

Digitalizaciski kruh, 06.11.2025

Spóznawanje řeče Delnjoserbščina

Staw / Postupowanje / Strategija

Jan Měškank, Astrid Schmiedelowa

- pilotowa faza 2023 + přihotowanska faza 2024 → jako pódlanski projekt
- aktualnje: wutwarjenska faza 2025 → nětko jako hłowny projekt
- klětu: hłowna faza I

Generelnje:

- orientacija na hornjoserbskim projekće z cilom přirunujomne delnjoserbske systemy wuwic
- profitujemy wot zakładow z projekta za předčitansku funkciju, wosebje hladajo na delnjoserbsku fonetiku a fonologiju
- projekt dale zmóžnja wobšěrne fonetiske/fonologiske slědženja z modernej metodologiju; poslednje přirunujomne prócowanja běchu slědženja za Serbski rěčny atlas w počatku 20. lětstotka

Prezentacije

• wědomostny poster „Rěčne technologie we wuwicu – Rěčna synteza a spóznananje rěče w přirunanju“

• 29.11.2023: Institutny dzeń

• 20./21.02.2025: Serbske žiwjenske rupy w digitalnej dobje – konferenca a wiki

Rěčne technologie we wuwicu

Rěčna synteza a spóznananje rěče w přirunanju

abc →   → abc

 Serbski Sorbisches
institut Institut

RĚČNA SYNTEZA – TEXT-TO-SPEECH (TTS)

Wudawanje informacijow na akustiskim rěčnym puću na př.:

- předčítanski hlós: mj. dr. podpěra za špatnje widžacych, unafabetow, luźu z čožemi při rěčenju
- informaciske systemy, nawigaciske systemy, připowědženje zastanišćow w buse, tramwajow abo čahu
- podpěra při wurknerju abo wučerju rěča

 STROMU PŘEDČÍTAC

DATOWY ZAKŁAD A TECHNISKI PRINCIP

Leksikon

- lěšćina abychow formow z antografiskej formu na jednym boku a plecto JEDNEJ fonetisce transkribowarjow formu na druhim
- ma zakryć wšěle dle STANDARDNEJE rěče

Zwukowy inventar

- lěšćina wšěch wurkowacych zwukow a zwukowych wjacow kaž tež wšěle, wšěch dźeržowacych zwukow

Awdijowy material

- nahradowanje JEDNEHO rěčnika za kódu rěči jako zakład za zapřewarjenje tutej přiradowanych rěčnych elementow za syntetizowane wudawanje
- po možnosti HOMOGENY nahradowani material nastupajo synstetizowane, intenciju, hlasowu kwalitu

Přewobdženje tekstow

- w zapřetajm tekstach se ličby, dety, symbole a skřiwěnski do wurkowacych abychow přewobdženja



SPÓZNAWANJE RĚČE – AUTOMATIC SPEECH RECOGNITION (ASR)

Ekstrakcija lingwistiskeho wobsoha wurpajnja (wocnam):

- diktowaniske systemy
- wudženje z rěču: poslušowanje aparatow na př. mobilny telefon



DATOWY ZAKŁAD A TECHNISKI PRINCIP

Leksikon

- lěšćina abychow formow z antografiskej formu na jednym boku a WJACORYM wotpowědowacymi fonetisce transkribowacymi formami na druhim
- ma po možnosti wšěle PRŮHODNEJE, WYSTUPOWACE wurkowanske formy zapřijet (na př. standard, dialekt, wocchadne rěče)

Zwukowy inventar

- lěšćina wšěch zwukow a zwukowych wjacow, kutaž po možnosti optimalne spóznanowanje zmogly
- podobne zwuki so ewd. do klasow zjednotě, zo by so přepřeznanowanje minimehowalo

Awdijowy material

- HETEROGENE nahradowanje WŠELAKICH rěčnikow (starob, spěch, wurkowowanje) za trenowanje rěčneho modelu, zo by potom mohl po možnosti široko spektrum hlasow a hlasowych kwalitow derje spóznać



PROJEKTOWY RAMIK

Wuwicu předčítanskeje funkcije za hornjo- a dolnosorbščinu (naprawa za přepřeznanje komunikaciskeje bjezbarměće w ramiku Słěšćiny za inkluziwu SŁWŲK).

Projektowy wotběh

- koncepciska faza 2018, pilotowa faza 2019, hłowna faza 2020-2023, kónčnica faza 2024

Zaměr

- wuwicu dweju programow (z mjenom Juro resp. Matej) za předčítanje dolnosorbščinskich resp. hornjosorbščinskich tekstow wšěch družinow na zrozumliwost wažnje a z dobrym wurkowowanjom
- produkt swobodnje přistupny, lochko wužiwajomy, da so na internetowych stronach integrowat

Techniske wuwicu

- Open Source Software: MaryTTS
- ma dźělowny grat za nahradowanje z nowymi rěčmi
- wšěle syntazy (Unit-Selection-zwizawowanje přiradowanych rěčnych elementow) skłó za předčítanski funkciju wažnje spěšne přewobdženje a wudawanje

 Live Demo Juro & Matej

PROJEKTOWY RAMIK

Tworjenje prototypa za awtomatisko spóznanowanje rěče za dolnosorbščinu (najprjedy jako pilotarski projekt, kutaž spěchowanje Zalcba za serbski lud).

Projektowy wotběh

- pilotowa faza 2023 – tworjenje prototypiskeho přikazoweho systoma za dolnosorbščinu
- pod wuměnenjom schwalenja srěčkow: wutwarjenska faza 2024 – cělowawice a přepřeznanje datoweho zakłada (dale jako pilotarski projekt, rozšěrjanje na wšěle rěče)

Zaměr

- wuwicu dolnosorbščinskeho spóznanowanja rěče za domenu „smart lamp“ (komandowy system za wšěleke wužiwowanje w bydlenskim wobutku)
- tworjenje zakładow za daleke wuwicu a wužiwowanje na polu spóznanowanja rěče

Techniske wuwicu

- wuchadźacy z wotpowědneho hornjosorbščinskeho komandoweho systoma (wuwicu 2020 w kooperaciji Zalcba za serbski lud z Fraunhoferowym institutom KITS Dresden) BILU Cottbusa wobkřežowanje so wšěleke dlele (skryty, moduly, materialje, a.d.) a přiměrowe so na kmanym mjejsnje za dolnosorbščinu
- Zalcba za serbski lud póló a straja hardware k dispoziciji

SŁĚDZENJE

W ramiku aktualneho projektow Serbskeho instituta z. t. za rěčnu syntezu a spóznanowanje rěče wotmowa so tež **woblične fonetiske a fonologiske sčlědženje**. Při wuwicu předčítanskeje funkcije wobkřežowanje so JEDNE wurkowowanje, w normalnym padnje je to **standardne wurkowowanje** cěneje rěče, z derje zakładowanym wurkowowanym znanim. Jako najprjedy nadležat, zo pak nimatej ani hornjo- ani dolnosorbščina tuteje standardy, čožowidla dycha wotmo so – přiměrowe na techniske polje a **množstwo syntetizoweho programow** – najprjedy nadležat. Zo bychu so težne ransady we wurkowowanach wariantach z rěčowstojneho zjednotěni z dobrym zakładow a kompetencijow tworje mohl, je **zakładne sčlědženje nuznje trěbne**. Sčlědži so tohorunja na lěto přisupnym materialu a při pólnych sčlědženjach zběranych nahradowanjach. Wuslědki so pódla direktneho načatowanja w projektovej zběrce a krok po kroku publikowanje. Dětne dopřeznanje so zodobom zakładow za přiměrowe w lěksikonje a rěčnym modelu za awtomatisko spóznanowanje rěče.

 **SACHSEN**
Landesministerium für Kultur und Medien
Zusammen mit dem Institut für die deutsche Sprache in der Sorben-Region

 **Fraunhofer IAI**
Institut für Arbeitswissenschaftliche Informationssysteme

 **Zalcba za serbski lud**
Institut für die deutsche Sprache in der Sorben-Region

 **Serbski Sorbisches Institut**
Institut für die deutsche Sprache in der Sorben-Region

- "dLabPro" (OSS) za jednore systemy
 - za zakładne slědženje
- "recIKTS" (FHG) tradicionelny system (akustiski model, fonetiski model, statistiski řečny model)
 - za "wobmjezowane domeny"
- "whisper" (OSS) transformer modele (end to end)
 - za powšitkowne wužiwanje
 - <https://huggingface.co/Korla/whisper-large-v3-turbo-dsb>
- "wav2vec2" (OSS) transformer modele (end to end + statistiski řečny model)

- interny korpus (nahrawanja DNW awdijo, předčitanska funkcija, nowe studijowe nahrawanja)
- wozjewjenje džěla korpusa na Huggingface (HF) po přihłosowanju řečnika
 - https://huggingface.co/datasets/Serbski-institut/dsb_audio_corpus
- Common Voice
 - <https://commonvoice.mozilla.org/dsb>



Fraunhofer-Institut für Keramische Technologien
und Systeme IKTS



- Fraunhofer IKTS (recIKTS, dLabPro, whisper)
- Założba za serbski lud (dLabPro, Hugging Face)
- K. Baier (whisper, wav2vec2)
- META "Language Technology Partner Program"

- Smart home prototyp (aktualne):
https://github.com/ZalozbaDev/raspberry_pi_demo_spoznawanje/tree/dolnoserbski
- transkripcije a podtitle wutworić: <https://spoznawanje.serbski-inkubator.de/> → dsb model transkripciskeje demo

Perspektiwisce:

- simultany přełožk (jako tekst abo zwuk) za
 - gmejnske posedženja (na př. Domowinska župa Dolna Łužyca)
 - zarjadowanja Serbskeho instituta abo druhich institucijow z delnjoserbskim podžělom

- analognje k strategiji za hornjoserbšćinu
- dale nahrawanja a teksty zběrać a předžěłać
- dalewuwiće technologie wobkedźbować a wuspytać

Rozšěrjenje zwukoweho a tekstoweho korpusa

- dalše nahrawanja
- sonděrowanje dalších žórłow:
 - "Forced Alignment" nahrawanjow z transkriptami → <https://huggingface.co/Korla/Wav2Vec2BertForCTC-hsb>
 - nahrawanjow bjez transkriptow
 - material ze skerje ćežko přistupnych žórłow (medijowe archiwy, nahrawanja posedženjow)
- korektury awtomatiskich transkriptow (zarjadowanja, stare filmy a słuchoknihi, ...) → po pruwowanju žórłow/material
- Common Voice
- zběrka tekstow

- problem rozdžěla mjez maćernorěčnymi a noworěčnikami, wosebje wulki w delnjoserbšćiny
- "Code switching" (Němčina, Jendželšćina)
- simultany přełožk do Němčiny
 - segmentacija po duktusu, nic po sadach
 - multiplikacija zmylkow spóznowanje + přełožk
 - problemy dla rozdžělnych řečných strukturow, na př. negacija

Prašenja?