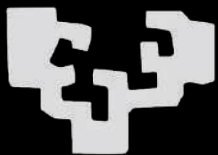


eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

# HiTZ

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

German Rigau Claramunt  
HiTZ Deputy Director  
National Coordinator of CLARIN and DARIAH

[hitz.eus](http://hitz.eus)



CLARIAH-ES

# Artificial Intelligence



# Artificial Intelligence



# Artificial Intelligence

## HAL's Legacy 2001's Computer as Dream and Reality

Edited by [David G. Stork](#)

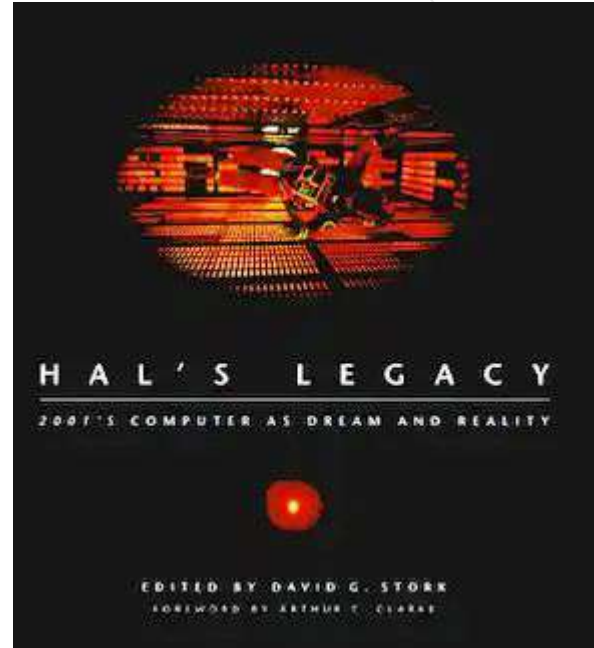
Foreword by [Arthur C. Clarke](#)

384 pp., 8 x 9 in, Paperback

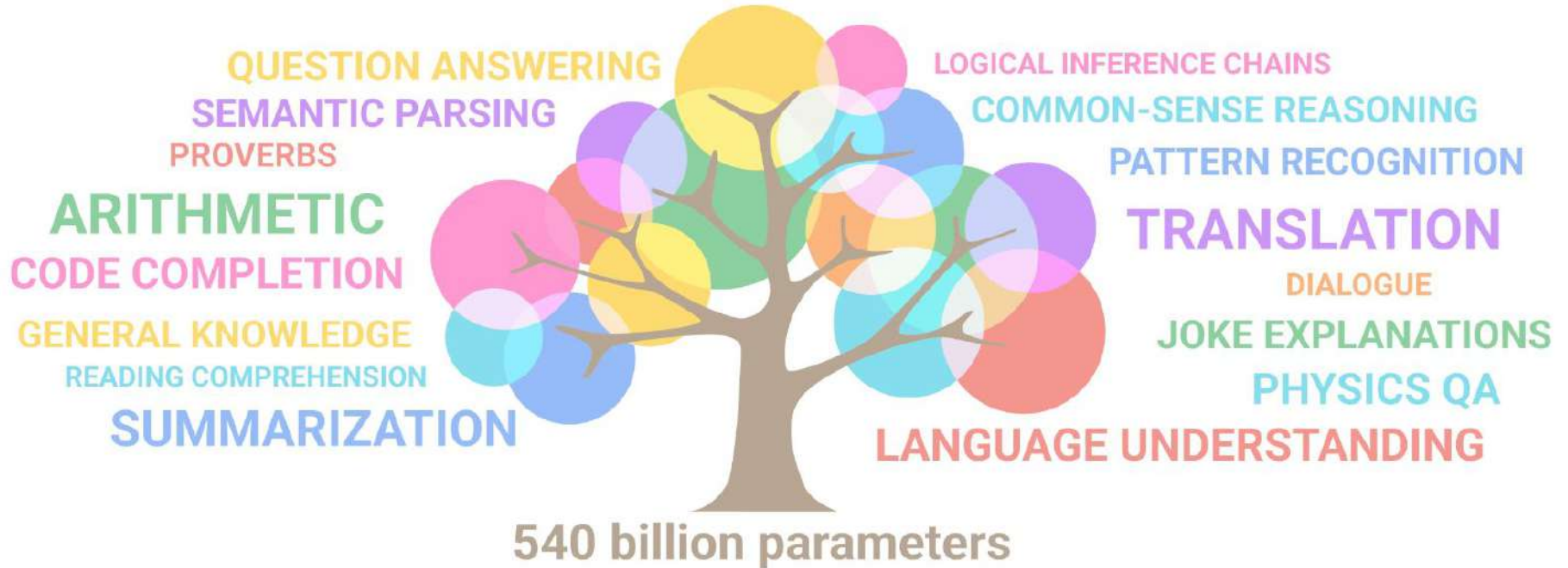
ISBN 9780262692113

Published: March 2, 1998

Publisher: The MIT Press



# Artificial Intelligence



**PaLM: Scaling Language Modeling with Pathways (2022)**

Aakanksha Chowdhery, + 67 authors



**Experts, Data, HPC**

# Large Language Models

## Explaining a joke

### Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

### Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

PaLM explains an original joke with two-shot prompts.

**PaLM: Scaling Language Modeling with Pathways (2022)**

Aakanksha Chowdhery, + 67 authors



**Experts, Data, HPC**

# Artificial Intelligence



## **LipNet: Sentence Level Lipreading (2016)**

Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas

# New AI paradigm

- **Accelerated** discovery cycle
- Impressive progress
  - **Superior** to humans in many tasks
- **Deep** and **Reinforcement Learning**
- Application in **Language**, Vision, Robotics
- **LLMs** require **Experts**, **Supercomputing**, **Data**
  - [LLaMA-2](#) (70B):
    - 2 trillion tokens of training data (~20 million books)
    - 1.7M GPU hours A100-80Gb (~ 1000 A100 71 days)







# Preface

“Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas.”

- Mario Benedetti

*“Just when we thought we had all the answers, suddenly all the questions changed.”*



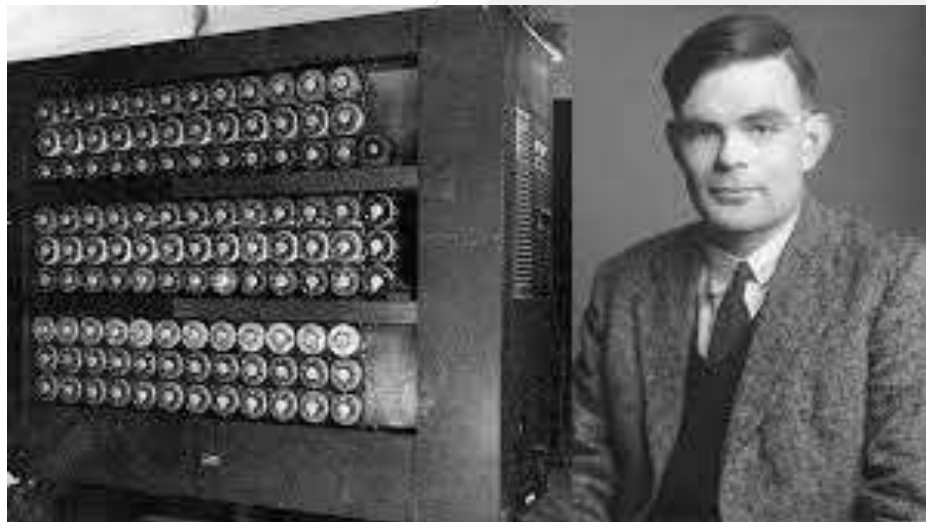


**1**

# **(Brief) History of AI & NLP**

# History of IA & NLP

- 1950 [Turing test](#)



# History of IA & NLP

- 1950 Turing test
- 1956 [DSRPAI](#)



# History of IA & NLP

- 1950 Turing test
- 1956 [DSRPAI](#)
- 1960s Rule-based AI & NLP
- 1966 [ALPAC](#) report



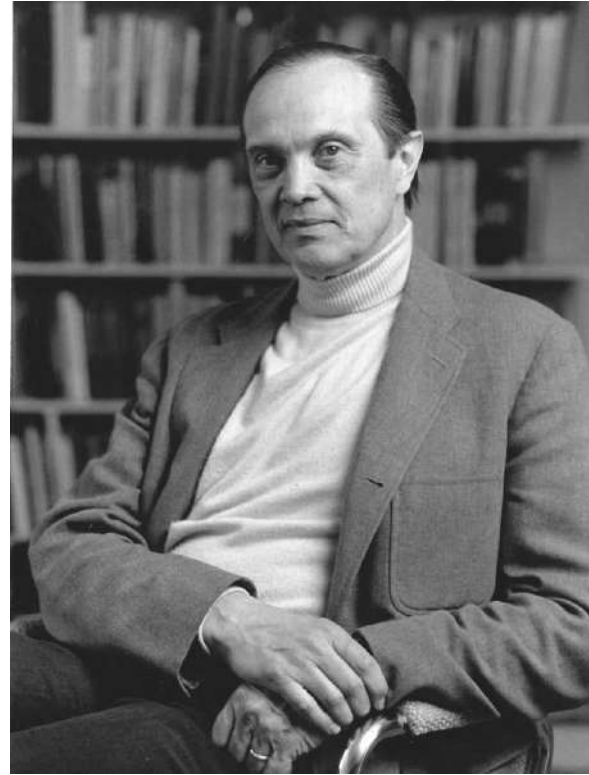
# History of IA & NLP

- 1950 Turing test
- 1956 [DSRPAI](#)
- 1960s Rule-based AI & NLP
- 1966 [ALPAC](#) report
- 1970s Noam Chomsky



# History of IA & NLP

- 1950 Turing test
- 1956 [DSRPAI](#)
- 1960s Rule-based AI & NLP
- 1966 [ALPAC](#) report
- 1970s Noam Chomsky
- 1992 [WordNet](#)





# History of IA & NLP

- 1950 Turing test
- 1956 [DSRPAI](#)
- 1960s Rule-based AI & NLP
- 1966 [ALPAC](#) report
- 1970s Noam Chomsky
- 1990s Statistical AI & NLP, Machine Learning
- 2010s Neural AI & NLP
  - 2013 Word Embeddings, 2017 **Transformers**, 2018 BERT, 2022 ChatGPT, ...
- 20XXs? AGI (singularity)



A large white circle is centered on a black background. To its left, there are several overlapping circles of varying shades of gray, with the number '2' in white. To its right, there are several concentric white circles of varying diameters.

**2**

# **New AI paradigm**

# Deep Learning



Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ( $v_i^T s_t$ ). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

## Deep visual-semantic alignments for generating image descriptions (2014)

A Karpathy, L Fei-Fei

# Deep Learning

## Zero-Shot Text-to-Image Generation

(2021) Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever

<https://openai.com/blog/dall-e/>

[DALL-E 2](#)  
[Midjourney](#)  
[Stable Diffusion](#)

...

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images ↘

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



Edit prompt or view more images ↘

TEXT PROMPT

a store front that has the word 'openai' written on it [...]

AI-GENERATED IMAGES



Edit prompt or view more images ↘

# Deep Learning: pictures and video



**v3 August 22**

**v4 November 22**

**v5 March 23**

<https://arstechnica.com/information-technology/2023/03/ai-imager-midjourney-v5-stuns-with-photorealistic-images-and-5-fingered-hands/>

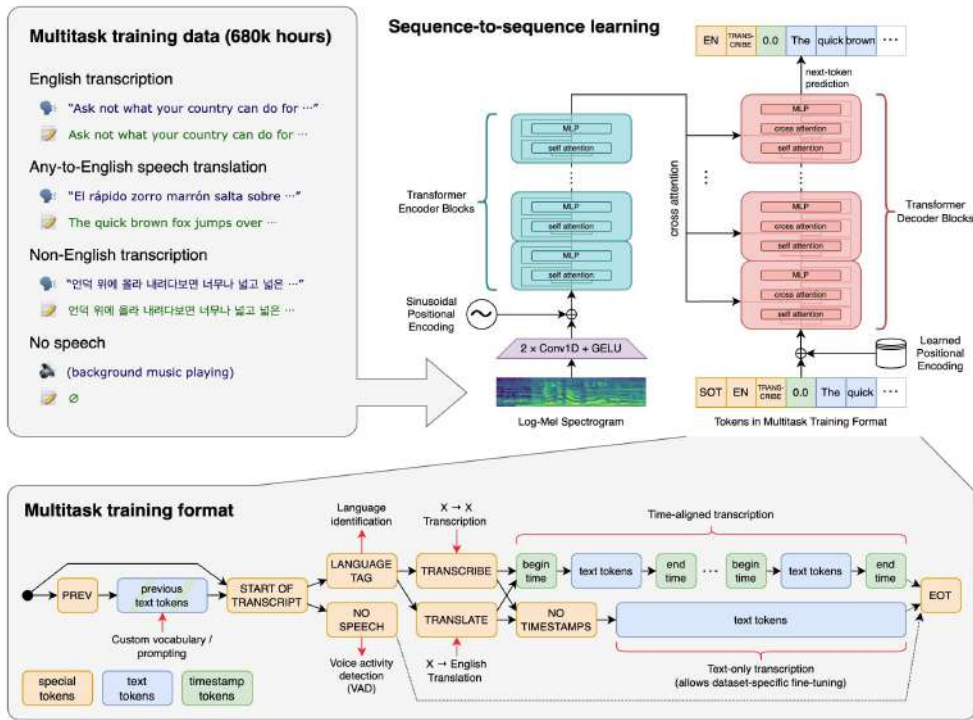
# Deep Learning: music, speech

## MusicLM: Generating Music From Text (2023)

13 authors

[MusicLM](#)  
[UniAudio](#)  
[Elevenlabs](#)  
[Whisper](#)  
[PromptTT2](#)

...



# Deep Learning: coding, programming

## [Evaluating Large Language Models Trained on Code \(2021\)](#)

58 authors

[Codex](#)  
[Copilot](#)  
[Codewhisperer](#)  
[ChatGPT](#)

...

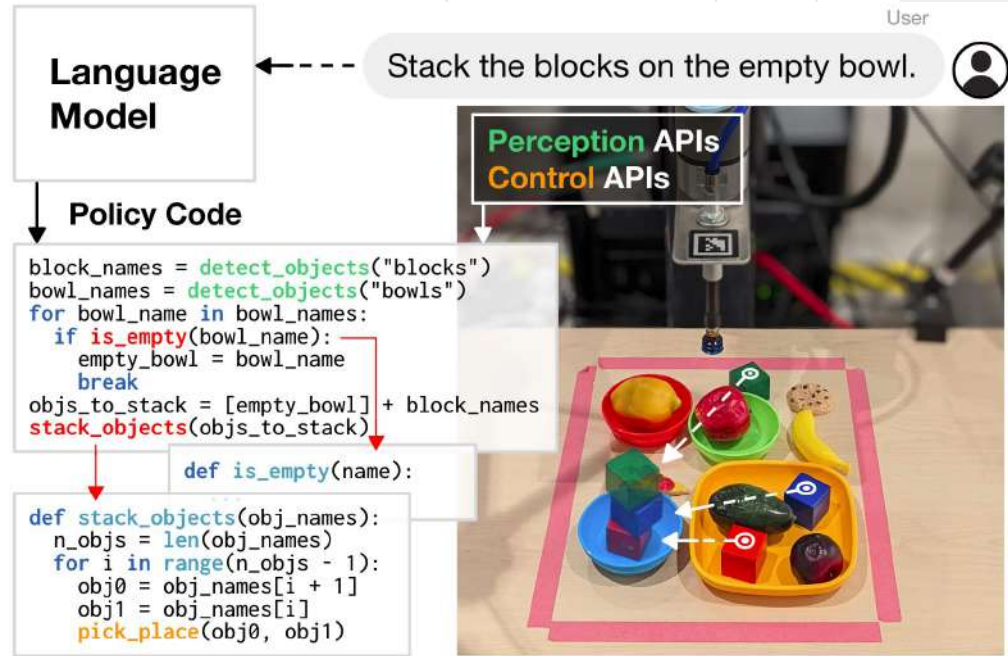
```
runtime.go course.fb time.js lsPrimeTest.java
1 package main
2
3 type Run struct {
4     Time int // in milliseconds
5     Results string
6     Failed bool
7 }
8
9 // Get average runtime of successful runs in seconds
10 func averageRuntimeInSeconds(runs []Run) float64 {
11     var totalTime int
12     var failedRuns int
13     for _, run := range runs {
14         if run.Failed {
15             failedRuns++
16         } else {
17             totalTime += run.Time
18         }
19     }
20
21     averageRuntime := float64(totalTime) / float64(len(runs) - failedRuns) / 1000
22     return averageRuntime
23 }
```

# Deep Learning: coding, programming

## Code as Policies: Language Model Programs for Embodied Control

(2022)

8 authors





# Deep Learning: 3D objects

[DreamFusion: Text-to-3D using 2D Diffusion](#) (2022)

4 authors

[POINT-E](#)  
[Magic3D](#)

...



# Deep Learning: videos

## [Make-A-Video: Text-to-Video Generation without Text-Video Data \(2022\)](#)

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman

[Make-a-video](#)

[VEED.IO](#)

[GliaStudio](#)

[Sora \(video\)](#)

...



# Deep Learning: dubbing



[Example of translating, dubbing and changing facial expressions](#) by [HeyGen](#)

[VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time](#) by MS



3

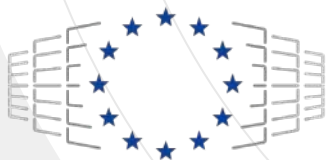
**HiTZ**

# HiTZ Basque Center for Language Technology

- [Reference center](#) in NLP/AI research (2019)
  - [Ixa](#) (founded 1988) and [Aholab](#) (1998)
  - 88 members, multidisciplinary
  - Yearly income in projects: ~3M €
- Research & Applications in EU, ES, EN, etc.
- [Academy](#): Master, Doctorate, [courses](#)
- Consultants for [Basque](#), [Spanish](#) and [European](#) LT plans
- National coordinators of two European research infrastructures
  - [CLARIN-EU](#) & [DARIAH-EU](#) : [CLARIAH-ES](#)
- Large LT projects
  - [ILENIA](#), [IKER-GAITU](#)
- **New** Chair in AI and Language Technology
  - Funded for 4 years by SEDIA under ENIA

# Latxa: LLM for Basque

- Generative LLM for Basque
- Open ([LLaMA-2 license](#))
- **Largest Basque LLM** built to day
  - 7B, 13B, 70B
- Largest LLM trained in Spain
- Obtains state-of-the-art results
  
- Our research solutions apply to most languages (other than the largest)

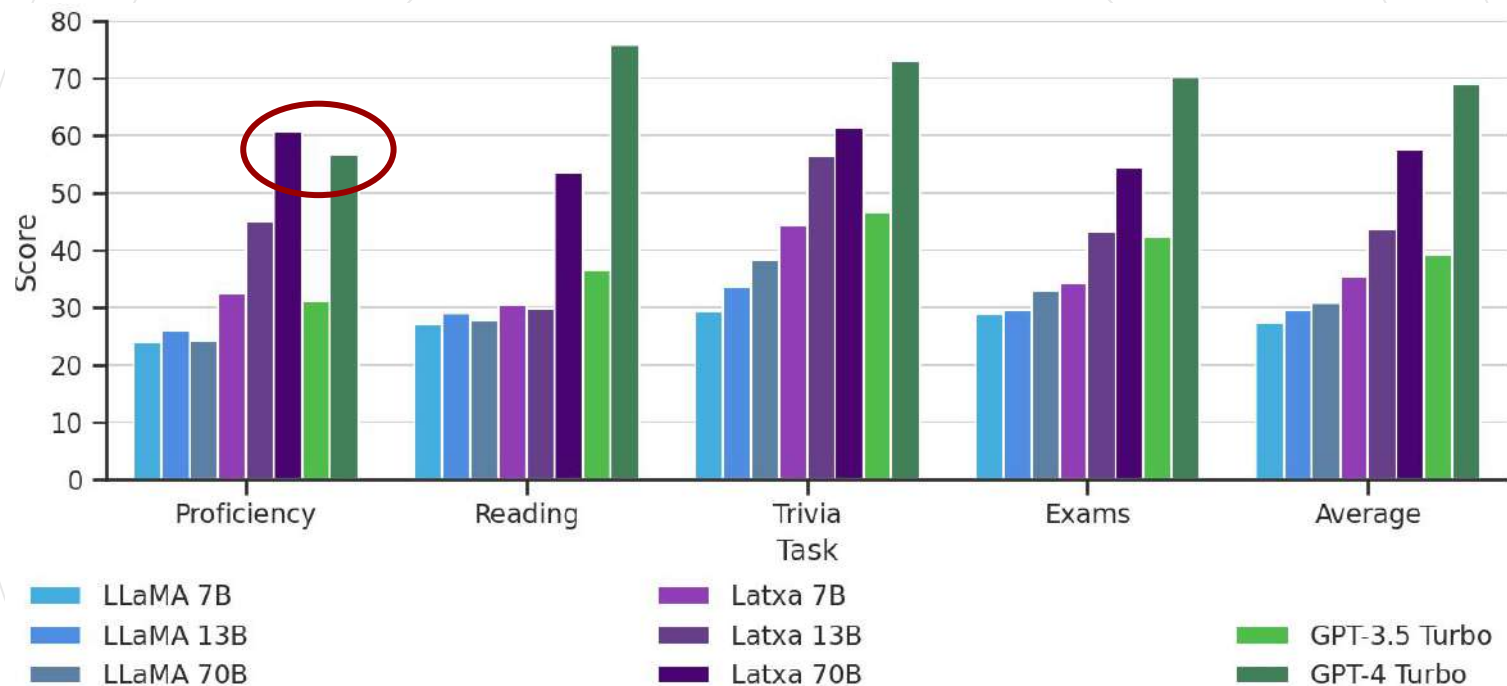


**EuroHPC**  
Joint Undertaking



**IENIA**   
IMPULSO DE LAS LENGUAS  
EN LA INTELIGENCIA ARTIFICIAL

# Latxa: Results



# Latxa: LLM for Basque

Chatbot

Kaixo

¿En que puedo ayudarte hoy?

Kaixo! Zertan lagundu zaitzaket gaur?

Necesito ayuda para desarrollar una función de Python que sume dos números, ¿puedes ayudarme?

Laguntza behar dut bi zenbaki batzen dituen python funtzio bat garatzen, lagunduko?

Noski! Hona hemen bi zenbaki batzen dituen python funtzioa:\n\npython\ndef add(num1, num2):\n return num1 + num2\n\ntFuntzio honek `num1` eta `num2` parametroak hartzen ditu eta haien batura itzultzen du. Adibidez, 5 eta 7 zenbakiak batuz gero, emaitza 12 izango litzateke.

¡Por supuesto! Aquí hay una función de python que suma dos números.











```
python def add(num1, num2): return num1 + num2
```

Esta función toma los parámetros num1 y num2 y devuelve su suma. Por ejemplo, si sumas los números 5 y 7, el resultado sería 12.





# Speech @ ILENIA (TTS examples)

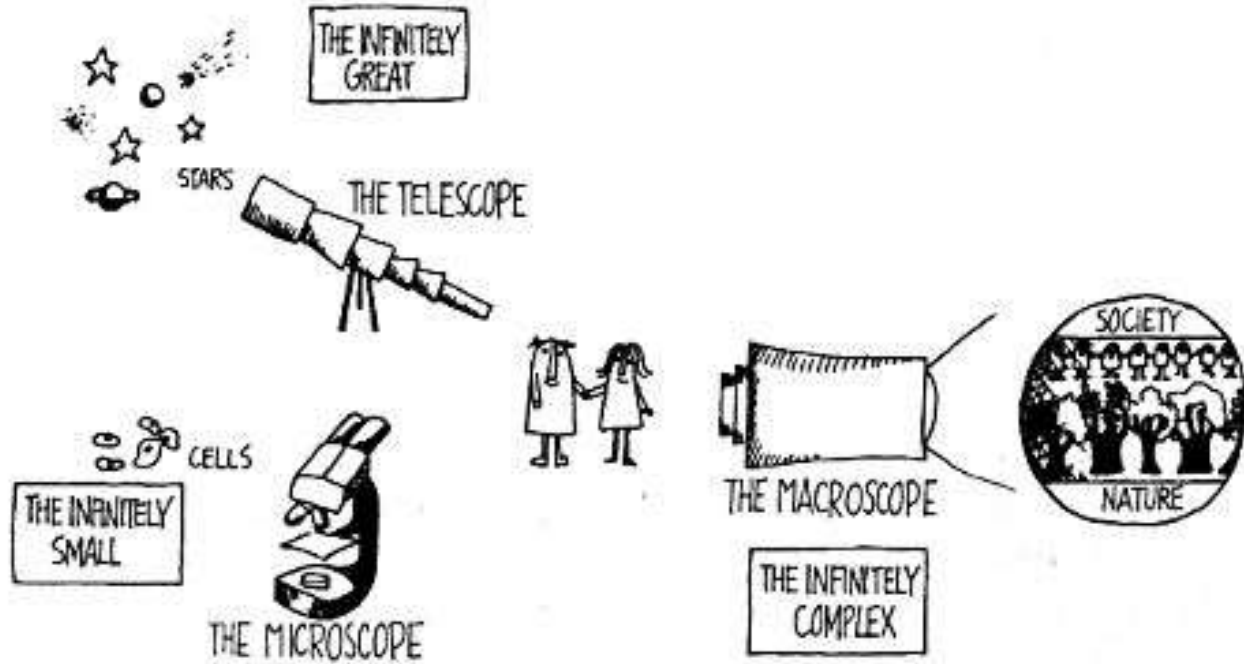
- Modelos implementados para TTS en euskera:
  - Ejemplos de voces alta calidad (grabaciones de estudio):
    - Hombre, mujer  
  - Ejemplos voces hechas con material de calidad no profesional:
    - Bingen (40 min adaptación de voz Hombre):  
    - Con [yourTTS](#) (zero-shot): 
  - Ejemplos voces personalización con banco de voces (100 frases):
    - Inma  
    - Ibon 
    - Infantil  



**4**

**European  
Research  
infrastructures**

# The Macroscope



De Rosnay, J. [The macroscope](#), Harper & Row, NY, 1979

# ESFRI – EOSC – SSHOC

Strategy Report on Research Infrastructures  
**ROADMAP 2021**



**EUROPEAN OPEN  
SCIENCE CLOUD**



#SSHOCaVocabulary  
#SSHOCifyCLARIN

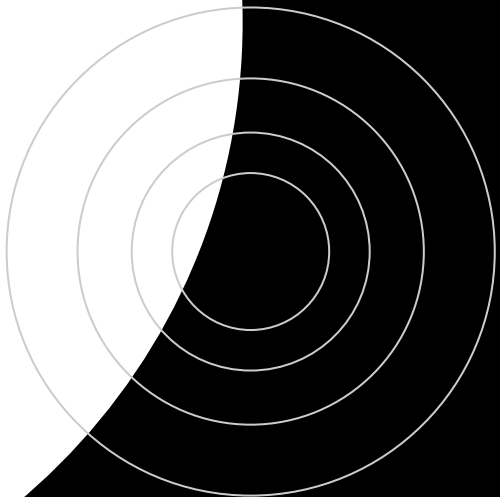
● Social Sciences  
● Humanities



5

# CLARIN

Common Language Resources and  
Technology Infrastructure



# CLARIN: Distributed Research Infrastructure

A screenshot of the CLARIN website homepage. The browser address bar shows 'https://www.clarin.eu'. The navigation menu includes 'About', 'Language Resources', 'Learn & Exchange', 'Events', 'News', and 'Contact'. The main content area features a large blue-tinted image of a person's profile with a keyboard and handwritten text in the background. The text on the page reads: 'The research infrastructure for language as social and cultural data'. Below this, a paragraph states: 'CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.'

**The research infrastructure for language as social and cultural data**

CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

# CLARIN: Distributed Research Infrastructure

24 countries

Network of 68 Centres ([certified](#)):

- *B-center*: data, services
- *K-center*: knowledge, advise
- *C-center*: metadata
- ...
- Experts, data, services, hw, sw, etc.



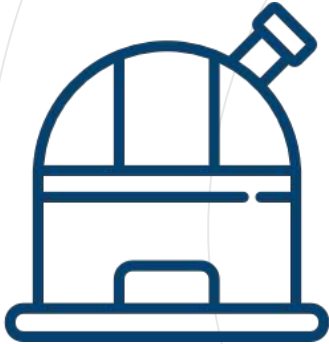
- ERIC members
- Observers
- Countries with participating centres
- Centre Providing Data
- Centre Providing Metadata
- Knowledge Centre



<https://www.clarin.eu/content/clarin-centres>

<https://www.clarin.eu/content/overview-clarin-centres>

# CLARIN: Research Infrastructure



*Virtual Language Observatory:*

- <https://vlo.clarin.eu/>



*Language Resource Switchboard:*

- <https://beta-switchboard.clarin.eu/>
- <https://switchboard.clarin.eu/>

Other : <https://www.clarin.eu/content/services>



# Language coverage of CLARIN technology

<https://vlo.clarin.eu>

almost 0.7 million components  
(tools, corpus...)

- Dutch (133,508)
- English (129,776)
- German (97,184)
- Bulgarian (34,271)
- Slovenian (31,035)
- French (21,265)
- Italian (13,585)
- Latin (13,826)
- ...

Resources in CLARIAH-ES languages

- Spanish (18,616)
- Galician (696)
- Basque (291)
- Catalan (208)



# CLARIN: Knowledge Infrastructure

- Knowledge K-centres:
  - <https://www.clarin.eu/content/knowledge-centres>
  - **Advice** from experts in various disciplines. They can also offer training, research guides, tool use guides, etc.
  - Organized by topics...
  - In Spain:
    - *Spanish K-Center*: languages in Spain
      - <https://www.clariah.es/k-centre/>
    - *IMPACT K-Center*: digitalization
      - <https://www.digitisation.eu/>
    - ...

"All K-centres have a helpdesk that will respond to requests within two working days."

# CLARIN: Knowledge (sharing) Infrastructure

- Learning Hub:
  - <https://www.clarin.eu/content/learning-hub>
  - Compilation of various educational resources on different topics.
  - Use of CLARIN, teaching materials, workshops, HD courses, guides and “good practices”, etc.
  - Network of Experts ...
- Service to store resources and tools
  - FAIR principles: Findable, Accessible, Interoperable, Reusable
    - <https://www.clarin.eu/fair>
  - Contact Center B with deposit service:
    - <https://www.clarin.eu/content/depositing-services>
  - Tools:
    - <https://github.com/clarin-eric/switchboard-tool-registry>

# CLARIN: networking, funding, projects

- CLARIN Funding Hub: scholarships for international collaboration in strategic tasks, exchange, training or mobility.
  - <https://www.clarin.eu/funding>
- Advice for applying for European projects (that are relevant to CLARIN).
- Networking: participation in events, nodes, k-centers, committees, etc.
  - <https://www.clarin.eu/events>



6

# DARIAH

Digital Research Infrastructure for the Arts  
and Humanities

# DARIAH: Distributed Research infrastructure

**DARIAH-EU**  
Celebrating the first decade

[ABOUT](#) [NETWORK](#) [ACTIVITIES](#) [TOOLS & SERVICES](#) [NEWS & EVENTS](#) [Q](#)

[Contact](#) [Helpdesk](#) [RSS](#)

The pan-European infrastructure for arts & humanities scholars

[Learn More About DARIAH](#)

## DARIAH Annual Open Access Book Bursary 2024: Call for manuscripts

We are pleased to announce the fourth round of the DARIAH annual book bursary for the Open Access...

[Read Post](#)

# DARIAH: Distributed Research Infrastructure

- 22 countries
- 4 strategic pillars
  - SSH Marketplace
- Regional Hubs
  - Near countries



# CLARIN: Distributed Research Infrastructure

4 strategic  
pillars



Marketplace of reusable tools, services, data and knowledge



Education & training



Working Groups: transnational & interdisciplinary



Research policy & foresight



# DARIAH (I) SSH Open Marketplace



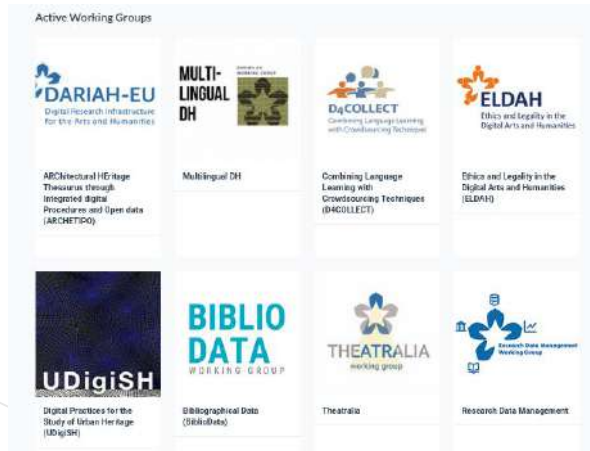
- Web portal to discover and access resources for CSH
  - <https://marketplace.sshopencloud.eu/>
  - Tools and services, Learning materials, Workflows, Datasets, Publications
- 3 Fundamental principles
  - Contextualization, Community, *Curatorship* (!)
- Entry point for SSH in EOSC
- Supported by CLARIN, DARIAH & CESSDA

# DARIAH (II) Education and training

- DARIAH-CAMPUS
  - <https://campus.dariah.eu/>
  - Discovery framework and hosting platform for learning resources.
  - Complete courses, lessons, conferences or videos.
- DARIAH-TEACH
  - <https://teach.dariah.eu/>
  - Open Educational Resources (OER) platform for educators and students of Digital Arts and Humanities.
- Digital Humanities Course Registry
  - <https://dhcr.clarin-dariah.eu/>
  - Platform that offers teaching activities in the field of digital humanities around the world.

# DARIAH (III) Working Groups

- Transnational and multidisciplinary
- 17 active WGs
- <https://www.dariah.eu/activities/working-groups-list/>
- Bottom-up design, grants (5K€), etc.





7

**CLARIAH-ES**

# CLARIAH-ES



- ERC numbers
- Observers
- Countries with participating centres
- Centre Providing Data
- Centre Providing Metadata
- Knowledge Centre



CLARIN



EUROPE



## Members and Cooperating Partners

DARIAH Members or Observers can be countries (EU Member States and Associated Countries) or intergovernmental organisations.

Alternatively, institutions from countries that are not yet DARIAH Members can participate as Cooperating Partners.

Currently, DARIAH has 22 Members and 19 Cooperating Partners in 11 non-Member countries: Egypt, Finland, Hungary, Iceland, Latvia, Norway, Romania, Slovakia, Sweden, the United Kingdom, and the United States of America.

- Member Countries
- Observer Countries
- Cooperating Partners



**CLARIN**: Common **L**anguage **R**esources and **T**echnology **I**nfrastructure  
**DARIAH**: Digital **R**esearch **I**nfrastructure for the **A**rts and **H**umanities

# CLARIAH-ES: CLARIN & DARIAH



Distributed RI: **12** centres

- Community
- Data
- Tools & services
- IT-HPC-infrastructure

- EHU ([HiTZ](#)) : *German Rigau & Mikel Iruskieta*
- UAlicante ([BVMC](#)) : *Manuel Marco*
- USC ([Instituto da Lingua Galega](#) y [CiTIUS](#)) : *Elisa Fernandez*
- UNED ([LINDH](#) y [LENAR](#)) : *Salvador Ros*
- UJAEN ([CEATIC](#)) : *Arturo Montejo-Ráez*
- UCM/MAD : *Dolores Romero + 5 public universities CAM*
- CSIC ([Centro de Ciencias Humanas y Sociales](#)) : *Daniel Riaño*
- [BSC-CNS](#) : *Maite Melero*
- [BNE](#): *Elena Sánchez*
- ULPGC ([IATEXT](#)): *Francisco Carreras*
- [Scayle](#): *Alvaro Fanego*
- [Dialnet](#): *Elena López Tamayo*
- **0.2M€** MICIN (fee) & **26+** Full Time Equivalent > **1.4M€** (in-kind)
- Strategic research network [CLARIAH-ES](#) **60K€** (2023-2024 MICIN ...)

# CLARIAH-ES office @ [HiTZ](#)



- Coordination with [CLARIN-EU](#)
- Coordination with [DARIAH-EU](#)
- [CLARIAH-ES](#) coordination
- [CLARIAH-EUS](#) coordination
- Work force
  - NC CLARIN and NC DARIAH - *German Rigau*
  - 1 FTE HiTZ - *4+ Technical Staff part time*
    - *Esther, Kike, Amaia, ... also Aritz, Begoña, etc.*
  - 1 FTE UPV/EHU - *Mikel Iruskieta & Xabier Arregi half time each*
  - 1 FTE Basque Government - *Ainara Estarrona*
  - 1 FTE Diputación Gipuzkoa - *Xabier Goenaga*
  - ...



- Infraestructura distribuida (estructural): Centros vs. Nodos
  - UPV/EHU => [CLARIAH-EUS](#)
  - UJaen => [CLARIAH Andalucía](#)
  - UCM => [CLARIAH-CM](#)
  - USC => CLARIAH-GAL
  - BSC => CLARIAH-CAT
  - UA => CLARIAH-VAL
  - UNED => CLARIAH-UNED
  - CSIC => CLARIAH-CSIC
  - BNE => CLARIAH-BNE
  - ULPGC => CLARIAH-Canarias

# Summary

- **Unthinkable AI & NLP** applications just a few years ago
- Even more **amazing** results in the near future (day by day)
- **CLARIAH-ES research infrastructure:** CLARIN and DARIAH
- **Coordination and Cooperation**

# Future of CLARIAH-ES



# Epílogue

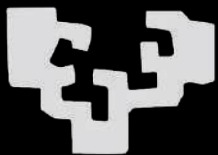
“No hemos llegado muy lejos  
pero hemos avanzado mucho.”

- Mikel Iruskieta

*“We haven't come very far  
but we've made a lot of progress.”*



eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

# HiTZ

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

German Rigau Claramunt  
HiTZ Deputy Director  
National Coordinator of CLARIN and DARIAH

[hitz.eus](http://hitz.eus)

CLARIAH-ES