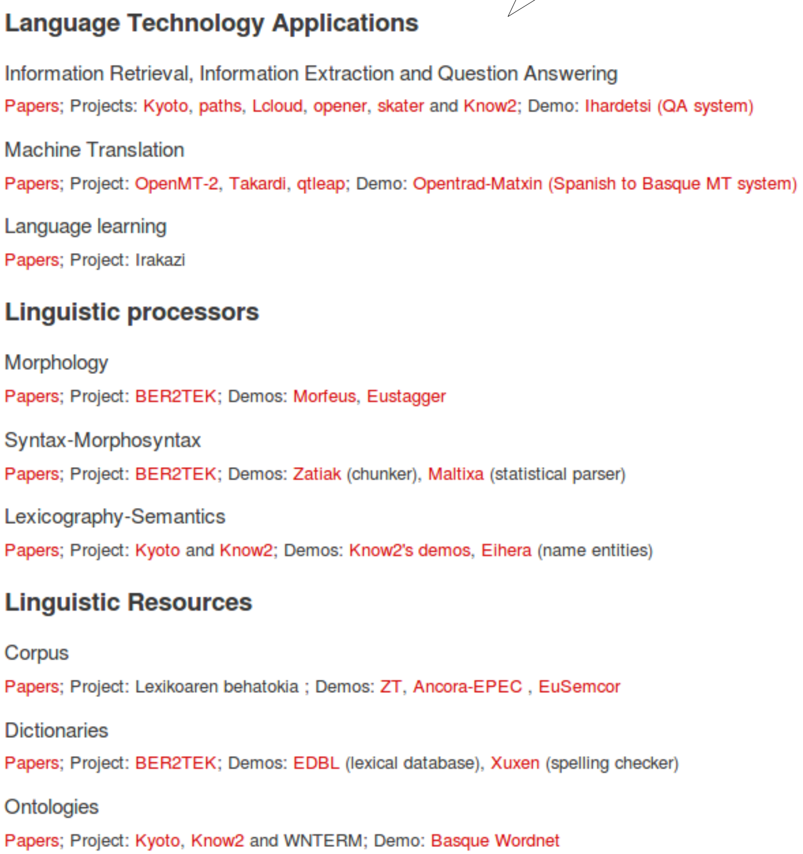
Ixa group (1988) - HiTZ Center (2020)

- **32 years** working on Language Technology
- **Basque**-centred research group but also other languages
- **Multidisciplinary**: computer scientists, linguistics...
- **Text**-based resources and apps (speech with Aholab group)
- **3 levels: r**esources, basic tools, applications (with Elhuyar)
- **Local** and **Global**
- Basque **community** and International research **community**
- **Collaboration**: Basque academy, Elhuyar Foundation, publications
- **Alternative forums**: Basque Summer University, NGOs...

Ixa taldea

People
Members

Products



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

Education
Masters

Master Ofiziala
Official Master's Degree
Hizkuntzaren Azterketa eta Prozesamendua (HAP)
Language Analysis and Processing (LAP)
http://ixa.si.ehu.es/master

eman ta zabal zazu
Universidad del País Vasco
Euskal Hei
Unibertsita

Successful applications since 2018

- Machine translation
- **Use of Basque in Health services**

- Digital humanities
- Speech synthesis
- Speech recognition
- Conversational interfaces, chatbots
- ...

Successful applications since 2018

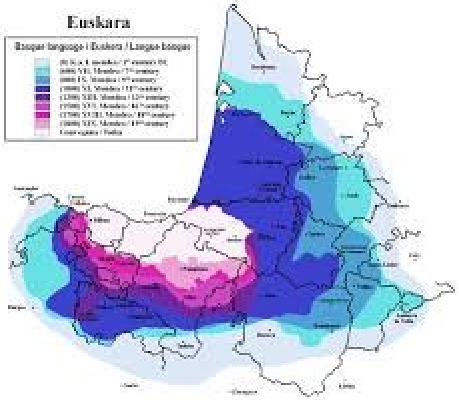
- Machine translation
- Use of Basque in Health services
  - **Olatz Perez de Viñaspre**
    Basque, NLP and Clinical domain

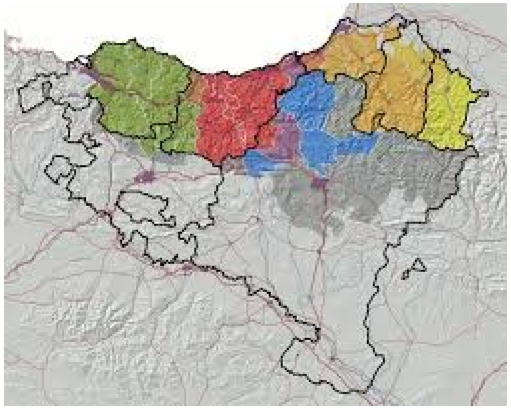
Olatz Perez de Viñaspre

- Digital humanities
- Speech synthesis
- Speech recognition
- Conversational interfaces, chatbots
- ...

Basque

- Old language (pre-Indo-European)
- ~ 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

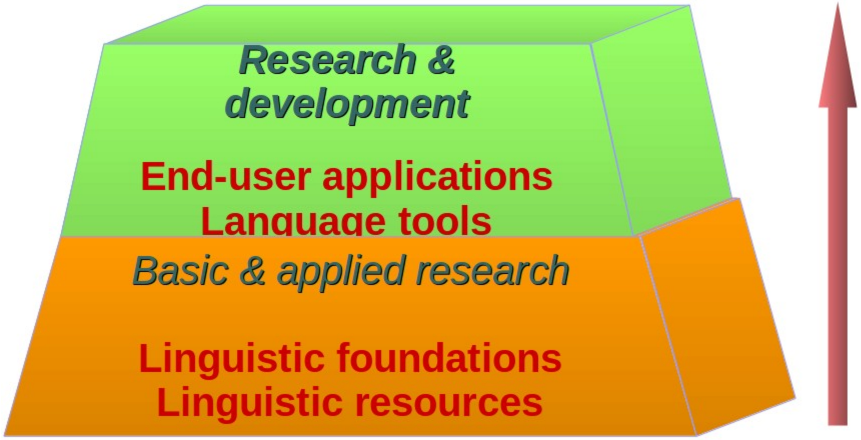




Foundations

Collecting 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



Research & development

End-user applications
Language tools

Basic & applied research

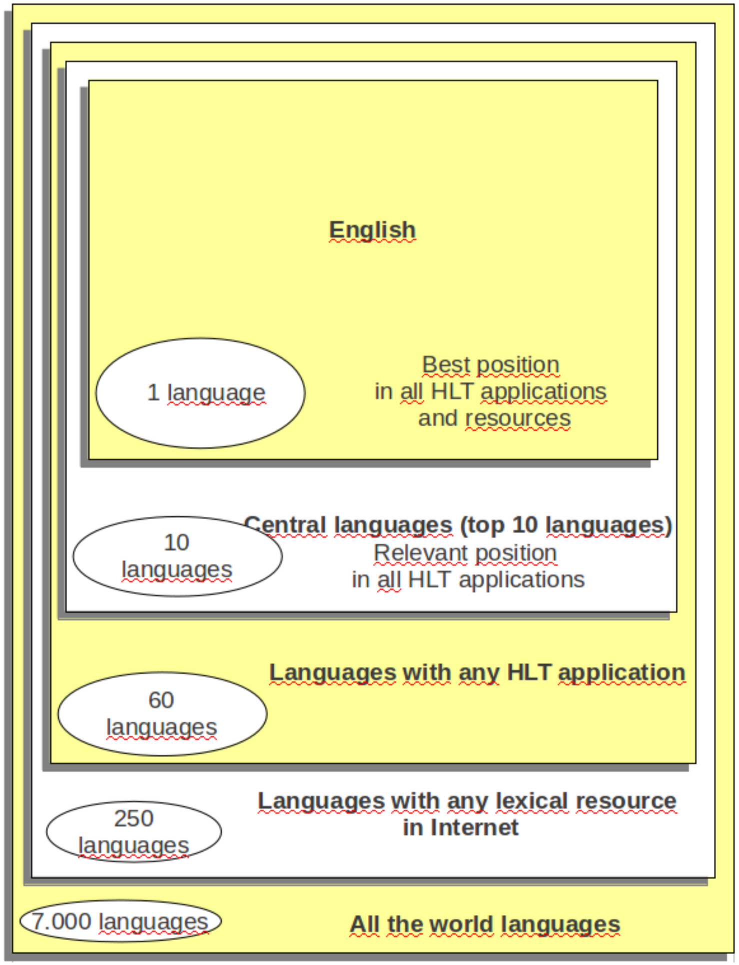
Linguistic foundations
Linguistic resources

Under-resourced languages (levels)

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

  Special Interest Group:
  Under-resourced Languages (SIGUL)

Foundation for Endangered Languages

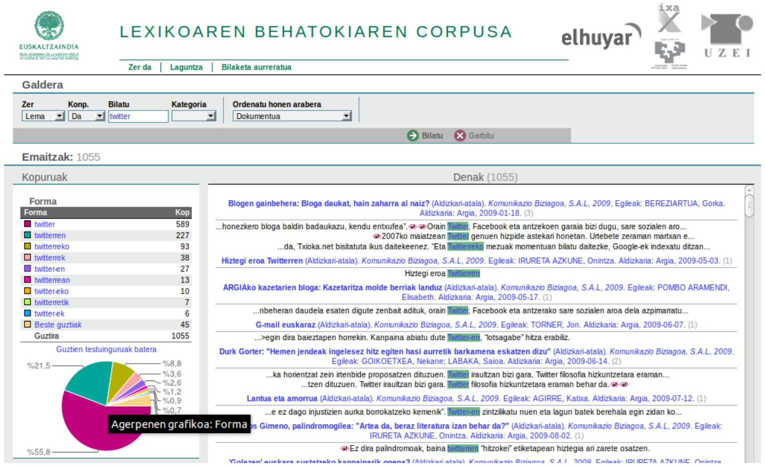


Basic resources

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

Corpora (digital texts)

- Collecting corpora is not easy fpr a under esourced language
- 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
- Applications : enriching the dictionary, creating the spelling checker, learning language models...



LEXIKOAREN BEHATOKIAREN CORPUSA

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...

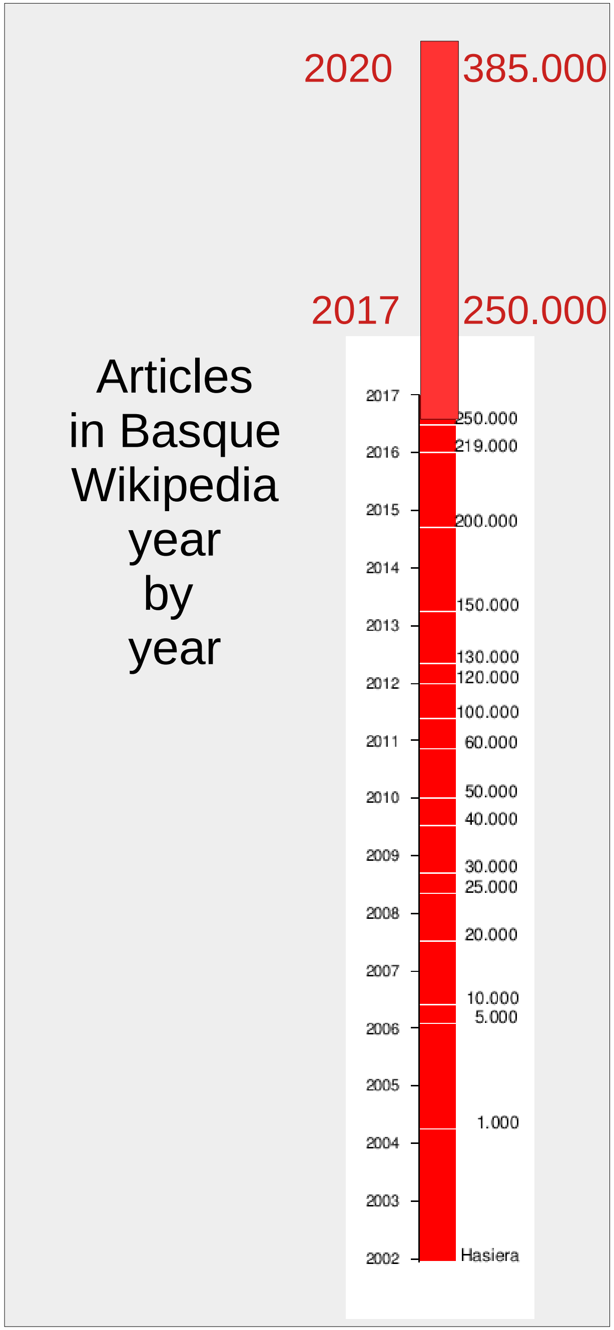
s Gimeno, palindromogilea: "Artea da, beraz literatura izan behar da?" (Aldizkari-atala). *Komunikazio Biziagoa, S.A.L, 2009*. Egileak: IRURETA AZKUNE, Onintza. Aldizkaria: Argia, 2009-08-02. (1)

↵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

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

ixa

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

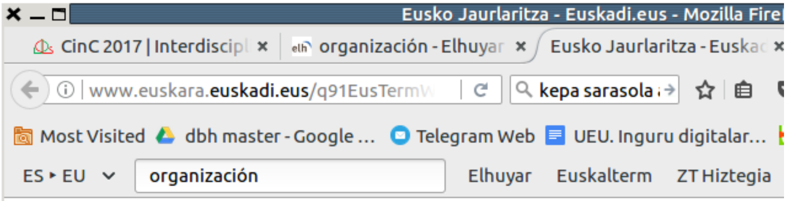
Basic tools and applications

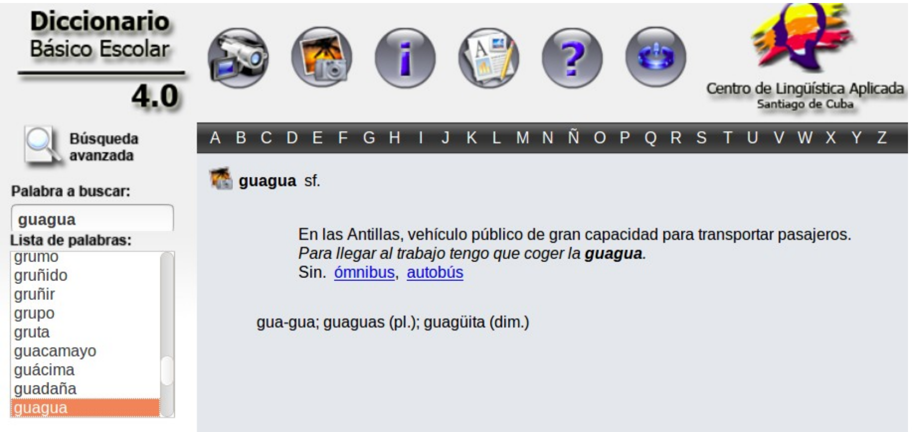
- On-line dictionary → Games
- Morphology → Spelling corrector
- Lemmatizer/POS_tagger→ Search engine
- Machine translator or Normalizator

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




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

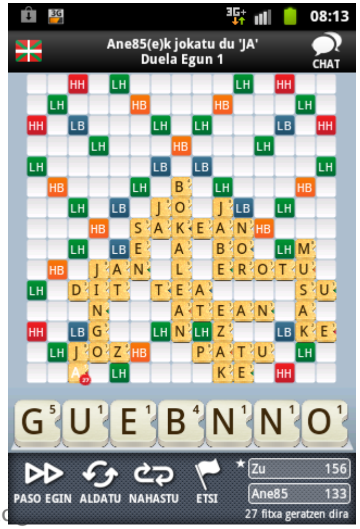
Palabra a buscar: guagua

Lista de palabras:
grumo
gruñido
gruñir
grupo
gruta
guacamayo
guácima
guadaña
guagua

guagua sf.

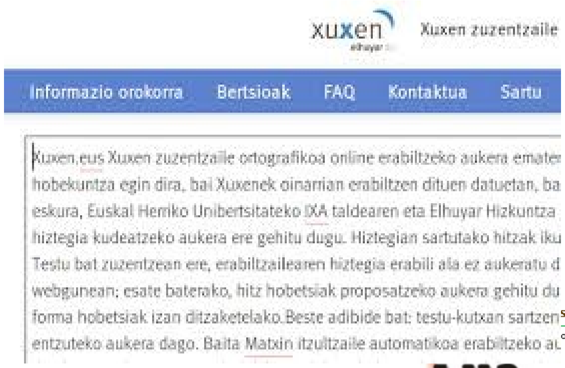
En las Antillas, vehículo público de gran capacidad para transportar pasajeros.
Para llegar al trabajo tengo que coger la ***guagua****.*
Sin. ómnibus, autobús

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

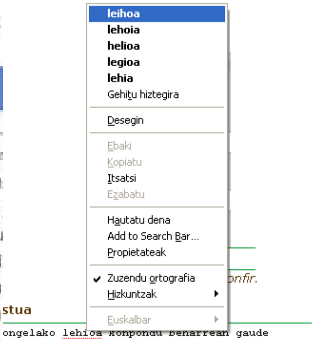


Morphology / Spelling corrector

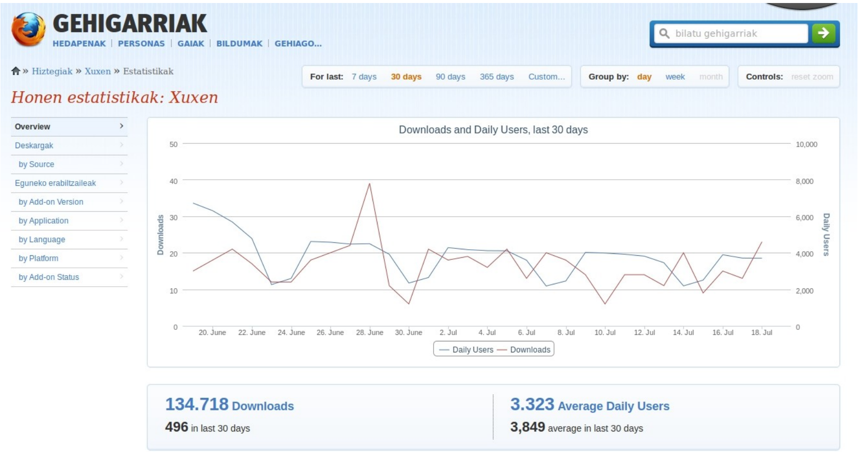
- Computational morphology is compulsory for most of the languages:
  - Dictionary + word-grammar
- The spelling corrector is a key application (only with big soft companies??)
  - Basic tool for students, journalists and writers
  - Key for standardization
- Integration/online: Microsoft, LibreOffice,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



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



Honen estatistikak: Xuxen

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

- Not perfect, it needs postedition,
  but the quality is very high
- Something incredible five years ago.
- From Basque to Spanish but also to French, English…
  … in both senses
- We have 4 free translation-services via web for Basque
- We are in a new world
  where the use of translation can be enormous
- We need human translators and with their help translation could be extended to many new fields and situations
- New horizons for under-resourced languages

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.

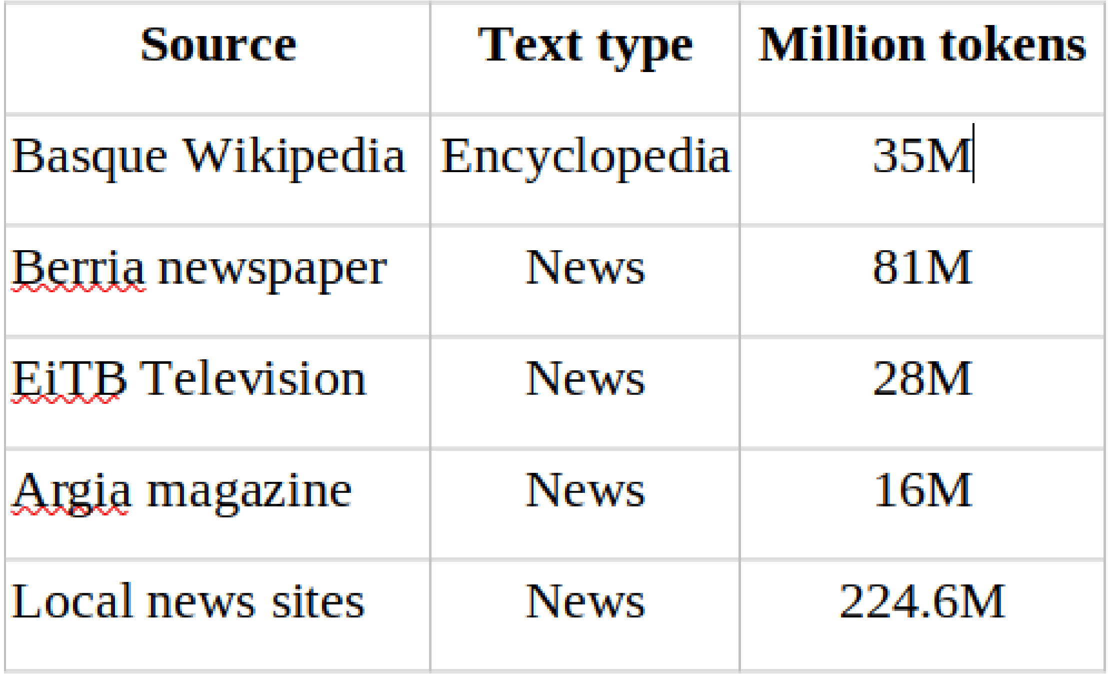
Eventhough they have been created:

- 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).

Spanish Plan for Language technology

- Plan for the Advancement of Language Technology
  2015-2020   90 M€

Spanish Plan for Artificial Intelligence

- 2021-2027
- Now being designed by the Spanish Government.

Basque Plan for Language technology

- 2021-2025 (?)
- Now being designed by the Basque Government.

PRIFYSGOL
BANGOR
UNIVERSITY

Symposiwm Academaidd Technolegau Iaith Cymru 2020
Wales Academic Symposium on Language Technologies 2020

Language Technology for Language Communities: An Overview based on Basque Experience 2020

Kepa Sarasola, Olatz Perez-de-Viñaspre. Iñaki Alegria
Ixa group. HiTZ Center
University of the Basque Country

Bangor 2020

ixa

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology