

Intelligente technologie: hoe staat het met het Nederlands?

Frieda Steurs
*Wetenschappelijk Directeur/
Bestuurder*

KANTL
Colloquium ChatGPT
21 juni 2024



Onze werkagenda: mails, kalender, meetings, transcripts, etc.

LLM's zijn bijna allemaal getraind op het Engels

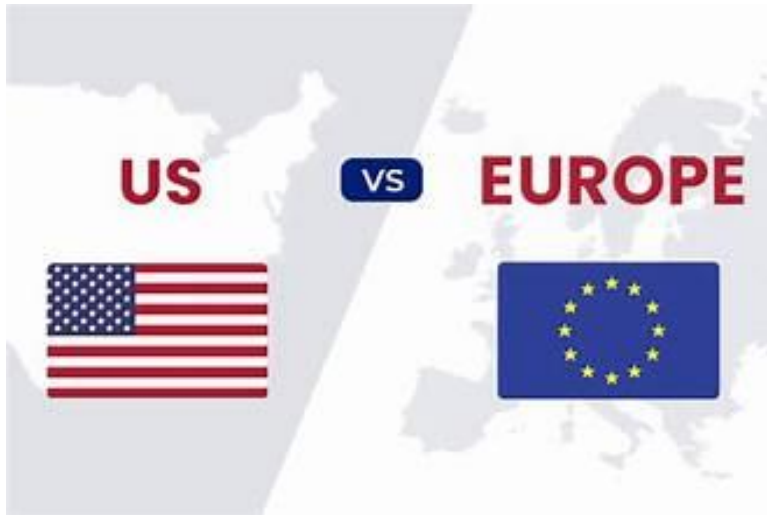
Microsoft ondersteunt 8 grote talen

Microsoft Bing



Andere AI ontwikkelaars

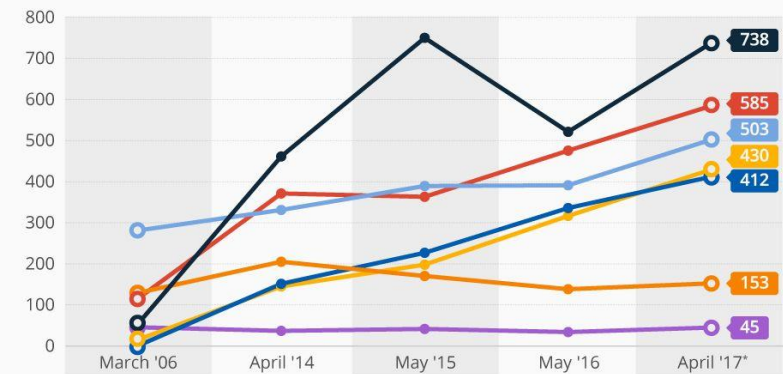




U.S. Tech Giants Still Growing

Market cap of U.S. tech and internet companies (in billion U.S. dollars)

● Apple ● Microsoft ● IBM ● Google/Alphabet ● Facebook ● Amazon ● Yahoo



CC BY-SA
@StatistaCharts

* As of April 20
Source: Yahoo! Finance

statista

BUSINESS

ChatGPT Shows Just How Far Europe Lags in Tech

Analysis by Lionel Laurent | Bloomberg

February 21, 2023 at 2:12 a.m. EST



Comment 1



Gift Article



Share

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.

Wie betaalt wie?



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.



Meer data, graag!

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](#).



Europees parlement 2018



11 september 2018

Taalgelijkheid in het digitale tijdperk

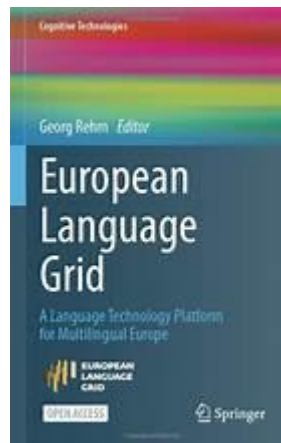
Nieuw beleidsterrein “**meertaligheid en taaltechnologie**”

De lidstaten worden aangemoedigd om een *allesomvattend taalgerelateerd beleid* te ontwikkelen en daartoe middelen toe te wijzen, dit om de taalkundige verscheidenheid en meertaligheid in de digitale sfeer te bevorderen.

Het is een gedeelde verantwoordelijkheid van de EU én de lidstaten om een bijdrage te leveren aan *het behoud van hun talen in de digitale wereld*. Het belang van digitale mogelijkheden voor vertaling, en de *open toegang* tot de gegevens die nodig zijn voor technologische vooruitgang.

Europese projecten

META  NET



Het delen van taaldata in Europa



LEARN • DISCOVER • PARTICIPATE • CONNECT • LOGIN

Search & exchange language resources

META-SHARE is an open and secure network of repositories for sharing and exchanging language data, tools and related web services

Share your own resources!

[JOIN OUR NETWORK NOW](#)

Already a member? [Log in](#)

32 % text corpora

27,630 number of downloads



CLARIN

Common Language Resources and Technology Infrastructure



ELRC-SHARE Repository

Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are of European interest and considered useful for feeding the CEF Automated Translation (CEAT) platform.

If you want to contribute resources, all you have to do is [register](#) (new user) or [login](#) (returning user) and go on to describe your resources.



Language Technologies

Discover, try out, use and download LT services and resources for all European languages.

Browse ELG and find the LT services, resources, developers and providers you are looking for.

Search the catalogues

8000 Corpora

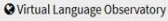
3884 Tools & Services

2812 Conceptual Resources

510 Models & Grammars

1775 Organizations

513 Projects



Search Contributors Help

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or [continue](#) to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

[See all records](#) [Take a quick tour](#)

Search through 1,030,321 records



Search

page 1 of 55

Order by: Resource Name A-Z

Arabic & Czech

Arabic & Czech consists of dependency treebanks used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The Conference on Empirical Natural Language Processing (CoNLL) is accompanied every year by a shared task intended to promote natural language processing research.

Resource	Access	License
Arabic & Czech (CoNLL 2006)	Available	CC-BY
Arabic & Czech (CoNLL 2006)	Available	CC-BY

Ten Languages

Arabic, Japanese, Portuguese, Slovenian, Spanish, Croatian, Swedish, Turkish

Europese projecten

ELE



European Language Data Space

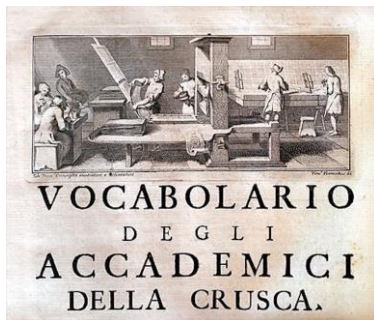


EUROPEAN
LANGUAGE
EQUALITY





/instituut
voor de
Nederlandse
taal/



Språk-
BANKEN



INSTITUT FÜR
DEUTSCHE SPRACHE



Eesti Keele Instituut

Dansk
Sprognævn



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

GPT-NL



Nederland gaat een eigen open taalmodel ontwikkelen: GPT-NL. Dit model is nodig voor het ontwikkelen, versterken en bestendigen van de **digitale soevereiniteit**.

Toetsbaar gebruik van AI **naar Nederlandse en Europese waarden** en richtlijnen en met **respect voor het eigenaarschap van data**.
Financiering : ministerie van EZK.

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)

	 OpenAI	 cohere	 stability.ai	 ANTHROPIC	 Google	 BigScience	 Meta	 AI21 labs	 ALEPH ALPHA	 EleutherAI
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

SURF

TNO innovation for life

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

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)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALPHA
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude 1	PaLM 2	BLOOM	LLaMA		
Data sources	● ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ○	●			
Data governance	● ● ● ○	● ● ● ●	● ● ● ○	○ ○ ○ ○					
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○						
Compute	○ ○ ○ ○	○ ○ ○ ○							
Energy	○ ○ ○								
Capabilities & limitations									
Disinformation									
Machine bias									
Model provenance									
Downstream documentation									
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48

"Impossible": OpenAI admits ChatGPT can't exist without pinching copyrighted work



TRUSTLLM

A benchmark of trustworthy large language models.

Europese grote taalmodellen (LLM's) ontwikkelen die de meest betrouwbare zijn in Europese AI, voor een reeks ondervertegenwoordigde talen.

Deens, Duits, IJslands, Nederlands, Noors en Zweeds

De ontwikkeling van een open, betrouwbare en feitelijke LLM, in eerste instantie gericht op de Germaanse talen.

De basis voor een geavanceerd open ecosysteem voor modulaire en uitbreidbare Europese betrouwbare, duurzame en gedemocratiseerde LLM's van de volgende generatie.

<https://trustllm.eu/>

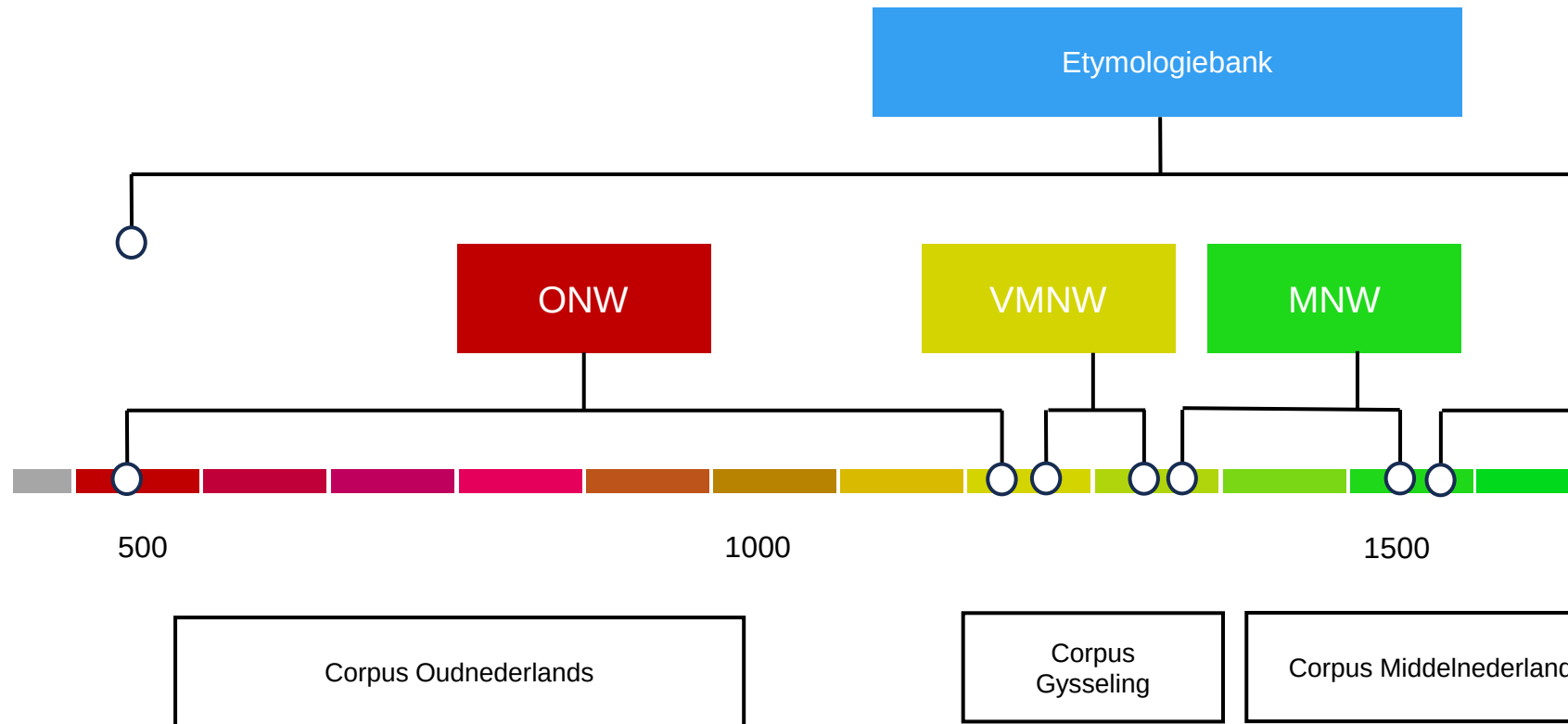
Het Instituut voor de Nederlandse Taal

- o **De schatkamer van het Nederlands**
- o De productie, koppeling en ontsluiting van **bronnenmateriaal** voor het Nederlands in de vorm van historische en moderne tekstcorpora, woordenboeken, lexicale databanken, grammatica's en de ontwikkeling van specifieke **technologische hulpmiddelen**.

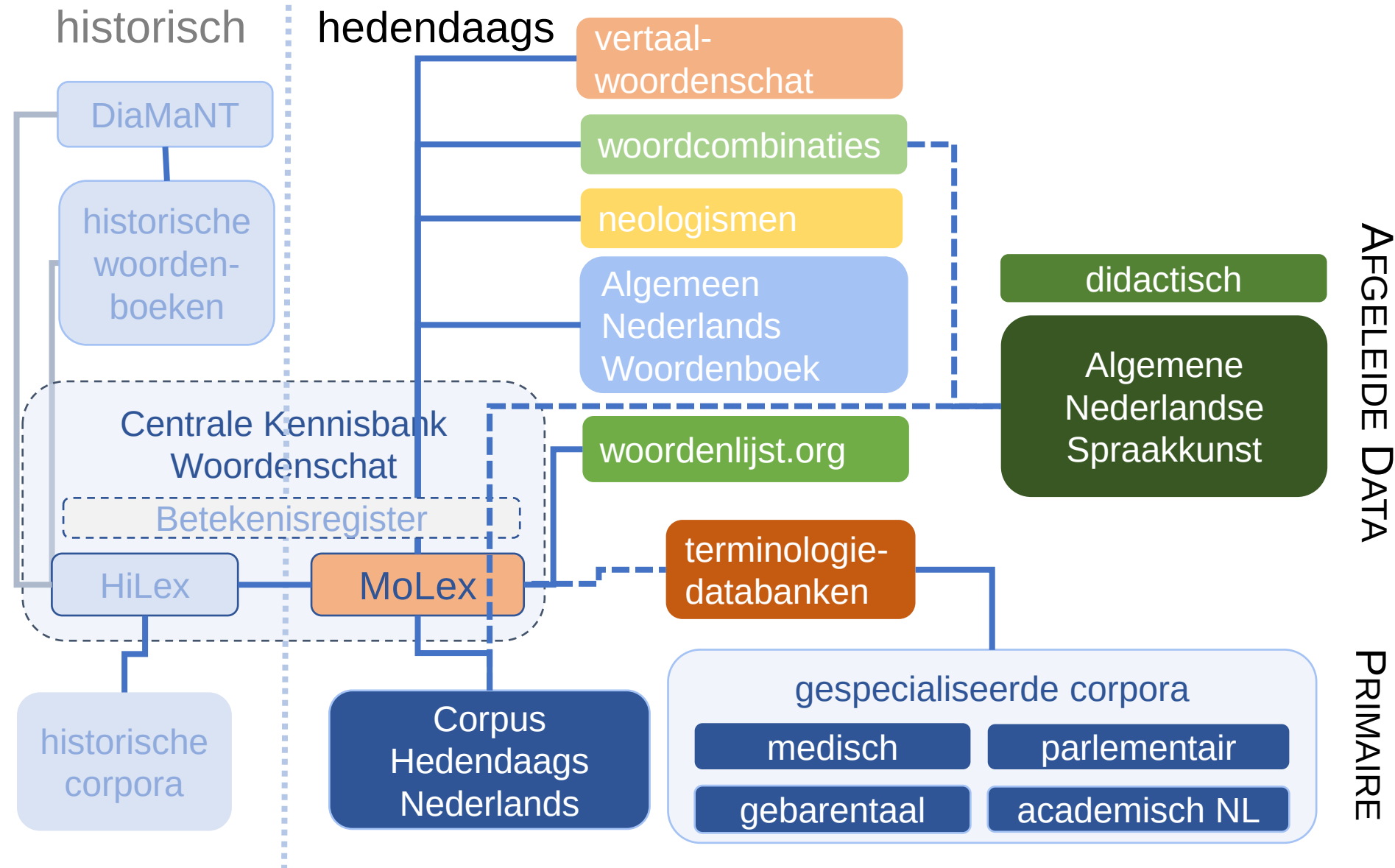


Woordenboeken en tekstcorpora

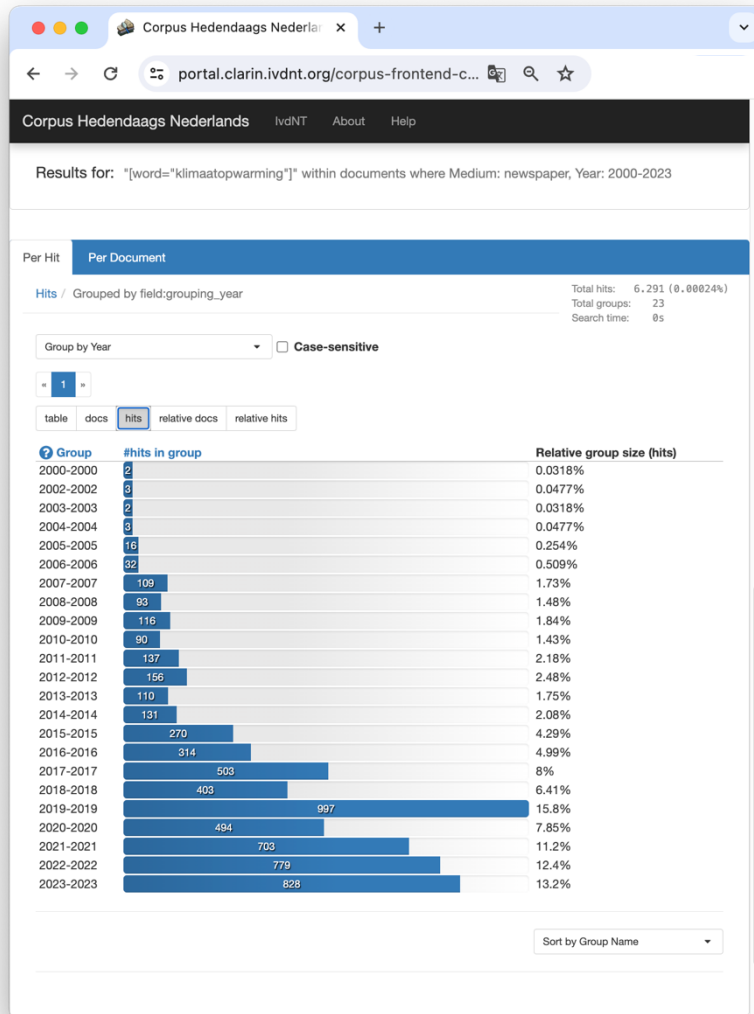
historisch en modern



Databanken hedendaags NL



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

Wat moet er gebeuren?

Een lijst met beschikbare taalmaterialen

<https://ivdnt.org/taalmaterialen/>

Bekijken hoe we bepaalde licenties moeten interpreteren:
commercieel gebruik of niet?

In onze standaardlicentie staat het volgende:

INT verleent hierbij aan Gebruiker [...] het niet-exclusieve recht om: Product te gebruiken ten behoeve van het ontwikkelen van nieuwe producten [...], mits (de inhoud van) Product, een gedeelte van Product of een kwalitatief of kwantitatief substantieel gedeelte daarvan niet herkenbaar in de Nieuwe Producten is opgenomen, verveelvoudigd of overgenomen.

Participate!

Data Providers

To train GPT-NL we need enormous amounts of data to ensure a diverse and inclusive language model which is widely applicable. 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.

Wat betekent dit de nabije toekomst?





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

Schrijf je in voor onze nieuwsbrieven:

- https://ivdnt.org/over-ons/nieuwsbrieven/#aanmelden_nieuwsbrief
- algemene nieuwsbrief (4x per jaar)
- terminologie nieuwsbrief (4x per jaar)