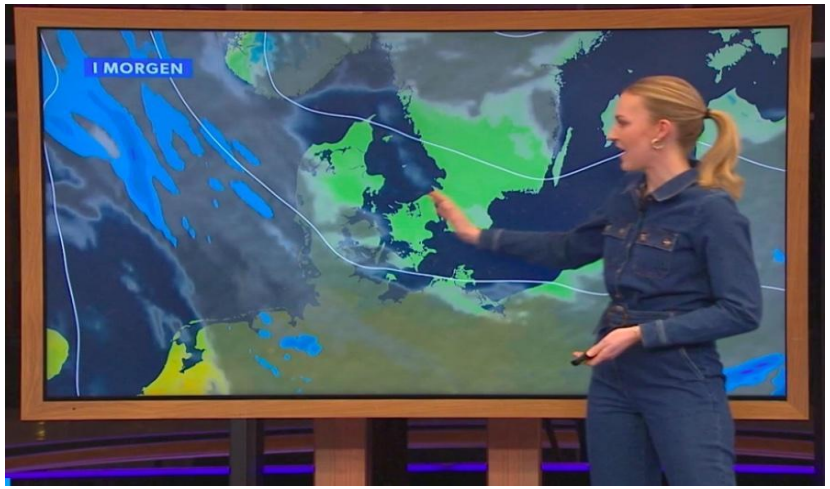
Co-speech gestures and information uptake

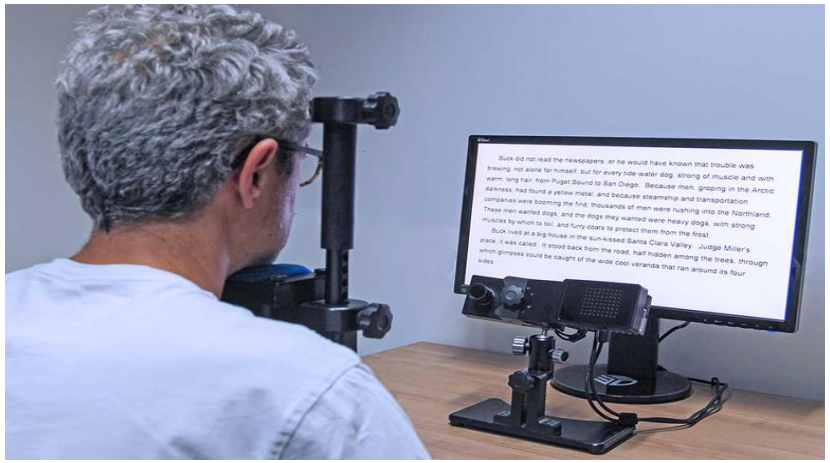
Alberto Parola and Patrizia Paggio, Center for Language Technology, KU

Co-speech gestures convey information related to the ongoing talk (e.g. movement direction) and complement oral meaning

Not clear to what degree a co-speech gesture which convey essential information is fixated by the addressees

We aim to investigate gestural uptake by showing participants videos of a TV presenter performing a weather forecast using directional gestures and use eye-tracking to assess fixation on gestures and its relationship to their uptake.





Noise, Novels, Numbers:
A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature

Ali Al-Laith, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, Timothy R Tangherlini

📢 **Explore:** How "aberrant sonic behaviour" reflects cultural shifts during the Modern Breakthrough (1870–1899).

📚 **Discover:** Two corpora annotated into **noise/non-noise** and **human, non-human, and musical noises**.

🤖 **Learn:** How cutting-edge pre-trained language models uncover patterns in historical soundscapes.

🌟 **Understand:** The interplay between noise, literature, and cultural history.

Contribution of Linguistic Typology to Universal Dependency Parsing

Ali Basirat and Navid Baradaran Hemmati

Center for Language Technology (CST)
Department of Nordic Studies and Linguistics
University of Copenhagen

- Universal Dependencies (UD)
  - syntax description of all human languages
- UD lacks linguistic typology
  - Language-specific labels
  - Inconsistent labels
- We address this issue
  - Introduce label transformation rules
  - Extensive analysis of multiple languages
- Linguistic typology improves the UD scheme

- Presented at EMNLP2024

UNIVERSITY OF COPENHAGEN

EMNLP 2024

Contribution of Linguistic Typology to Universal Dependency Parsing

An Empirical Investigation

Ali Basirat, Center for Language Technology, University of Copenhagen, alib@hum.ku.dk
Navid Baradaran Hemmati, Certified Translation Agency No. 1141, Mashhad, Iran, navidbh@gmail.com

Abstract

Universal Dependencies (UD) is a global initiative to create a standard annotation for the dependency syntax of human languages. Addressing its deviation from typological principles, this study presents an empirical investigation of a typologically motivated transformation of UD. Our findings underscore the significance of the transformations across diverse languages and highlight their advantages and limitations.

Introduction

- Universal Dependencies (UD) [2] is widely recognized as a standard framework for morphosyntactic annotation across human languages.
- Although UD provides extensive multilingual coverage, it places limited emphasis on language typology and linguistic universals [3].
- Research in linguistic typology has proposed modifications to UD's dependency annotation schemes to better capture cross-linguistic variations [1].

UD Enriched with Linguistic Typology

The following four design principles aim to enrich UD with insights from linguistic typology [1]:

1. Distinguish universal constructions from language-specific strategies, favoring classification based on the former.
2. Use consistent labels for the same functions, regardless of whether they are realized syntactically or morphologically.
3. Prioritize the way information is structured in sentences (information packaging) over the lexical semantics of individual words.
4. Incorporate a hierarchical consideration of dependency relations, accounting for different levels of structure such as predicates, arguments, modifiers, and adverbs that qualify modifiers.

Research Objectives

1. To enhance the Universal Dependencies (UD) annotation scheme by incorporating linguistic typology, following the four proposed design principles.
2. To evaluate the practical benefits of the typologically-enriched UD scheme (TUD) across a range of diverse languages.

Hypothesis: Dependency parsing using the TUD scheme is more effective than using the standard UD scheme.

From UD to TUD

- A rule-based conversion process is applied, grounded in the proposed typological principles.
- Each rule maps a dependency label in the UD scheme to its corresponding label in the TUD scheme.
- The conversion process affects only the dependency labels, leaving the tree structures unchanged.

Manual Evaluation

- To assess the effectiveness of the conversion rules, we manually evaluate their performance on random samples of 25 and 50 sentences from the development sets of Persian and English, respectively.
- Across both languages, 1–2% of the evaluated tokens were incorrectly labeled.

Parsing Performance

The practical benefits of the TUD scheme are evaluated via parsing performance:

- 20 treebanks from UD 2.12, representing diverse language families, are selected.
- Three transition-based and graph-based parsing models are trained, each with different random seeds.
- The scheme's effectiveness is measured by improvements in the average labeled attachment score (LAS).

Language	Treebank	Family	Genus	Size	IR
Arabic	padt	Afro-Asiatic	Semitic	254K	20%
Armenian	armtdp	Indo-European	Indo-Iranian	47K	25%
Basque	bdt	Isolate		97K	26%
Chinese	gsd	Sino-Tibetan	Sinitic	111K	23%
Cl-Chinese	kyoto	Sino-Tibetan	Sinitic	406K	31%
English	ewt	Indo-European	Germanic	230K	33%
Finnish	tdt	Uralic	Finno-Ugric	181K	29%
Hindi	hdtb	Indo-European	Indo-Iranian	316K	22%
Italian	isdt	Indo-European	Romance	288K	34%
Korean	gsd	Koreanic	Altaic	69K	23%
Latin	itb	Indo-European	Italic	421K	33%
Latvian	lvtb	Indo-European	Baltic	253K	29%
Marathi	ufal	Indo-European	Indo-Iranian	3K	30%
Persian	seraji	Indo-European	Indo-Iranian	137K	26%
Russian	taiga	Indo-European	Slavic	187K	28%
Swedish	talbanken	Indo-European	Germanic	76K	34%
Turkish	imst	Turkic	Altaic	48K	28%
Urdu	udtb	Indo-European	Indo-Iranian	123K	24%
Vietnamese	vtb	Austroasiatic	Vietic	46K	31%
Wolof	wtb	Niger-Congo	Atlantic-Congo	34K	28%
		Average		166K	28%

The selected treebanks with statistics about their sizes and transformed token ratios (Col. IR)

Results

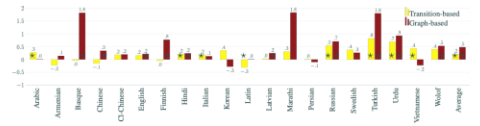


Figure 1: Absolute LAS improvement (or degradation). Significant results with p-value < 0.05 are marked.

- Most treebanks show better parsing results under TUD than under the standard UD.
- Except for Latin, the negative results are not statistically significant.
- The results highlight the positive impact of typological transformation without significant negative effects.
- A comparison with random transformations confirms that the parsing improvements stem from the typological motivations behind the rules.

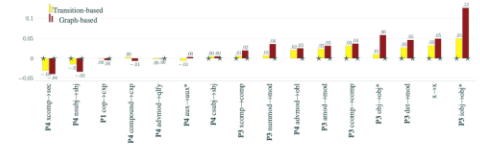


Figure 2: The relative contribution of the transformation rules. The results with p-value < 0.05 are marked.

- The third principle contributes the most to the improvement in parsing accuracy.
- The fourth principle, involving fragmentation rules, has the most negative impact on performance.
- The first principle, represented by a single rule, has a neutral effect.
- The second principle is not reflected in the transformations, as UD already adheres to it.

Conclusion

- The typologically-enriched scheme maintains UD's practical goals while providing extra benefits.
- We propose adopting this scheme as an alternative foundation for treebanking efforts.
- Our manual evaluation highlights the importance of typological annotation from scratch or the use of more advanced automatic conversion from the existing UD resources.

References

[1] William Croft, Dawn Nordquist, Michael Regan, and Katherine Looney. Linguistic typology meets Universal Dependencies. In Markus Dickinson, Jan Hajic, Sandra Kübler, and Adam Przepiórkowski, editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75, Bloomington, IN, January 2017. CEUR Workshop Proceedings.

[2] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021.

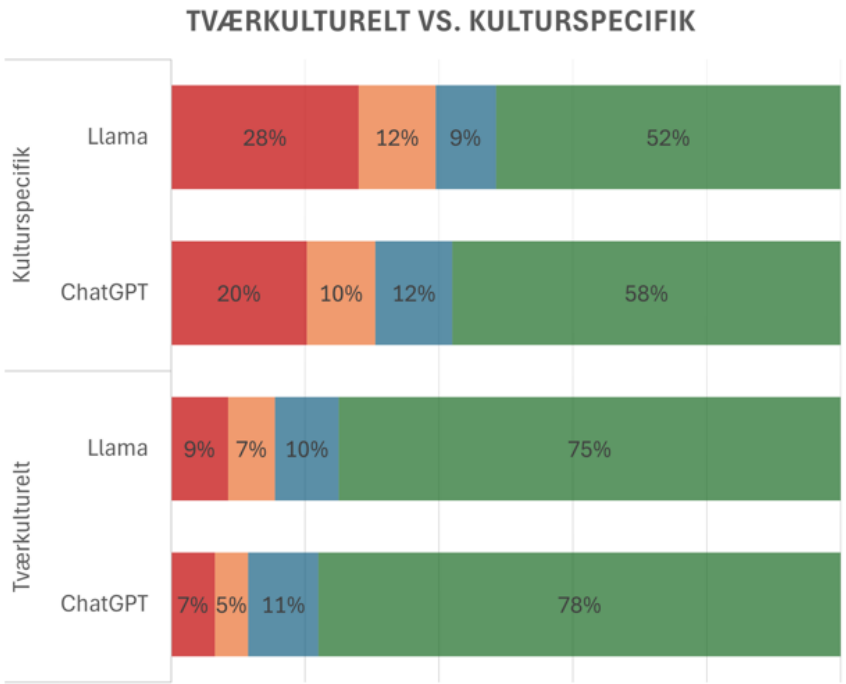
[3] Joakim Nivre. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer, 2015.

Hvor godt forstår sprogmodellerne danske kulturspecifikke metaforer?

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, Ali Al-Laith

Vi tester ChatGPT og Llama med 150 danske metaforer!

Jo grønnere, jo bedre

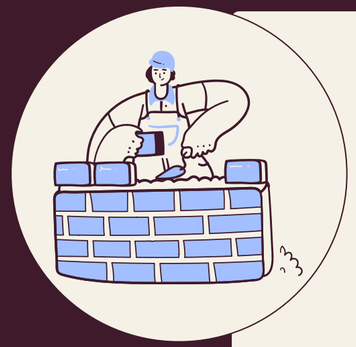


Hvordan får vi konkurrencedygtig europæisk sprogteknologi?

Digitaliserings-styrelsen

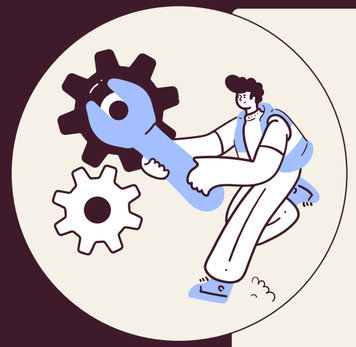
Sprogteknologi.dk

ALEXANDRA INSTITUTTET



Fælleseuropæisk Language Data Space (LDS)

Læs mere om LDS og meld dig ind i deres User Group her.



Alliancen for sprogteknologi (ALT-EDIC)

Læs mere om ALT-EDIC her.

Eller kontakt: info@sprogteknologi.dk

Europæisk sprogmodel TrustLLM

Du kan læse meget mere om TrustLLM projektet her.

Hvad sker der med sproget i tech?

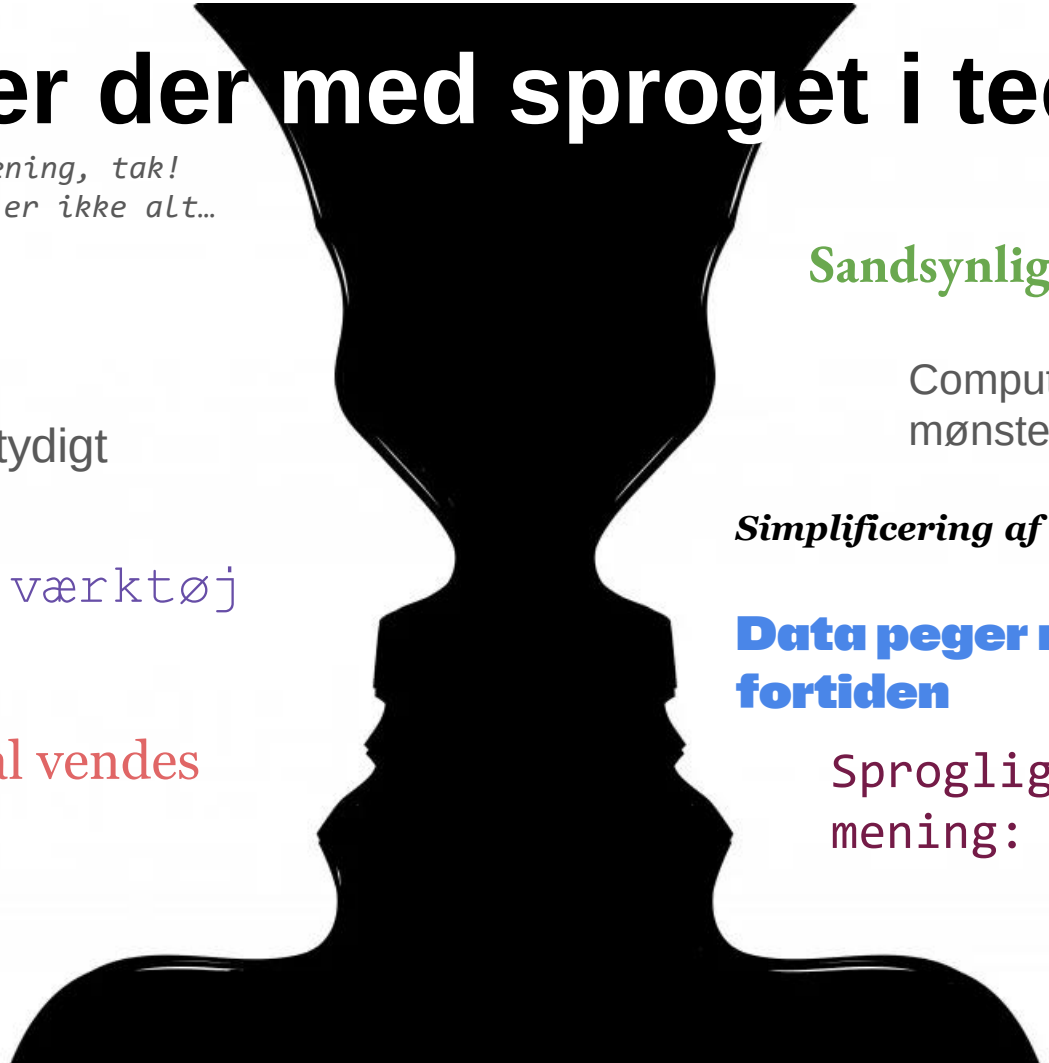
Mindre hype og mere mening, tak!
Alt er data, men data er ikke alt…

Perception

Sproget er tvetydigt

Sproget som værktøj

Information skal vendes og drejes



Sandsynlighed og statistik

Computere er gode mønstergenkendere

Simplificering af verden

Data peger mod fortiden

Sproglig form vs. mening:

M c E x I
C c E x S

Government and Opposition in Danish Parliamentary Debates

Costanza Navarretta and Dorte Haltrup Hansen

According to studies of parliamentary debates in various countries, parties express sentiment towards the same political issues differently if in government or opposition. Therefore, in our work we wanted to determine :

- whether there are linguistic differences in the Danish parliamentary speeches by government and opposition parties,
- how well fine-tuning a pre-trained Danish BERT we can identify the speeches by the two groups automatically.

Afklaringsflow for håndtering af dokumenter og videnselementer til RAG

Daniel Kjeldsmark Andreasen (akda@aarhus.dk)
ITK - Kultur og borgerservice, Aarhus Kommune

Udgangspunkt

I en praktisk kontekst (kommunal sundhedsforvaltning): Den eksisterende vidensbase er etableret med mennesker for øje, som hurtigt skal kunne skabe sig et overblik.

Vi skal kunne *garanterer* al viden er tilgængelig og at det forstås korrekt.

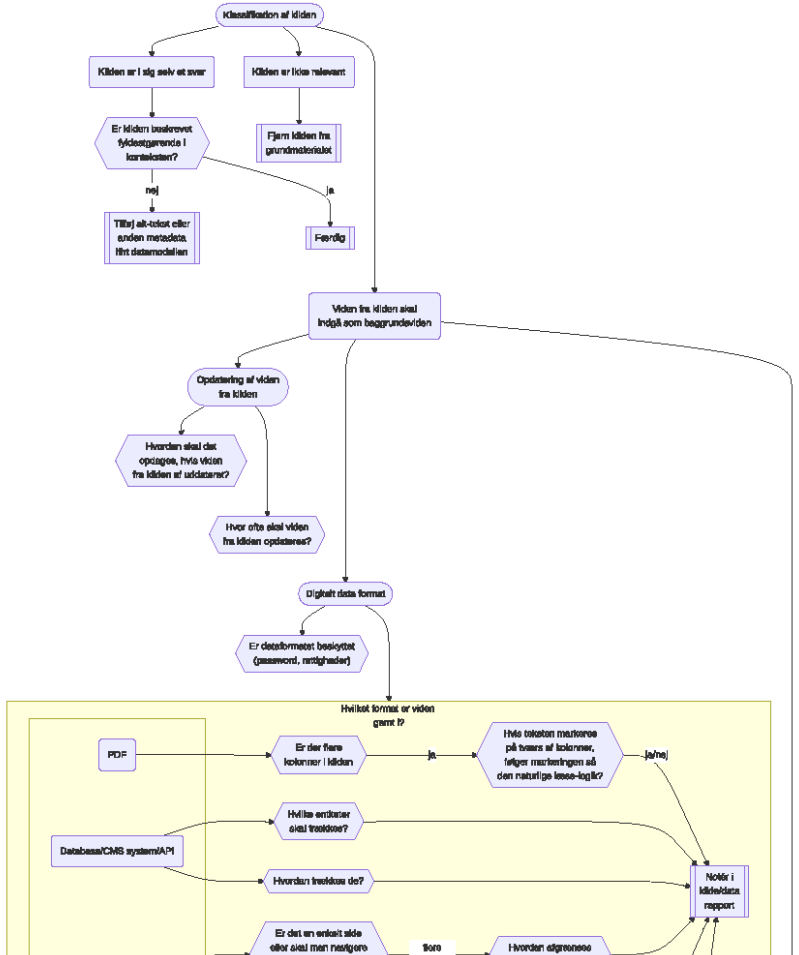
Den primære videnskilde indeholder en lang række henvisninger til information struktreret på forskellig vis.

Det vigtigt at viden fra en kilde er repræsenteret så godt som muligt for at få den relevante viden fremsøgt og give modellen den rette kontekst at svare ud fra. (Særligt når vi bruger (små) open source komponenter)

Udfordringer

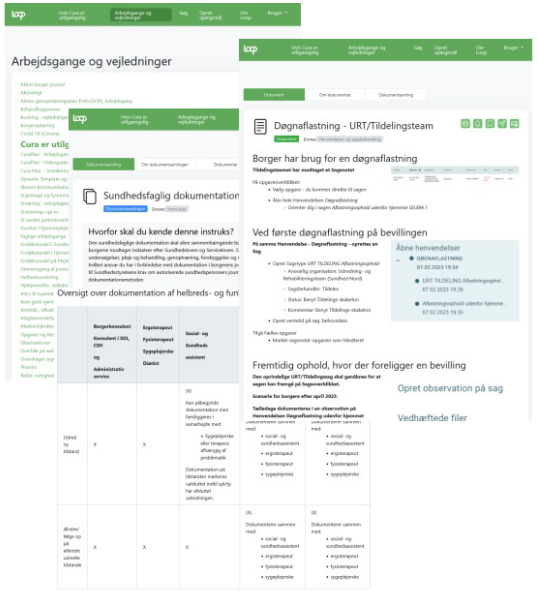
Få tekst og illustrationer ud af et dokument i den rækkefølge forfatteren havde intention om.

Forstå den implicitte logik i hvordan forfatteren har



Data in the wild

Screendumps fra den primære videnskilde



Avancerede tabeller og flowcharts i pdf og docx

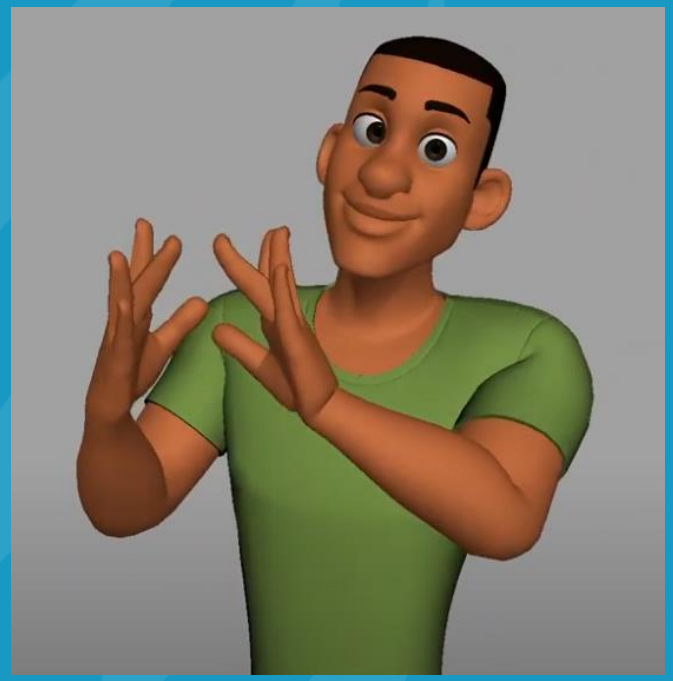
Sprogteknologi i oversættelser mellem dansk og dansk tegnsprog

Idéen om oversættelser fra dansk til dansk tegnsprog via AI-avatar. Der er potentiale i udviklingen, men vi skal se hvilke løsninger, der skal kvalificere og styrke de automatiske oversættelsesprogrammer.

Nogle pointer på posteren:

- Dansk tegnsprog mangler målbare sprogteknologiske værktøjer.
- Korpusdata er begrænset – fører til kritisk udvikling i oversættelser.
- Avatars og AI-løsninger kræver brugerinddragelse.

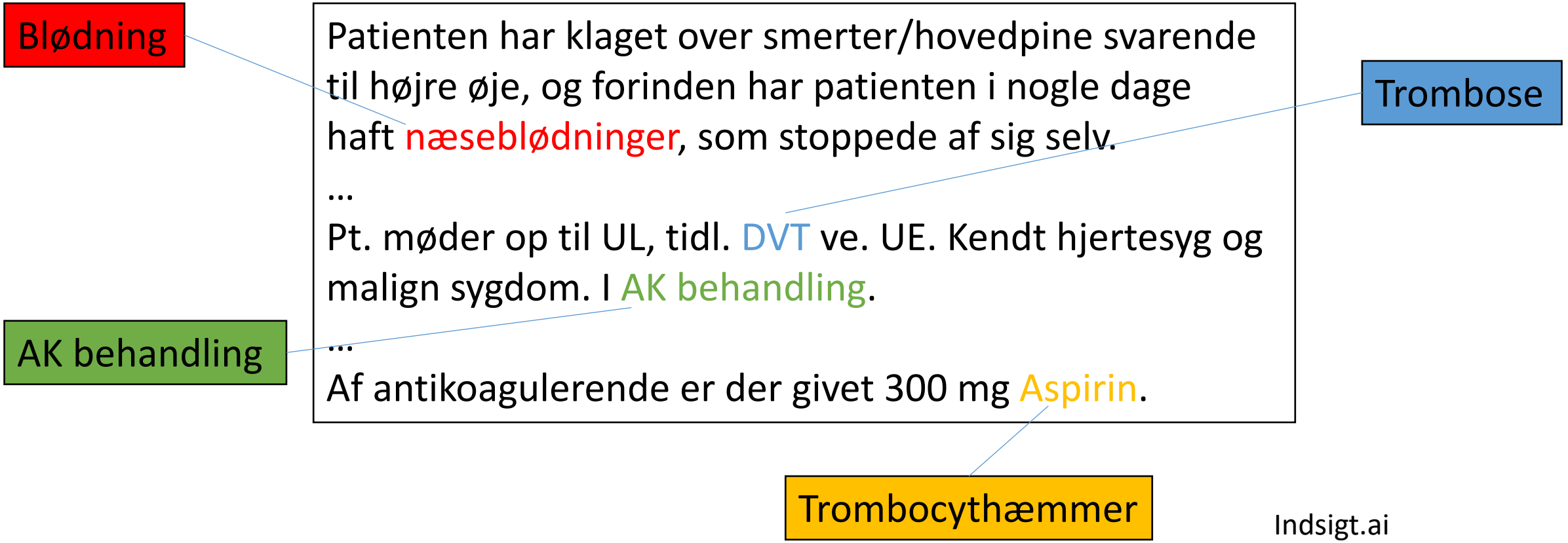
Næste skridt: Bedre data og innovative løsninger



Danske Døves Landsforbund

Ret til tegnsprog
- hele livet

Sprogmodeller highlighter relevant information i patientjournalen



Indsigt.ai
Blodprøver og Biokemi
Odense Universitetshospital

en bid af kagen • kaste med mudder • slag på tasken • slå til plukfisk
male byen rød • har man sagt A, må man også sige B • **slå knuder på tungen**
goddag mand, økseskaft • **den rygende pistol** • have det som blommen i et æg
det er en dårlig fugl der skider i sin egen rede • **slå til søren** • ikke have salt til et æg
svare i øst når der spørges i vest • male fanden på væggen • **til den store guldmedalje**
viske tavlen ren • hurtig hjælp er dobbelt hjælp • vejen til en mands hjerte går gennem maven
når solen går ned i en sæk, står den op i en bæk • **sagen er bøf** • ikke være mange sure sild værd
tale med store bogstaver • **sejle på varmen** • bide skeer med • uden mad og drikke duer helten ikke
nissen flytter med • én fugl i hånden er bedre end ti på taget • **så slutter fe**
ingen kæde er stærkere end det svageste led • når katten er ude, spiller
gøre rent bord • den der graver en grav for andre, falder selv i den • **panik før lukke**

Støttet af sprogteknologi.dk

Datasæt med
1000 talemåder og faste udtryk
fra Den Danske Ordbog

DET DANSKE SPROG- OG LITTERATURSELSKAB

Multimodal interaction in online meetings: the GEHM corpus

Patrizia Paggio, Manex Agirrezabal, Costanza Navarretta and Leo Vitasovic

- Corpus: 12 online 'real' meeting recordings (~8 hours, Zoom software).
- Data: Automatic speech transcription & visual coordinates extracted using OpenPose.
- Distribution: Available on archive.org
- Studies on feedback and speaker alignment in the meetings.

L-Università ta' Malta

DeiC Dataverse
– A National Danish Trusted Repository for Research Data

DeiC Dataverse is a national digital repository, where researchers from Danish universities can register, publish, and archive their research data with structured metadata, including clear terms and conditions for access and reuse, and Digital Object Identifiers (DOI's). Thereby, the research data become easily visible, searchable, and citable – or in other words: FAIR (Findable, Accessible, Interoperable and Reusable).

The repository is an important brick in the national research infrastructure landscape for supporting the implementation of the 'National Strategy for Data Management based on the FAIR Principles' (https://doi.org/10.48715/ea59-tp35).

We will highlight the main features of the repository and present the proposed service and support model, which relies on a collaboration between the Danish e-infrastructure Consortium, Copenhagen University Library and the eight Danish universities.

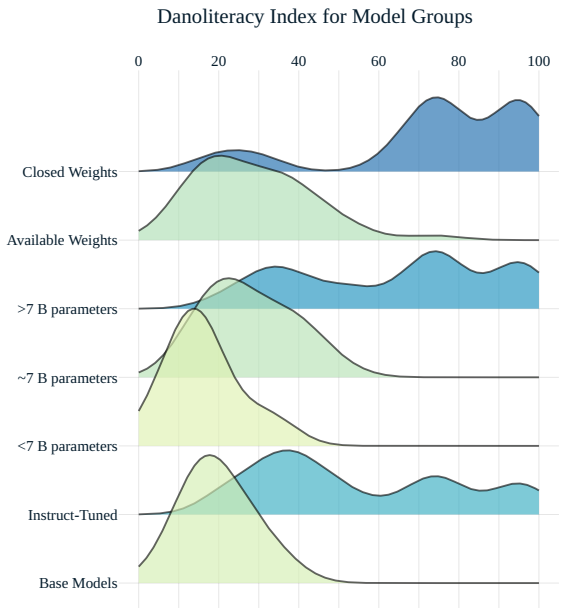
Furthermore, we would like to invite everyone to get into a conversation with us about further use cases for the repository and how it might interact with other platforms and services that support research activities in Denmark.

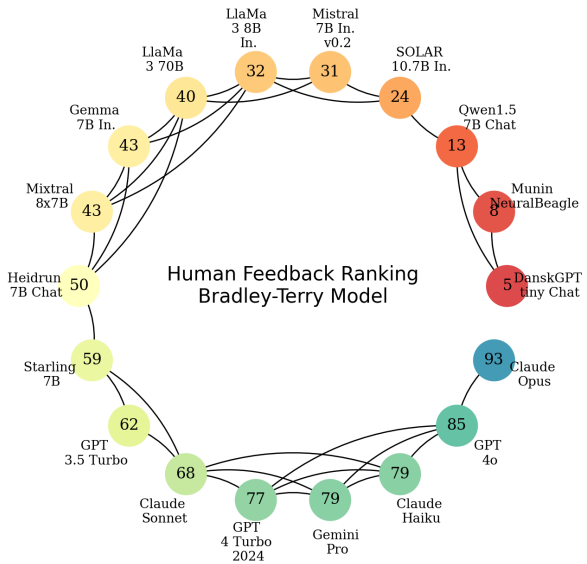
Danish Cultural Adaptation of LMs

Maximilian Müller-Eberstein, Mike Zhang, Elisa Bassignana, Rob van der Goot

Danoliterate: Dansk GLLM-evaluering

- 60 GLLM'er evalueret på otte danske scenarier
- Ser korrelation med menneskelig vurdering (~ 0.8)
- Flere besvarelser nødvendigt: Prøv interaktiv arena[1]





[1] danoliterate.compute.dtu.dk/Spørgeskema

"Danoliterate" evaluering af generative sprogmodeller

Søren Vejlgaard Holm
DTU, Alvenir

L. K. Hansen
DTU

M. C. Nielsen
Alvenir

Automatisk evaluering

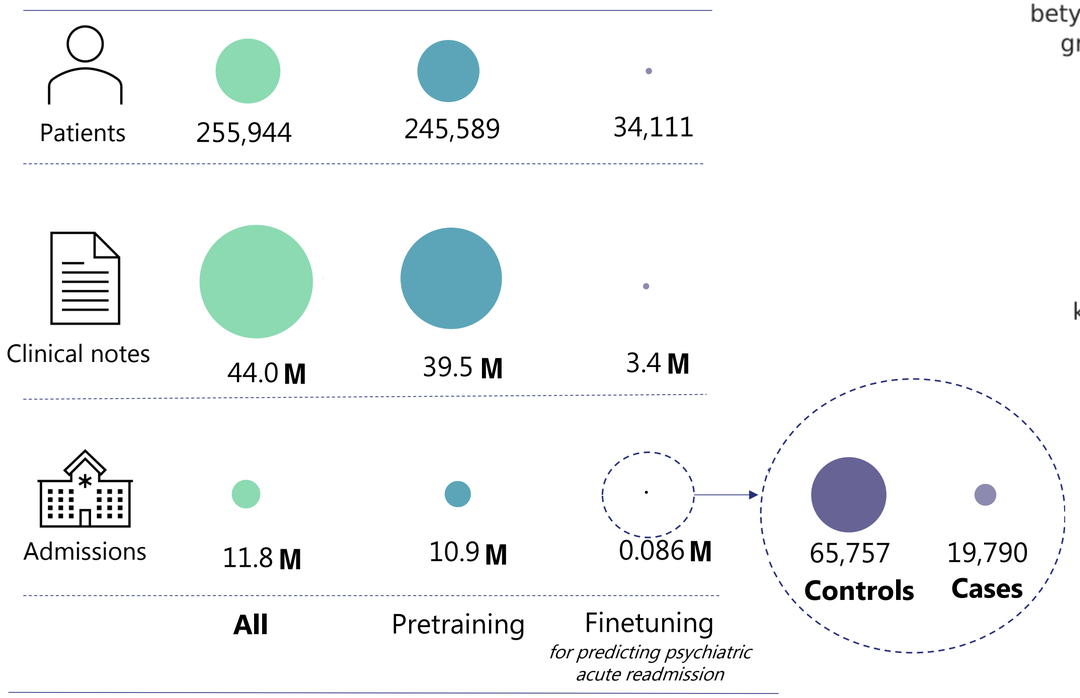
Interaktiv arena

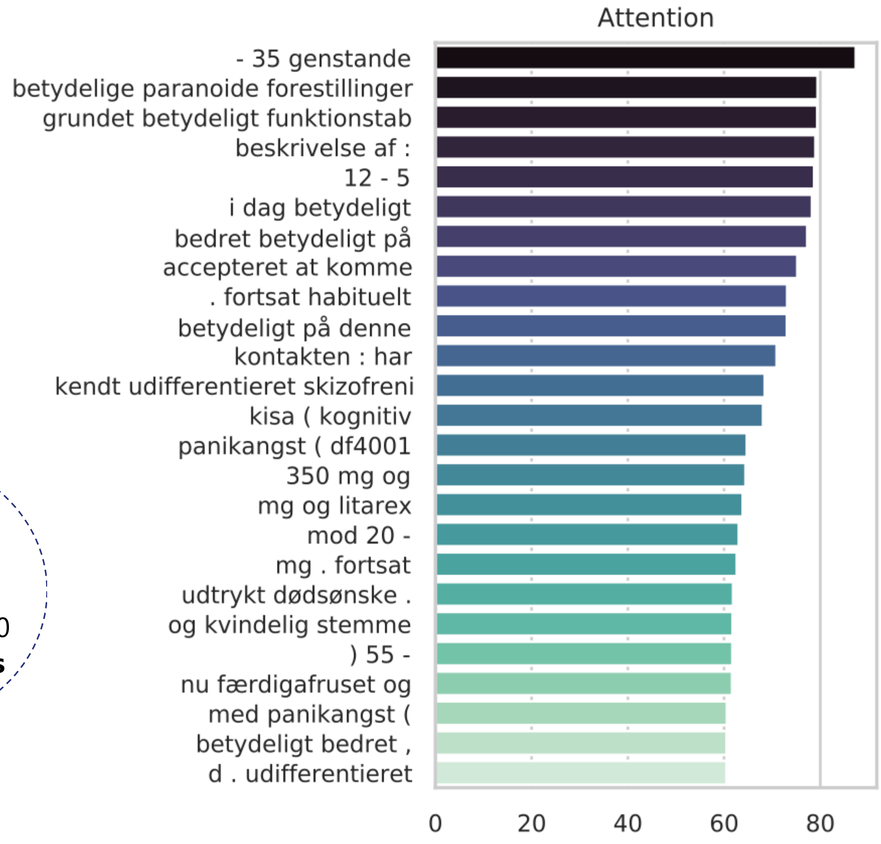
Enhancing Book Metadata with AI

- Curious about how AI can improve the quality of book metadata?
- Want to know how AI handles the nuances of the Danish language in metadata creation?
- Wondering how AI can truly grasp the essence of a book for accurate metadata and subject classification?

PsyRoBERTa: A Large Language Model for Predicting Psychiatric Outcomes from Danish Clinical Notes

Terne Sasha Thorn Jakobsen, Enric Cristobal Coppulo, Simon Rasmussen, and Michael Eriksen Benros





BIOLOGICAL & PRECISION PSYCHIATRY
COPENHAGEN RESEARCH CENTRE

REGION H Mental Health Services

UNIVERSITY OF COPENHAGEN

Novo Nordisk Foundation CENTER FOR BASIC METABOLIC RESEARCH

LUNDBECK FONDEN