



Spóznawanje hornjoserbšćiny staw / postupowanje / strategija

Daniel Zoba, 06.11.2025



Stawizny

- 2020: studija přewjedžomnosće (njeje so wozjewiło)
- 2021: [ryzy hornjoserbski system + předłoha za dalše řeče](#)
- 2022: [technologiske polěpšenja](#)
- 2023: [wusměrjenje na podtitle božej mšě + strategiske poradźowanje](#)
- 2024: [wozjewjenje a wužiwanje za live podtitle na youtube](#)



Prezentacije



SPECOM 2024 Běłohród

○ rozprawa



LT4ALL 2025 Paris

○ rozprawa



ESSV 2025 Halle

○ rozprawa





Technologije

- **"dLabPro"** (OSS) za jednore systemy
 - za zakładne slědženje
- **"recIKTS"** (FHG) tradicionelny system (akustiski model, fonetiski słownik, statistiski řečny model)
 - za "wobmjezowane domeny"
- **"whisper"** (OSS) transformer modele (end to end)
 - za powšitkowne wužiwanje
- **"wav2vec2"** (OSS) transformer modele (end to end + statistiski řečny model)
 - za "wobmjezowane domeny" (a powšitkowne wužiwanje?)



Korpora

- interny korpus (hlej rozprawy)
- wozjewjenje [džěla korpusa na Hugging Face](#) po přihłosowanju řečnika
- [Common Voice](#)



Hugging Face





Kooperacije

- **Fraunhofer IKTS** (dLabPro, recIKTS, whisper)
- **Korla Baier** (whisper, wav2vec2)
- **HTW Drježdźany** (whisper)
- **Dowera GmbH** (whisper)
- **META** "Language Technology Partner Program"
(wav2vec2 + ?)



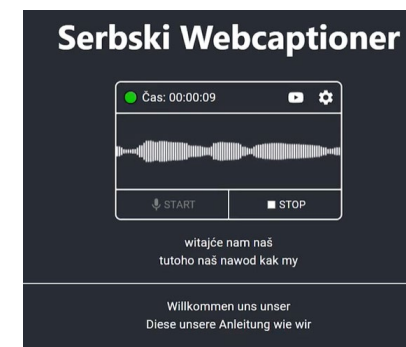
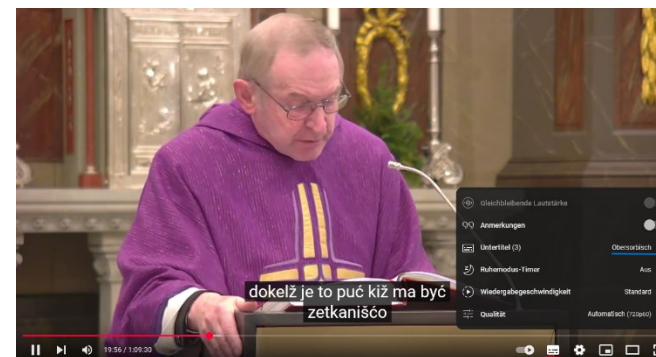
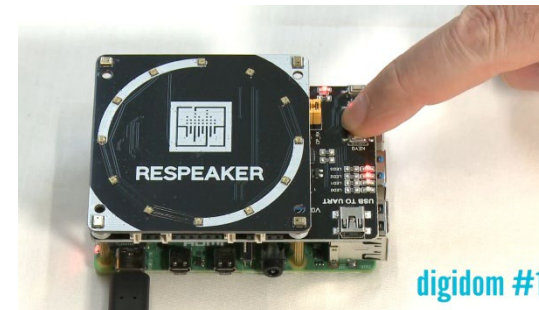
Fraunhofer-Institut für Keramische Technologien
und Systeme IKTS





Aplikacije

- [Smart home prototyp](#)
- [Lingufino](#)
- [Transkripcije a podtitle wutworić](#)
- [Webcaptioner](#)
- **Jitsi Meet:**
přidawk za simultany přełožk so přihotuje





Strategija za dalewuwiće

- [Strategy for Future Development of Upper Sorbian Speech Recognition](#)
- dale nahrawanja a teksty zběrać a předžělać
- dalewuwiće technologije wobkedźbować a wuspytać



Rozšěrjenje zwukoweho corpora

- Common Voice
- nahrawanja w studiju (lucija)
- korektury awtomatiskich transkriptow
(zarjadowanja, stare filmy a słuchoknihi, ...)
- "Forced Alignment" nahrawanjow z transkriptami
- naprašowanja pola MDR a NSLDŽ
- "Pseudo labels" z nahrawanjow bjez kontrole (po dowěry filtrować)



Rozšěrjenje tekstoweho corpora

- transkribowanie gmejnskich posedženjow Chróšćicy
- naprašowanje pola zwjazkowneho předsydstwa Domowiny dla transkribowanja posedženjow
- korektura transkriptow božeje mšě Chróšćicy
- zběrka tekstow



Dalše wuwiće aplikacijow

- simultany přełožk (jako tekst abo zwuk) za:
 - gmejnske posedženja
 - staršiske zhromadźizny
 - podtitle za youtube
 - dalše zarjadowanja
 - spřistupnjenje systema za zajimowanych
- integracija do šmóratka
 - za Android eksistuje OSS projekt





Wužadanja

- wobchadna řeč
- „code switching“ (Němčina, Jendželšćina)
- NER ("named entity recognition")
 - na přikład "18:30 hodź." město "wosomnaće hodžin třiceći"
- simultany přełožk do Němčiny
 - segmentacija po duktusu, nic po sadach
 - multiplikacija zmylkow spóznowanje + přełožk
 - kontekst



Wozjewjene KI-modele (wuběrki)

- <https://huggingface.co/Korla/whisper-large-v3-turbo-hsb>
- <https://huggingface.co/Korla/Wav2Vec2BertForCTC-hsb>
- <https://huggingface.co/DILHTWD/whisper-large-v3-turbo-hsb>
- <https://huggingface.co/DigitalLabs42/whisper-large-hsb-v1/tree/version2>



Kónc / Prašenja

- https://github.com/ZalozbaDev/language_exchange