DEN DANSKE BEGREBS-ORDBOG

Frame Net

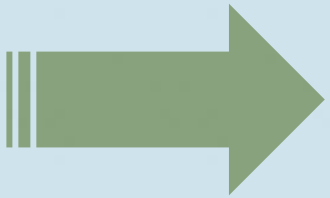
DanNet

COR
DET CENTRALE ORDREGISTER

Udvikling af danske benchmarkdata for sprogforståelse

med udgangspunkt i semantiske ordbøger

Finkornede datasæt kræver **ekspertviden** og er **tidskrævende** at indsamle.



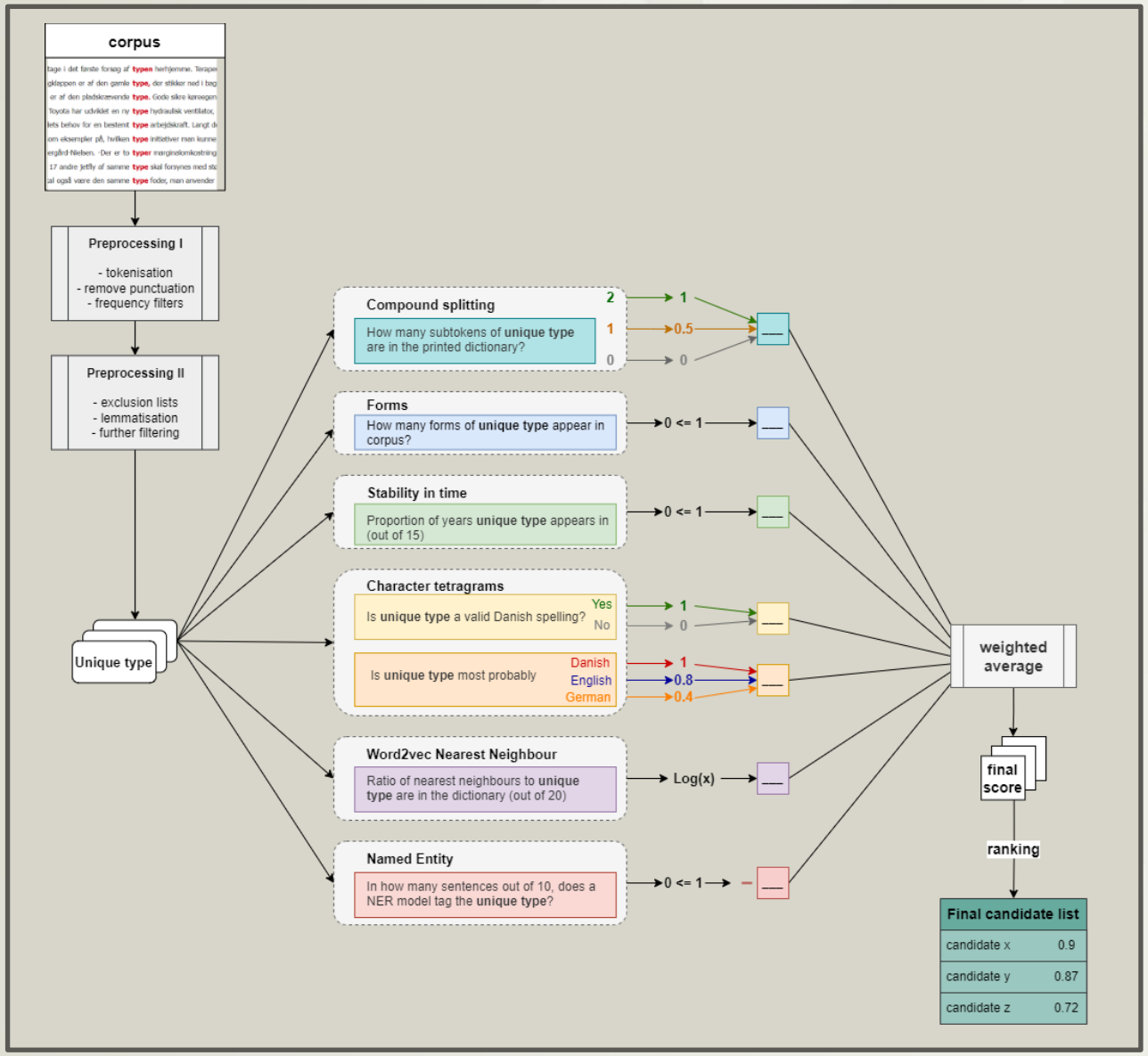
Kan vi genbruge ordbøger til at teste LLM's sprogforståelse?

Trawling the Corpus for the Overlooked Lemmas

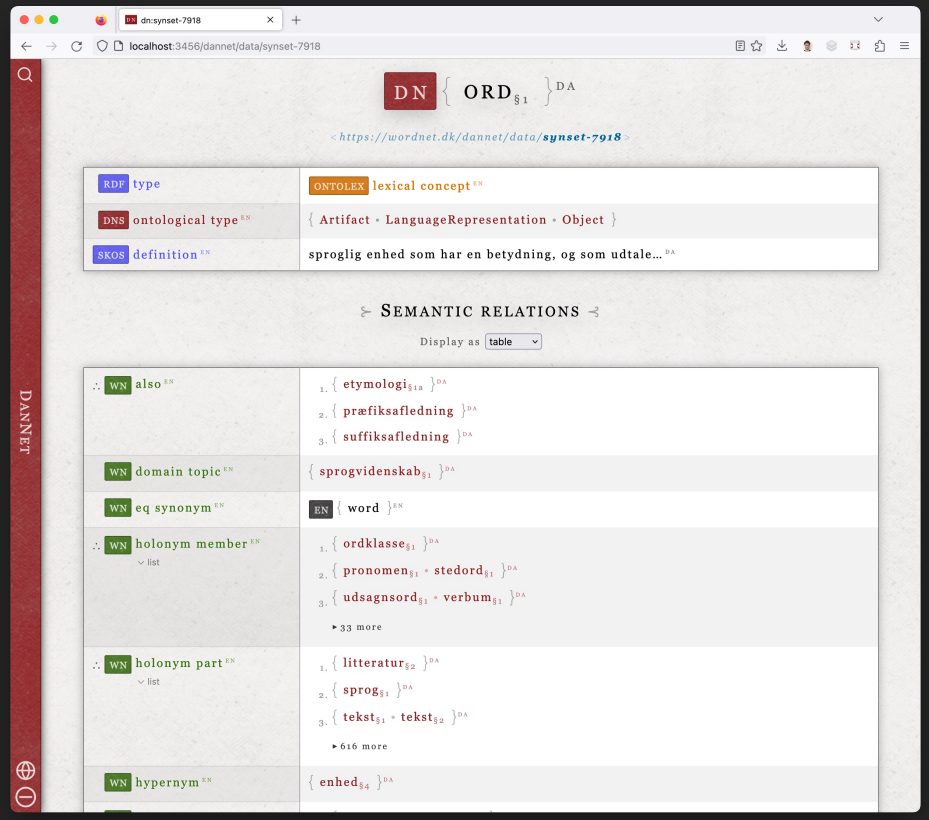
The Danish Dictionary (DDO)

- A corpus-based online monolingual dictionary
- Continuously updated with new words and senses

How do we find the mundane **lemmas** that our previous lemma selection methods have **overlooked**?



The new wordnet.dk/dannet DanNet




```
SELECT DISTINCT ?slang
WHERE {
  ?sense lexinfo:register      lexinfo:slangRegister .
  ?word  ontolex:sense         ?sense ;
         ontolex:canonicalForm ?form .
  ?form  ontolex:writtenRep    ?slang .
}
```

Danish Hyperbole Corpus: A Discourse-Level Approach

Nina Skovgaard Schneidermann

- Objective: Build and annotate a Danish hyperbole corpus to advance linguistic and NLP research.
- Hyperbole: An expression exceeding what's justified by the context, used to exaggerate for emphasis or effect.
- Example: "The Real Madrid players ran **3000 kilometers an hour**".
- Methodology: Use Danish articles, whole-text annotation, and a specialized two-step annotating process.
- Key Contribution: Support advancements in figurative language identification, sentiment analysis, and clickbait detection.
- Novelty: Address the limitations of existing corpora by providing contextual and comprehensive coverage of Danish hyperbole.

Efficient Task Adaptation for Transformer Encoders

Ali Basirat – Center for Language Technology (CST)
University of Copenhagen

- Fine-tuning is effective but costly, especially for structure prediction
- A new encoder adaptation method based on
  - Aggregation of the encoder's intermediate representation across token
  - Tailoring the aggregation to diminish task differences
- Experiments on a range of structure prediction
  - Most of the fine-tuning performance is retained at a fraction of the training cost
  - Consistent performance across different types of transformer encoders
  - Acceptable performance on other tasks (document classification)

According to BERTopic, what do Danish Parties Debate on when they Address Energy and Environment?

Costanza Navarretta and Dorte Haltrup Hansen (NorS, UCph)

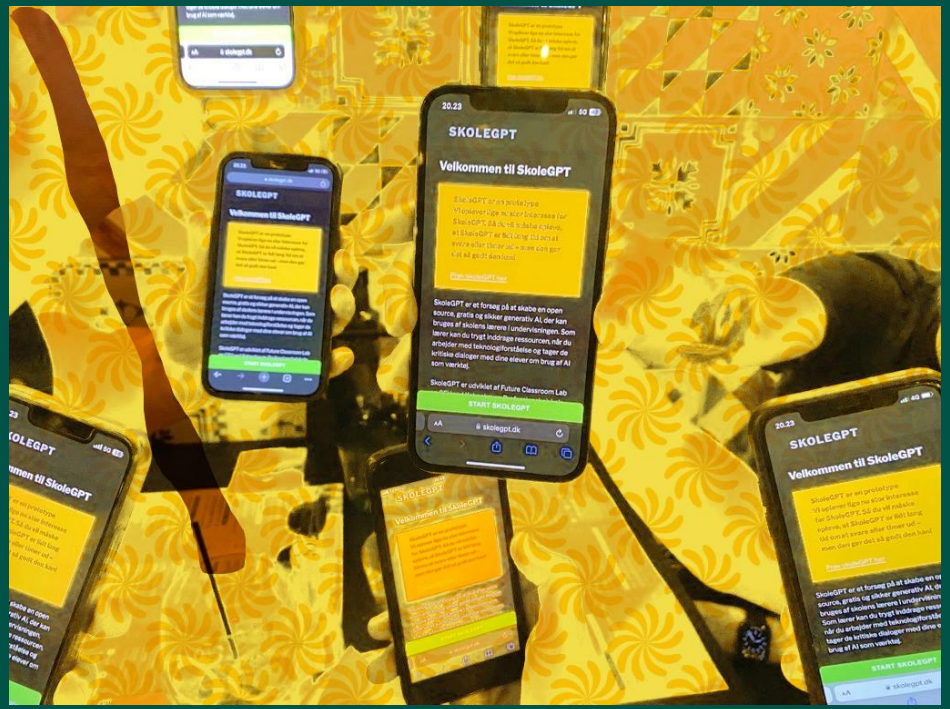
- We investigate how two policy areas, *Environment* and *Energy,* are dealt with by seven Danish left- and right-wing parties in their electoral manifestos (2007-2019) and in parliamentary debates (2009-2020), both encoded with policy areas.
- We use quantitative and qualitative analyses based on the subtopics generated by BERTopic on the parliamentary debates. BERTopic is trained with the multilingual word embeddings provided by the system and Danish word embeddings (https://certainly.io/blog/danish-bert-model/).

SKOLEGPT

Dansk generativ AI prototype

Åben, sikker og gratis

Til lærere og elever i grundskolen



Future Classroom Lab KP CFU

Louise Pram Nielsen (PhD, MSc)

PILOTFORSØG: Brugervendt Data Glossary

Forskningsspørgsmål: *Kan simple brugerforsøg omsættes til systematisk terminologi, som dels kan beriges med kunstig intelligens, dels kan udstilles i et datakatalog?*

Brugerforsøg (3 x 15 minutter):

1) **Skriv** de vigtigste fagudtryk
2) **Beskriv** betydningen
3) **Sorter** fagudtrykkene

TermLab

ChatGPT

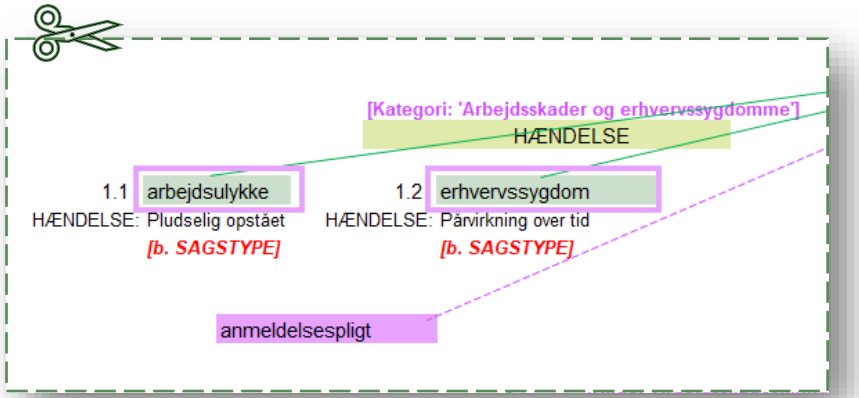
Prompt Engineering:

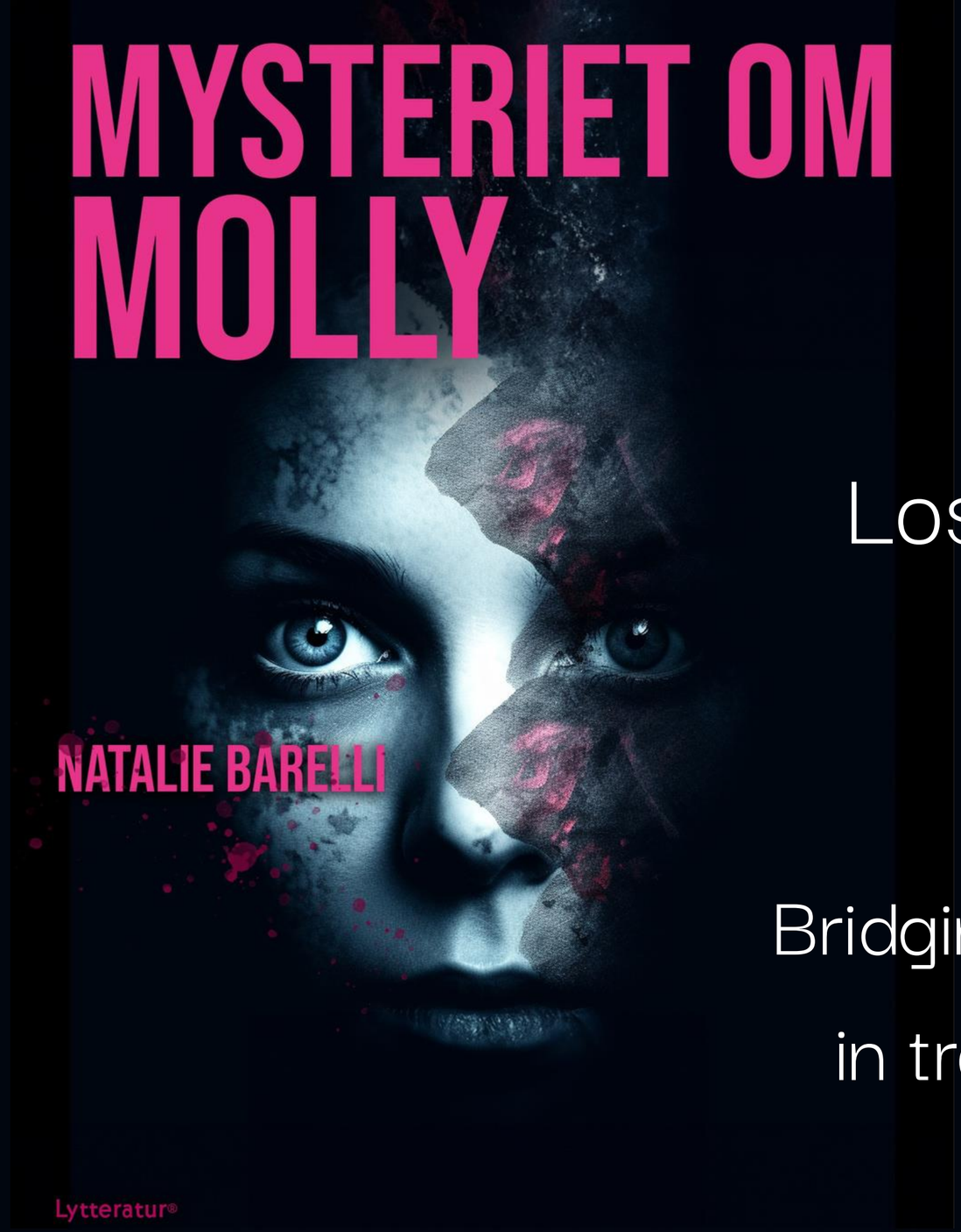
i. Kontekst Fagtekst (kilde: aes.dk)
Ii. Eksempel: 10 fagtermer med definition
iii. Instruktion: Persona + 2 opgaver: 1. Find flere termer + 2. Strukturer termerne

Knowledge engineering:

Bruger: 11 termer + 5 inddelingskriterier
Chat-GPT: 10 termer + 3 kategorier
Berigelse: 2 nye termer (uændret struktur)

Terminologisk ontologi





Lost in Literary Machine Translation?

Bridging the Nordic-English Divide in translated fiction with MTPE

Sprogteknologisk Konference 2023 | Center for Sprogteknologi (CST)

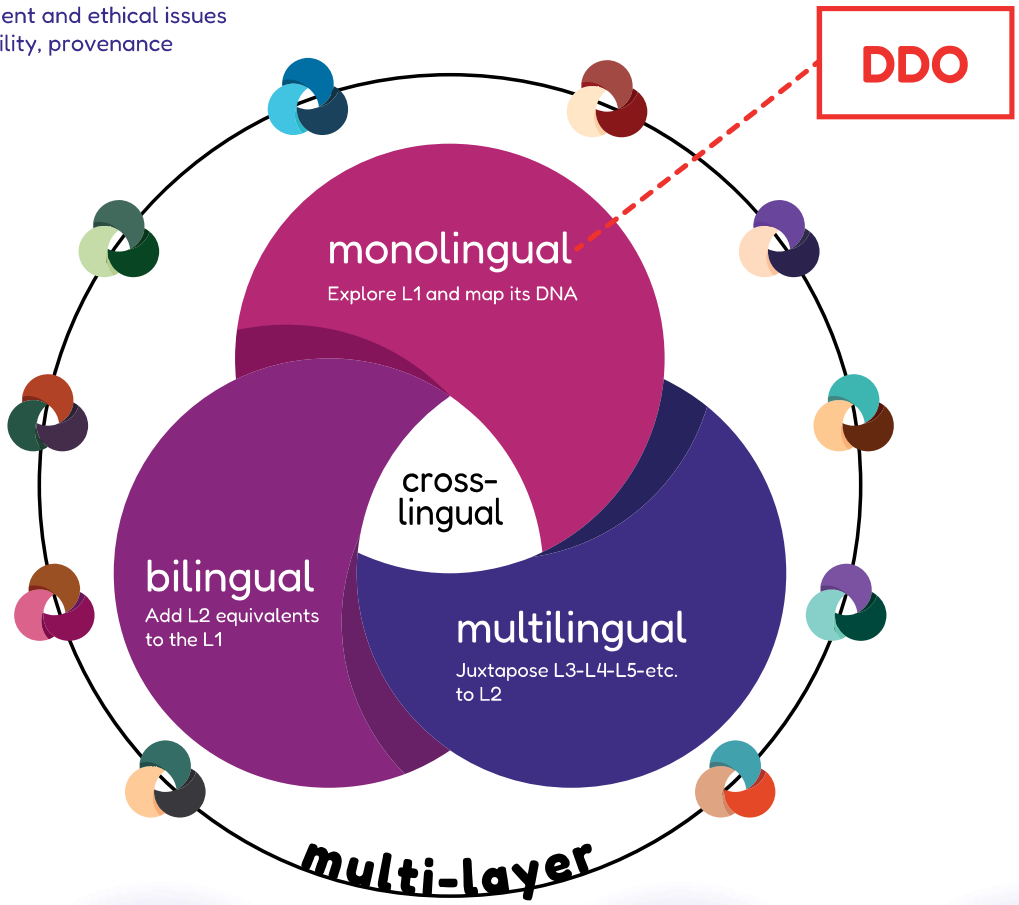
Integrating Multi-layer Lexical Data in Language Model Training

LLM Flaws

- expensive to train and run
- GPU environmental pollution
- undecipherable black box
- non-interoperable across knowledge bases
- web-crawled data noise and bias
- hallucination – deceptive fluency
- inconsistency
- long tail on under-resourced languages
- copyright infringement and ethical issues
- lacking trust, reliability, provenance

Examples of Usage

- typical language patterns
- short phrases and full sentences
- (foreign) language learners
- reception and production
- expert training data
- prime parallel corpora

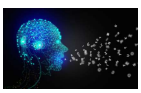


Mapping the Language DNA

- word sense disambiguation
- multiword expressions
- register and domain classification
- synonyms and antonyms
- phonetic transcription
- alternative spelling and scripts
- inflected forms
- cross-reference

Outcomes/Advantages

- **Precision. Efficiency. Compactness. Savings.**
- **Fine-Tuning.** domain/task-specific data
- **Specialized Vocabulary.** terminology, jargon
- **Quality Control.** evaluate and improve performance
- **Language-specific Nuances.** better translations
- **Ethical and Bias Mitigation.** promote inclusivity, diversity, fairness
- **Intellectual Property Rights.** ownership and branding



Ilan Kernerman · Lexicala by K Dictionaries · ilan@lexicala.com · https://lexicala.com

lexicala by K DICTIONARIES

Sprogteknologi og EU

Language Data Space

HVAD KAN DU FÅ UD AF DETTE?

Gode data er fundamentet for god sprogteknologi. Men det er ofte en udfordring at finde gode data for særligt lavressourcesprog som dansk, ligesom at det kan være svært at gennemskue anvendelsesvilkår for specifikke data. I EU er man derfor påbegyndt etableringen et fælles-europæisk *Language Data Space* (LDS) for sprogdata.

LDS har til formål at udgøre en funktionel platform og markedsplads, hvor multilinguale og multimodale sprog-data kan indsamles, skabes, udveksles og handles på tværs af både den private og den offentlige sektor. Alt dette på dataejernes vilkår.

Med et data space vil man altså sikre decentral datadeling, datasuverænitet samt transparens i både dataudveksling og dataøkosystemer. Derudover vil man sikre, at der i anvendelsen af sprogdata tages højde for EU's værdier, herunder beskyttelse af persondata.

DANSK DELTAGELSE

Der er to styrende organer i LDS. Det ene består af myndighedsrepræsentanter fra hvert medlemsland. Dette organ skal bl.a. udarbejde regelsæt og understøtte videndeling omkring LDS og dets aktiviteter.

Det andet styrende organ skal bestå af aktører fra sprog-teknologisk relevante virksomheder. Dette organ får til opgave at koordinere udviklingen af de arkitektoniske og infrastrukturelle byggeblokke for platformen.

Alliancen for sprogteknologi

ET DEDIKERET FLERLANDEPROJEKT

Alliancen for sprogteknologi er et 'digitalt infrastruktur-konsortium', dvs. en sammenslutning af medlemslande, som dedikerer sig til at fremme et område af Europas digitale infrastruktur, og som er finansieret delvist af de deltagende medlemslande og delvist af EU.

HVAD KAN DU FORVENTE?

Fem handlingspunkter skal sikre alliancens formål ved;

1. At tilvejebringe højkvalitetssprogdata i tæt samarbejde med LDS og herunder udarbejde et rammeværktøj for 'kunstig datagenerering' for lavressourcesprog.
2. At indsamle, finjustere, reducere og optimere eksisterende sprogmodeller med henblik på at øge modellernes anvendelighed.
3. At pulje ressourcer, herunder computerkraft, til at understøtte udviklingen af nye sprogmodeller.
4. At tilvejebringe metodologier for evaluering, certificering og normalisering af sprogmodeller med særligt fokus på potentiel diskrimination og bias.
5. At styrke det europæiske sprogteknologiske økosystem gennem facilitering af samarbejder mellem industri og forskning, inkubation af sprogteknologiske start-ups, understøttelse af den europæiske strategiske retning for sprogteknologi, mv.

SKAL DANMARK VÆRE MED?

Danmark er for nu meldt ind som observatør i alliancen, men der skal stadig tages en beslutning om, hvorvidt Danmark skal være medfinansierende deltager i alliancen.

Sprogteknologi.dk

DANMARKS NATIONALE INDSATS FOR SPROGTEKNOLOGI

Som led i Digitaliseringspagten og Økonomiaftalerne for 2020 mellem Regeringen, KL og Danske Regioner blev det besluttet at igangsætte et fællesoffentligt samarbejde om udviklingen af en 'fælles digital dansk sprogressource'.

Arbejdet har siden taget sit udgangspunkt i nogle af de anbefalinger, som Sprogteknologiudvalget, nedsat af Dansk Sprognævn under Kulturministeriet, kom med i rapporten 'Dansk Sprogteknologi i Verdensklasse' i april 2019. Digitaliseringsstyrelsen udgør indsatsens sekretariat og varetager formandskabet for indsatsens styregruppe.

Indsatsen fokuserer dels på at 1) indsamle og udstille metadata om danske sprogressourcer på portalen sprogteknologi.dk, 2) tilvejebringe nye danske sprogres-sourcer, 3) vidensdisseminere om dansk sprogteknologi og 4) facilitere et netværk for det danske sprogteknologiske aktørlandskab.

EKSEMPLER PÅ DELPROJEKTER

- 166 metadatabeskrivelser på sprogteknologi.dk.
- 700+ følgere på LinkedIn.
- Årlig sprogteknologisk konference.
- Det Centrale Ordregister.
- CoRaL – et danskssproget taledatasæt.
- Sundhedsfaglig tekstkorpus.
- Europæisk samarbejde.
- Gå-hjem-møder og workshops.

DIGITALISERINGSSTYRELSEN

MORE DATA, FASTER

- Semi-automatic construction of acceptability corpora

- *Using SpaCy, DaCy*
- *and a method so simple*
- *it's almost crazy*

Mikkel Niclasen, BA in linguistics
Studying MA in IT & Cognition @UCPH