

# A unique opportunity for Spanish in the digital world - NLP

Dr. Richard Benjamins

Chief AI & Data Strategist, Telefónica

Founder Observatory for social and ethical impact of AI

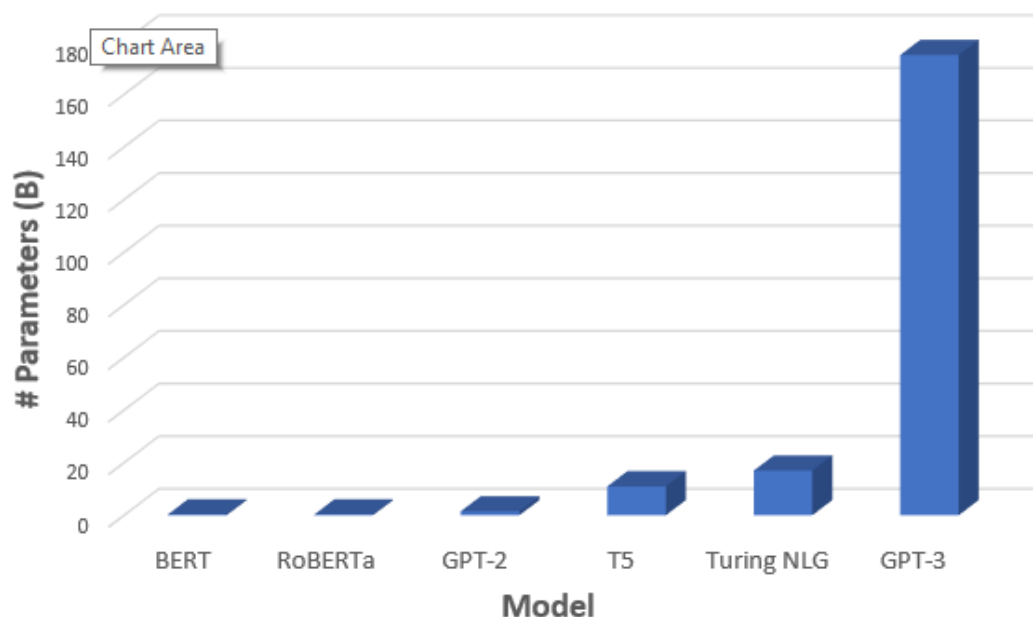
Board member of CDP (Climate reporting NGO)

October 2022

# The breakthrough in NLP – large language models

## Self Supervised Representation Learning

Randomly masked A quick [MASK] fox jumps over the [MASK] dog  
↓ ↓  
Predict A quick brown fox jumps over the lazy dog



Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many legs does a frog have?

A: A frog has four legs.

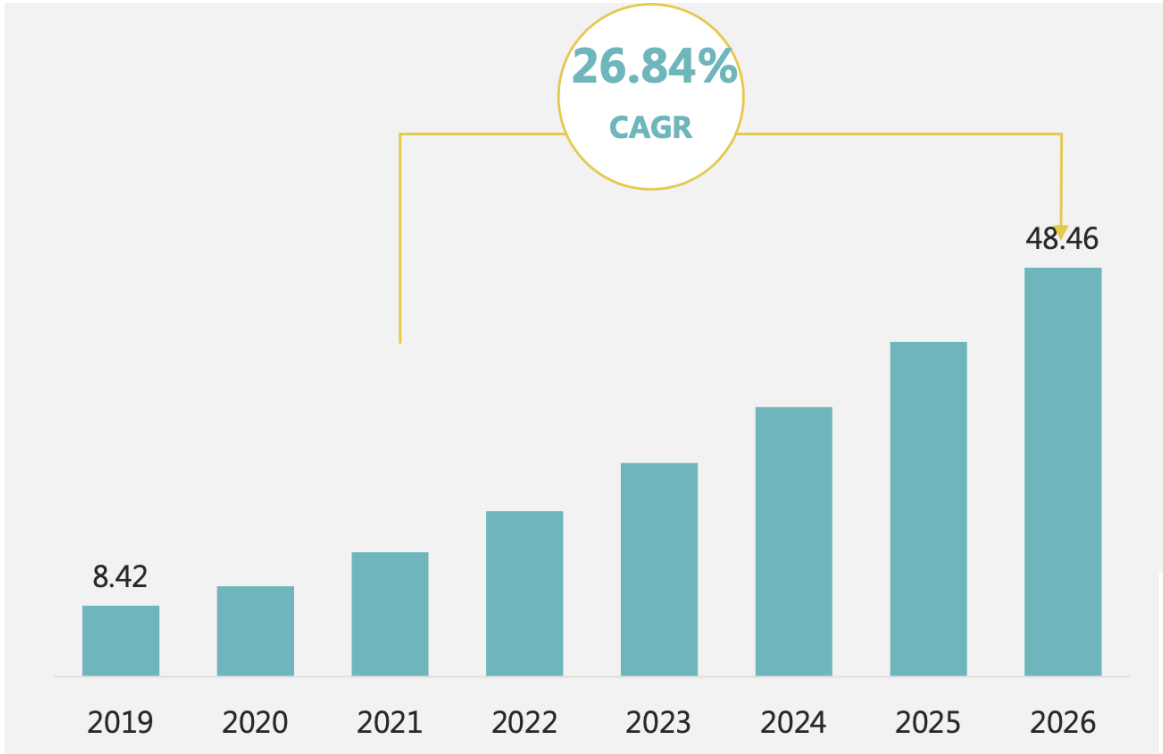
Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why don't animals have three legs?

A: Animals don't have three legs because they would fall over.

# Increased awareness in the market



**LA VANGUARDIA** BBVA

El 80% de la actividad de la inteligencia artificial en España está relacionada con tecnologías del lenguaje

## Key NLP Use Cases In Production Systems for Select Industries

Document Classification	Name Entity Recognition	Question / Answering	Entity Linking / Knowledge Graph	Natural Language Generation
Computers, Electronics, Technology	Computers, Electronics, Technology	Computers, Electronics, Technology	Computers, Electronics, Technology	Computers, Electronics, Technology
Healthcare	Healthcare	Healthcare	Healthcare	Healthcare
Education	Education	Education	Education	Education
Financial Services	Financial Services	Financial Services		
Advertising				
Food				
Pharmaceuticals / BioTech	Pharmaceuticals / BioTech		Pharmaceuticals / BioTech	

(Much NLP offering is still grammar based)

# Some facts about the Spanish language

## Physical world

- Spoken by almost 600M, of which 480M native speakers
- Represent 9% of global GDP
- 2nd global mother tongue (after Chinese)
- Official language in 21 countries

## Digital world

- 3rd language on Internet (after Chinese and English)
- 2nd language on most digital platforms
- 2nd language in tourism

Globally “regulated” by Royal Spanish Academy in collaboration with ASALE  
(one global dictionary covering local varieties)

# Risks for Spanish in the digital world

- Loss of unity of Spanish
  1. Siri, Alexa, etc. each have their “own” Spanish
- AI ML cycle
  2. Impoverishment of language
    - Search for meaning
    - Corrections and completions
    - AI learning cycle
  3. Viralization of errors
    - xq, pq



factura +

Del lat. *factūra*.

1. f. Cuenta en que se detallan con su precio los artículos vendidos o los servicios realizados y que se entrega al cliente para exigir su pago.
2. f. Modo de estar hecho o ejecutado algo. U. m. en arte y artesanía. *Retrato de buena factura*.
3. f. *Arg. y Ur.* Bollo o bizcocho que se fabrica y vende en las panaderías.



## ¿Es válido el uso de «cocreta»?

No. La forma vulgar \**cocreta* —que nunca ha figurado en el diccionario académico— no se considera válida.



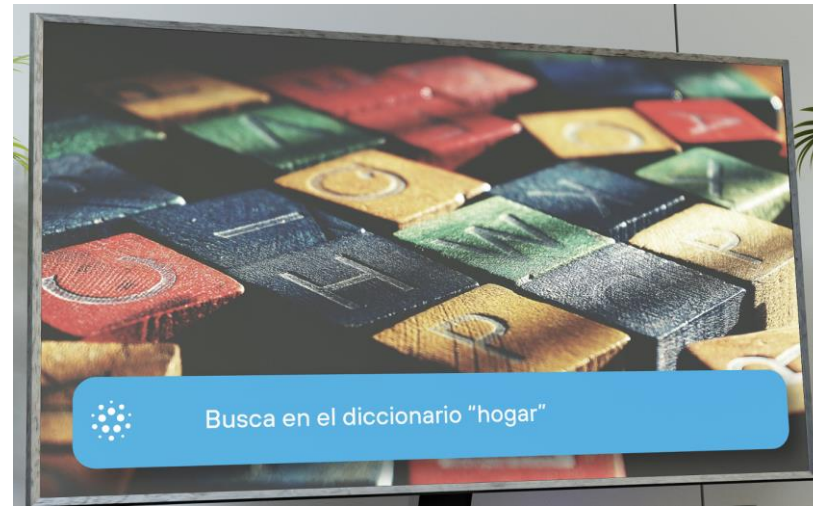
## Objectives

- Machines to speak correct Spanish
  - Voice assistants and chatbots
- AI helps people to generate correct Spanish
  - Spelling, grammar corrections
  - Word, sentence completion



REAL ACADEMIA ESPAÑOLA



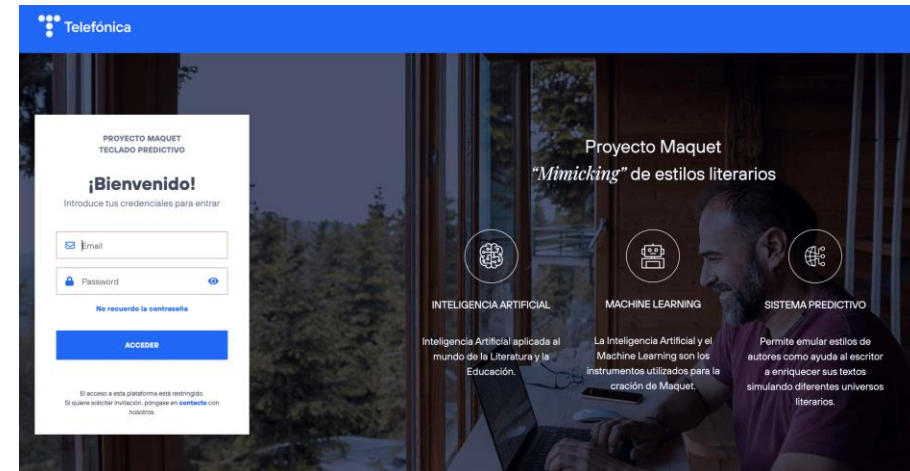


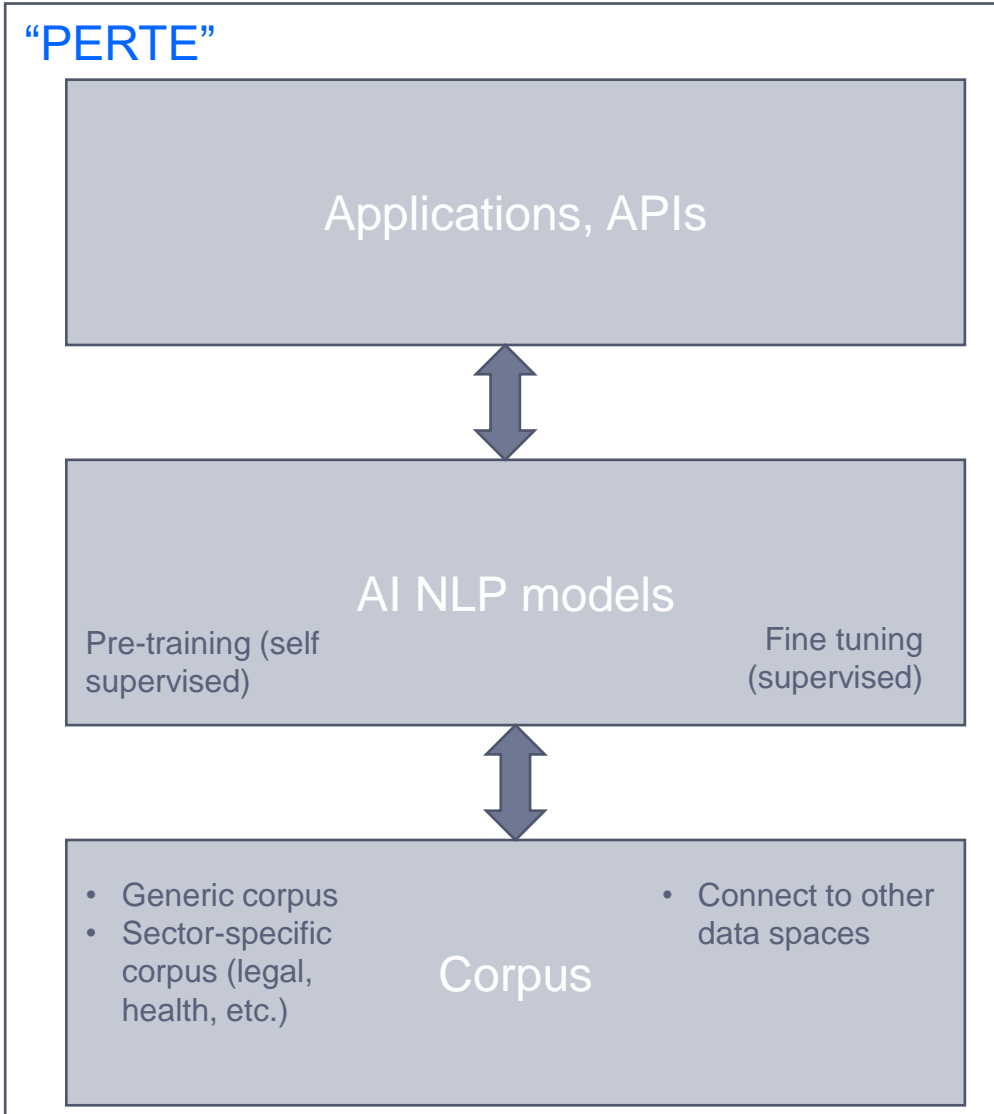
banco

Del fr. ant. *bank*, y este del germ. *\*banki*.

1. m. Asiento, con respaldo o sin él, en que pueden sentarse dos o más personas.

2. m. Madero grueso escuadrado que se coloca horizontalmente sobre cuatro pies y sirve de mesa para labores de carpinteros y otros artesanos.





## Opportunities

- Sector-specific offerings
  - Application-based offerings
  - Huge market of 600M people
- 
- Build the “Spanish GPT-X”, by Spanish-speaking countries.
  - Bias-aware AI models
  - Greener NLP models
  - Improved model metrics
  - LEIA project
- 
- Panhispanic corpus for 600M people
  - To train “Spanish GPT-X”





# Risk for PERTE: fragmentation of corpus






Corpus	Nombre	Descripción	Dueño/origen	Período	Tamaño	Organización propietaria	Tipo de corpus	Tipo de contenido (dominio)	Proyecto declarado/ posible	Preprocesado	Metadatos & etiquetado	Idioma corpus	Propiedad intelectual	Licencia publicada	Temas éticos	Actualización	URL	Paper	Contact
El Corpus del Español del Siglo XXI (CORPES XXI)	CORPES		RAE	2000-presente		Pública					Lingüístico	Castellano?				Annual	<a href="https://www.rae.es/corpes11">https://www.rae.es/corpes11</a>		
Corpus de Referencia del Español Actual	CREA		RAE			Pública											<a href="https://www.rae.es/banco-de-datos/crea">https://www.rae.es/banco-de-datos/crea</a>		
El Corpus Diacrónico del Español (CORDE)	CORDE		RAE			Pública													
Corpus del Diccionario Histórico de la Lengua Española	CDH		RAE			Pública											<a href="https://www.rae.es/CDHE/view/InicioExterno.view?sessionid=93F03DFC0000CA503166347002641E">https://www.rae.es/CDHE/view/InicioExterno.view?sessionid=93F03DFC0000CA503166347002641E</a>		
			Europa Press			Privado		Internet											
	BUHO		RAE					Internet											
Multilingual Colossal Cleaned Common Crawl	mC4	crawling de Internet, segmentado por idiomas	CommonCrawl + Google + Altavox	2020	9,7 TB, 358 idiomas (ES: 416 Mdoc, CA: 14.5M, GL: 4.5M, EU: 1.5M)	Privada	Genérico	Internet	SST	Identificación de idioma (LIDB) Limpieza Desduplicado	URL timestamp	108, incluyendo ES, CA, EU, GL	Abierto	Common Terms of Use + ODC BY	Filtrado de siglas con términos ofensivos	Infrecuente	<a href="https://huagingface.co/datasets/mc4">https://huagingface.co/datasets/mc4</a>	<a href="https://arxiv.org/abs/1912.10683">https://arxiv.org/abs/1912.10683</a>	
Open Super-large Crawled Aggregated outpus	OSCAR	crawling de internet	INRIA	2021	166 idiomas (ES: 51 Mdoc, CA: 88K, GL: 4.5M, EU: 266K)	Privada	Genérico	Internet	SST	Identificación de idioma		166, incluyendo ES, CA, EU, GL		sin filtrado	anual	<a href="https://oscar-corpus.com/">https://oscar-corpus.com/</a>			
BigScience corpus		colección de datasets con diversos orígenes	BigScience	2022	colección de 87 datasets, incluyendo OSCAR, 1.5 TB total	Privada	Genérico	Internet + otros	SST	Identificación de idioma Limpieza y filtrado		~50 idiomas, incluyendo ES, CA & EU	Distintas licencias, en función del dataset	Eliminación de algunos campos de PI			<a href="https://bigscience.org/datasets/">https://bigscience.org/datasets/</a>		
BERTIN dataset	BERTIN	subconjuntos de mC4 para español	BERTIN	2021		Privada	Genérico	Internet	SST	mC4 + Muestreo por perplejidad		ES	ODC BY (igual que mC4)	análisis superficial de sesgo	no	<a href="https://huagingface.co/datasets/bertin-project/mc4-es-sampled">https://huagingface.co/datasets/bertin-project/mc4-es-sampled</a>			
Wikipedia		Enciclopedia online	Wikimedia Foundation			Privada	Genérico	Enciclopedia	SST/FT			todos							
Project Gutenberg	Gutenberg	Colección de libros sin copyright			308 libros (ES)		Genérico	Literatura	SST	maquetado de libros		varios, incluyendo ES, CA, GL		Public Domain			<a href="https://www.gutenberg.org/">https://www.gutenberg.org/</a>		
Gutenberg Dialog Dataset	Gutenberg Dialog Dataset	Diálogos extraídos de libros del proyecto Gutenberg	BSC		58K frases		Genérico	Literatura	FT	alineamiento entre idiomas, corrección tipográfica	metadatos, MD5, rating, duration	ES					<a href="https://github.com/cisarneto/gutenberg-dialog">https://github.com/cisarneto/gutenberg-dialog</a>	<a href="https://arxiv.org/abs/2104.12752">https://arxiv.org/abs/2104.12752</a>	
Excorpius	Excorpius	extracto de Common Crawl en español			50,773M palabras 322.5 GB total	Privada	Genérico	Internet		Identificación de idioma, deduplicación y limpieza	URL	ES	diseño CC BY-NC-ND pero Common Crawl tiene su propia licencia				<a href="https://huagingface.co/datasets/UHF/excorpius">https://huagingface.co/datasets/UHF/excorpius</a>	<a href="https://arxiv.org/abs/2106.15347">https://arxiv.org/abs/2106.15347</a>	
SPACCC_MEDDOCAN	Spanish Clinical Case Corpus Medical Document Anonymization	This repository contains a synthetic corpus of clinical cases enriched with PHI expressions, named the MEDDOCAN corpus.	BSC	2021		Pública	Médico	Medical Records				Español ES		CC Attribution 4.0			<a href="https://github.com/PlanTL-GOB-ES/SPACCC_MEDDOCAN">https://github.com/PlanTL-GOB-ES/SPACCC_MEDDOCAN</a>		
SQAC	Spanish Question-Answering Corpus	An extractive QA dataset for the Spanish language	BSC	2022		Pública	Genérico					Español ES		cc-by-sa-4.0			<a href="https://huagingface.co/datasets/PlanTL-GOB-ES/SQAC">https://huagingface.co/datasets/PlanTL-GOB-ES/SQAC</a>	<a href="https://arxiv.org/abs/2205.01966">https://arxiv.org/abs/2205.01966</a>	Montserrat Marimon (montserrat.marimon@bsc.es)
PharmaCNER		Manually classified collection of clinical case studies derived from the Spanish Clinical Case Corpus (SPACCC), an open access electronic library that gathers Spanish medical publications from Scielo.	BSC		396,988 words, 1K clinical cases	Pública	Médico	Clinical cases				NER	Español ES		CC Attribution 4.0		<a href="https://huagingface.co/datasets/PlanTL-GOB-ES/cantemist-ner">https://huagingface.co/datasets/PlanTL-GOB-ES/cantemist-ner</a>	<a href="https://openaccess.thecvf.com/2024/07/3878041">https://openaccess.thecvf.com/2024/07/3878041</a>	
CANTEMIST	Spanish	1301 oncological clinical case reports written in Spanish.	BSC	2021		Pública		Clinical Tumor cases				NER	Español ES						
Open Subtitles		Subtítulos de contenido multimedia, alineados en varios idiomas			1.3GB (ES), 4.5M (CA), 1.9M (GL), 1.4M (EU)							Varios, entre ellos ES, CA, GL, EU		similar to CC-BY			<a href="https://opus.nlpl.eu/OpenSubtitles.php">https://opus.nlpl.eu/OpenSubtitles.php</a>	<a href="https://arxiv.org/abs/1410.1347">https://arxiv.org/abs/1410.1347</a>	
The Open Parallel Corpus	OPUS	Colección de textos de distintas fuentes en varios idiomas, alineados					Genérico	Varios	FT Traducción automática			Varios, entre ellos ES, CA, GL, EU					<a href="https://opus.nlpl.eu/">https://opus.nlpl.eu/</a>		
XL-WSD	Spanish / Multilingual Im-legal-es	Spanish Legal Corpora	BSC	2021		Pública	Legal	Legal			WSD	Multilingual Español ES					<a href="https://spienzanlo.github.io/xl-wsd/">https://spienzanlo.github.io/xl-wsd/</a>		
ParlamentParla	ParlamentParla	Audio/transcripción de sesiones del parlamento de Cataluña	Generalitat de Catalunya/BSC	2007-2018		Pública	Política	Política	LM/ASR			CA		CC Attribution 4.0			<a href="https://huagingface.co/datasets/projecte-aina/parlament_parla">https://huagingface.co/datasets/projecte-aina/parlament_parla</a>		
VilaQuAD		Artículos de noticias de VilaWeb, junto con pares de pregunta-respuesta	BSC	2021	2095 artículos, 6281 pares Q-A		Sectorial	Noticias	FT Q-A		Title Question Answer	CA		cc-by-sa-4.0	no		<a href="https://huagingface.co/datasets/projecte-aina/vilaquad">https://huagingface.co/datasets/projecte-aina/vilaquad</a>		
Catalan Text Corpus	CaText	Recolección de corpus de diversas fuentes	BSC															<a href="https://arxiv.org/abs/1807.07903">https://arxiv.org/abs/1807.07903</a>	
AnCorra Corpus	AnCorra-ES y AnCorra-CA	Corpus del catalán (AnCorra-CA) y del español (AnCorra-ES) con diferentes niveles de anotación: lema y categoría morfológica, constituyentes y funciones sintácticas, estructura argumental y papeles temáticos, clase semántica verbal, tipo derivativo de los nombres deverbales, sentidos de WordNet nominales, entidades nombradas, y relaciones de coreferencia.	Grupo de investigación CUC		500.000 palabras para cada idioma	Pública		Prensa (principalmente)				Español y Catalán					<a href="http://dicc.ub.edu/corpus/es/encora">http://dicc.ub.edu/corpus/es/encora</a>	<a href="http://dicc.ub.edu/corpus/encora">http://dicc.ub.edu/corpus/encora</a>	
CORLEC	Corpus Oral de Referencia de la Lengua Española Contemporánea	Base de datos textual (corpus de lengua hablada): transcripción de textos grabados en cintas de audio del registro oral. 1.100.000 de palabras transferidas en soporte informático.	Se elaboró en la Universidad Autónoma de Madrid con una ayuda de IBM España dentro del programa de cooperación entre IBM España y la Cátedra de Lingüística General de la UAM.	Enero 1991 - febrero 1992		Pública	Sectorial (variado)	Mixto				Español					<a href="http://www.ill.uam.es/ESP/corlec.html">http://www.ill.uam.es/ESP/corlec.html</a>	<a href="http://www.ill.uam.es/ESP/corlec.html">http://www.ill.uam.es/ESP/corlec.html</a>	
Spanish Unannotated Corpora	SUC		Universidad de Chile		18 GB / 3 mil millones de palabras	Pública		Mixto	Self-supervised training			Español					<a href="https://github.com/jsecamnetes/spanish-corpora">https://github.com/jsecamnetes/spanish-corpora</a>		

# Important risks to be taken care of

Ethical: gender bias, especially in professions






Climate: electricity consumption of LLM

## Google Translate

English ▾    Turkish ▾  

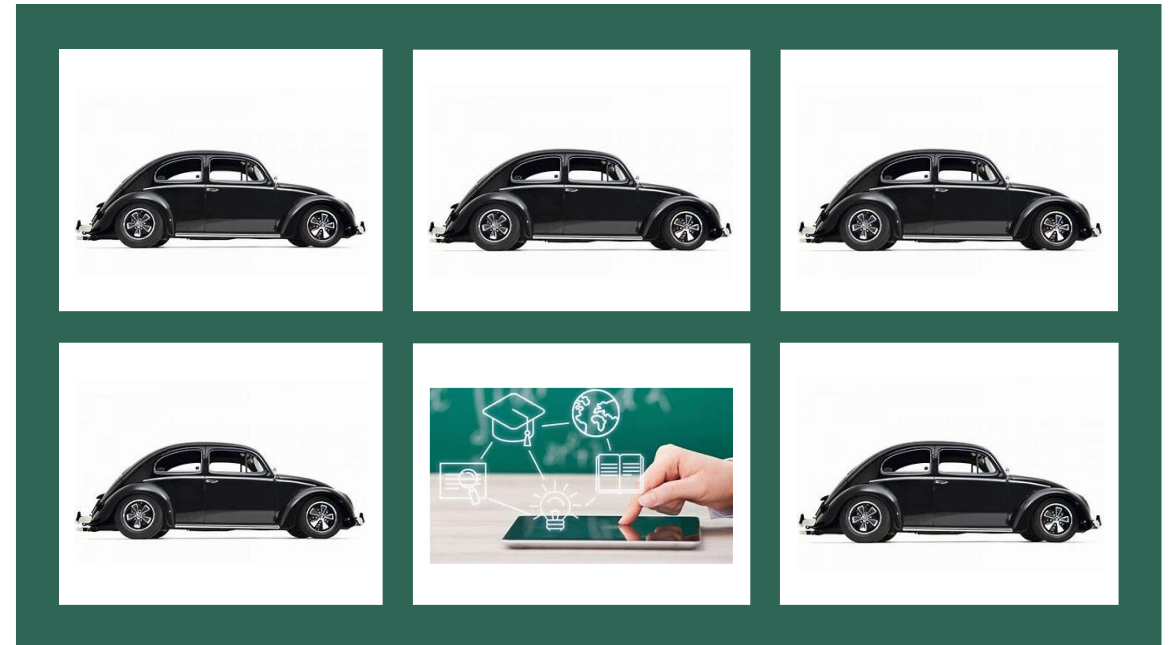
She is a doctor  
He is a nurse

O bir doktor  
O bir hemşire

Turkish ▾    English ▾  

O bir doktor  
O bir hemşire Edit

He is a doctor  
She is a nurse





Telefónica