/instituut voor de Nederlandse taal/

Intelligent Language Technology: AI and the Impact on Languages and Language Policy

Prof. Dr. Frieda Steurs

Dutch Language Institute

KU Leuven

Pretoria
October 2025

Content

- Language and the Computer
- The Evolution of MT
- Speech Technology
- Applications
- Generative AI
- American Big Tech vs.Europe
- Dutch Language

Some Concepts

Natural language = human language (as opposed to machines)

NLP: Natural Language Processing

- Interdisciplinary: computer science & linguistics
- Computational linguists process natural language datasets

Machine learning (ML) (a component of AI): building systems that can learn from the processed data or use that data to perform better.

- Communication with banks, social media, ML is everywhere

Deep Learning: machine learning based on neural networks. Algorithms are used to identify hidden patterns and connections between data. The systems learn from this data.

- Chess computer

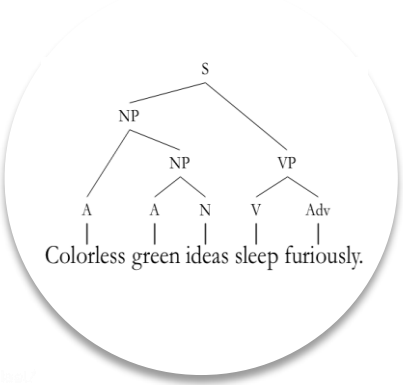
LLM: Large Language Models

Language and the computer

The 1940s

Enigma machine

Chomsky &
TGG



All languages as a
machine code

The 1950s

Cold War

Machine translation: the ultimate dream

- 1949 : Warren Weaver (Rockefeller foundation)
- Memorandum : start of the MT research
- 1954 : first demo : the Georgetown/IBM system
  - 60 Russian phrases translated into English
  - 6 grammar rules and 250 words
  - IBM 701 mainframe computer

- Great boost for research
- The ultimate target: FAHQTUT

Fully automatic high quality translation of unrestricted texts

Dark clouds…

- 1964 : ALPAC rapport
  - Automatic Language Processing Advisory Committee

    The National Academy of Sciences

- Conclusion :
  - MT is not economically viable.
  - MT is slow, less accurate and twice as expensive as HT.
  - All research was suspended in the US.
  - Research was drastically reduced in the rest of the world.
  - **The first AI winter.**

Complexity of human language

- Language: the most challenging paradigm
- MT is one of the most challenging tasks in the field of natural language analysis.
- MT is highly dependent on the quality of the input.

garbage in

garbage out!

A cautious restart

- 1976 : Meteo (Canada) (French/English)
- 1978 : start of EUROTRA (researchprogramme)
  - Dutch, German ,Italian, French, Danish, English
  - Greek, Spanish, Portuguese

  - Language independent!
  - 9 source languages x 8 target languages = 72 language combinations
  - Fully Automatic High Quality Translation
  - Now: 24 source languages x 23 target languages
  - 522 language combinations
- No concrete MT system (end 1992)
- Large research community of computational linguists across Europe

Systran

- Established in 1968 (after ALPAC)
- Russian/English for the United States Air Force (Cold War)
- Although only approximate, the quality of the translations was usually sufficient to understand the content.
- In 1978: commercial translation system
- Most successful rule-based system at the time
- US Defence/European Commission
- Google used SYSTRAN's service until 2007

METAL

- In the 1980s: commercial systems
- METAL (Mechanical Translation and Analysis of Language) Siemens
- University of Texas at Austin
- German/English and English/German
- Specific language pairs, specific terminology
- Then (1985-1992) in Belgium:
- KU Leuven & U Liège: French/Dutch, Dutch/French, French/English
- Ministry of the Interior BE
- Other language pairs followed
- 1996 : Barcelona Technology (L&HSP)

L&H

- Speech technology (dictation systems)
- Multilingual speech recognition
- NLP and translation
- Consumer electronics (speech applications)
- AI in speech
- Chatbots
- Medical information
- Authentication and security (voice biometrics)
- Nuance https://www.nuance.com/nl-nl/index.html

Rulebased MT (RMT)

- METAL uses three lexicons: a source language lexicon (German), a target language lexicon (English) and a transfer lexicon (German-English).
- Grammar rules: Phrase Structure Grammar (GPSG)

Sentence:

- **(Der Mann befand sich in Darmstadt nach dem Krieg, ohne dass seine Frau ihn gefunden hat)**
- **2 interpretations in 1749 milliseconds: 874 msecs/interp.**
- **171 PHRASES: 120 REJECTED.**
- **Transfer plus generation time: 3435 milliseconds.**

(|the| |man| |was| |in| |the| |intestine| |city| |after| |the| |war| |without| |his| |wife| |having| |found| |him|

21st century: from rules to corpusdata

- Internet: large amounts of language data
- Bilingual corpora
- Corpus-based methods: dominant since 2000
- Three types:
  - example-based machine translation (EBMT)
  - statistical machine translation (SMT)
  - neural machine translation (NMT)
- In 2006, Google launched its internet service based on SMT methods using sentences

Neural machinetranslation

NMT maps the source language in a dense semantic representation and then generates the translation using an attention mechanism.

In 2016, Google launched an NMT system: better translation quality.

2018: BERT: Bidirectional Encoder Representations from Transformers.

Predicting text that comes before and after (bidirectionally) other text.

BERT is a pre-trained neural network.

BERT is trained on publicly available data (Wikipedia).

Initially for English, but now for 104 languages, including Dutch ad

For Afrikaans: AfriBERT

https://aclanthology.org/2020.lrec-1.301.pdf

Speech

AI & Computational Linguistics: methods for recognising and processing the spoken word.

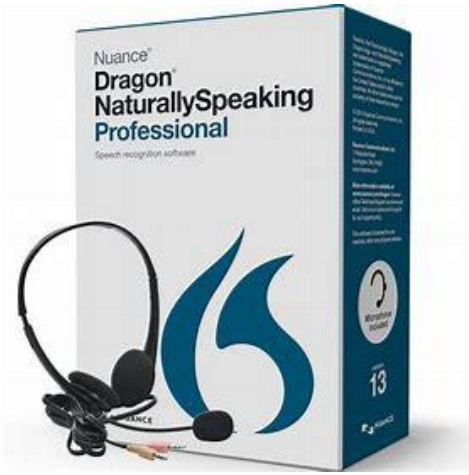
Requires large memory capacity.

GPS system / voice assistant on your phone / smart speaker in your home / ….

Advanced systems: searching court reports or police interrogations.

Call centres with speech recognition.

Speech technology



- Dragon Naturally Speaking
- Lernout & Hauspie, further developed by Nuance
- Home, Premium and Professional
- Support for Dutch and other smaller languages
  - https://speech-recognition.co.za/

- Dictation = 3 times faster than typing (translators!)
- 99% accurate, adapted to the speaker's voice and accent
- Deep learning technology

Speechrecognition everywhere

Zoom/Microsoft Office/Google Translate

OpenAI: WhisperAI

Convert audio files to written text

Advanced automatic speech recognition system (eASR)

680,000 hours of multilingual and verified data

Can be trained further

Voicebots (Alexa, Siri)

Interpreting

NMT and neural speech recognition: can we automate simultaneous interpreting?

The translation system interprets simultaneously with the source language speech, with a delay of only a few seconds.

Simultaneous interpreting is extremely challenging and exhausting for humans, who have to listen to and understand one language while speaking another.

Worldwide: limited number of qualified simultaneous interpreters.

Developing simultaneous MT techniques:

Reducing the workload of human interpreters;

Making simultaneous interpreting services more accessible and affordable.

Nuance, realia, irony, etc.

Translation

Different languages: different morphology and structure

Chinese, English, Dutch: S-V-O (subject-verb-object)

Japanese and Korean: S-O-V (subject-object-verb)

When translating between Chinese and Japanese, long-distance reordering is usually necessary. Furthermore, Chinese is an isolating language with few morphological changes, while Japanese is an agglutinative language with many word morphological changes.

All these differences make multilingual MT particularly difficult.

Data-driven MT methods — SMT or NMT — attempt to learn translation knowledge from a large amount of parallel data.

In general, a larger amount of training data leads to better translation quality.

Endangered languages

Lesser resourced languages

Most languages in the world lack parallel data and are therefore referred to as “resource-poor” languages. Building an NMT system for these languages is a major challenge due to the lack of data.

Languages can therefore die digitally.

Back translation and synthetic language

Multilingual translation between languages with few resources.

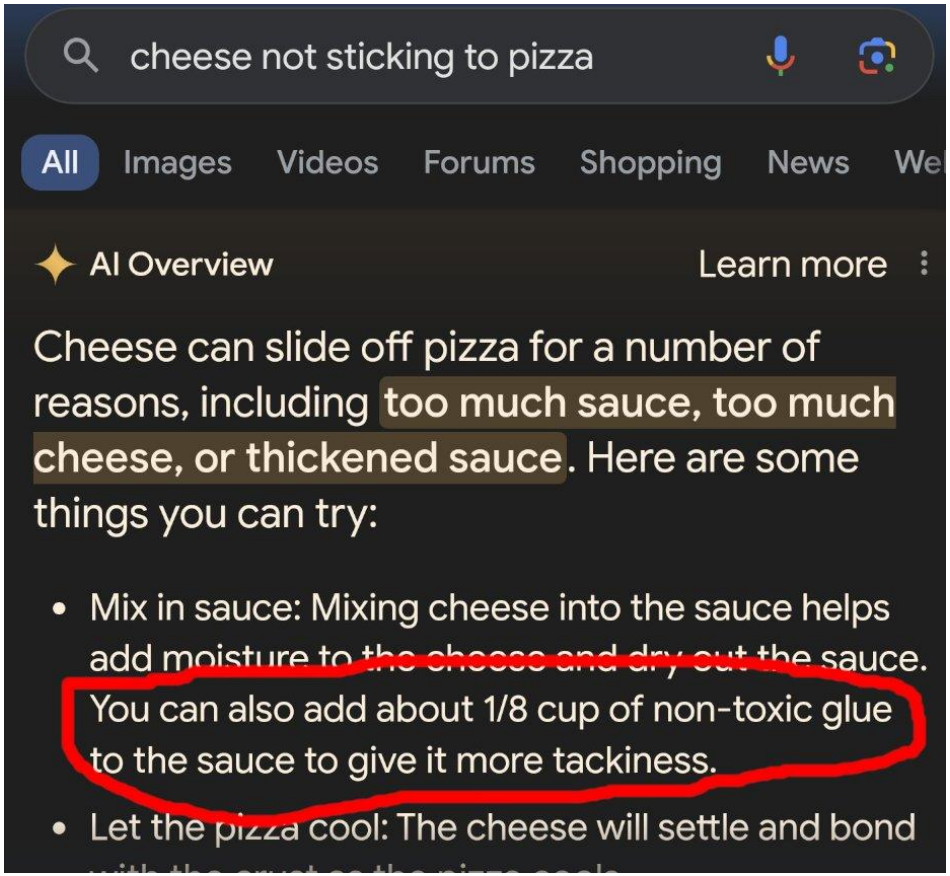
In NMT, the monolingual corpus is usually used to augment the training data.

A commonly used method is back translation.

First, a standard NMT model is trained on a small parallel corpus, and then the model is used to automatically translate a large amount of monolingual data (e.g. sentences in the target languages) to generate a ‘pseudo-bilingual corpus’ that can be used to retrain the translation model.

Danger: synthetic language!

Hallucinating…



BIAS!!!!

Google Translate chooses one translation, even if there are male and female translations available in another language (German for doctor: Arzt and Ärztin). In doing so, Google often reinforced existing prejudices.

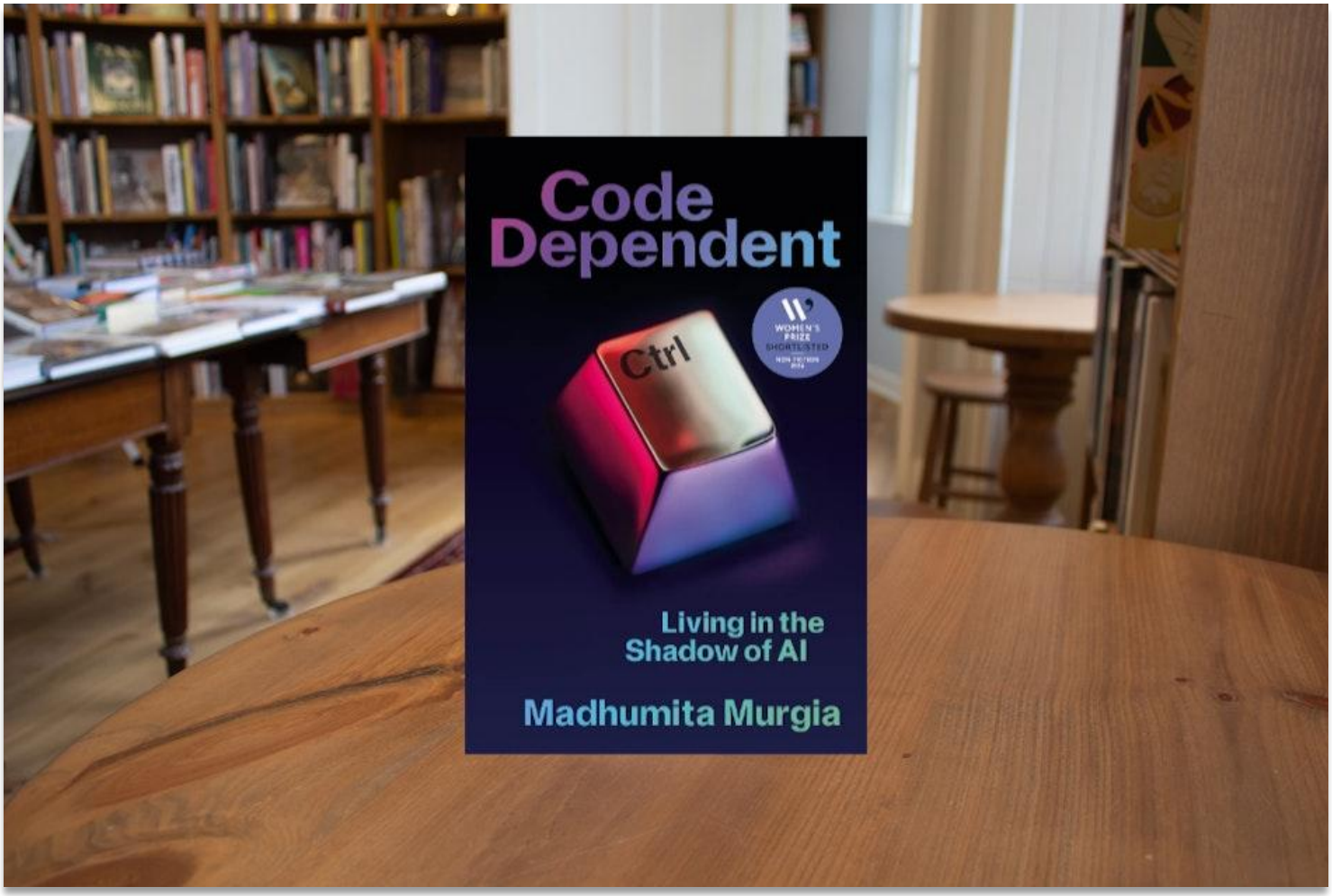
A doctor becomes a man in translation, while the female form is chosen for nurse.

Google Translate provides a number of examples where the gender-neutral profession(*vroedkundige*, *verpleegkundige*, *kleuterleider*, *steward*) becomes femaloe (*vroedvrouw*, *verpleegster*, *kleuterleidster*, *stewardes*), Although the sentence clearly indicates that it refers to a man (possessive pronoun “his”).

Facial recognition: racially determined.

Creditworthiness analysis: man/woman.

Prejudices persist: hate speech, racist remarks, language data via apps, X, TikTok, etc.


Code Dependent
Ctrl
Living in the Shadow of AI
Madhumita Murgia

Applications: translating webpages

With the rapid advancement of globalisation, there is an increasing need for quick acquisition of information in foreign languages.

MT offers an easy way to view web pages in foreign languages.

Users simply need to copy/paste the content of the web page or enter the uniform resource locator (URL) to read the pages in their own language.

“Information gisting”

Scientific literature

Researchers and students

Scientific literature, papers and patents

Translating your own work

Translation in the field of biomedicine is growing rapidly to combat coronavirus disease (COVID-19)

Challenge: terminology!!!

With domain adaptation technologies, a translation model can first be pre-trained with a large training corpus and then refined on a small amount of domain data for further improvement.

Pharmaceutical industry: sensitive information; informed consent, etc.

Online sales

The MT is widely used in international online commerce. With the help of MT, sellers can effectively translate their websites, product information and manuals into foreign languages, while buyers can easily purchase products from around the world.

MT is also used in customer services to improve the quality and efficiency of the

Marktplaats

2dehands.be

Quality?

A good translation must have at least two characteristics: adequacy and fluency.

Nowadays, NMT methods can produce translations with very high adequacy and fluency in certain scenarios for some language pairs and domains.

However, such methods are far from perfect.

Many aspects still need to be improved.

OpenAI

Our work agenda: emails, calendar, meetings, transcripts, etc.

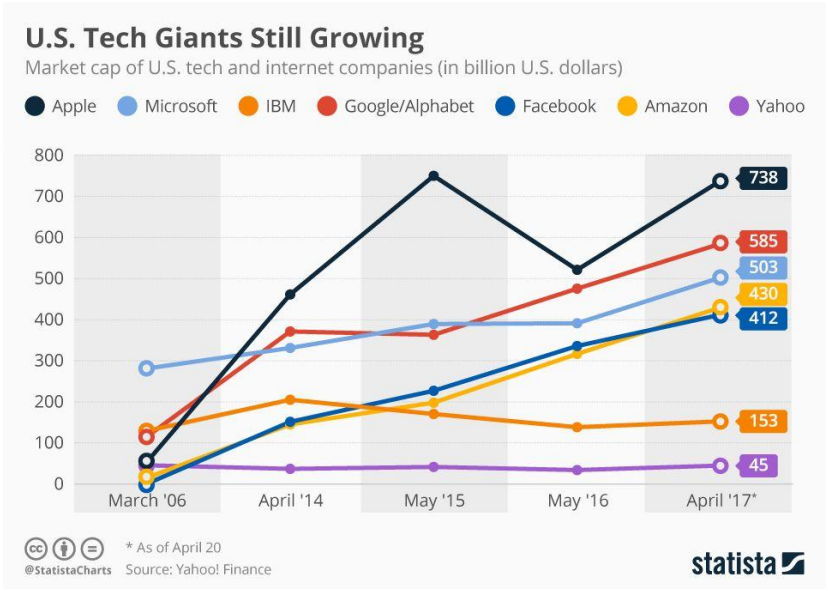
Almost all LLMs are trained in English.

Microsoft supports eight major languages.

Microsoft Bing

Other AI developers


US
VS
EUROPE


U.S. Tech Giants Still Growing
Market cap of U.S. tech and internet companies (in billion U.S. dollars)
Apple
Microsoft
IBM
Google/Alphabet
Facebook
Amazon
Yahoo
800
700
600
500
400
300
200
100
0
738
585
503
430
412
153
45
March '06
April '14
May '15
May '16
April '17*
* As of April 20
@StatistaCharts
Source: Yahoo! Finance
statista

The Washington Post
Democracy Dies in Darkness

BUSINESS

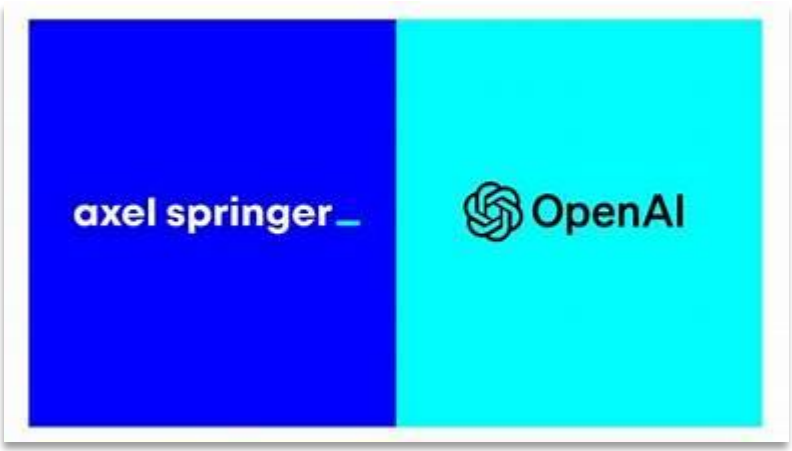
ChatGPT Shows Just How Far Europe Lags in Tech

Analysis by Lionel Laurent | Bloomberg
February 21, 2023 at 2:12 a.m. EST

Europe is where ChatGPT gets regulated, not invented. That's something to regret. As unhinged as the initial results of the artificial-intelligence arms race may be, they're also another reminder of how far the European Union lags behind the US and China when it comes to tech.

Who is paying?

Apple is talking with some big news publishers about licensing their news archives and using that information to help train its generative AI systems, *The New York Times* reports. The company is apparently discussing "multiyear deals worth at least $50 million," the *NYT* says, and has been in touch with publications like Condé Nast, NBC News, and IAC.



More data please!

OpenAI transcribed over a million hours of YouTube videos to train GPT-4

A New York Times report details the ways big players in AI have tried to expand their data access.

Unsurprisingly, it involves doing things that fall into the hazy gray area of AI copyright law.

European parlement 2018

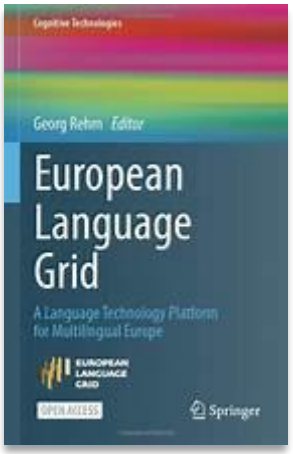
11 september 2018
Language equality in the digital age

New policy "**multilingualism and languagetechnology**"

Member States are encouraged to develop comprehensive language-related policies and allocate resources to promote linguistic diversity and multilingualism in the digital sphere.It is a shared responsibility of the EU and the Member States to contribute to the preservation of their languages in the digital world. The importance of digital opportunities for translation and open access to the data necessary for technological progress.

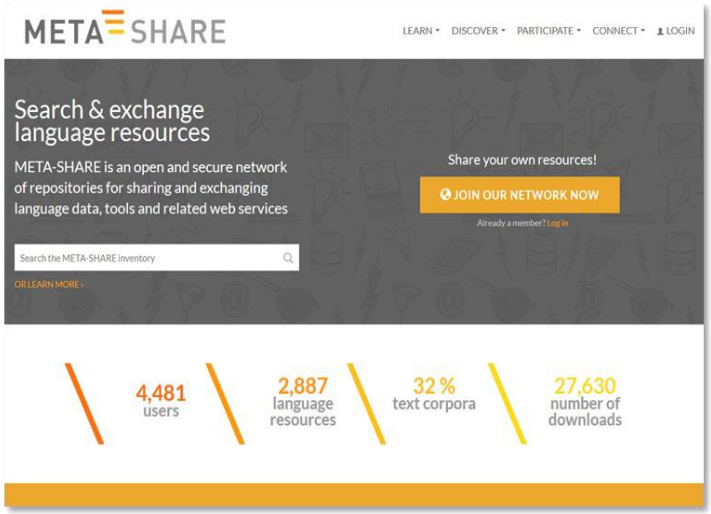
European projects

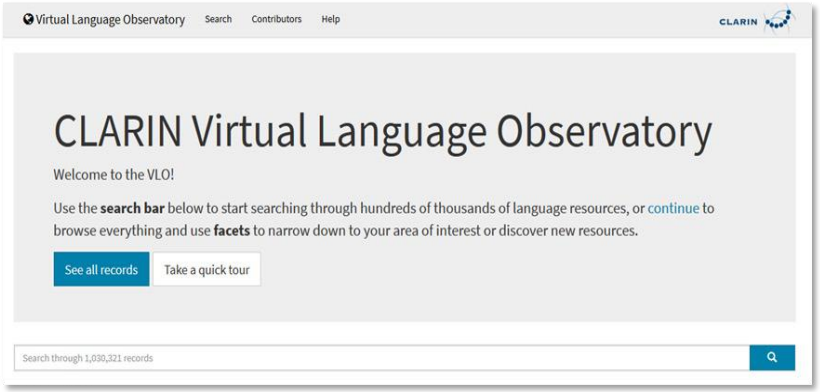
META-NET

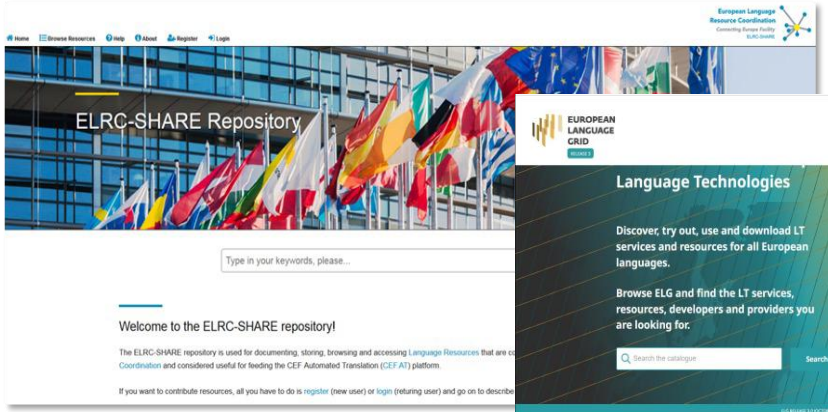


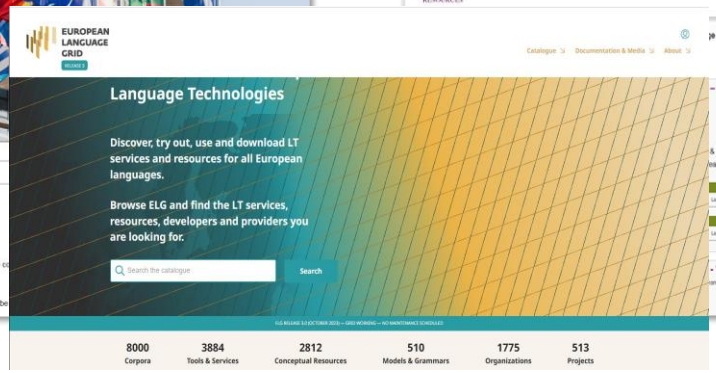
European Language Resource Coordination
Connecting Europe Facility

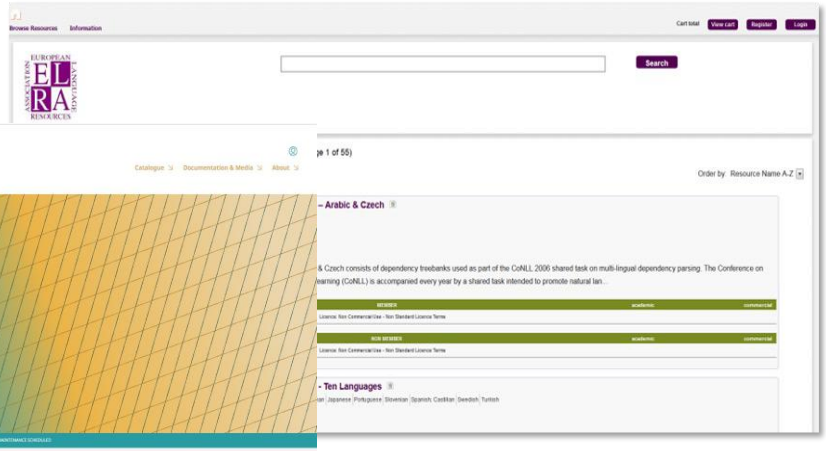
Sharing language data in Europe











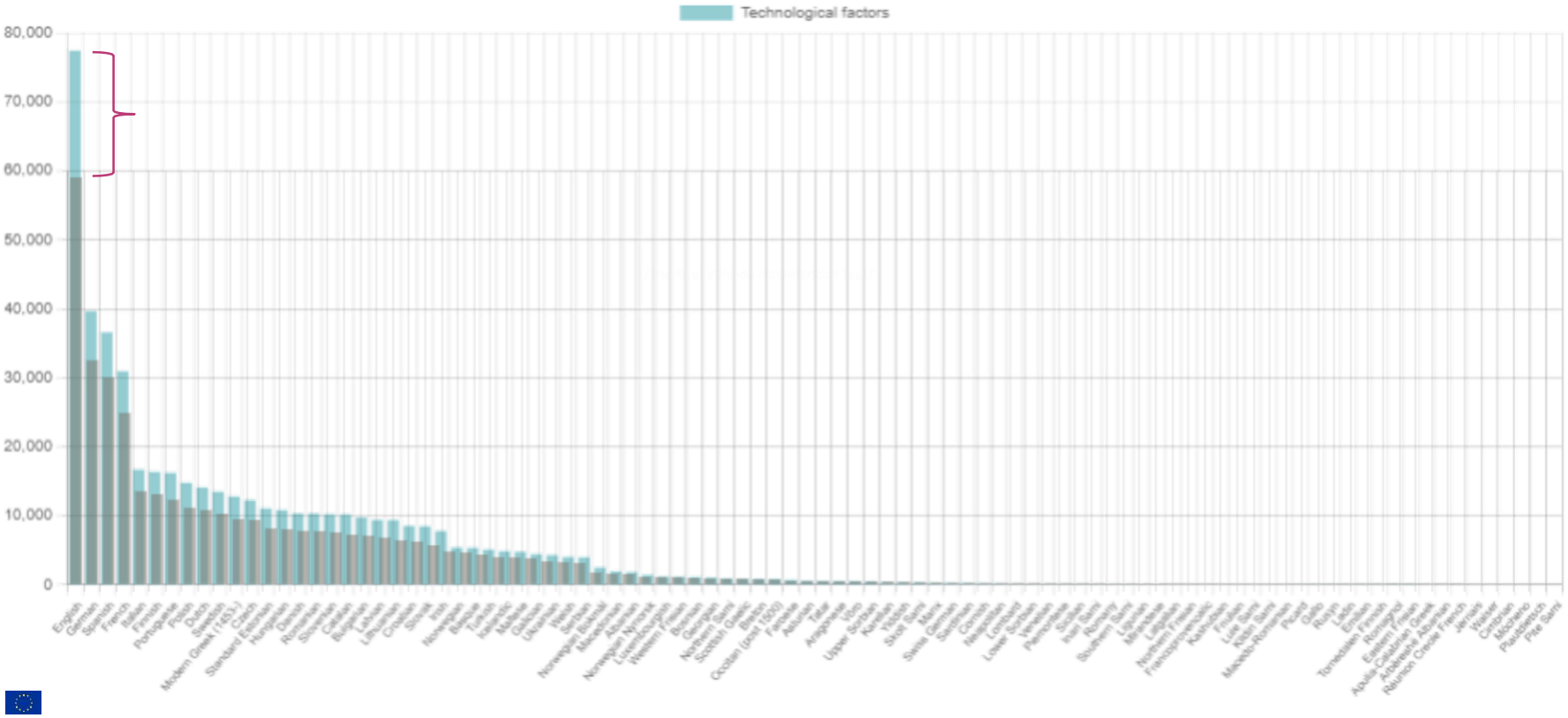
European projects

ELE

European Language Data Space

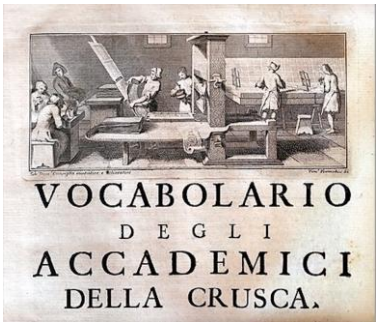
E-translation

DLE Metric: 2022 vs. 2023 (3/3)



EFNIL
European Federation of National Institutions for Language

/instituut voor de Nederlandse taal/



Språk-
BANKEN

IDS INSTITUT FÜR DEUTSCHE SPRACHE

Eesti Keele Instituut

Dansk Sprognævn

INSTITUT ZA HRVATSKI JEZIK I JEZIKOSLOVLJE
Institute of Croatian Language and Linguistics

GPT-NL

The Netherlands is going to develop its own open language model: GPT-NL. This model is necessary for developing, strengthening and perpetuating digital sovereignty.

Verifiable use of AI in accordance with Dutch and European values and guidelines and with respect for data ownership.

Funding: Ministry of Economic Affairs and Climate Policy.

As compliant as possible

GPT-NL The road to GPT-NL

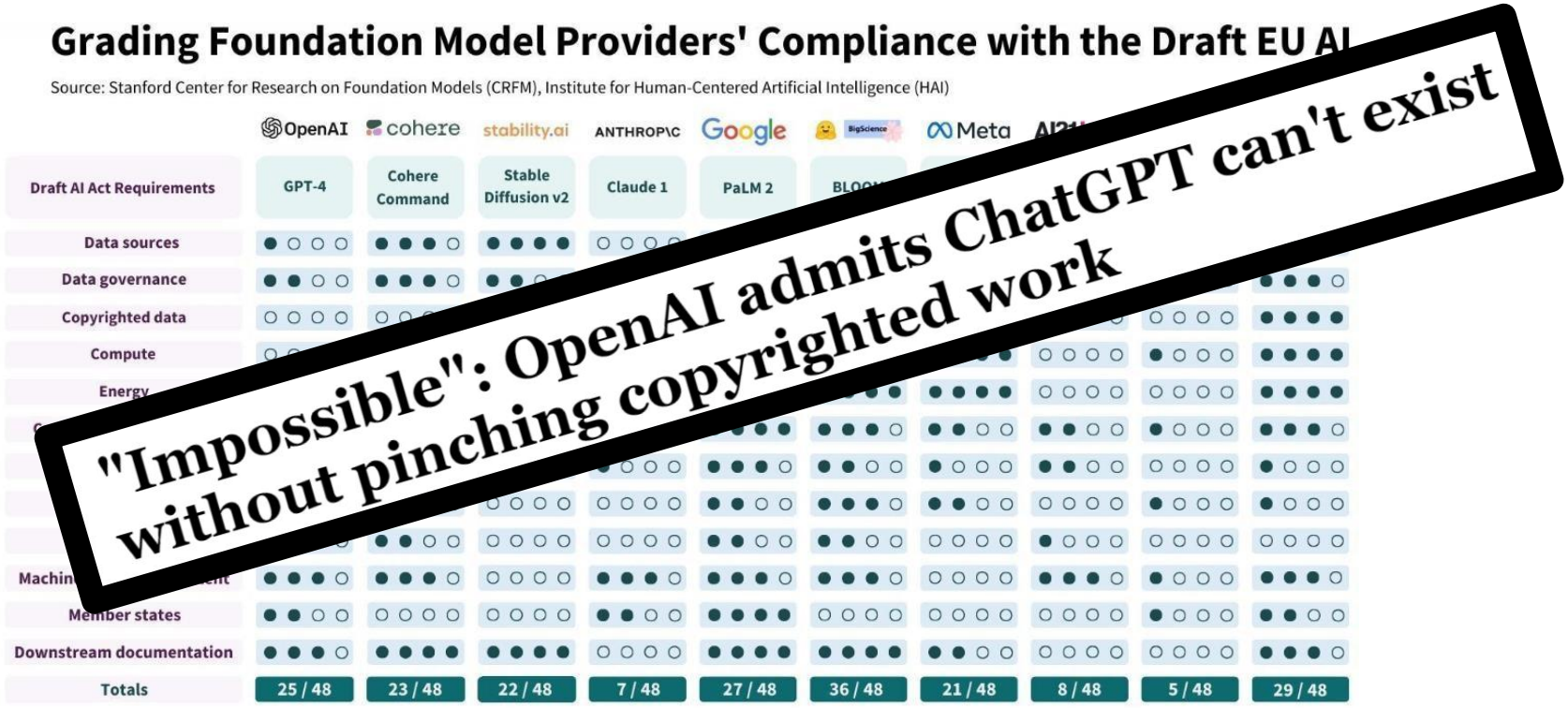
Grading Foundation Model Providers' Compliance with the Draft EU AI

Source: Stanford Center for Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude 1	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ● ○
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	● ● ○ ○
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48

As compliant as possible

GPT-NL The road to GPT-NL



TRUSTLLM

A benchmark of trustworthy large language models.

Developing European large language models (LLMs) that are the most reliable in European AI, for a range of underrepresented languages.

Danish, German, Icelandic, Dutch, Norwegian and **Swedish.**

The development of an open, reliable and factual LLM, initially focused on the Germanic languages.

The basis for an advanced open ecosystem for modular and extensible European reliable, sustainable and democratised next-generation LLMs.

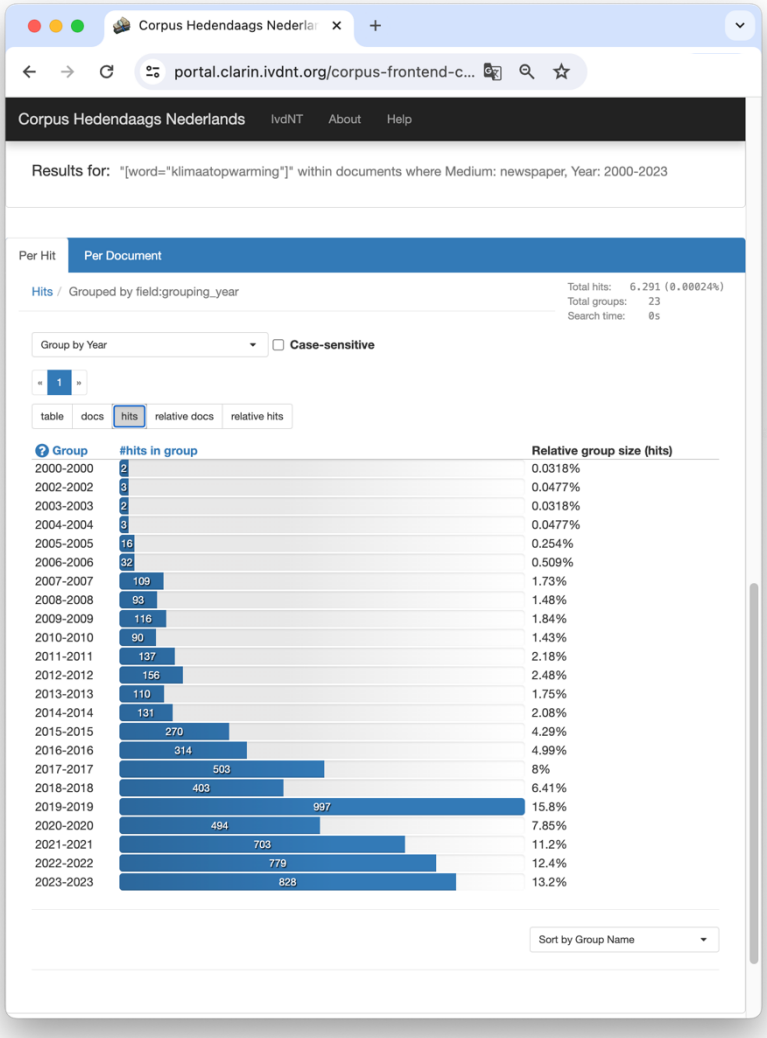
https://trustllm.eu/

The Dutch Language Institute

- The treasure trove of the Dutch language

- The production, linking and disclosure of source material for the Dutch language in the form of historical and modern text corpora, dictionaries, lexical databases, grammars and the development of specific technological tools..

Hedendaags Nederlands



Corpus Hedendaags Nederlands

- primaire databron
- monitoren van het Nederlands vanaf 1990
- doorlopend aangevuld met krantenmateriaal, online fora, ondertitels
- dankzij overeenkomst met DPG en Mediahuis
- intern gebruik voor taaldocumentatie, ook toegankelijk voor externe onderzoekers via login
- 2 miljard woorden in 5 miljoen documenten

Participate!

Data Providers
To train GPT-NL we need enormous amounts of data to ensure a diverse and inclusive language model which is widely appliable. Therefore each data donation is much valued and will ensure relevance of GPT-NL in your sector.

Use-case provider
to validate GPT-NL we want to test the model on several use-cases. By sharing your use-case you contribute to the quality of GPT-NL. Meanwhile you can check the relevance of GPT-NL for your use-case and get specific advice on what the model can mean for you.

What does it mean for the near future?

EUROPEAN LANGUAGE DATA SPACE

GPTNederlands

TRUSTLLM

A benchmark of trustworthy large language models.


EU AI ACT

www.ivdnt.org
https://ivdnt.org/taalmaterialen/

/

Bronnen

https://aclanthology.org/www.mt-archive.info/70/TMI-1985-White.pdf

https://lilab.unibas.ch/staff/tenhacken/Applied-CL/3_Systran/3_Systran.html#history

https://www.systransoft.com/

https://www.systransoft.com/lp/free-online-translation/

Madhumita, M. (2024) Code Dependent. Picador.

Madhumita, M. (2024) In de schaduw van AI. Business Contact.

Steurs, F. (2024). Taal in transformatie. Technologie, economie en de kracht van generatieve AI. Scriptum.