Anteckningar från NoDaLiDa 2025, The 25th Nordic Conference on Computational Linguistics, 2 mars till 5 mars 2025 i Tallinn, Estland

av Martin Hansson och Thomas Vakili

Highlights

Paper 1: Dan Saattrup Nielsen, Kenneth Enevoldsen and Peter Schneider-Kamp
Paper 7: Mike Zhang, Max Müller-Eberstein, Elisa Bassignana and Rob van der Goot
Paper 8: Jenny Kunz
Paper 9: Samia Touileb, Vladislav Mikhailov, Marie Ingeborg Kroka, Lilja Øvrelid and Erik Velldal

Resourceful 2025-03-02

33 reviewed papers, fler än tidigare iterationer av workshopen.

Keynote 1: Beáta Megyesi

Unlocking Hidden Histories: AI and Expert Collaboration in Deciphering Rare Scripts
Manuscripts written in rare or unknown scripts represent a largely untapped reservoir of historical and cultural knowledge, yet their study is frequently sidelined due to the multifaceted challenges they present. These texts, characterized by unique linguistic structures and diverse symbol sets, demand an interdisciplinary approach that spans linguistic analysis, paleography, cryptanalysis, and cultural studies. While recent advancements in artificial intelligence have introduced promising tools for automating tasks such as identification and transcription, the nuanced interpretation and verification of these manuscripts remain firmly in the realm of human expertise. In this talk, I will explore the inherent complexities of working with rare scripts, discuss the current state of automation in manuscript analysis, and argue for the development of hybrid systems that combine AI efficiency with expert intervention. By enabling minimal corrective inputs and adapting models to various handwriting styles and script idiosyncrasies, such systems have the potential to bridge the gap between computational capabilities and the specialized domain knowledge required for meaningful historical interpretation.

Based on DECRYPT (completed in 2024) and DESCRYPT (project from 2025-2032). Reasoning to understand the past, difficult with rare writing systems with few examples of text. Many AI models do not adapt to such small datasets. The goal is to create digital annotated corpora from rare languages, build recognition models for specific scripts, and build frameworks for interpretation by historians and linguists. Linguistic challenges such as uninterpreted or speakerless languages, and lack of standardized writing systems. Little data, lacks systematic collection and distributed collections. Global collaboration for collection, open-access research infrastructure. Annotation lacks standardization for symbols within languages, lacks metadata. The languages do not work with existing models, models lack cultural context and data. Want to build adaptive model with experts, experts correct few outputs. HTR: Unsupervised (Chen, F, Souibgui, M.A., Fornés, A. & Megyesi, B., 2020),

HTR: few-shot (Souibgui, M.A. et al., 2020). Few-shot better for in-domain data. Frequency analysis and random start to automatically decipher text (https://www.cryptool.org/en/).

Paper 1: Ilya Afanasev och Olga Lyashevskaya

The Application of Corpus-based Language Distance Measurement to the Diatopic Variation Study (on the Material of the Old Ngorodian Birchbark Letters)
The paper presents a computer-assisted exploration of a set of texts, where qualitative analysis complements the linguistically-aware vector-based language distance measurements, interpreting them through close reading and thus proving or disproving their conclusions. It proposes using a method designed for small raw corpora to explore the individual, chronological, and gender-based differences within an extinct single territorial lect, known only by a scarce collection of documents. The material under consideration is the Novgorodian birchbark letters, a set of rather small manuscripts (not a single one is more than 1000 tokens) that are witnesses of the Old Novgorodian lect, spoken on the territories of modern Novgorod and Staraya Russa at the first half of the first millennium CE. The study shows the existence of chronological variation, a mild degree of individual variation, and almost absent gender-based differences. Possible prospects of the study include its application to the newly discovered birchbark letters and using an outgroup for more precise measurements.
https://dspace.ut.ee/server/api/core/bitstreams/0c81f710-19a3-44ff-9e13-82e2c0a1f960/content

Birchbark are mostly Old Novgorodian, short (mostly 100 tokens), written around 1000-1500 CE. One of the first examples of Baltic/Finnish languages. More are found every year, digitization is ongoing. Want to discover individual variation, chronological variation and variation between sexes. Some tokens cannot be recreated, some tokens can be recreated but become biased (researcher interprets), unbalanced dataset. Clears out the tokens that cannot be recreated, some of those recreated by researchers, too long or too short characters, cluster the remaining into individual, chronological and cluster on gender. Splits into 3-shingles, type n-gram splitting. Gives points on alphabet entropy and frequency rank for 3-shingles. Combination of different metrics: Mean DistRank for coinciding 3-shingles, Mean DistRank / string similarity, Dærensen-Dice coefficient. vector-based string similarity. UPGMA classification, statistical analysis through PCA and DHBSCAN and qualitative analysis. Individual variation appears to be some shared innovation and noise in the dataset. Chronological clustering hierarchical. There were large chronological differences, few gender-based differences.

Keynote 2: Joshua Wilbur

Digitizing Pite Saami: Making the most of limited resources
Pite Saami is a critically endangered Uralic language spoken by only a few dozen individuals originating from areas in and around Arjeplog in Swedish Lapland. Due to the exceptionally small number of native speakers, a very limited amount of language data is available; nonetheless, there is a surprisingly diverse set of language resources available, both in digital and in analogue form. In this talk, I will explore the perhaps extraordinary state of Pite Saami language data and digital tools, including how this came about, what potential the data holds in the context of current technological advances, and the challenges involved in

this. In doing so, I hope to provide a starting point for a discussion on both the realities and realistic prospects of developing NLP for seriously under-resourced languages.

Pite Saami a Uralic language, around 30 people who speak the language around Arjeplog, almost all of them are 50 years old, almost not taught to younger people, like no media that uses the language. The language is almost extinct. Quite complex stem variations. Quite a lot of collections of text by Pite Saami at ISOF in Uppsala. Mostly handwritten, a little digital. A few other text collections such as books. The Pite Saami Syntax Project created a digital corpus of 60,000 tokens. ELAN for direct annotations for videos, data in XML. Some research of the Pite Saami is done. Also crowsourcing that creates its own glossaries "Collection of Pitesami words". Published a dictionary "Pitesamisk dictionary". In collaboration with Giellatekno, Finite state transducer (FST) for morphological parsing, and Constraint grammar (CG) for syntatic disambiguation have been created. Quite impressive resources for such a small language that is almost extinct. For a century speakers, linguists have worked with language, need language technologists. Nothing new to combine NLP and endangered languages. CARE data princples with original population. Challenges in making language technology accessible, useful and valuable to the population.

Paper 2: Nina Hosseini-Kivanani, Christoph Schommer, and Peter Gilles

Voices of Luxembourg: Tackling Dialect Diversity in a Low-Resource Setting
Dialect classification is essential for preserving linguistic diversity, particularly in low-resource languages such as Luxembourgish. This study introduces one of the first systematic approaches to classifying Luxembourgish dialects, addressing phonetic, prosodic, and lexical variations across four major regions. We benchmarked multiple models, including state-of-the-art pre-trained speech models like Wav2Vec2, XLSR-Wav2Vec2, and Whisper, alongside traditional approaches such as Random Forest and CNN-LSTM. To overcome data limitations, we applied targeted data augmentation strategies and analyzed their impact on model performance. Our findings highlight the superior performance of CNN-Spectrogram and CNN-LSTM models while identifying the strengths and limitations of data augmentation. This work establishes foundational benchmarks and provides actionable insights for advancing dialectal NLP in Luxembourgish and other low-resource languages.
https://dspace.ut.ee/server/api/core/bitstreams/c0dfd02f-a0b3-424c-af33-a0782d82c5a6/content

Luxembourgish is complicated, several dialects with specific word choices. 600,000 inhabitants and four major dialects (west, north, south, east). Dialects are important for understanding cultural variations, especially for a language like Luxembourgish that has been influenced by both German and French. Challenges to classify for models due to their variety. First approach to try to classify dialects in Luxembourgish. Little annotated data, traditional phonetic techniques have difficulty with dialectal variations. Created the dataset by having participants translate French and German audio files into Luxembourgish. Feature extraction through Mel-Frequency Ceprstral Coefficients (MFCCs), Spectrogram features (used within CNN), Deep speech embeddings (Wav2Vec2, Whisper, XLSR-Wav2Vec2). Speech data augmentation to increase the diversity of the data set, more robust to real linguistic variations (time-stretching and pitch shifting). Training through 5-fold CV, RF (Optuna), DL (Adam optimizer, categorical cross-entropy loss), early stopping (patience 10 epochs). Evaluates on Accuracy, Precision, Recall. Results show between 55%-73%

accuracy to identify dialects, Random Forest had trouble classifying. CNN-spectogram and CNN-LSTM perform best. The Eastern dialect most difficult to classify. Optimized models with augmentation gave 3-6% better results, compared to the same models that were strong.

Paper 3: Mena Hernández et al.

Automatic Validation of the Non-Validated Spanish Speech Data of Common Voice 17.0

Mozilla Common Voice is a crowdsourced project that aims to create a public, multilingual dataset of voice recordings for training speech recognition models. In Common Voice, anyone can contribute by donating or validating recordings in various languages. However, despite the availability of many recordings in certain languages, a significant percentage remains unvalidated by users. This is the case for Spanish, where in version 17.0 of Common Voice, 75% of the 2,220 hours of recordings are unvalidated. In this work, we used the Whisper recognizer to automatically validate approximately 784 hours of recordings which are more than the 562 hours validated by users. To verify the accuracy of the validation, we developed a speech recognition model based on a version of NVIDIA-NeMo's Parakeet, which does not have an official Spanish version. Our final model achieved a WER of less than 4% on the test and validation splits of Common Voice 17.0. Both the model and the speech corpus are publicly available on Hugging Face.
https://dspace.ut.ee/server/api/core/bitstreams/918ec35c-a079-4258-b20d-07275ea28ae4/content

Mozilla Common Voice is crowdsourcing where people donate data with their voice where they speak their language. Spanish lacks evaluation of the language compared to other languages. There are 2220 hours of spoken language and 562 hours of evaluation. Categories are validated (divided into train-test choices, at least two more positive votes than negative), unvalidated (at least two more negative votes than positive), reported (controversial content) and other. Validated 35.31% of data (784 hours out of ~1500 hours). Using Whisper out-of-the-box, looking for perfect matches (if the model produces exact results like the transcripts). Normalize the transcriptions (lower case, remove punctuation and remove characters that are not in the Spanish alphabet). Evaluating with Nvidia Parakeet, indirect validation. There is no official model in Parakeet in Spanish. Better results than Whisper if they combine validated and other data. Argues that an ASR system still does not invalidate results for other ASR systems, normalized data enables comparison and even though Whisper was probably trained on Common Voice, their model performs better. Created new model, can be applied to other datasets in the future.

Paper 4: Jenna Kanerva et al.

OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches

Optical Character Recognition (OCR) systems often introduce errors when transcribing historical documents, leaving room for post-correction to improve text quality. This study evaluates the use of open-weight LLMs for OCR error correction in historical English and Finnish datasets. We explore various strategies, including parameter optimization, quantization, segment length effects, and text continuation methods. Our results demonstrate that while modern LLMs show promise in reducing character error rates (CER) in English, a practically useful performance for Finnish was not reached. Our findings

highlight the potential and limitations of LLMs in scaling OCR post-correction for large historical corpora.
https://www.arxiv.org/pdf/2502.01205

Many historical texts through scanning and OCR, different levels of noise. OCR error post-correction has been proposed to improve quality but is expensive, does not have access to the original images or OCR output but only text. English data, Eighteenth Century Collections Online (ECCO). Finnish data through OCR ground truth from the National Library of Finland containing human annotation and OCR engine output. ECCO TCO 301,937 pages, NLF GT 449 pages. Using Relative CER reduction (how many % of residual error is reduced), CER% is weighted average of example CER%. Normalizes systematic differences between spellings in ancient texts to modern versions. Post-processing by removing overgeneration from LLM, align LLM input to character-level output. Most models positive on both metrics in English, LLama 3.1 70B improved ~40% of errors. GPT-4o better than open models CER% 58.1. In Finnish, most models are negative, they make the quality worse. GPT-4o is positive but with low level CER% 11.9. Results degrade if the segments are too short. LLMs can be used to OCR post-correct historical texts in English but quite poorly in Finnish. Details have meanings (post-process, segment length, etc). Evaluation is hard! Want to use in the future on ECCO OCR 180,000 books in English, currently evaluating. Better models are needed for Finnish.

Keynote 3: Jussi Karlgren

What are the most sustainable and valuable resources that language technologists should develop for training language models?
The current generation of generative language models exhibits impressive behaviour in many language processing tasks, thanks to their capacity to estimate a probability distribution over linguistic elements by observing linguistic data. These successes have been achieved through training models on very large data sets, which may be difficult to establish for languages with less digital footprint than the largest ones. There will be new architectures, memory models, and processes to train models as technical development advances, and the amount of data needed to train models is likely to change in the near future making it possible to train models at less cost. What resources should language technology research focus on to address the likely needs of future generations of representations?

What does one need to learn a language? A basic understanding of the world (L1), understanding of situation and social contexts, curriculum, opportunism and anomalies, qualitative input. Movement from first language (L2), cultural understanding. Loose understanding of situation and context needs context in data collection. Solve curriculum, opportunism and anomalies by thinking about how one learns (progression). Solve qualitative input, unlock archives and libraries! To translate between languages, focus on the specifics, don't automatically translate. What do we want to check to ensure that the machine is doing the right thing? What is unusual and strange in one language may be common in another. We all want to be nice but niceness is different. Be careful with automatic translation of tests. Short todo list: collect real data, keep track of metadata, collect situations, identify norms and preferences specific to that language, use test development (start with scenarios, build evaluation mechanisms from first principles to meet scenarios).

Paper 5: Annika Simonsen, Dan Saattrup Nielsen, and Hafsteinn Einarsson

FoQA: A Faroese Question-Answering Dataset
We present FoQA, a Faroese extractive question-answering (QA) dataset with 2,000 samples, created using a semi-automated approach combining Large Language Models (LLMs) and human validation. The dataset was generated from Faroese Wikipedia articles using GPT-4-turbo for initial QA generation, followed by question rephrasing to increase complexity and native speaker validation to ensure quality. We provide baseline performance metrics for FoQA across multiple models, including LLMs and BERT, demonstrating its effectiveness in evaluating Faroese QA performance. The dataset is released in three versions: a validated set of 2,000 samples, a complete set of all 10,001 generated samples, and a set of 2,395 rejected samples for error analysis.
https://arxiv.org/pdf/2502.07642

Why is QA needed for Faroese? Few annotators for languages with few speakers. Can the cost of producing data be lower with the LLM? Single annotator methodology for creating extractive QA datasets semi-automated, valuable for RAG applications. Created the first Faroese QA dataset. QA systems are divided into extractive and abstractive types, deal only with extractive QA, give examples of a text and ask questions about that text. Created by generating QA pairs from text corpus, rewrite the questions but keep the answers. The answers are rewritten because the models tend to use the same entities as from the example. More sophisticated recognition through synonyms and connections between entities. Evaluates through grammar, semantics and classification. Based on Faroese Wikipedia, selected only articles with more than 1000 tokens. Created 10,001 samples and manually annotated 4130 of which 1759 were correct. Most questions were classified with the people label (33.95%). Did a qualitative error analysis to see what types of grammatical errors GPT-4o-turbo makes. GPT-4 is better on the examples that were changed.

Paper 6: Anna Lindahl

Annotating Attitude in Swedish Political Tweets
There is a lack of Swedish datasets annotated for emotional and argumentative language. This work therefore presents an annotation procedure and a dataset of Swedish political tweets. The tweets are annotated for positive and negative attitude. Challenges with this type of annotation is identified and described. The evaluation shows that the annotators do not agree on where to annotate spans, but that they agree on labels. This is demonstrated with a new implementation of the agreement coefficient Krippendorff's unitized alpha.
https://dspace.ut.ee/server/api/core/bitstreams/e9b57f6a-c25e-4410-95ec-38e01e27dc99/content

Around 4500 tweets from Swedish politicians, between 2018-2022. 4 annotators, annotated positive and negative classes. Ingot tweets that were judged too difficult. Two rounds of annotations. Objective to identify positive or negative attitude, the object of an attitude. Uses spans as annotation. The objects could be a word or phrase, sometimes the entire tweet. 6 pages of guidelines. Few tweets were rejected. 80-95% of tweets were considered to contain attitudes. More positive spans than negatives. Kirppendorff's alpha was 0.41 at the token level. Evaluation at the token level does not take span into account. Used unitized alpha

(created implementation in python) takes into account both labels and spans. The annotators do not agree on which attitude occurs in the text but when they do, they agree on which attitude is expressed, identifying different objects but the same class. The limit of span is problematic. The class can change depending on how many words are included in a span. Possibly annotate in steps?

Paper 7: Mariia Fedorova et al.

Multi-label Scandinavian Language Identification (SLIDE)
Identifying closely related languages at sentence level is difficult, in particular because it is often impossible to assign a sentence to a single language. In this paper, we focus on multi-label sentence-level Scandinavian language identification (LID) for Danish, Norwegian Bokmål, Norwegian Nynorsk, and Swedish. We present the Scandinavian Language Identification and Evaluation, SLIDE, a manually curated multi-label evaluation dataset and a suite of LID models with varying speed–accuracy tradeoffs. We demonstrate that the ability to identify multiple languages simultaneously is necessary for any accurate LID method, and present a novel approach to training such multi-label LID models.
https://arxiv.org/pdf/2502.06692

Multi-label Scandinavian language identification tries to identify similar languages at the sentence level, which is difficult. Many language models are trained on web ceawl where text of different lengths can be included in the training data. Code switching, text can contain sentences in several different languages. The LID corpus expects the source of the text to define the language. Inspected dev and test splits manually. Annotators spoke the language. Machine translation silver label. Different scores, loose accuracy (intersection between predictions and gold labels), strict accuracy and F1 (macro F1) per language. Various BERT models. Do not classify semantics as language. Normalized URÖ, e-mail, number. Changed various alphabetic variations (eg Danish characters in Swedish sentences by name). Results show that SLIDE-base (123M) performs best in all metrics, although one of the slowest models (38.41 ms/sample). Challenges with Norwegian due to spelling variations and ambiguity between Nynorsk and Bokmål. Argues that silver multi-label seems to work. Not clear how much preprocessing makes the model robust.

Panel discussion: Jussi Karlgren, Joshua Wilbur, Danila Petrelli, Hafsteinn Einarsson, and Beáta Megyesi

People use AI to generate annotations (silver-label) are you pro or against?
Danila: See the need of them, do not see any possibility to be a good alternative to gold-standards.
Jussi: Pro using but alot of risks. Use same stuff over and over again. Do that with human generated data aswell. Scores go up, but don’t go through the samples and understand why.
Beata: What do we use the generated data for? If its used control for bias and errors. How do we generate data? Some control is needed when generating aswell.
Joshua: Linguist are particularly skeptical because it should be authentic. Hard of where to draw the line.
Hafsteinn: Evaluating errors, using same model to generate and evaluate. Could several language models be used to generate and not evaluate the model on data it created? What

is best practice? Training LLMs might be good, it has been quite successful. Annotating output might be used as more training data.
Danila: Don't you think the current standards focus so much on evaluating errors, instead of focusing on generating authentic data which is the hardest part?
Hafsteinn: I still think we need to review but I don't know to what extent. Review is important so we don't just retrain on those things.
Jussi: It is harder to get information know because alot of it is generated AI slop. Then models are trained for this wordiness. Overtraining models to fulfill some sense of language competence without use.

How do we conduct our research? How much should we focus on peoples needs rather than improving metrics?
Jussi: Some are humanities more than others and technology more than others. Frequently papers are not explanatory or tooling, they are worthless. Either we build something that is of use that's the technology goal or then we explain something so we understand the world better, that's the humanities goal.
Beata: I always have the user in mind but its not the everyday user in the world but scholars, historians, archivists. We try to model problems they have, step-wise. We learn about language on the way, in this iterative process. It depends what you want, what do we want them to do?
Danila: If we do applications then usability and users are key. Participatory design could be community based approaches that makes sure you build something meaningful. Multidisciplinarity. You need to have the users first in mind.
Hafsteinn: One answer to get native speakers engaged is getting the models to have impact. Demand to improve performance of their native workforce to get better language models for those languages.
Joshua: CARE framework, standards and expectations with endangered langauges. Responsibility to support their communities if we develop these tools. Even if the communities they could produce the data but someone in the Language community needs to care. I'm not sure how LLMs would help certain communities with the problems because they are not relevant to them. Multilingual communities usually can get help in the majority language so why wouldn't they use it? Then that makes the minority language less used and it disappears.

How much benefit does our resources give them? Should our focus be focused on preserving languages?
Joshua: Even if it is a recording it is such a small aspect of language. It is not the same as being there. All contextual knowledge is missing.
Danila: It is extremely dangerous in building technology that is not needed by a specific community and the ownership becomes controversial. The importance of institutions to be mediators so researchers can focus on needs. Local dialect of mine in Greek was protected by UNESCO and now we can do research on it.
Hafsteinn: People should get something back from giving away their data. Some compensation by how much their data contribute to these models.
Jussi: We like to pretend that we build something for people. We do stuff because its their and fun, and they we shoehorn the need afterwards. Typically we haven't created technology for the need of people. The investment in LLMs probably won't be paid back. Technological advancements are fun to do. I think we worry too much of ownership of intangibles.

Collective data with humans come with variations. Now we have LLMs that are approximate models. What should we take into account to deal with variational context?

Beata: User-friendly platforms, add data, describe data, promote cooperation, less property right. Make resources available. Funding agencies don't see it as resources but infrastructure.

Hafsteinn: One important thing in gathering data is getting incentives to participate unless they get something for it. What we will need to focus on going forward is to give something back to the people. How can we do it in a sustainable way?

Danila: If someone can come up with the best platform to gather data. If you want a more crowdsourced effort then it is harder.

Joshua: Humanities is variety. These models are approximated is a shame. You don't have the same variety as with humans. How can we capture that variety?

Jussi: Best crowdsourcing we have is Wikipedia and the editors really want to help, but they aren't always nice people. Personality bias. It seems to work at times and not other times.

If we compensate, is it because they want to help or because they get money? Are there differences in intent? What is the future of annotation?

Beata: We need multimodality. We need to work on it in a more systematic way. How shall we represent all variety that relates to language?

Hafsteinn: If there was enough money and governments so the reasons to collect data. We had some sort of bounty program run by the government maybe that would work? Maybe this will be a job in the future.

Danila: One thing I think needs to change is disagreement. Moving forward we really need to embrace people who disagree. Multijudgement to take into account all perspectives. How do we weight different annotations?

Sometime ago it was so hard to synthethize voice. Now you can do it easily. Maybe it is possible to solve it in different languages. Maybe it is just about crossing barriers between languages?

Beata: I agree with you. Four major areas of machine learning and AI have similar problems and need to collaborate better.

Hafsteinn: Crosslingual transfer is really interesting. Some models that are trained on multiple languages can be good at languages they have not been trained on. We probably need to find the gaps in the models and methods and improve from there.

Joshua: If we can develop that do a decent job on speech synthetics then that's a great application in learning languages, pronunciation and such.

Is there any way that other than abolishing itself that LLMs can be a vaccine against themselves?

Hafsteinn: To detect when they make mistake is very hard. We can try to gauge how well knowledge is gained by asking multiple models. We need some outside information to fix these gaps.

Most data online is not paid for. We create data and other people benefit from them.

Danila: It was shown that people tend to love to correct other people.

What are our roles in creating benefit for communities? The community is often not ready to accept our tools. Then Google comes and takes our data and creates something easier to use. What should we do?
Hafsteinn: If you make data available I think it is good that other tools become better as a result. These big companies also invest in creating data but don't share it. We would be better of if everyone shared it.
Danila: I think we should keep on building resources, even if they are not competitive.
Joshua: If I hadn't done this, I don't know if anyone else would have done this. If there is nothing there, there is also nothing to take.

NoDaLiDa 2025-03-03

NoDaLida 2027 kommer att vara på University of Copenhagen, Danmark med Centre for Language Technology (CST). Förmodligen i maj 2027. BalticHLT 2027 förmodligen i Vilnius på Mykolas Romeris University, Litauen. Förmodligen på hösten. Machine Translation Marathon (MTM) 2025 25-29 augusti i Helsinki.

Keynote 1: Arianne Bisazza

Not all Language Models need to be Large: Studying Language Evolution and Acquisition with Modern Neural Networks
NN as tools to study human language. Simulation can help answer questions of language science: Why do languages look the way they do? What makes us, as humans, so good at learning languages from little input? Many human processes can't be studied experimentally. NNs can help us understand how general-domain statistical learning mechanisms in our minds interplay with factors like cognitive constraints. LMs display human-like behavior, and they are inhuman in how they learn. Started merging language evolution and language technology. Artificial language learning is one way of studying the origins of language universals. Introduced the NeLLCom framework, small NN agents, learning pre-defined language via supervised learning + playing a meaning reconstruction game via RL (two agents paired that try to optimize the language). Architecture linear-to-sequence / sequence-to-linear NNs models "meaning." A speaking agent and listening agent. Reconstruction game inspired by neural-agent emergent communication. Speaking agent conveys a meaning *m* by generating an utterance *u*. The listening agent tries to map *u* to its respective meaning *m*. Communicative success = listener prediction matches speakers' intended meaning. Boldt & Mortensen 2024. Artificial language design borrowed from human experiments. Results show supervised learning shows no "drift" and no regularization. RL shows that the accuracy of ambiguous languages increases, word order regularizes, and marking decreases in non-ambiguous languages. Can look at individual trajectories and compare them to human results. Extending to group communication, population-level behavior is not simple to model. Presented NeLLCom-X (https://arxiv.org/pdf/2407.13999), an extension where agents can alternate between roles (speaker, listener). Increased group sizes show lower variability and converge in fixed-order. Managed to simulate emergence, a well-studied language universal without hard-coding ad-hoc pressures. Moving on to simulating language acquisition, BabyBERTa trained on 5 million word child-directed languages (CDL) obtained similar syntactic knowledge as its counterpart trained on 30 billion tokens. Studying the salient properties of CDL. Trained GPT-2-small *from scratch* on babyLM datasets augmented with synthetic VSs. NNs have been used to simulate human

language learning since the early days of connectionism. Modern NNs have the opportunity to revise the simulations, make them more realistic, and combine tasks and learning objectives to integrate traditionally separated approaches. A controlled setup and small-scale research remain essential to understanding how LMs learn; otherwise, we risk anthropomorphizing LLMs. Inspiration for reading:
Language as shaped by the brain,
https://www.cambridge.org/core/services/aop-cambridge-core/content/view/EA4ABB50915417A1A10569707F574F5E/S0140525X08004998a.pdf/language-as-shaped-by-the-brain.pdf
A metatheory of Classical and Modern Connectionism
Provocations from the Humanities for Generative AI Research,
https://arxiv.org/pdf/2502.19190

Paper 1: Dan Saattrup Nielsen, Kenneth Enevoldsen and Peter Schneider-Kamp

Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks

This paper explores the performance of encoder and decoder language models on multilingual Natural Language Understanding (NLU) tasks, with a broad focus on Germanic languages. Building upon the ScandEval benchmark, initially restricted to evaluating encoder models, we extend the evaluation framework to include decoder models. We introduce a method for evaluating decoder models on NLU tasks and apply it to the languages Danish, Swedish, Norwegian, Icelandic, Faroese, German, Dutch, and English. Through a series of experiments and analyses, we also address research questions regarding the comparative performance of encoder and decoder models, the impact of NLU task types, and the variation across language resources. Our findings reveal that encoder models can achieve significantly better NLU performance than decoder models despite having orders of magnitude fewer parameters. Additionally, we investigate the correlation between decoders and task performance via a UMAP analysis, shedding light on the unique capabilities of decoder and encoder models. This study contributes to a deeper understanding of language model paradigms in NLU tasks and provides valuable insights for model selection and evaluation in multilingual settings.
https://dspace.ut.ee/server/api/core/bitstreams/5f7142c4-1029-4253-b811-1f4e58c4dc28/content

Pretrained vs. fine-tuned encoder model. Pretraining to get model weights and fine-tuning to get labels. The decoder language model can generate text directly and be fine-tuned, but it is slightly different because you use the same pretraining paradigm during fine-tuning. Usually, people evaluate encoder and decor decoupled; there are benchmarks for each. EuroEval is a robust multilingual benchmarking framework. To evaluate encoder models, they fine-tune the model on the training split and evaluate it on val with early stopping. For decoders, they phrase tasks as text-to-text tasks. Get a few-shot examples from the training split, and evaluate the model on the test split with the few-shot examples. One difference is that they try to keep the level of quality equal between the languages; they try not to include machine-translated datasets. When evaluating, there are several noise sources: the choice of training examples and performance can vary significantly between the few-shot examples

and the choice of test examples. Training and test examples are bootstrapped 10 times, yielding a more reliable estimation of the true mean. Natural Language Understanding has four tasks: sentiment classification, linguistic acceptability, reading comprehension, and named entity recognition. They identify the first token of the label for decoder text classification. If logprobs are available, they get the generation logprobs for each of these “label first tokens” and return the label whose “label first token” has the highest logbprob. If they don’t have probabilities, they generate 5 tokens and return a label whose word edit distance is closest. In evaluating reading comprehension, they have the model output at most 32 tokens and use the output as-is. In evaluating NER, they ran experiments and found that structured generation generates structured JSON. Smaller models struggle to generate a structured format, forcing them to make outputs equal. They have used different prompts for base and instruction models where base is structured as auto-completion, and the instruction is structured as user/assistant dialogue.

Paper 2: Emilie Marie Carreau Francis

Language of the Swedish Manosphere with Swedish FrameNet
The manosphere is a loose group of online communities centralised around the themes of anti-feminism, misogyny, and hetero-masculinity. It has gained a reputation for violent extremism, particularly from members of the incel community. Sweden sees one of the highest volumes of online traffic to well-known incel forums in all of Europe. In spite of this, there is little information on manosphere/incel culture in Swedish. This paper uses posts from Flashback's manosphere subforum automatically annotated with Swedish FrameNet to analyse the language community in a Swedish context. To do so, a lexicon for the Swedish manosphere was created and terms of interest were identified in the Swedish discourse. Analysis of prominent semantic frames linked to these terms of interest presents a detailed look into the language of the Swedish manosphere.
https://dspace.ut.ee/server/api/core/bitstreams/3f1b39d5-43d5-4559-8941-0ee17bb4dba5/content

The manosphere in Sweden is among the top traffic to major incel forums. Feminism is used to legitimize the dehumanization of women and personal attacks on female scholars of the Swedish manosphere. Posts claim that Swedish women are privileged and have power over Swedish society. Posts discuss immigration and gender imbalance in Sweden. Semantics Frames and SweFN are developed in line with Berkeley Frame. Data from Flashback forum between 2012-2024. There are multiple manospheres and men's movements, and posts have increased. I created a lexicon (based on English) and looked at the log ratio to compare terminology in Swedish with no English equivalent or translations. Narrowed down the TOI list based on previous literature. Manually analyze the context of TOIs and frames, ff-icf. 63 TOIs (incel, feminist, svensk, usa, osv.). Five themes: inceldom and mental health, feminism and LGBTQ+, race and origin, immigration and male surplus, power and violence. The results of TOIs match the performance of previous literature. Chad/Stacy hypermasculine/-feminine terms. Blackpill comes to believe with employed (extended pill theory), blackpill is an extreme version of the red pill from Matrix. Incel often refers to origin. Feminism talks about the origin and point of dispute. LGBTQ+ often contains terms of intoxicants. Race and origin, often terms of becoming with Middle Eastern and Hispanic and Latin American, are referred to by color, people of origin, or death. Immigration and surplus refer to measurable attributes and change position scale. Power is often attributed to women

and violence; when correlated to killings, then women are victims, but with suicide, then men are victims. The idea of immigration and male surplus is very specific to Sweden.

Paper 3: Emil Nuutinen, Iiro Rastas and Filip Ginter

Finnish SQuAD: A Simple Approach to Machine Translation of Span Annotations
We apply a simple method to machine translate datasets with span-level annotation using the DeepL MT service and its ability to translate formatted documents. Using this method, we produce a Finnish version of the SQuAD2.0 question answering dataset and train QA retriever models on this new dataset. We evaluate the quality of the dataset and more generally the MT method through direct evaluation, indirect comparison to other similar datasets, a backtranslation experiment, as well as through the performance of downstream trained QA models. In all these evaluations, we find that the method of transfer is not only simple to use but produces consistently better translated data. Given its good performance on the SQuAD dataset, it is likely the method can be used to translate other similar span-annotated datasets for other tasks and languages as well. All code and data is available under an open license: data at HuggingFace TurkuNLP/squad_v2_fi, code on GitHub TurkuNLP/squad2-fi, and model at HuggingFace TurkuNLP/bert-base-finnish-cased-squad2.

Machine translation (MT) has become mainstream, enabled by progress in MT quality. This work deals with QA, but techniques could be applied to similar tasks. Produced a Finnish SQuAD2.0 dataset from the English one. Trained BERT extractive QA model.. Number of documents with spans and span being a subset of document. Translate context, questions, and answers separately and then locate the translated answers in the translated context. The answer is out of context, and its translation does not necessarily match the translated context. Approach to try and realign using some heuristics. Separately trained alignment model on automatically generated data in 10 languages (Masad et al., 2023). Another approach is to tag with tokens around the answer you want to preserve and rely on the MT system to preserve these tokens. No matter which approach, it is a tedious process that is not easily transferred to other languages. Commercial MT systems (Markup-based transfer) are meant for translators, and they offer the ability to translate formatted documents. Direct application of the formatting-based transfer methodology, DeepL, as the translation engine. 90233 QA pairs out of 92749 in the original data, cost of ~20€. Evaluating MT is a challenge, there is no manually made test set in the target language. Percentage of QA pairs recovered, numerical comparison of observed scores to other versions in other languages, back-translation, and evaluation, manual inspection of a sample. Not all QA pairs can be recovered, recovering a higher percentage of the original annotation is beneficial for model training. Observation is close to 100% recovery rate, much better than previous works. Finnish SQuAD has the highest reported scores among machine-translated datasets in both EM and F1, a bit of a gap between the original but is expected. Backtranslation is pessimistic since errors increase with concurrent MT. A drop of 8.4 EM and 5.1 F1, unknown how this drop distributes across the two translation rounds. IF roughly evenly, we could expect ~5 EM and ~3 F1, and we can argue whether this is good or bad. Training on our version of Spanish leads to better F1 on both test sets, yet EM metric prefers a match of training and test data source hinting at subtle systematic span boundary differences. Sample and focus on span boundaries. Over-extension by a small margin (typically one token) the most common issue. Evaluation indicates the dataset is of good quality compared to similar machine-translated

versions of SQuAD. MT is, however, consistently lower quality, but usually, the choice is between MT or nothing, and then MT is better.

Paper 4: Rishabh Shastry, Patricia Chiril, Joshua Charney and David Uminsky

Entailment Progressions: A Robust Approach to Evaluating Reasoning Within Larger Discourse

Textual entailment, or the ability to deduce whether a proposed hypothesis is logically supported by a given premise, has historically been applied to the evaluation of language modelling efficiency in tasks like question answering and text summarization. However, we hypothesize that these zero-shot entailment evaluations can be extended to the task of evaluating discourse within larger textual narratives. In this paper, we propose a simple but effective method that sequentially evaluates changes in textual entailment between sentences within a larger text, in an approach we denote as "Entailment Progressions". These entailment progressions aim to capture the inference relations between sentences as an underlying component capable of distinguishing texts generated from various models and procedures. Our results suggest that entailment progressions can be used to effectively distinguish between machine-generated and human-authored texts across multiple established benchmark corpora and our own EP4MGT dataset. Additionally, our method displays robustness in performance when evaluated on paraphrased texts a technique that has historically affected the performance of well-established metrics when distinguishing between machine generated and human authored texts.

Encoding entailment as a feature. Word entropy is a way of identifying a human-authored text or an LLM. Human text tends to have a higher average word entropy. What words do humans use that LLMs do not? Paraphrasing can replace low-entropy words with high-entropy words to confuse MGT detectors. Mutual fund reports are used as a structured way of evaluating words. Interchanging between sentences is generally aligned with textual entailment. Positive entailment between sentences 1 and 2, 2 and 3, and a negative entailment between sentences 3 and 4. A pattern from the same author has similar entailment patterns, at least in the context of mutual funds. Introduce entailment progressions as a framework for comparing human and model written text. Propose a dataset called EP4MGT 70 158 machine-generated across eight SOTA LLMs. Cycle through the premise and hypothesis pairs and generate an entailment progression matrix. Can entailment progressions capture patterns exhibited across texts from the same author? Can entailment progressions capture differences in patterns across different authors? Three key datasets: MULTITUDE (a subset of 9109 English records), Ghostbuster (6256 records), and EP4MGT (70158 records). Convert text into premise hypothesis pairs using title-sentence or sentence-sentence approaches. Generate entailment progressions from pairs and classify them. Visually compare entailments in sentence progression. It outperforms previous metrics like word entropy. Entailment progression and EP4MGT can provide insight into entailment as a feature, Extending entailment progressions into potential identifiers of authorship and evaluating entailment progressions across different languages, tasks, and genres.

Paper 5: Hele-Andra Kuulmets, Taido Purason and Mark Fishel

How Well do LLMs know Finno-Ugric Languages? A Systematic Assessment
We present a systematic evaluation of multilingual capabilities of open large language models (LLMs), specifically focusing on five Finno-Ugric (FiU) languages. Our investigation covers multiple prompting strategies across several benchmarks and reveals that Llama-2 7B and Llama-2 13B perform weakly on most FiU languages. In contrast, Llama 3.1 models show impressive improvements, even for extremely low-resource languages such as Võro and Komi, indicating successful cross-lingual knowledge transfer inside the models. Finally, we show that stronger base models outperform weaker, language-adapted models, thus emphasizing the importance of base model in successful language adaptation.

Multilingual LLMs are getting better and better. Proprietary models are better than open models. Open models are catching up, but officially supported languages remain limited. It has been shown that Llama 2 7B could answer 14% and 40% of Finnish and Estonian even though only sub 0.01% of those languages were used in training data. 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro, and Komi. 7 models: 5 models from Llama 2 and 3.1, Mistal NeMo (Mistral AI), and Llamas (TartuNLP). LLMs are shown to perform better if English is used as a pivot language (prompts are translated). 5 tasks: MT, multiple choice QA, text classification, extractive QA, and commonsense reasoning. In general, Bigger models perform better. Almost no model could do commonsense reasoning. There is some improvement in few-shot prompting, but it is not consistent. CoT seemed to improve commonsense reasoning. Inconsistent with multiple choice QA and lower in extractive QA. Mistral NeMo outperforms Llama in Finnish and Estonian but not on very low-resource languages tested.

NoDaLiDa 2025-03-04

Keynote 3: Dirk Hovy

The Illusion of Understanding – Unpacking the True Capabilities of Language Models
Learning comparison: you read ~9000 words (0.4 GB) in your lifetime, models training on ~10 TB of data (25 000 lifetimes, 450 000 years to read). Humans learn from vast inputs too: audio ~280 GB/year, visual input ~4.5GB/hour (22TB/year), social context immeasurable quantity. Knowing vs. speaking vs. understanding language. Language is a social construct to facilitate social environments. Models' current limitations are detecting and predicting language patterns. Models don’t use socio-cultural aspects. Models don’t model context; equations could give counter-arguments. Knowledge does not equal predictions. Limits of knowledge: models don’t know representations of the world. Learning from text alone was never sufficient. Does it matter? LLMs are a Chinese room experiment; they mimic knowledge. Social understanding: embedding in the social environment. RL has a source of feedback, horribly inefficient. Understanding social behavior is critical for the next leap. Moravec’s paradox: simple tasks baffle current AI. Social factors: speaker and receiver have a social relation, context matters, social norms, culture and ideology, and surrounding it all communicative goals. Demographics matter, context matter, norms matter. Socio-economic status (SES) in the models: higher socio-economic classes tend to get better results from models. They ran different movies and shows through text-to-speech models and compared

errors to the SES of the character. SES also affects how people see GenAI: potential disparity in utility. If people could use these models better simply due to language ability, it would create a rift in society. SOA speech recognition: performance gaps for non-binary speakers: gender encoded in model representations during training. Modeling groups and individuals: group attributes are coarse but useful priors. Individual traits need complex modeling but are more personal. LLMs must adapt to education level: new metrics are needed. People treat machines as if they have the same feelings. Emotions are demographically stratified in languages: LLMs reflect gender stereotypes. Gender emotion stereotypes. LLMs also have religious emotional stereotypes. LLms can be tired out with long lists of harmful prompts (Many-shot Jailbreaking): safe in English but can be broken down in other languages. Tested safety behaviors: identified model compliance in the balance between harmful and harmless prompts. Metrics stop benign useful as soon as they become a target. How you ask models allows the models to take positions as you want. Realistic evaluation is crucial for accurate models. Loss of language is not loss of personality, empathy, motion, planning, etc. Understanding beyond words. Socio-demographic factors crucial to understanding language, necessary to balance safety and utility in LLMs, require new tests and metrics (behavioral evaluation).

Paper 6: Erik Henriksson, Otto Tarkka and Filip Ginter

FinerWeb-10BT: Refining Web Data with LLM-Based Line-Level Filtering
Data quality is crucial for training Large Language Models (LLMs). Traditional heuristic filters often miss low-quality text or mistakenly remove valuable content. In this paper, we introduce an LLM-based line-level filtering method to enhance training data quality. We use GPT-4o mini to label a 20,000-document sample from FineWeb at the line level, allowing the model to create descriptive labels for low-quality lines. These labels are grouped into nine main categories, and we train a DeBERTa-v3 classifier to scale the filtering to a 10B-token subset of FineWeb. To test the impact of our filtering, we train GPT-2 models on both the original and the filtered datasets. The results show that models trained on the filtered data achieve higher accuracy on the HellaSwag benchmark and reach their performance targets faster, even with up to 25\% less data. This demonstrates that LLM-based line-level filtering can significantly improve data quality and training efficiency for LLMs. We release our quality-annotated dataset, FinerWeb-10BT, and the codebase to support further work in this area.
https://dspace.ut.ee/server/api/core/bitstreams/0fcb2b41-903c-472e-9e38-fe113ac4dd19/content

Data quality matters! Can we leverage an LLM to do line-wise quality annotation in web data? Exploratory analysis: what are low-quality lines like? How many low-quality lines are there? Improved data cleaning: less data -> shorter training time -> grinder models. FineWeb 15 trillion tokens from CommonCrawl. Prompt GPT-4o-mini to label 20 000 documents. Clean, high-quality documents suitable for training LLM. Low quality “junk” are given descriptive label (why are they junk?), labels form a dynamically growing taxonomy. Model prompt to evaluate lines suitable for LLM training, evaluate in context and retain valuable and diverse linguistic content. 83% of ~270k lines are labeled as clean. 547 unique junk labels. If label is just one line, they were labeled as clean, manual inspection also led to 23 clean labels, resulting in 382 unique junk label. UMAP projection show some clusters of junk labels. Create label categorize with o1-preview. Labelling verified by two human annotators

(IAA: 0.70, Cohen's kappa). Can't scale GPT-4o-mini to millions of documents, testesd three encoder models and DeBERTa-v3-base got best result (F1: 0.81) and mos missclassifications were clean. Label FineWeb-10BT using DeBERTa-v3-base and give each line a quality score. 75% of lines scored > 0.9 and 8% of lines < 0.5. Removing 25% lowest quality lines results in a better model in less time. Currently labeling full FineWeb dataset (~62 000 GPU hours). Aim to make the pipeline automatic, aim to apply to the multilingual data, refine method.

Paper 7: Mike Zhang, Max Müller-Eberstein, Elisa Bassignana and Rob van der Goot

SnakModel: Lessons Learned from Training an Open Danish Large Language Model
We present SnakModel, a Danish large language model (LLM) based on Llama2-7B, which we continuously pre-train on 13.6B Danish words, and further tune on 3.7M Danish instructions. As best practices for creating LLMs for smaller language communities have yet to be established, we examine the effects of early modeling and training decisions on downstream performance throughout the entire training pipeline, including (1) the creation of a strictly curated corpus of Danish text from diverse sources; (2) the language modeling and instruction-tuning training process itself, including the analysis of intermediate training dynamics, and ablations across different hyperparameters; (3) an evaluation on eight language and culturally-specific tasks. Across these experiments SnakModel achieves the highest overall performance, outperforming multiple contemporary Llama2-7B-based models. By making SnakModel, the majority of our pre-training corpus, and the associated code available under open licenses, we hope to foster further research and development in Danish Natural Language Processing, and establish training guidelines for languages with similar resource constraints.
https://dspace.ut.ee/server/api/core/bitstreams/ea4db214-46ee-431c-8745-fe275336de07/content

To what extend can we further pre-train an LLM for Danish? Mid-resource language, typologically related to english and overlapping character sets, sufficient data. Started from Llama2-7B checkpoint. 13.6B words from Danish sources. 8928 GPU hours to train. More language identification and deduplication of documents, started with 24.6B and after preprocessing had 13.6B words. 97% data for training and 3% validation. ValidationLLM for 3D parallelism (not updated anymore). Learning rate was tricky, warmed up and decayed, suggestions are to re-warm and re-decay the learning rate (.5 * original value of peak). Instruction tuning with language modeling, they used Danish OpenHermes, SkoleGPT, and AyaColelction. Ideas on how to get truly native instruction tuning data? Evaluation using EuroEval. SnakModel-7b-instruct performed better than Llama-2-7B. If you use a better base model are already performing better. Drastic improvements on Danish specific tasks. NER and QA performance drops over steps for base. NER is attributed to forcing the model to produce JSON, QA doesn't answer but continues the question. Instruction tuning recover QA and NER performance; language specific tasks improve over training steps. Performance doesn't increase after around 2000-5000 steps, could stop training early (good for smaller datasets). SSAs measure the weight changes (cosine-similarity for weight matrices). Higher concentration of change to the later ⅓ of the model. Change per parameter type: embedding layer and Language Model Head changes alot, and alot is happening in Gate and Weight

forward layers. Not a lot changes in the attention layers. Hypothethize the attention layer doesn't need to change due to syntactic similarity with English. Training shows several improvements on language-specific tasks, likely generalizes for any language. Instruction tune for 1 epoch for efficient, focus on specific parameters in LoRA-based fine-tuning.

Paper 8: Jenny Kunz

Train More Parameters But Mind Their Placement: Insights into Language Adaptation with PEFT

Smaller LLMs still face significant challenges even in medium-resourced languages, particularly when it comes to language-specific knowledge -- a problem not easily resolved with machine-translated data. In this case study on Icelandic, we aim to enhance the generation performance of an LLM by specialising it using unstructured text corpora. A key focus is on preventing interference with the models' capabilities of handling longer context during this adaptation. Through ablation studies using various parameter-efficient fine-tuning (PEFT) methods and setups, we find that increasing the number of trainable parameters leads to better and more robust language adaptation. LoRAs placed in the feed-forward layers and bottleneck adapters show promising results with sufficient parameters, while prefix tuning and (IA)^3 are not suitable. Although improvements are consistent in 0-shot summarisation, some adapted models struggle with longer context lengths, an issue that can be mitigated by adapting only the final layers.

https://dspace.ut.ee/server/api/core/bitstreams/31c3ac9d-0cd9-425c-8294-392ca65d4c99/content

[Missade första minuterna…] LoRA and bottleneck adapters show improvements especially in the zero-shot setup. Simply adding target language task demonstrations also improves the score. Higher number of parameters was also generally better. LoRA in the feed-forward layers are the best performing setup, followed by bottleneck adapters. Not much performance increase in attention layers. Prefix tuning hues the models capabilities. Not suitable architecture for CPT? LoRA in feed-forward is better than both in attention and the combination. LoRA in attention with few trainable parameters got worse when context increase. The issue can be mitigated by training only the last layer. Are fine-grained language capabilities really improved? Sis the model acquire the language specific knowledge than is missed in adaptation with translated data? We need evaluation data that explicitly tests for such capabilities and human evaluation.

Paper 9: Samia Touileb, Vladislav Mikhailov, Marie Ingeborg Kroka, Lilja Øvrelid and Erik Velldal

Benchmarking Abstractive Summarisation: A Dataset of Human-authored Summaries of Norwegian News Articles

We introduce a dataset of high-quality human-authored summaries of news articles in Norwegian. The dataset is intended for benchmarking of the abstractive summarisation capabilities of generative language models. Each document in the dataset is provided with three different candidate gold-standard summaries written by native Norwegian speakers and all summaries are provided in both of the written variants of Norwegian – Bokmål and Nynorsk. The paper describes details on the data creation effort as well as an evaluation of

existing open LLMs for Norwegian on the dataset. We also provide insights from a manual human evaluation, comparing human-authored to model generated summaries. Our results indicate that the dataset provides a challenging LLM benchmark for Norwegian summarisation capabilities.
https://dspace.ut.ee/server/api/core/bitstreams/797af9a2-13d3-4c36-987b-15b2c4101c2f/content

LLMs are used to generate consensed summaries of texts in Norway, no gold standard. Hired three annotator, strong academic backgrounds related to journalism, all Norwegian speakers.Make short and precise summary, summary should be bulleted list, language must be clear, precise and concise, journalistic integrity must be mainstained, summary must be engaging, must answer 5 Ws (who, what, where, when, why), maximum 700 characters. Simple text editing platform, several meetings to discuss process and progression, no alignment. No unique gold summary version, benchmark with diversity. Two preferred Bokmål and one Nynorsk, and they wrote a summary each and then they translated so in total 3 summaries each of Bokmål and Nynorsk. 189 summaries in total (63 for each annotator summary). Diverse length in summaries between the annotators. Different strategies were used: highlighting key elements or reading articles twice. Bulleting-like news articles, sports articles, disaster-related new, injuries and investigations were easiest. Complex articles required more time. Translation was easy. 6 prompts each for Bokmål and Nynorsk. Tested on 9 LLMs. Computed the maximum of the output from the LLM compared to the annotators, then averaged. Human evaluation for comparing human and LLM summaries. Three criteria for human evaluation: relevance, consistency and fluency. 146 responses from human evaluation, 138 preferred human authored summaries. Issues related to relevance: often miss important information, cut off mid-sentence, copy-pase, repetitive or too short with incomplete contexts and unnatural sentences. Issues related to consistency: summaries generally consistent with the source, but identified some issues like phrase petition minor text alterations, invented quotes and entity confusion. Issues related to fluency: some summaries perpetuated sentences excessively and occasionally missed function words, affecting clarity. First freely available dataset of human-authored summaries of Norwegian new articles for benchmarking abstractive summarization. Comprehensive evaluations with human evaluators and generative models demonstrating robustness and complexity.

Paper 10: Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal and Lilja Øvrelid

A Collection of Question Answering Datasets for Norwegian
This paper introduces a new suite of question answering datasets for Norwegian; NorOpenBookQA, NorCommonSenseQA, NorTruthfulQA, and NRK-Quiz-QA. The data covers a wide range of skills and knowledge domains, including world knowledge, commonsense reasoning, truthfulness, and knowledge about Norway. Covering both of the written standards of Norwegian – Bokmål and Nynorsk – our datasets comprise over 10k question-answer pairs, created by native speakers. We detail our dataset creation approach and present the results of evaluating 11 language models (LMs) in zero- and few-shot regimes. Most LMs perform better in Bokmål than Nynorsk, struggle most with

commonsense reasoning, and are often untruthful in generating answers to questions. All our datasets and annotation materials are publicly available.
https://dspace.ut.ee/server/api/core/bitstreams/b535f086-e14d-4d2b-9231-c1c184c93cfb/content

There is no QA benchmark for Norwegian. Created four novel datasets for the least addressed QA directions for Norwegian: Norwegian specific and world knowledge, commonsense reasoning, truthfulness. Evaluating 11 open Norwegian LMs in zero- and few-shot prompting. 21 annotators, native Norwegian speakers (BA/BSc/MA/MSc students in linguistic and computer science). Provided dataset-specific annotation guidelines. Adaptation of English dataset: human annotation and translation of OpenBookQA, CommonsenseQA and TruthfulQA. Filter out low-quality examples and make minor edits (local examples). Adaptation of NRK Quiz Data: targeted adaptation of quizzes with temporal adjustment, content filtering, data cleaning. Multiple Choice questions, generation task, quiz task. 50 prompts in Norsk Bokmål (NB) and Nynorsk (NN) are integrated into NorEval. K-shot evaluation (0-shot on NRK-Quiz-QA, NorCommonSenseQA, NorTruthfulQA and 0, 1, 4, 16 on NorOpenBookQA). Multiple-choice QA: probability based scoring; Generation: Rouge-L. In general no single LM performs best on all datasets, smallest models performs on par with a random classfiier. All the data are publicly available.

NLP4CALL 2025-03-05

Keynote: Andrew Caines

The Potential and the Pitfalls of Very Large Language Models for Language Learning Applications
Use of LLMs for black-box CALL. Probably not the case that LLMs are ready for CALL out of the box. Human-machine hybrid applications, use LLMs are judges. (Dynamic) benchmarking for CALL / SLAM, we could do with some better benchmarking. “Baby” LMs trained on high quality, domain specific data. Plots of Large Langauge model changes (https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/). 57 subjects in MMLU, LLMs seems to be pretty good on the task. A lot of targeted instruction tuning probably the reason behind the optimal performance. Problems with Multiple choice questions: LLMs are sensitive to order of the options. The pattern isn’t persistent between models. LLMs are sensitive fot prompt format, e.g. symbols instead of numbers (Alzahrani et al. 2024) or modified spacing/saparators (Sclar et al. 2024). LLMs are sensitive to paraphrasing, calls to evaluate on multiple prompts (Mizrahi et al 2024). LLMs are very large, training is costly (Llama 2 70B required 1.7 million GPU hours, emitting 291 tons of CO2). The cost of training sucha model in popular cloud computing was ~2.6 million USD. Training has become more secretive. LLM leaderboards should be treated with caution, they are static, prone to targeting. The data could be leaked and be included in LLM training. Another kind of LLM evaluation? Chatbot arena, give two LLM and give prompts and then rate them (https://lmarena.ai/). LLMs are good at certain taska dn getting better, really optimized for QA and chat. ALTA Institute (Automated Language Teaching & Assessment) at the University of Cambridge. An SLA / CALL benchmark, wanted to investigate how LLM had coded knowledge about the CEFR from pre-training. Evaluated 6 LLMs on 3 tasks. They couldn’t reliably perform on CEFR tasks. Two version for prompts,

requirest for JSON format. GPT-4o performed the best on the task. The different LLM have different output patterns, required many different regex patterns to parse the outputs (as much as 30). All three tasks were challenging. Readability classification had high level of error, no benefit from including CEFR descriptors. Essay scoring some predict on CEFR predominantly, failing to make predictions for all essays. Learner simulation of specific CEFR-levels does not work well in zero-shot setting. Supervised models offer more controllable and viable solutions. Grammatical Error Correction with LLMs. The task is to correct grammatical errors in learner text. Could be more than just minimal edits toward grammaticality, such as fluency correction. Evaluate 10 models (7 open weights, 3 commercial). 4 English GEC benchmarks. 7 zero-shot prompts, 3 for few-shots. Takeaway: size does not always get increasing performance. Performance tended to decrease as you went up the CEFR levels. Evidence from other papers suggests that humans prefer LLM corrections to reference & non-LLM corrections. It appears they are making fluency corrections. Not a pedagogical perspective but a stylistic one. Like to continue by evaluating which ones are more useful for the learner. aLOA: *adative* Learning Oriented Assessment. Previous work by Gladys Tyen on adaptive chatbots. With known learner level ensure that chatbots do not feature language that is too difficult. Apply penalties for words which are above level. Filters out-of-vocabulary words. Potential model for teacher chatbots in the Teacher Student Chatroom Corpus, 1 hour long communication between teacher and student. Task to predict teacher response from some dialogue context. 8 teams participated using LLMs with zero- and few-shot learning, fine-tuning and RL. Human judges were preferring some of the LLM based systems in terms of student teacher chats (compared to human annotated references). Automated metrics could be gamed, DialogRPT preferred complete responses which gave answers too easily to the students. Education-specific auto-metrics needed. Can all the featured research be improved with smaller “baby” LMs pre-trained on relevant data? Both int erms of cost and performance (and security). Even in English it is hard to find sufficient (relevant) data. A data processing problem. Other issues with LLMs. Environmental concers regarding use of GPU/TPU, ethical concerns regarding commercal LLMs, safety & security concerns, copyright issues with training data collection, bias, misinformation, offensive language, stereotypes, prejudice. ALERT benchmark to circumvent safety measures through adversarial prompting. GEC in languages other than English. Work on English has dominated NLP for a long time, exacerbated by LLM era (Matthew effect). LLMs appear to have some multilingual capabilities. It’s clear that GEC remains challenging cross-linguistically, and for many languages LLMs are poor.