Stockholms universitet

TREDJE SYMPOSIET FÖR SVENSK DIALEKTOLOGI

ASR-UD

Automatic Speech Recognition – Upper Dalecarlian

Automatisk taligenkänning för ovansiljanmål

Nathan J. Young · Stockholms universitet
Piotr Garbacz · Universitetet i Oslo

Uppsala · 2 oktober 2026

Bakgrund och syfte: behovet av transkription

- Ovansiljanmålen har begränsade digitala språkresurser.
- Få personer har den språk- och lokalkännedom som transkription kräver.
- Transkription blir en flaskhals, liksom för många andra lågresursspråk.
- Syftet är att ge ett första textförslag för nya inspelningar, att granska och rätta.



Programredaktør Andor Birkeland og

hustru Halinas legat STI

Bakgrund: vad är ovansiljanmålen?

- Nordiska folkmål i Siljansbygden och dess omgivningar i övre Dalarna (Garbacz et al., 2026).
- Sju varieteter (med inbördes variation): Moramål, Oremål, Orsamål, Sollerömål, Venjansmål, Våmhusmål och Älvdalsmål.
- Utgör ett av de mest särpräglade dialektområden i hela Norden (pga sin unika blandning av arkaismer och novationer)



Material och insamling

- Befintliga dialektinspelningar med transkriptioner (gjorda i regi av Universitetet i Oslo i 2010)
- Nuvarande underlag: Älvdalsmål, Moramål och Orsamål.
- Samtalsmaterial med en eller flera talare.

Korpusuppbyggnad: insamling och urval

- Inventera inspelningar och befintliga transkriptioner.
- Dokumentera ort, bymål, talare, inspelningsår och samtalssituation.
- Beskriva transkriptionernas tillkomst och lokala skrivkonventioner.
- Granska materialet och sammanställa täckningen per varietet.

Korpusuppbyggnad: träningsunderlaget

- 53 ljud–textpar; flera par kan komma från samma inspelning.
- 5 542 korta träningsavsnitt och cirka 38 000 transkriberade ord.
- 616 avsnitt används för jämförelser och ingår inte i modellträningen.
- ELAN-exporterna innehåller text, talare och tidsangivelser (Wittenburg et al., 2006).
- Enhetlig teckenkodning i transkriptioner och ordlistor.

Metod: bootstrapping från svenska

- Bootstrapping, eller överföringsinlärning: återanvända en tränad modell i stället för att börja från noll.
- Vi utgår från svenska VoxRex och tränar vidare med ljud och transkriptioner från ovansiljanmål (Malmsten et al., 2022).
- Svenska och ovansiljanmålen är besläktade. Gemensamma ljudmönster kan ge ett försprång.
- Målet är att behöva mindre transkriberat material i målvarieteterna. Studier med wav2vec 2.0 visar överföring mellan språk (Conneau et al., 2021).

Metod: från ljud till text

- Ljudmodellen bygger på wav2vec 2.0 (Baevski et al., 2020).
- PyTorch och Hugging Face Transformers används för träning och transkription (Paszke et al., 2019; Wolf et al., 2020).
- Silero VAD (Voice Activity Detection, talaktivitetsdetektion) hittar talavsnitt i nya inspelningar (Silero Team, 2024).
- pyannote.audio med Community-1 grupperar avsnitten efter talare (pyannote, 2025).
- Transkriptören granskar textförslaget mot inspelningen.

Metod: ord och texter som stöd

- En separat n-grammodell väger in ord och ordföljder vid valet av textförslag.
- Ett n-gram är ett ord eller en följd av ord: ”ja” är ett 1-gram, ”ja just” ett 2-gram.
- Vi använder ordlistor, andra texter och träningsdelens transkriptioner.
- Textunderlaget behöver passa den varietet och det tal som ska transkriberas.
- Effekten prövas med samma ljudmodell, med och utan ordstöd.

Resultat: två textförslag

- Referens 1: ”menn war- og nų byddjer du į remað ja- just eð- sagd du”.
- Förslag 1: ”men n wa o nu byd d jer du įr em a ja ju st eð s aįgd du ?”.
- Referens 2: ”ig fikk luv so att dä wart inggan match”.
- Förslag 2: ”ig fik k lö v so at t eð war t ing ga ma t j”.

Resultat: den aktuella modellen

- Modellen från 25 augusti 2026, utan ordstöd: 616 provavsnitt.
- Ordfelkvot (Word Error Rate, WER): 156,4 %.
- Teckenfelkvot (Character Error Rate, CER): 68,9 %.
- Teckenfelkvot utan mellanslag: 61,0 %.
- Det prövade n-gramstödet har inte än gett någon stabil förbättring, men modelljusteringar pågår.

Fortsatt arbete

- Bredda materialet till fler ovansiljanmål och redovisa täckningen per varietet.
- Förbättra ordstödet och pröva det på nya inspelningar och talare inom varje varietet.
- Jämföra granskningstid med tiden för transkription från början.
- Pröva varietetsspecifika skrivkonventioner och modellanpassningar.
- ASR-UD lanseras på autophon.org i december 2026.

Referenser (1/3)

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021* (pp. 2426–2430). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2021-329

Garbacz, P., Jacobsen, M., Østmo-Sæter, B. O., Mikkelsen, C. V., Slethei, J. E., Reksnes, S. L., Ullestad, Y., Rachek, A. R., & Emblem, D. B. R. (2026). Ovansiljanmålens syntax och morfologi. *Svenska landsmål och svenskt folkliv, 147–148*, 35–105. https://doi.org/10.69824/svlm.147.54605

Referenser (2/3)

Malmsten, M., Haffenden, C., & Börjeson, L. (2022). *Hearing voices at the National Library: A speech corpus and acoustic model for the Swedish language*. arXiv. https://doi.org/10.48550/arXiv.2205.03026

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

pyannote. (2025). *speaker-diarization-community-1* [Model card]. Hugging Face. https://huggingface.co/pyannote/speaker-diarization-community-1

Referenser (3/3)

Silero Team. (2024). *Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier* [Computer software]. GitHub. https://github.com/snakers4/silero-vad

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association. https://aclanthology.org/L06-1082/

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6