EUROPEAN LANGUAGE DATA SPACE

Taaldata en taaltechnologieën voor de landstalen in Nederland en België

Panel I

Walter Daelemans (University of Antwerp, Belgium) - Moderator
walter.daelemans@uantwerpen.be

23-01-2025 LDS Workshop, Breda, The Netherlands
https://language-data-space.ec.europa.eu

EUROPEAN LANGUAGE DATA SPACE

Participants

Moderator:

Walter Daelemans (Universiteit Antwerpen)

Matthias De Lange (Techwolf)

Carla Verwijmeren (Y.digital)

Peter Spyns (WEWIS)

Frieda Steurs (INT)

Taaltechnologie voor het Nederlands

- Degelijke **opleidingen**, goede **research funding**, een bloeiende ecologie van (kleine) **bedrijven.**
  - LLM's domineren (vooral GPT's).
  - Grote commerciële bedrijven domineren (vooral US).
  - Voor het Nederlands ontbreken taalmodellen, benchmarks, maar vooral **DATA.**
  - De kwaliteit van spraakherkenning voor accenten en dialecten blijft ondermaats.
- Initiatieven voor het Nederlands:
  - ***Leaderboard LLMs voor het Nederlands:*** https://scandeval.com/leaderboards/dutch/
  - Meertalige modellen als GPT-4-o en Llama 3 presteren redelijk goed voor het Nederlands.
  - GPT-NL als specifiek Nederlands taalmodel.
  - *Continued pre-training van* Mistral en Llama modellen met Nederlandse data (GEITje, fietje, ChocoLlama, …)
- Trends:
  - Open-source modellen, die lokaal worden ontwikkeld en ingezet, creëren nieuwe mogelijkheden.
    - Distillatie, *test-time computation*
  - Gebruik van **datasynthese** met LLM's als alternatief voor ontbrekende data.

Introductie: Carla Verwijmeren van Y.digital

carla@y.digital | 06-31991385

- Partner bij Y.digital
- Bestuur Nederlandstalige Spraakcoalitie
- AI-commissie Klantenservice Federatie

- Y.digital ontwikkelt en implementeert betrouwbare AI-oplossingen. We hebben diepgaande expertise op het gebied van kennismodellering & (generatieve) AI.
- We richten we ons op streng gereguleerde markten, waar bescherming van vertrouwelijke data essentieel is, zoals overheid, financiële sector & uitvoeringsorganisaties.
- Wij doen dit via ons in Europa gehoste AI-platform Ally. Hiervoor hanteren we de hoogste standaarden voor dataprivacy en datasecurity.

De praktijk van spraaktechnologie voor Vlaamse accenten

Goed, Ik weet niet wat Het is, dus Ik heb mijn mijn gaskachel lekker direct uitgelegd afgelegd. Ik ga nu de gasten voeren dicht te rijden, maar.

‘Gasten voeren dicht te rijden’ = Gas toevoer dichtdraaien

Ja, ik zit ermee, want God hoor ik toch. Het is een sterke patroon licht dat ik binnenkwam.

‘patroon licht’ = petroleum lucht

Mike geboortedatum alsjeblieft.

‘Mike’ = mag ik uw

Als messysteem meerveld kan ik dat snel voor u in orde brengen.

‘Als messysteem meerveld’ = als het systeem meewerkt

TECHWOLF

We help large enterprises maintain a continuous overview of the skills across their workforce.

- Use case driven
  — **L&D Optimization** — **Internal Mobility** — **Strategic Workforce Planning**
- Data centered
  — **Client data is unstructured** — **17+ billion data lake**
- AI powered
  — **Unbiased** — **50+ languages**

Matthias De Lange, PhD

Senior AI researcher

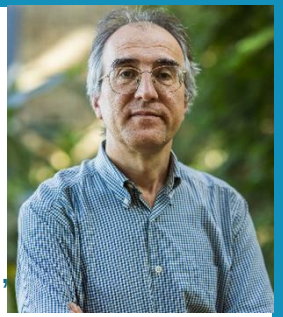
Global Investors

20VC felix ACADIAN NOTION SemperVirens STRIDE workday. servicenow SAP

Global Customers

ENGIE SYNOPSYS IQVIA KBC KLM HSBC UNITED SITA ucb bp workday. Booking.com Virgin Media O2 GSK Atlas Copco signify

Peter Spyns



- Works for the Flemish Public Administration – Dept. of Work, Economy, Science, Innovation & Social Economy (since 2006)
- Coordinator international R&I policy (since 2013) & senior AI (including HLT) advisor (since 2006)
- responsible for the research part of Flemish AI plan (since 2019)
- Participating in Flemish and federal AI steering boards
- Belgian delegate of the Horizon Europe AI, Data Robotics co-programmed partnership Member States Representatives Group (ADRA)
- Flemish observer at the Assembly of Members of the Alliance for Language Technology European Digital Infrastructure Consortium (ALT-EDIC)
- Belgian delegate at the General Assembly of the Common Language Resources and Technology Infrastructure European Research Infrastructure Consortium (CLARIN-ERIC)
- Former programme coordinator of STEVIN [joint FL-NL HLT programme] (2006-2012)
- Former researcher at UZ Leuven, UZ Gent, VUB & former employee at L&H Speech Products

PhD in Informatics (medical language processing), M.A. (Romance philology)

/instituut voor de Nederlandse taal/

Hét kennisinstituut voor het Nederlands

Hedendaags Nederlands | Historisch Nederlands | Taalmaterialen | Terminologie | Nieuws

frieda.steurs@ivdnt.org

- Directeur Instituut voor de Nederlandse Taal
- Prof.em.KU Leuven
- Secretary-general CIPL

- Nederlands-Vlaams instituut
- Focus op de digitale taalinfrastructuur voor het Nederlands
  - **Primaire data:** compilatie van linguïstisch verrijkte corpusdata uit ruwe taalgebruiksdata (geschreven, gesproken, gebaren)
  - **Secundaire data:** taaldocumentatie, beschrijving van woordenschat en spraakkunst, zowel algemeen als vaktaal (terminologie)
  - **Software en services** voor analyse, verwerking, ontsluiting, publicatie en ondersteuning van onderzoek en ontwikkeling
- Zowel ontwikkeling eigen data/software als depot voor anderen
  - vrij beschikbaar voor onderzoek en ontwikkeling
  - duurzaam beschikbaar na afloop van O&O-projecten

/instituut voor de Nederlandse taal/taalmaterialen

https://ivdnt.org; https://taalmaterialen.ivdnt.org/

Stellingen en debatpunten

- Stelling: Onafhankelijkheid van grote (voornamelijk Amerikaanse) commerciële taalmodellen is essentieel en urgent.
- Stelling: Het gebruik van auteursrechtelijk beschermde data voor de training van LLM's valt onder *fair use*, omdat het transformatief is en van groot maatschappelijk belang.

- Welke rol moeten de overheid en de bedrijven spelen bij de ontwikkeling van taalmodellen?
- Hoe kunnen datahouders (uitgeverijen, media, enz.) worden overtuigd om data beschikbaar te stellen voor de ontwikkeling van taalmodellen?
- Is het gebruik van synthetische data, gegenereerd door taalmodellen, een goed alternatief?
- Hoe waarborgen we inclusie van alle accenten en taalvarianten?

Common European Language Data Space

Thank you!

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Walter Daelemans (University of Antwerp, Belgium)
walter.daelemans@uantwerpen.be

23-01-2025, Breda, The Netherlands
https://language-data-space.ec.europa.eu