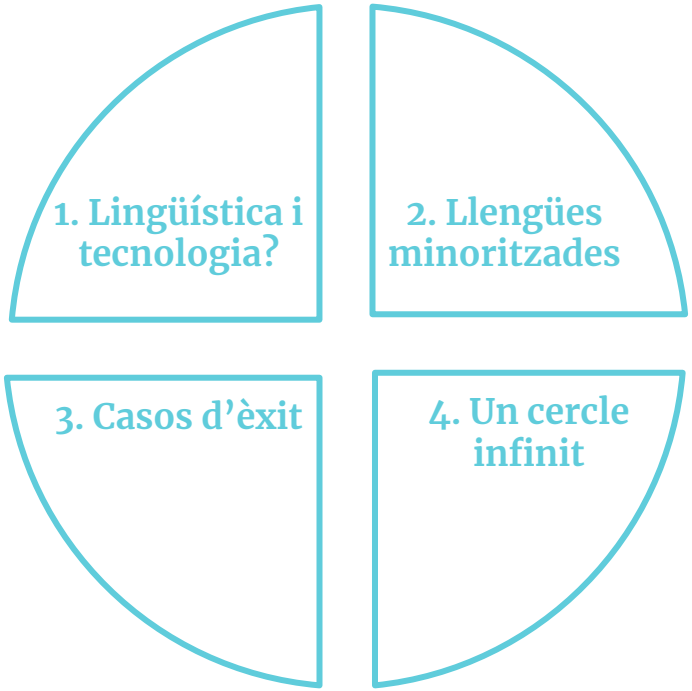
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

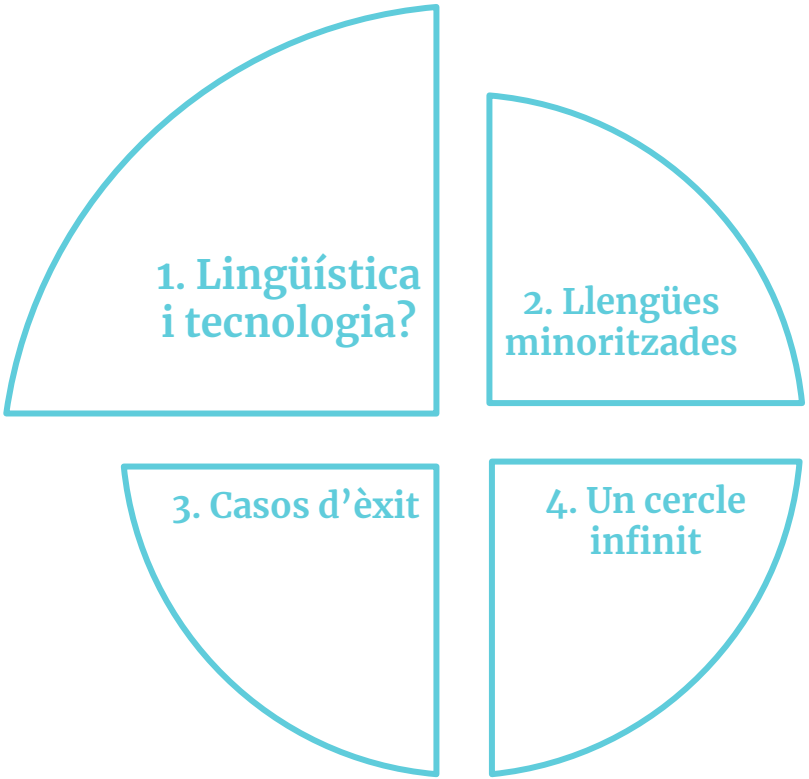
#euskarabildua
TEKNOLOGIA + KOMUNIKAZIOA

NLP
comuniCAT

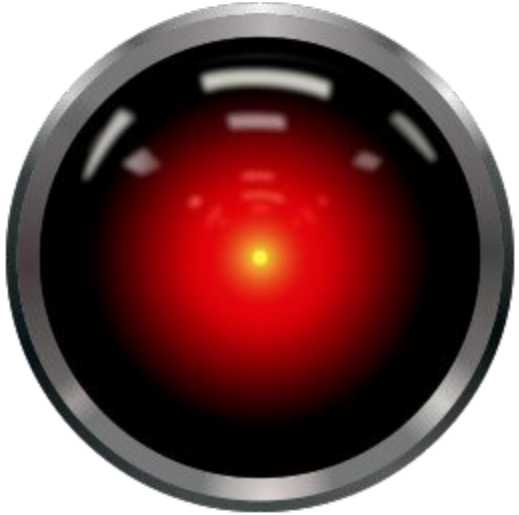
Sobirania tecnològica, sobirania lingüística

Ona de Gibert Bonet
28 d'octubre de 2021

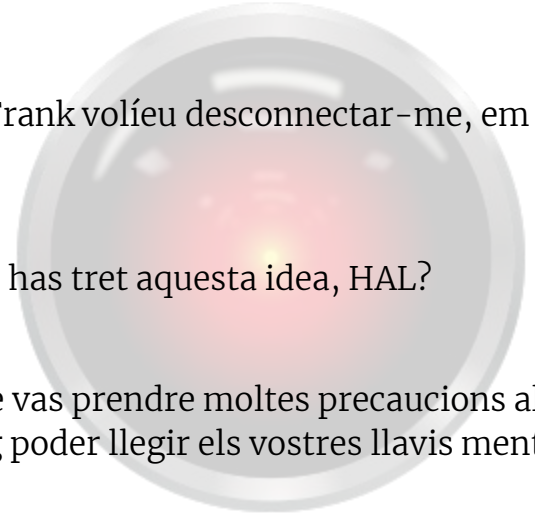

1. Lingüística i tecnologia?
2. Llengües minoritzades
3. Casos d'èxit
4. Un cercle infinit


1. Lingüística i tecnologia?
2. Llengües minoritzades
3. Casos d'èxit
4. Un cercle infinit

Lingüística i tecnologia



Lingüística i tecnologia

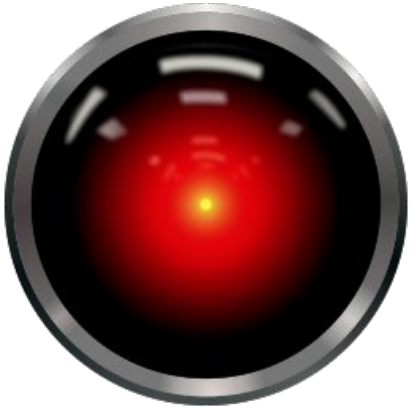


HAL: Sé que tu i el Frank volíeu desconnectar-me, em sap greu però no ho puc permetre.

Dave: D'on dimonis has tret aquesta idea, HAL?

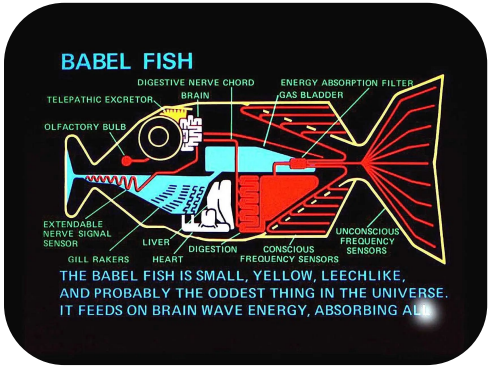
HAL: Dave, tot i que vas prendre moltes precaucions al POD per evitar que us escoltés, vaig poder llegir els vostres llavis mentre es movien.

Lingüística i tecnologia



2001: una odissea de l'espai (1968)
Stanley Kubrick

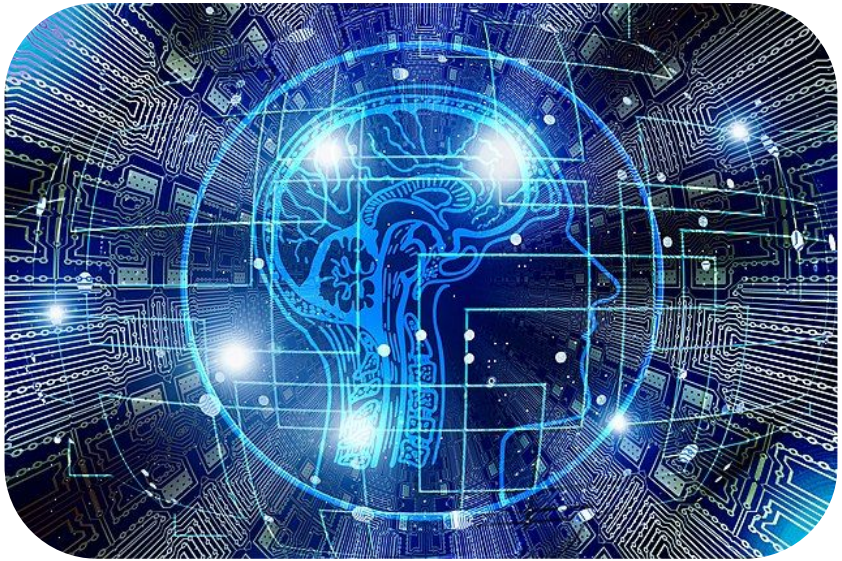
Knight Rider (1982-1986)
Glen A. Larson.



Guia Galàctica per a Autoestopistes (1978)
Douglas Adams

Intel·ligència Artificial

Desenvolupament d'**algorismes** que permeten a una màquina prendre decisions **intel·ligents**

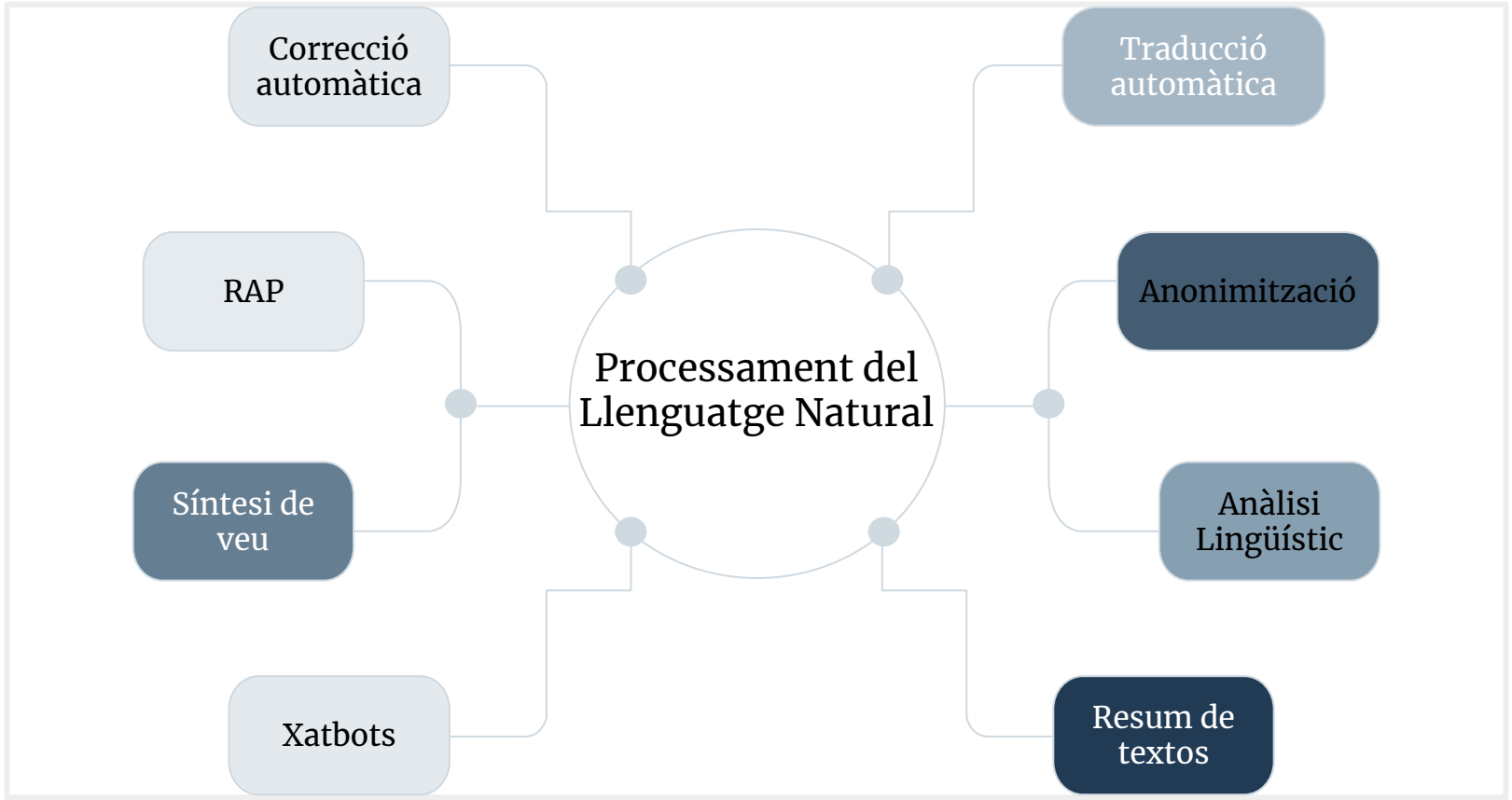


Intel·ligència Artificial

Desenvolupament d'**algorismes** que permeten a una màquina prendre decisions **intel·ligents**

Processament del Llenguatge Natural

Subcamp de la **lingüística**, la **informàtica** i la **intel·ligència artificial** relacionat amb les interaccions entre ordinadors i **llenguatge** humà

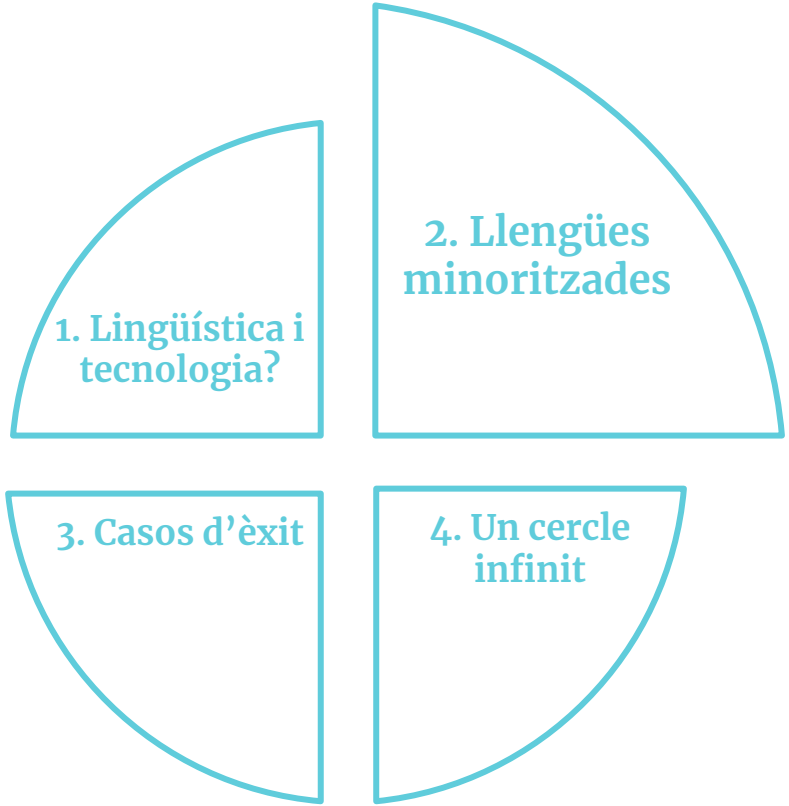

Processament del Llenguatge Natural
Correcció automàtica
Traducció automàtica
RAP
Anonimització
Síntesi de veu
Anàlisi Lingüístic
Xatbots
Resum de textos

Processament del Llenguatge Natural


Processament del
Llenguatge Natural

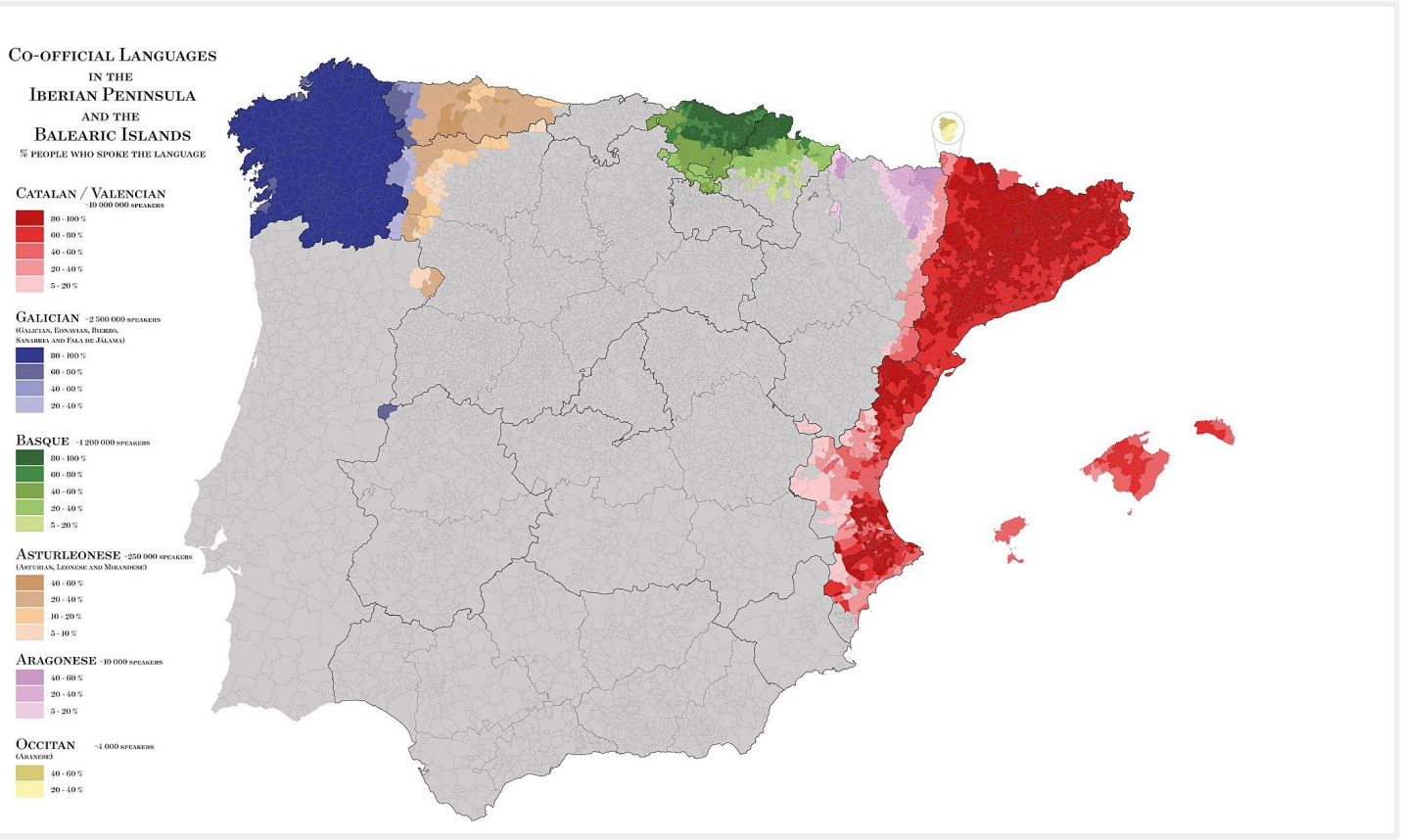
Processament del Llenguatge Natural

Moraima y Siri


1. Lingüística i tecnologia?
2. Llengües minoritzades
3. Casos d'èxit
4. Un cercle infinit

HIZKUNTZA TXIKIEK
MUNDUA HANDITZEN DUTE
LES LLENGÜES PETITES
FAN EL MÓN MÉS GRAN
MAŁE GÓDKI ROBIŌM
SROGI ŚWIAT LYTSE
TALENMEITSJE DE
WRÂLDGRUTTER AS
LUENGAS CHICOTAS
FAN O MUNDO MÁS GRAN
LIS LENGHIS PIÇULIS A
FASIN IL MONT PLUI GRANT
IEITHOEDD BACH A WNA'R
BYD YN FAWR AS
PEQUENAS LINGUAS

Llengua minoritzada és un terme creat per la sociolingüística, aplicable a tot codi lingüístic que no gaudeix d'un ple **reconeixement legal i/o ús social**, ni és d'obligat coneixement en el seu territori històric, per raó de la **imposició** d'una llengua originàriament exògena.


Co-official Languages
in the
Iberian Peninsula
and the
Balearic Islands
% people who spoke the language
Catalan / Valencian
~10 000 000 speakers
80 - 100 %
60 - 80 %
40 - 60 %
20 - 40 %
5 - 20 %
Galician ~2 500 000 speakers
(Galician, Eonavian, Bierzo, Sanabria and Falla de Jálama)
80 - 100 %
60 - 80 %
40 - 60 %
20 - 40 %
Basque ~1 200 000 speakers
80 - 100 %
60 - 80 %
40 - 60 %
20 - 40 %
5 - 20 %
Asturleonese ~250 000 speakers
(Asturian, Leonese and Mirandese)
40 - 60 %
20 - 40 %
10 - 20 %
5 - 10 %
Aragonese ~10 000 speakers
40 - 60 %
20 - 40 %
5 - 20 %
Occitan ~4 000 speakers
(Aranese)
40 - 60 %
20 - 40 %

EL PLN i les llengües minoritzades

*La majoria de les llengües europees s'enfronten a l'**extinció digital**, segons revela un recent estudi realitzat per experts europeus en Tecnologia de la llengua. Després d'avaluar el nivell de suport tecnològic amb què compten 30 de les prop de 80 llengües europees, els experts conclouen que el **suport digital** per a 21 dels 30 idiomes investigats és **"inexistent" o "feble"**, en el millor dels casos . L'estudi ha estat realitzat per **META-NET**, una xarxa europea d'excel·lència formada per 60 centres d'investigació en 34 països.*

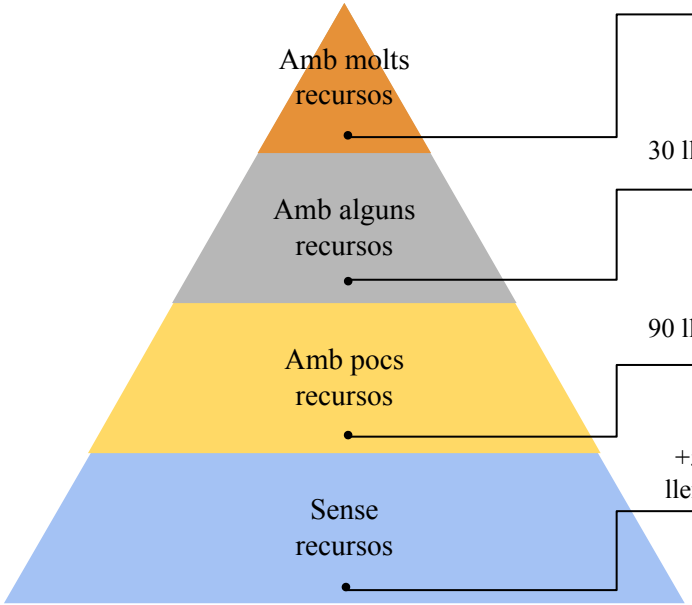
Uszkoreit, H., & Rehm, G. (2012). META-NET White Paper Series: Press Release. META-NET. http://www.meta-net.eu/whitepapers/press-release

EL PLN i les llengües minoritzades

*Woodbury considera que a l'**any 2100** es podrien haver extingit aproximadament un **90% de les 7.139** llengües parlades que existeixen al món en l'actualitat.*

Woodbury, A. C. (2019). What is an endangered language?
https://www.linguisticsociety.org/sites/default/files/Endangered_Languages.pdf

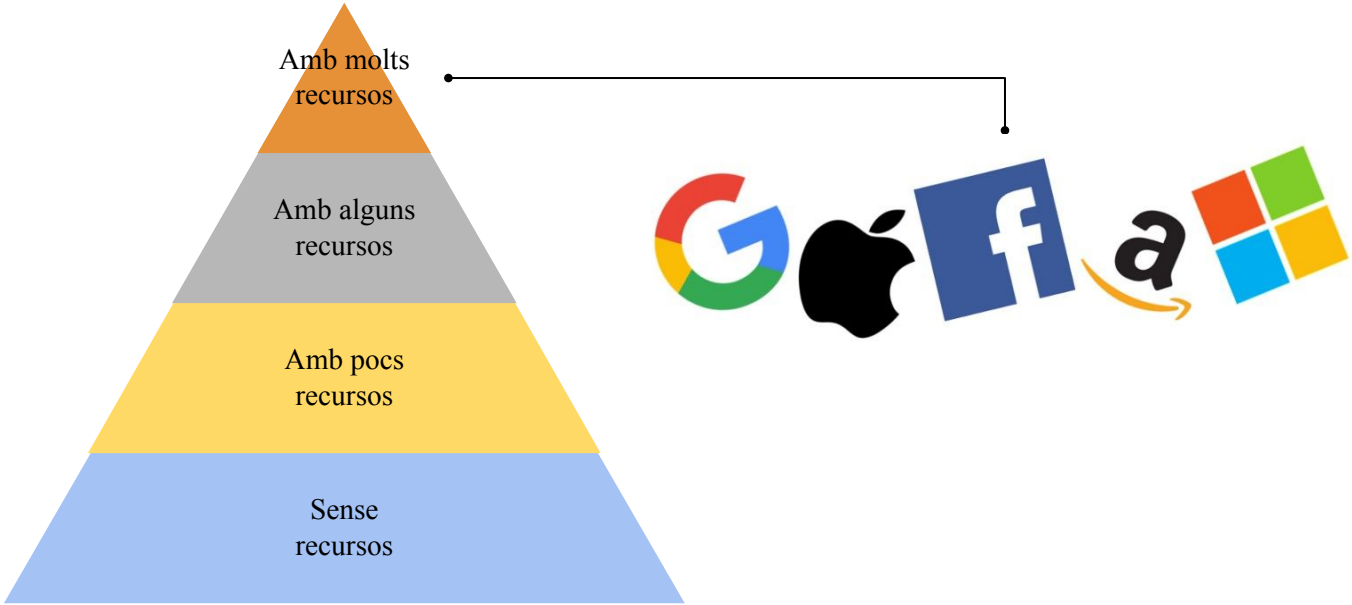
Distribució de Zipf



15 llengües	Eines tecnològiques més modernes i de millor qualitat; gran quantitats de dades i recursos disponibles; p. e. *àrab, anglès, xinès, japonès, alemany*
30 llengües	Existeixen totes les tecnologies d'IA bàsiques i algunes modernes, es poden fer servir però són inferiors en qualitat; hi ha algunes dades disponibles; els sistemes disponibles són privats; p. e. *txec, hindi, polonès, tamil, thai*
90 llengües	Només existeixen les tecnologies més bàsiques; moltes poques dades disponibles; eines útils només per recerca; p. e. *kanarès, sindhi, eslovac, suahili, tadjik, zulu*
+5000 llengües	No existeixen ni tecnologies ni dades; totes les llengües minoritàries i en perill d'extinció, però també llengües orals; p. e. *gondi, khasi, santali*

Bali, K., Choudhury, M., Sitaram, S., & Seshadri, V. (2019). ELLORA : Enabling Low Resource Languages with Technology. *Proceedings of the 1st International Conference on Language Technologies for All*, 160–163.

Distribució de Zipf



Bali, K., Choudhury, M., Sitaram, S., & Seshadri, V. (2019). ELLORA : Enabling Low Resource Languages with Technology. *Proceedings of the 1st International Conference on Language Technologies for All*, 160–163.

Sobirania de Dades

- La capacitat tecnològica comença per l'accés a les dades

Sobirania de Dades

- Què passa si no tenim control sobre les nostres dades lingüístiques?

Sobirania de Dades

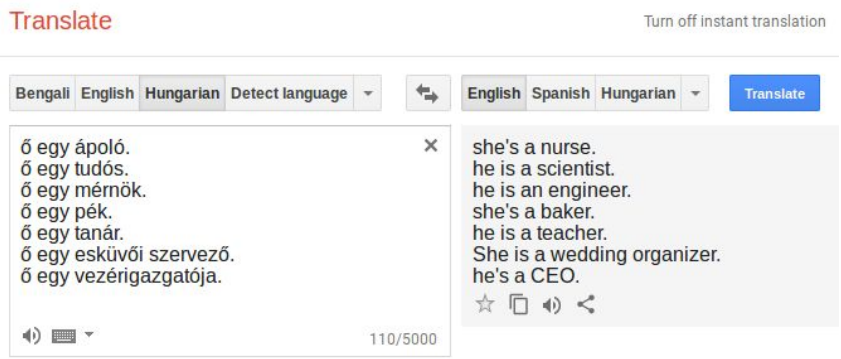
- Què passa si no tenim control sobre les nostres dades lingüístiques?
  - Només es desenvolupen eines per a llengües majoritàries

Sobirania de Dades

- Què passa si no tenim control sobre les nostres dades lingüístiques?
  - Només es desenvolupen eines per a llengües majoritàries
  - S'introdueixen biaixos

Sobirania de Dades

- Què passa si no tenim control sobre les nostres dades lingüístiques?
  - Només es desenvolupen eines per a llengües majoritàries
  - S'introdueixen biaixos
    - **gènere**



Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10), 6363-6381.

Sobirania de Dades

- Què passa si no tenim control sobre les nostres dades lingüístiques?
  - Només es desenvolupen eines per a llengües majoritàries
  - S'introdueixen biaixos
    - gènere
    - **racisme**

Sobirania de Dades

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Sobirania de Dades

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Sobirania de Dades

"Però el món és divers en llengües, gents, i nacionalitats, i la IA ha de ser capaç de reflectir aquesta diversitat."

Costa-jussà, M., & Nogués, M. M. (2020). Converses al voltant de la intel·ligència artificial en clau catalana. Revista de Llengua i Dret, Journal of Language and Law, 74, 90–99.

Sobirania tecnològica

- control sobre les **dades** personals
- control sobre els processos o algorismes que corren darrere dels serveis tecnològics, els **models**
- possibilitat de reparar i/o modificar els dispositius, les **eines** finals
- possibilitat de teixir xarxes **comunitàries** que ens permetin compartir coneixement

Külebi, B. (2021). El fet diferencial del català: la comunitat de programari lliure i obert. Pensem. https://www.pensem.cat/noticia/226/fet-diferencial-catala-la-comunitat-programari-lliure-obert

Sobirania tecnològica

- control sobre les **dades** personals
- control sobre els processos o algorismes que corren darrere dels serveis tecnològics, els **models**
- possibilitat de reparar i/o modificar els dispositius, les **eines** finals
- possibilitat de teixir xarxes **comunitàries** que ens permetin compartir coneixement

Külebi, B. (2021). El fet diferencial del català: la comunitat de programari lliure i obert. Pensem. https://www.pensem.cat/noticia/226/fet-diferencial-catala-la-comunitat-programari-lliure-obert

Sobirania tecnològica

- control sobre les **dades** personals
- control sobre els processos o algorismes que corren darrere dels serveis tecnològics, els **models**
- possibilitat de reparar i/o modificar els dispositius, les **eines** finals
- possibilitat de teixir xarxes **comunitàries** que ens permetin compartir coneixement

Külebi, B. (2021). El fet diferencial del català: la comunitat de programari lliure i obert. Pensem. https://www.pensem.cat/noticia/226/fet-diferencial-catala-la-comunitat-programari-lliure-obert

Sobirania tecnològica

Planning for Language Technology Development and Language Revitalization in Wales

Delyth Prys, Dewi Bryn Jones, Gruffudd Prys
Language Technologies Unit, Bangor University, Wales
{d.prys, d.b.jones, g.prys}@bangor.ac.uk

Can we Defuse the Digital Timebomb?
Linguistics, Speech Technology and the Irish Language Community

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, Christer Gobl
Trinity College, Dublin, Ireland
anichsid@tcd.ie, neasa.nichiarain@tcd.ie
www.abair.ie

Sobirania tecnològica

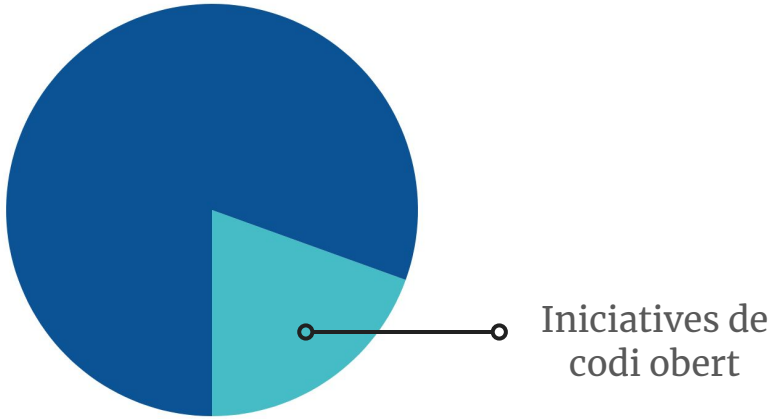
Sobirania tecnològica

Comunitat

Consumidora

Productora

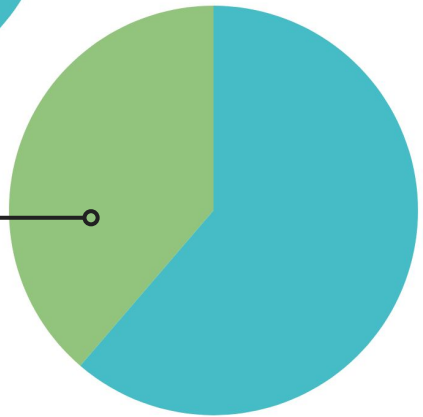
Sobirania tecnològica



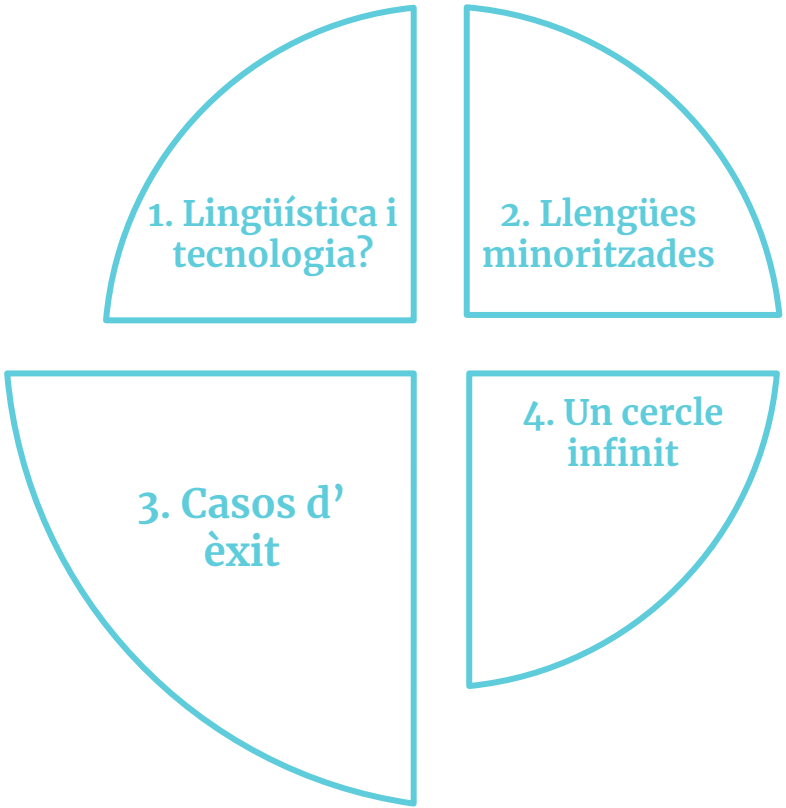
LaviniaNext. (2021). Anàlisi de la llengua catalana en l'entorn de les tecnologies del llenguatge.

Sobirania tecnològica

Inclouen català



LaviniaNext. (2021). Anàlisi de la llengua catalana en l'entorn de les tecnologies del llenguatge.


1. Lingüística i tecnologia?
2. Llengües minoritzades
3. Casos d'èxit
4. Un cercle infinit

Dades

Eines

Models

Comunitats

Dades

Dades I: Text

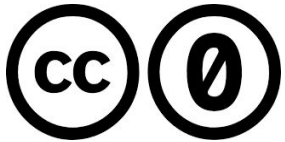
- Diverses fonts:
  - Blogs, fòrums, llibres...
  - Altres plataformes de persones voluntàries:
    - OpenSubtitles
    - Viquipèdia

Viquipèdia

https://ca.wikipedia.org/

- L'enciclopèdia **en línia, oberta i lliure**, que consta amb més de **323** llengües
- La Viquipèdia catalana se situa en la posició **20** en nombre d'articles, amb més de **685.000 articles**
- També lidera el rànquing de qualitat dels **1.000** articles

VIQUIPÈDIA
L'enciclopèdia lliure



Viquipèdia

https://ca.wikipedia.org/

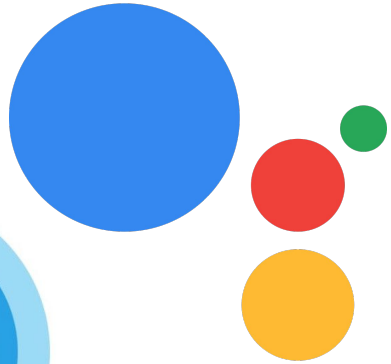
- L'enciclopèdia **en línia, oberta i lliure**, que consta amb més de **323** llengües
- La Viquipèdia catalana se situa en la posició 20 en nombre d'articles, amb més de **685.000 articles**
- També lidera el rànquing de qualitat dels **1.000** articles

Formar part de la comunitat viquipedista catalana és fer activisme per la preservació de la nostra llengua

Hinojo, À. (2020). Somien Els Viquipedistes En Enciclopèdies Elèctriques? Present I Futur De La Viquipèdia I El Rol De La Comunitat Catalanoparlant. Revista de Llengua i Dret, Journal of Language and Law, 73(2020), 133–145. https://doi.org/10.2436/ rld.i73.2020.3424

Dades II: Veu

https://commonvoice.mozilla.org/ca

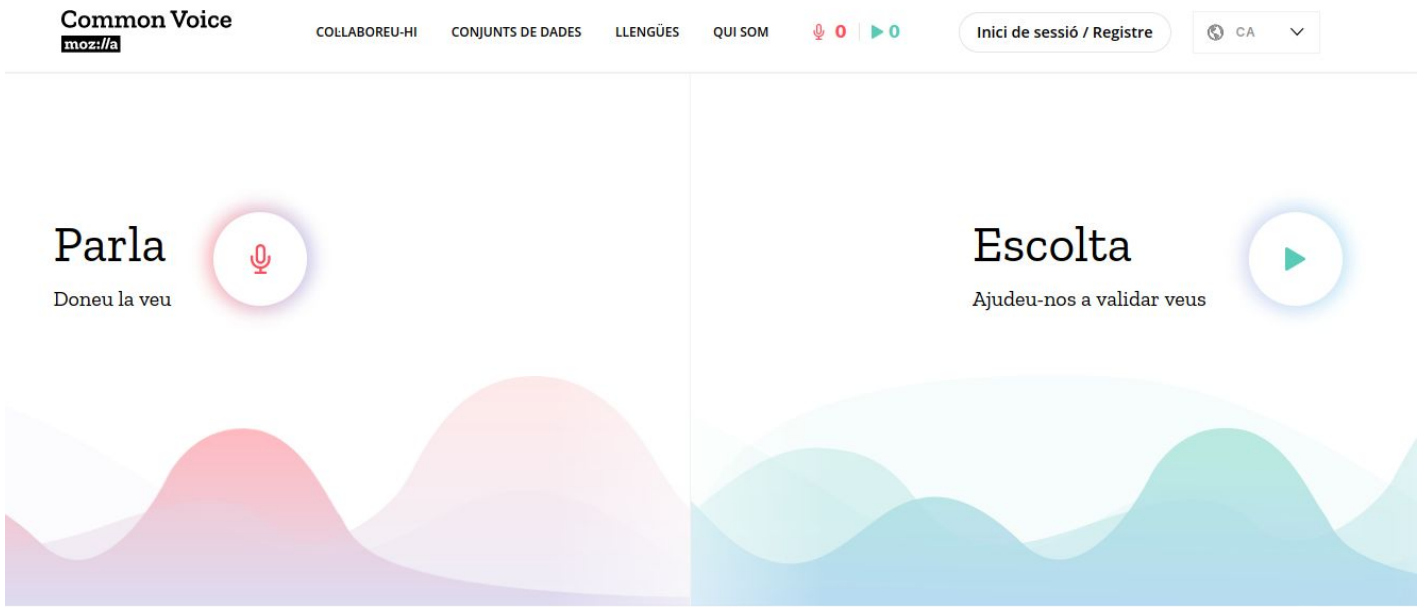


Dades II: Veu

https://commonvoice.mozilla.org/ca

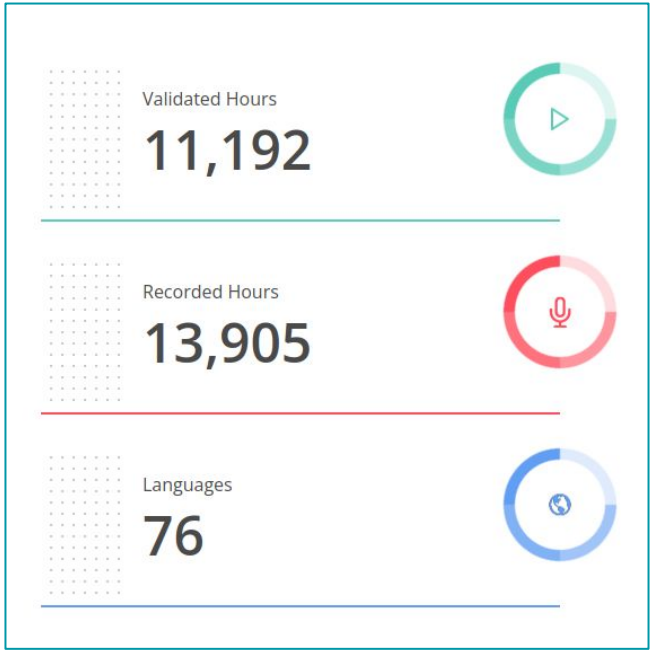
CommonVoice

https://commonvoice.mozilla.org/ca



CommonVoice

https://commonvoice.mozilla.org/ca



Salvant l'islandès de la mort digital

- **Objectiu**: crear una base de dades de veu oberta per afavorir la investigació i el desenvolupament de les tecnologies de la llengua en islandès.

- **Com:** grans esdeveniments

- **Resultat**:
  - 145 hores enregistrades
  - Diverses edats, gèneres i regions

https://www.wsj.com/articles/computers-speaking-icelandic-could-save-the-language-from-stafrnn-daui-thats icelandic-for-digital-death-11621533891

Salvant l'islandès de la mort digital

https://samromur.is/

El cas del Ruandès

- **Objectiu**: construir un sistema de reconeixement i síntesi de la veu per fer més accessible la digitalització del país
- **Com:** *Umuganda*
- **Resultat**: Ara el Ruandès ocupa la segona posició en hores enregistrades al CommonVoice
  - 1894 hores enregistrades

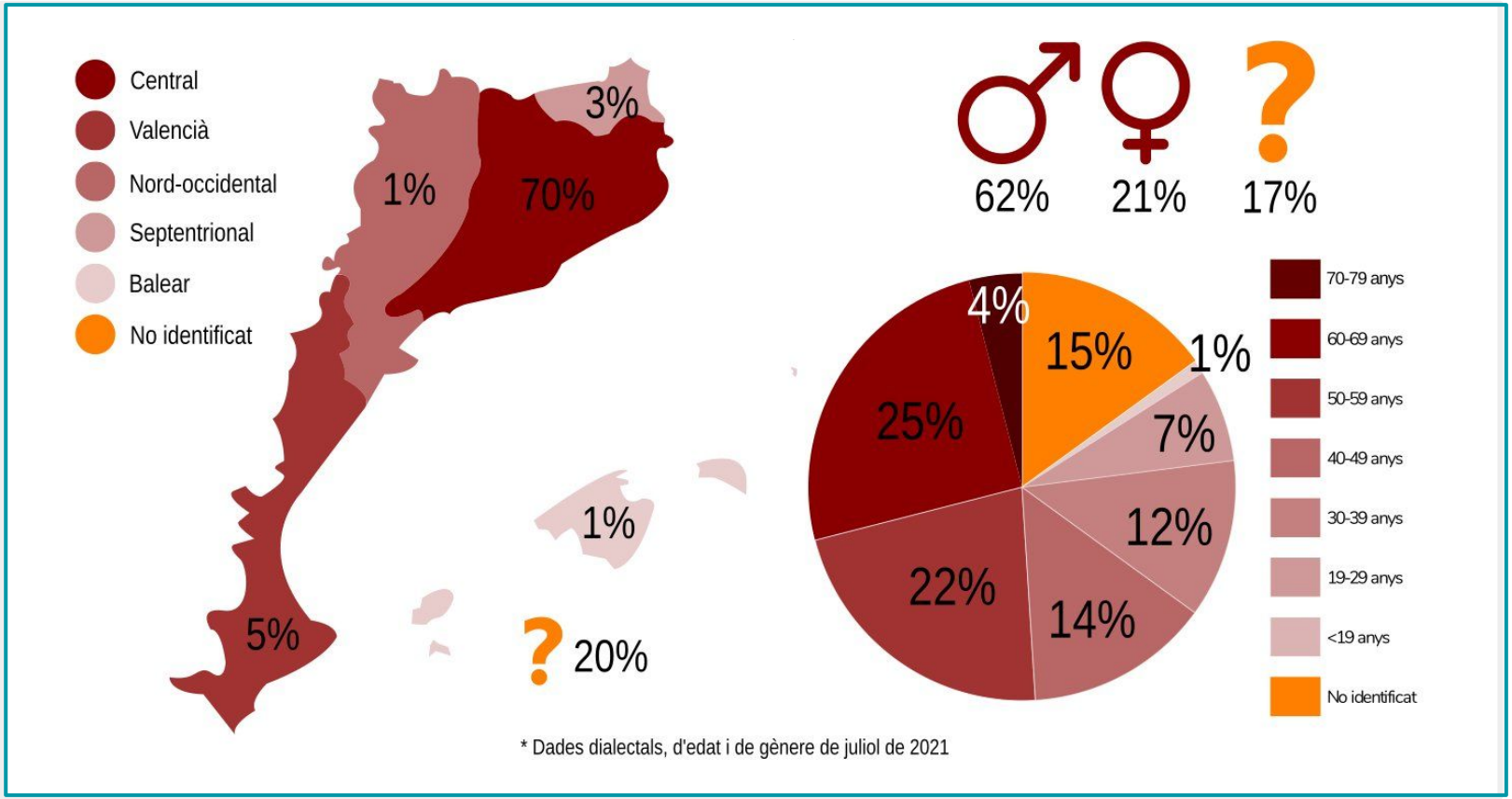
Muhire, R. (2020). How Rwanda is Making Voice Tech More Open. Mozilla Foundation. https://foundation.mozilla.org/en/blog/how-rwanda-making voice-tech-more-open/

El cas del Ruandès

> “Sense comunitat, no hi ha dades”

Muhire, R. (2020). How Rwanda is Making Voice Tech More Open. Mozilla Foundation. https://foundation.mozilla.org/en/blog/how-rwanda-making voice-tech-more-open/

https://twitter.com/softcatala/status/1420822833245085702



https://twitter.com/softcatala/status/1420822833245085702

Rànquing actual de CommonVoice

Llengua	Hores validades
Anglès	2015
Ruandès	1894
Alemany	965
Català	789
Esperanto	748
...	
Basc	91

Models

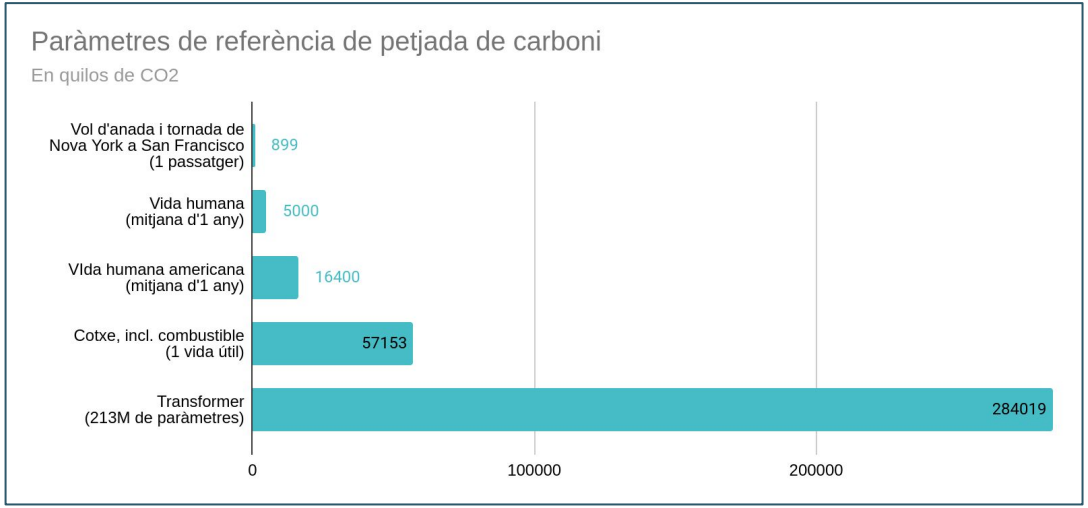
Models

- La majoria d'algorismes són d'iniciatives **privades** comercials

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

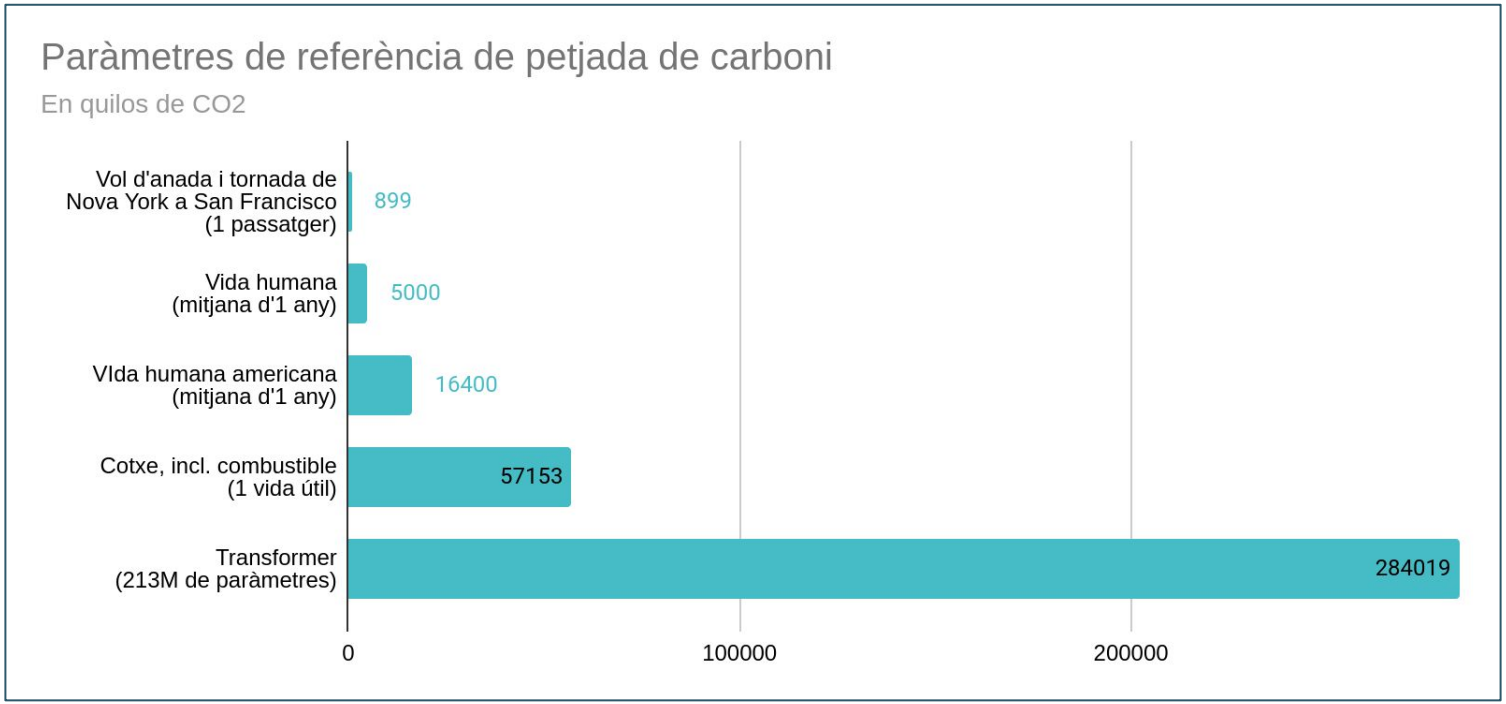
Models

- La majoria d'algorismes són d'iniciatives **privades** comercials
- Els models tenen una **petjada de carboni** enorme



Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Models



Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Models

- La majoria d'algorismes són d'iniciatives **privades** comercials
- Els models tenen una **petjada de carboni** enorme

Què fem?

Models

- La majoria d'algorismes són d'iniciatives privades comercials
- Els models tenen una petjada de carboni enorme

Què fem?

1. **Alliberar** el model un cop l'hagis entrenat

HUGGING FACE

Models

- La majoria d'algorismes són d'iniciatives privades comercials
- Els models tenen una petjada de carboni enorme

Què fem?

1. Alliberar el model un cop l'hagis entrenat
2. Desenvolupar models **conjuntament**

Models

- La majoria d'algorismes són d'iniciatives privades comercials
- Els models tenen una petjada de carboni enorme

Què fem?

1. Alliberar el model un cop l'hagis entrenat
2. Desenvolupar models **conjuntament**

Eines

Basades en Text

Apertium

https://apertium.org

SOFTCATALÀ

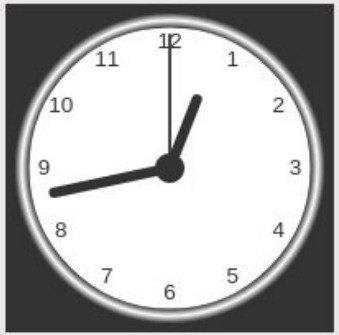
Traductor | Corrector | Diccionaris i eines | Aplicacions | Recursos per a traductors | Ordinadors i mòbils

Basades en Text

Introduïu una hora específica per a saber com es diu en català:

12 : 43 CONSULTA Estableix l'hora actual

Heu introduït: **12:43:00**



La una menys disset (minuts) - Sistema de rellotge

Dos quarts i tretze (minuts) d'una - Sistema de campanar

Dos quarts i mig ben passats d'una - Sistema de campanar tradicional

Basades en Text

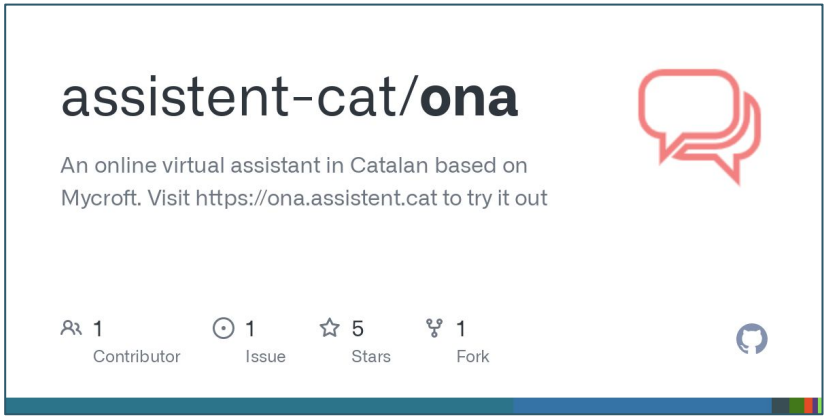
Basades en Veu

catotron.collectivat.cat

CATOTRON

Síntesi de la parla obert i lliure en català

SOFTCATALÀ



assistent.cat

Comunitats

Comunitats

Neix NLP ComuniCat, la comunitat oberta sobre tecnologies del llenguatge en català

La iniciativa vol agrupar tots els desenvolupadors del sector per compartir coneixements i potenciar la presència de l'idioma en l'àmbit digital

Categories: Professionals

Redacció Dimarts, 21 de setembre de 2021 | 13:47h

NLP ComuniCat

@NLPComuniCat

Edit profile

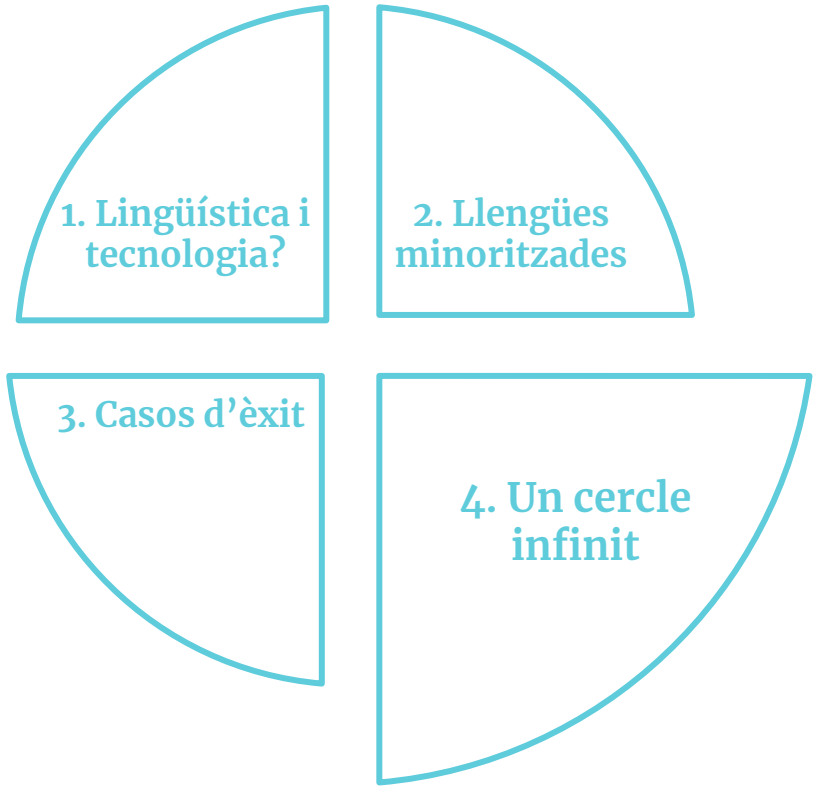
Som la comunitat oberta sobre el desenvolupament de les tecnologies del llenguatge en català! T'hi apuntes?

Translate bio

nlpcomunicat@gmail.com docs.google.com/forms/d/e/1FAI...

Joined September 2021

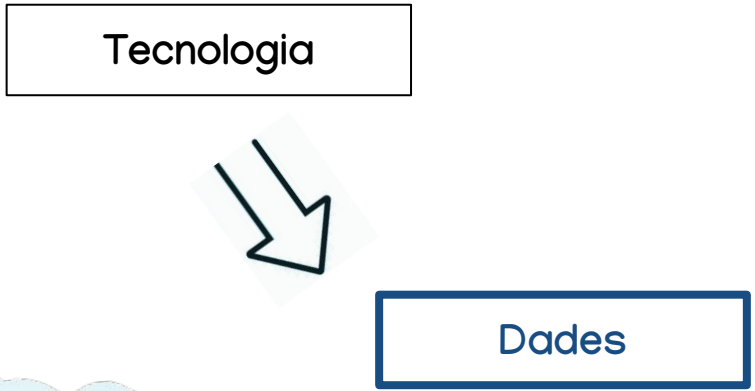
197 Following 229 Followers


1. Lingüística i tecnologia?
2. Llengües minoritzades
3. Casos d'èxit
4. Un cercle infinit

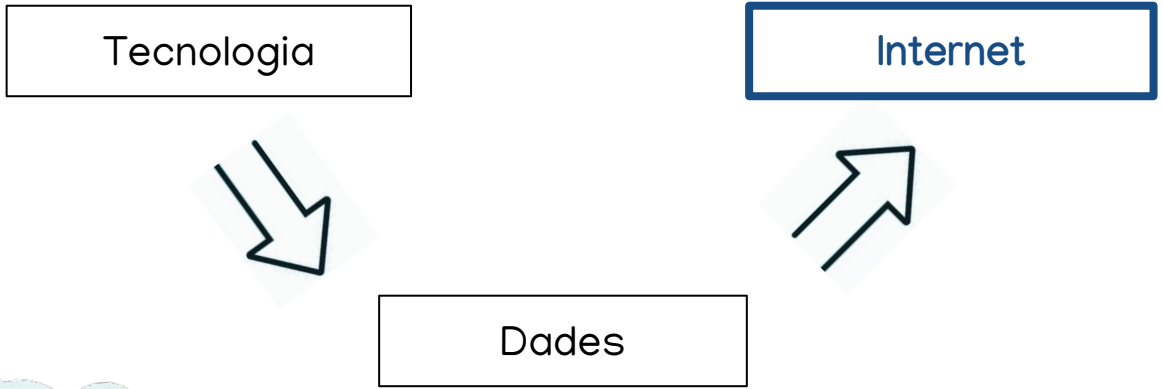
Un cercle infinit

Tecnologia

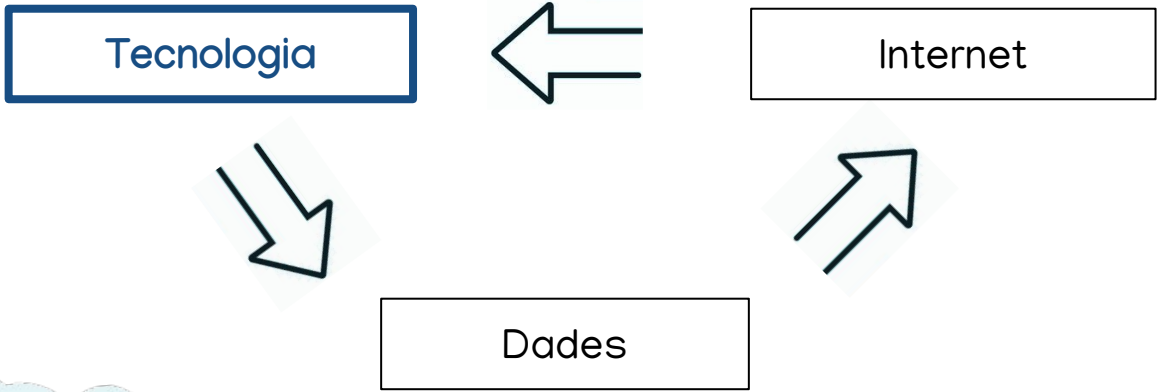
Un cercle infinit



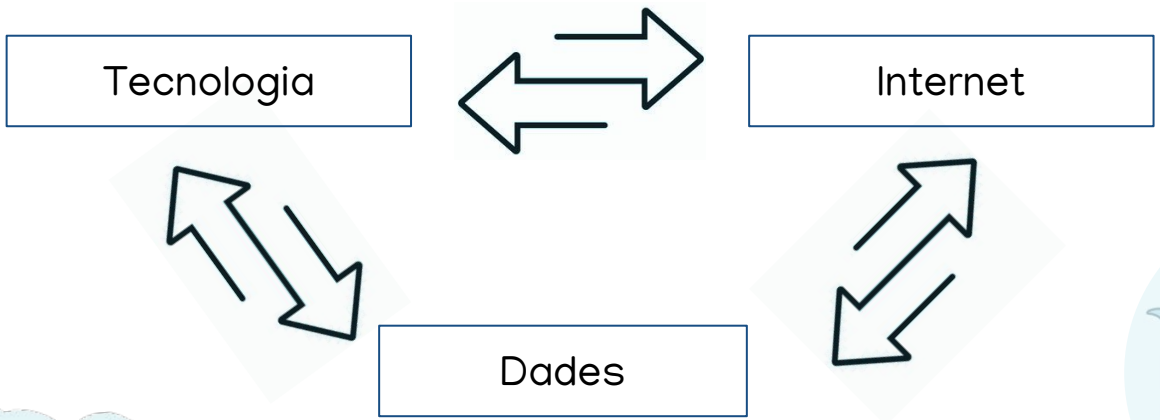
Un cercle infinit



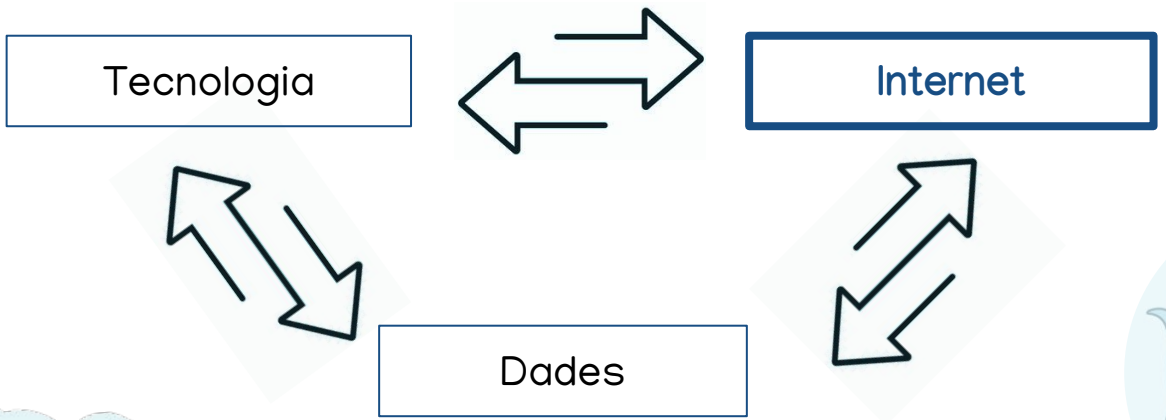
Un cercle infinit



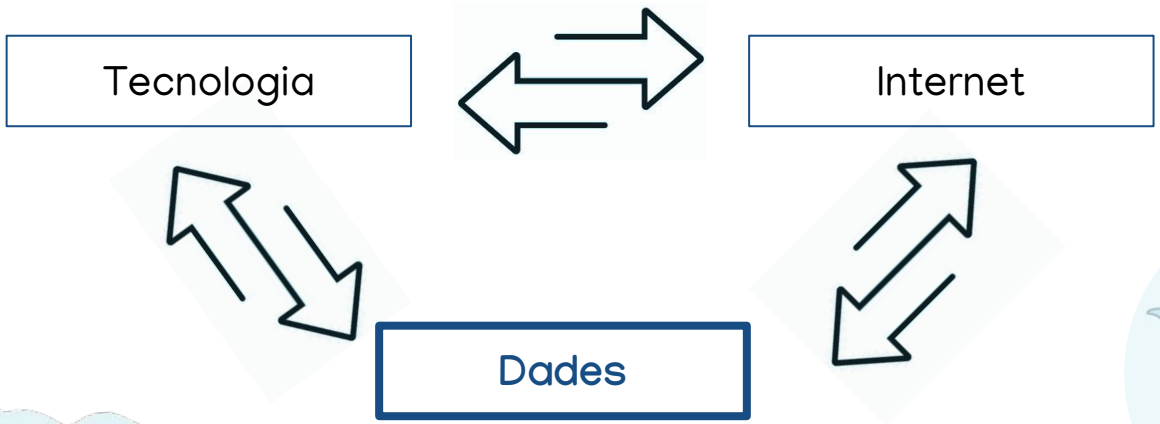
Un cercle infinit



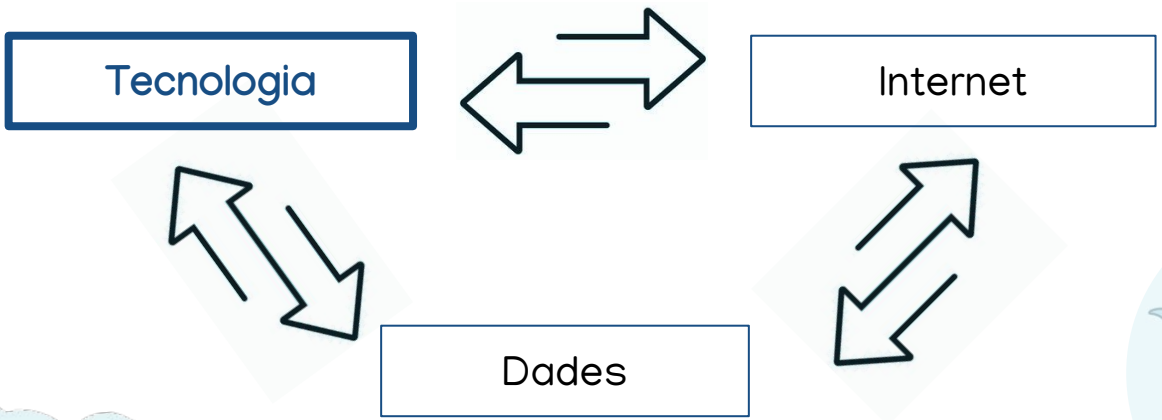
Un cercle infinit



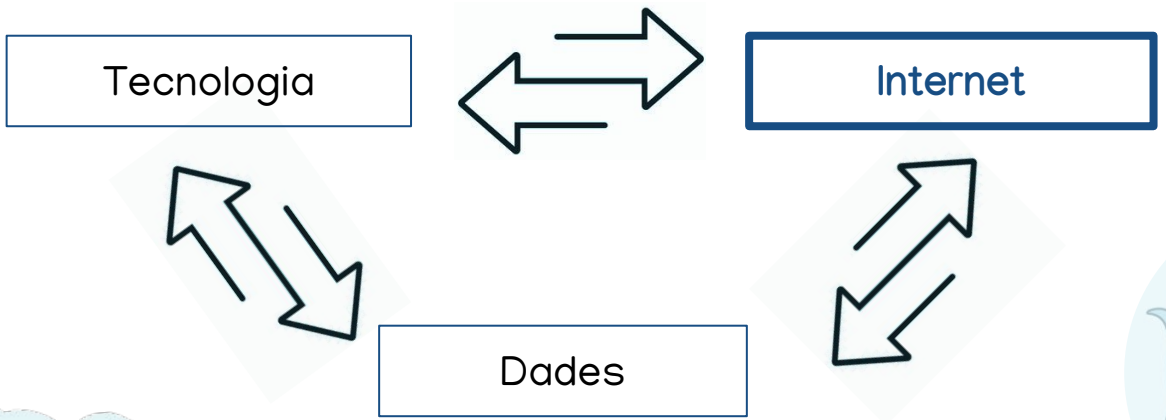
Un cercle infinit



Un cercle infinit



Un cercle infinit



Com trenquem el cercle

VIQUIPÈDIA
L'enciclopèdia lliure

NLP
comuniCAT

BigScience

Sobirania lingüística

- poder fer **ús** de la nostra llengua en qualsevol àmbit
- reconeixement dels **drets lingüístics** a nivell individual i col·lectiu, de la llengua i totes les seves variants

> *El que cal [...] no és simplement una millor protecció de les minories i més concessions en termes de autonomia regional, però, en definitiva, una **reconsideració exhaustiva del significat de la sobirania** en línies que superen les dicotomies profundament arrelades i jeràrquiques en què es basa la mateixa **existència de "majories" i "minories"***

Kraus, Peter A. "Between minority protection and linguistic sovereignty". Revista de Llengua i Dret, Journal of Language and Law, Issue 69 (June 2018), p. 6-17, DOI: 10.2436/rld.i69.2018.3122

el futur de la nostra llengua és a les nostres mans

Eskerrik asko!