Multilingvální rozpoznávač západoslovanských jazyků pro generování folklorní transkripce z audionahrávek

verze 1.0

Uživatelská příručka

a

Technická dokumentace



Jazyková paměť regionů České republiky.
Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka.

Základní údaje

Název výsledku

Multilingvální rozpoznávač západoslovanských jazyků pro generování folklorní transkripce z audionahrávek

Verze

1.0

Druh výsledku

R – software

Lokalizace výsledku

https://folkprepis.jamap.cz/

Lokalizace dokumentace

https://jamap.cz/folkprepis/dokumentace.pdf

Vlastník

Vysoké učení technické v Brně – Ústav pro jazyk český AV ČR, v. v. i.

Stručný popis produktu

Rozpoznávač má na vstupu audionahrávku a na výstupu folklorní transkripci, např. *„dęž mně bęlo šesnást let, tož jsem šla slóžęt k Rozkopalum za chuvọ k jednemọ sedlákovi*" (ukázka pro centrální středomoravskou nářeční podskupinu); standardní ortografický přepis by u téže věty nezachytil nářeční znaky: *„když mně bylo šestnáct let, tož jsem šla sloužit k Rozkopalům za chůvu k jednomu sedlákovi"*. Vzhledem k blízkosti češtiny s dalšími západoslovanskými jazyky (polština, slovenština) a k vzájemným jazykovým vlivům v některých oblastech ČR je systém natrénován pro češtinu, polštinu a slovenštinu, ke kterým má VUT dostatek dat díky účasti v EU projektu SpeechDat-East. Uživatel jazyk zvolí před spuštěním přepisu. Systém je adaptován na jednotlivá nářečí na základě existujících a nově sbíraných nářečních dat. Software je upraven tak, aby byl použitelný odborníky z ÚJČ a dalších partnerských institucí, které o něj projevily zájem, v ČR i v zahraničí, stejně jako pracovníky médií i laiky.

Technické parametry produktu

Systém je algoritmicky založen na moderní architektuře umělé neuronové sítě. Vychází z vícejazyčného rozpoznávače řeči Parakeet-TDT 0.6B v3 (NVIDIA, 2025). Jde o model s kodérem typu FastConformer a dekodérem typu Token-and-Duration Transducer o velikosti přibližně 600 milionů parametrů. Výchozí model je předtrénován na řeči 25 evropských jazyků, mezi nimi na češtině, polštině a slovenštině. Vícejazyčné pokrytí těchto tří západoslovanských jazyků navazuje na řečová data, která má VUT k dispozici díky účasti v projektu EU SpeechDat-East. Česká větev byla dále adaptována na nářečí podle nahrávek

dialektologického oddělení Ústavu pro jazyk český AV ČR. Trénovací data navazují na sadu použitou u rozpoznávače dialektologické transkripce (přibližně 16,5 hodiny) a jsou rozšířena o nově přepsané úseky; celková stopáž je přibližně 29 hodin. Nezávislá testovací data o stopáži 1 hodina vykazují úspěšnost rozpoznání 74.4%. Jde o nahrávky různého stáří, pořizované sběry od 60. let 20. století po současnost, včetně nově sbíraného materiálu. Nahrávky byly segmentovány, čištěny a lingvisticky anotovány a jejich cílový přepis má podobu folklorní transkripce. Polština a slovenština se přepisují vícejazyčným modelem. Jazyk se neurčuje automaticky: uživatel jej zvolí ve webovém rozhraní. Rozhraní umožňuje vyzkoušet rozpoznávač v internetovém prohlížeči jako cloudovou službu.

Ekonomické parametry výsledku

Software urychluje pořizování folklorní transkripce, tedy té podoby zápisu, která se užívá v etnologii a folkloristice i běžně v literatuře, tištěných a online médiích nebo titulcích k audiovizuálnímu obsahu a která je bezprostředně srozumitelná i laickému uživateli. Veškeré přepisy musely být dosud prováděny buď manuálně, nebo prostřednictvím systémů určených pro rozpoznávání spisovné řeči, které generovaly nepřesné a zkreslené výstupy. Nyní lze odborný proces ruční transkripce pomocí našeho nástroje výrazně zautomatizovat a zefektivnit, případně je možné jím nahradit starší automatickou transkripci, a dosáhnout tak mnohem vyšší přesnosti. V současnosti disponuje dialektologické oddělení 3 080 nahrávkami s celkovou stopáží 2 235 hodin (ke dni 15. 10. 2026), přepisy byly dosud pořízeny pouze u 23 % záznamů z celkového počtu nahrávek, avšak většinově na úrovni reprezentativních ukázek o délce cca 5 minut. Ruční přepis hodinového záznamu řeči je v průměru pořizován 5 až několik desítek hodin (Šimečková a kol., 2025, s. 58), možnost automatického přepisu tak představuje výrazné zjednodušení v procesu zpracování nářečního audiálního materiálu. Demonstrátor zvládne vygenerovat automatický přepis minutové nahrávky na stávajícím serveru za pouhých 10 sekund.

Druh možnosti využití výsledku jiným subjektem

A – k využití výsledku jiným subjektem je vždy nutné nabytí licence

Požadavek na licenční poplatek

N – poskytovatel licence na výsledek nepožaduje licenční poplatek

Licence

Software slouží k interní aplikaci pro potřeby projektu a Ministerstva kultury ČR. Po domluvě je možné využití i třetími stranami, a to při dodržení zásady uvedení zdroje a zachování licence. Celé znění licence:

Autorské právo © [2026]
[Vysoké učení technické v Brně – Ústav pro jazyk český AV ČR, v. v. i.] Licencováno podle Apache License, verze 2.0 (dále „Licence“); tento soubor není možné použít jinak než v souladu s Licencí. Kopii Licence můžete získat na adrese: http://www.apache.org/licenses/LICENSE-2.0
Pokud to nevyžaduje platný zákon nebo není uvedeno jinak, je software distribuován podle Licence na bázi jak stojí a běží, bez jakýchkoliv záruk a podmínek, ať výslovných nebo předpokládaných. Podrobnosti o Licenci a jejích omezeních jsou k dispozici v samotné Licenci.

Pro uvedení zdroje doporučujeme formulaci:

Multilingvální rozpoznávač západoslovanských jazyků pro generování folklorní transkripce z audionahrávek. Verze 1.0. Vysoké učení technické v Brně – Ústav pro jazyk český AV ČR, v. v. i., 2026.

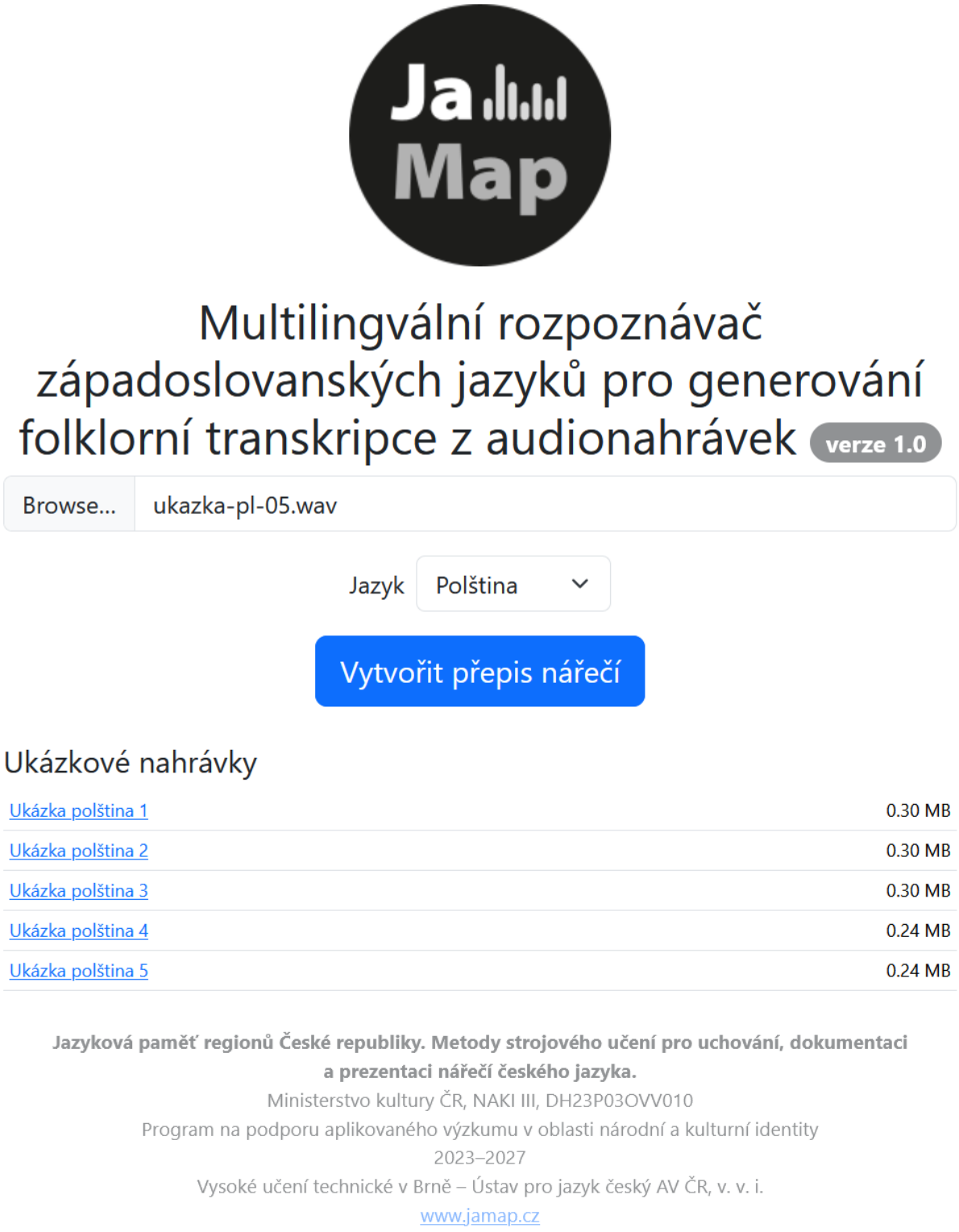
Anglická verze:
Multilingual West Slavic Speech Recognizer for Generating Folklore Transcription from Audio Recordings. Version 1.0. Brno University of Technology – Czech Language Institute, Czech Academy of Sciences, 2026.

Zdrojový kód

https://www.jamap.cz/folkprepis/zdrojovy-kod.tar.gz

Uživatelská příručka

Krok 1
Ovládání je intuitivní. Na stránce https://folkprepis.jamap.cz/ uživatel tlačítkem Vybrat soubor zvolí zvukový soubor WAV ze svého počítače. V nabídce Jazyk vybere Čeština, Polština nebo Slovenština podle jazyka mluvčího. Tlačítkem Vytvořit přepis nářečí přepis spustí. Po odeslání nahrávky se zobrazí sdělení „Probíhá zpracování souboru. Počkejte prosím…“. Po zpracování se zobrazí folklorní přepis prvních 60 vteřin zvukového souboru, segmentovaný na věty. Pod tímto přepisem je přehrávač s extrahovaným úsekem nahrávky, který si uživatel může přehrát pro kontrolu. Odkaz Vložit novou nahrávku vrátí uživatele na formulář. Chce-li uživatel software jen vyzkoušet, aplikace nabízí ukázkové nahrávky: dvacet českých, dvacet polských a dvacet slovenských. Zobrazený seznam se řídí zvoleným jazykem.
Poznámka: Vyšší úspěšnost přepisu je u záznamu jedné osoby mluvící tradičním nářečím. Směs několika mluvčích, překryv hlasů nebo projev blízký spisovnému jazyku snižují přesnost přepisu. Česká větev je připravena na nářeční materiál různého stáří, od historických sběrů po novější terénní nahrávky.



Obr. 1: Náhled rozhraní

Krok 2

Výstupem je automaticky pořízený text ve folklorní transkripci zvoleného jazyka. U češtiny přepis zachycuje nářeční podobu promluvy. Pro centrální středomoravskou nářeční podskupinu může mít například tuto podobu: *„dęž mně bęlo šesnást let, tož jsem šla slóžęt k Rozkopalum za chuvọ k jednemọ sedlákovi“* . V ortograficky standardní češtině by týž úsek zněl: *„když mně bylo šestnáct let, tož jsem šla sloužit k Rozkopalům za chůvu k jednomu sedlákovi“* . Striktně dialektologický přepis se zvláštní soustavou grafů poskytuje samostatný

rozpoznávač na adrese https://prepis.jamap.cz/. U polštiny a slovenštiny je výstupem přepis v příslušném jazyce.
Poznámka: Pro vyšší úspěšnost přepisu je nezbytné pracovat se záznamy tradičních nářečních promluv (nikoliv smíšených nebo nenářečních projevů), a to jedné osoby (nikoliv více osob), která byla v době pořízení nahrávky zástupcem starší generace, přičemž splňovala i další požadavky kladené na ideálního nářečního mluvčího (k tomu Šimečková, 2024). Česká větev pracuje s historickými i novějšími nářečními záznamy a není vázána na jeden typ nahrávací techniky.



Obr. 2: Náhled automaticky přepsaného textu

Charakteristiky výsledku

Multilingvální rozpoznávač západoslovanských jazyků pro generování folklorní transkripce z audionahrávek je další elektronický výstup z řady softwarů vyvíjených v projektu Jazyková paměť regionů České republiky. Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka. Ministerstvo kultury ČR, NAKI III, DH23P03OVV010. Software umožňuje u vstupní audionahrávky zachycující jakékoli teritoriální nářečí českého národního jazyka pořídit automatický přepis ve folklorní transkripci, a to v češtině, polštině nebo slovenštině podle volby uživatele.

Jedinečnost a novost softwaru

Jde o první veřejně dostupný automatický přepisovač regionálně rozrůzněné češtiny, a to jak po stránce architektury softwaru, tak trénovacími daty, na nichž byl nástroj vyvinut (viz technická dokumentace níže).

Naplnění očekávaných dopadů programu NAKI III a využití softwaru

Software usnadňuje práci s nářečními nahrávkami, umožňuje automaticky přepisovat nářeční zvukové projevy, a to v podobě standardizované folklorní transkripce. Je využitelný např. při digitalizaci fonoték a archivů mluveného slova, při zpracování záznamů z terénních výzkumů, též v běžné praxi lingvistů, etnologů, orálních historiků a dalších odborníků i při práci novinářů, tvorbě titulků nebo i laickém zpracování jakéhokoli mluveného nářečního projevu.
Jednoduché webové rozhraní, snadné ovládání a srozumitelná prezentace výsledků rozšiřují okruh možných uživatelů od odborné veřejnosti až po studenty, pedagogy či laické zájemce o jazyk. Software má také velký potenciál ve vzdělávání, protože umožňuje prezentovat živou podobu nářečí a její transformaci do psané podoby.
Těmito charakteristikami software vyhovuje důrazu, který je v programu NAKI III kladen na směřování výzkumu k výsledkům podporujícím rozvoj inteligentních specializací ve smyslu RIS3.
Odpovídá také následujícím bodům očekávaných dopadů programu NAKI III:
3) zpřístupnění kulturního dědictví široké obci zájemců a uživatelů,
4) posílení a integrace sociálně ekonomického uplatnění kulturního dědictví ve společnosti,
6) výzkum nástrojů a jejich ověření pro:

a) aktivní podíl na rozvíjení národní identity jako jednoho z elementů spoluvytváření evropské identity a kultury,

b) systematickou dokumentaci a prezentaci způsobu života a kultury menšin v minulosti i současnosti,

k) systematické a efektivní využívání prostředků technologií pro rozvoj národní kultury a kulturní identity obyvatelstva.

Technická dokumentace

Cíl výsledku

Hlavním cílem softwaru je automatizovaný přepis nestandardních řečových projevů (teritoriálních dialektů češtiny a vývojově příbuzných západoslovanských jazyků) ve folklorní transkripci. Cílem řešení je nahradit časově i odborně náročnou manuální transkripci a překonat zásadní limitaci stávajících ASR (Automatic Speech Recognition) systémů, které jsou trénovány výhradně na spisovném jazyce, čímž nářeční prvky nuceně normalizují nebo u nich vykazují vysokou chybovost (WER). Software dosahuje automatizovaného mapování akustického signálu nářeční mluvené řeči na foneticko-ortografické nářeční grafémy při zachování rychlé výpočetní odezvy a přístupnosti přes webové rozhraní.

Architektura

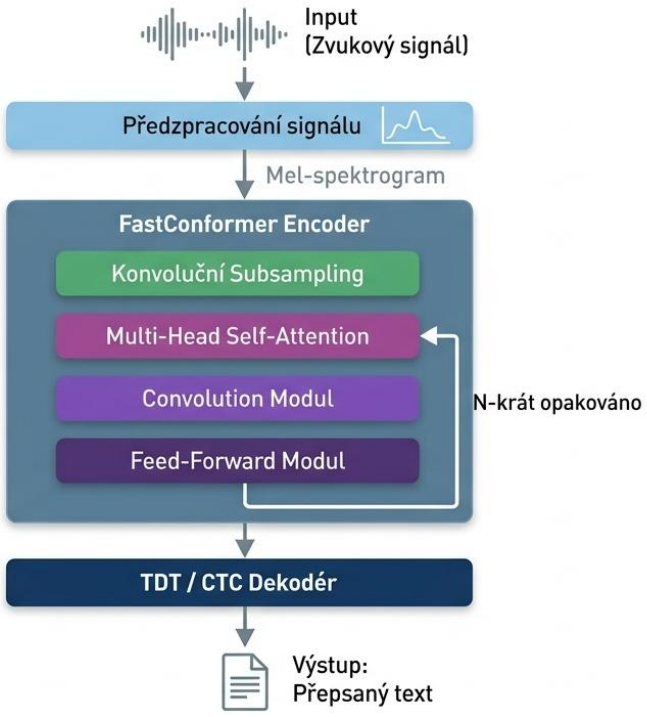
Architektura a její unikátnost

Architektura řešení je koncipována jako víceúrovňový end-to-end akusticko-jazykový model s navázaným mikroservisním cloudovým pipeline:
1. Akusticko-jazykové jádro (FastConformer + Token-and-Duration Transducer): Na rozdíl od konvenčních rozpoznávačů využívajících standardní architektury Transformer/Conformer a CTC či Attention mechanismy využívá náš systém architekturu FastConformer kombinovanou s dekodérem Token-and-Duration Transducer (TDT). TDT architektura umožňuje predikovat nejen tokeny, ale i jejich trvání v čase, což výrazně zrychluje inferenci a zvyšuje stabilitu rozpoznávání u nahrávek s kolísavým tempem či šumem na pozadí (typické pro historické nářeční nahrávky od 60. let 20. století).

2. Novost a unikátnost architektury:
- Jedinečnost úpravy dekodéru pro folklorní transkripci: Běžné modely mapují akustické featury na standardní ortografické tokeny (slovníkovou zásobu). V našem řešení byla výstupní tokenizační vrstva a dekodér přeadaptovány na speciální sadu nářečních fonetických znaků a morfemických variačních vzorů (folklorní transkripce).
- Cross-lingual transfer trénování: Unikátnost spočívá v paralelní integraci tří západoslovanských jazyků (čeština, polština, slovenština) s využitím přenosu znalostí (transfer learning) z vícejazyčného základu Parakeet-TDT 0.6B na vysoce specifická nářeční data (29 hodin anotovaných zvukových záznamů), čímž model získává odolnost vůči interdialektickým a pomezním jazykovým jevům.
Základ tvoří vícejazyčný model Parakeet-TDT 0.6B v3 (NVIDIA, 2025; viz též Rekesh a kol., 2023; Xu a kol., 2023). Kodér je FastConformer, dekodér je tzv. Token-and-Duration Transducer. Výchozí model je předtrénován na řeči 25 evropských jazyků včetně češtiny, polštiny a slovenštiny. Česká větev je adaptována na nářečí na datech o celkové stopáži přibližně 29 hodin s cílovými přepisy ve folklorní transkripci. Polština a slovenština používají vícejazyčný model; jazyk v demonstrátoru volí uživatel.

Výstup:
- Systém vrací čistý textový přepis.
- Pro účely výzkumu je možné doplnit i časové značky (timestamps).
- Přepis je optimalizován pro historické nahrávky, ale dokáže pracovat i s novými záznamy.

Architektura systému je zobrazena na obr. 3.



Obr. 3: Architektura NVidia Parakeet systému.

Trénovací data

Systém byl pro polštinu a slovenštinu trénován na řečových datech, která má VUT k dispozici díky účasti v projektu EU SpeechDat-East. Česká nářeční data byla poskytnuta dialektologickým oddělením ÚJČ AV ČR, v. v. i. Jedná se o unikátní data získávaná terénním výzkumem od 50. let 20. století do současnosti, která jsou uložena v Archivu zvukových

záznamů nářečních promluv a nyní převáděna do budované Databáze nářečních promluv pro odbornou veřejnost (k těmto zdrojům viz Šimečková, 2024).
Data sestávají z nářečních nahrávek a přepisů pořízenými dialektology v normalizované dialektologické transkripci. Pro adaptaci rozpoznávače na výstup ve folklorní transkripci bylo nutné tyto textové anotace převést do odpovídající cílové podoby. K tomuto účelu byl nejprve vytvořen paralelní textový korpus obsahující dvojice téhož textu v normalizovaném folklorním a dialektologickém přepisu. Jeho základem byly texty původně zapsané folklorní transkripcí, jejichž zápis byl sjednocen podle stanovených normalizačních pravidel a následně převeden do dialektologické transkripce pomocí soustav regulárních výrazů specifických pro jednotlivé dialekty (viz Stupňánek in: Šimečková et al. 2025, s. 63–171). Takto získané dvojice sloužily k učení s učitelem: vstupem modelu byla dialektologická podoba textu a cílovým výstupem její folklorní protějšek. Model se tedy učil převod v opačném směru, než v jakém byly paralelní texty vytvořeny, neboť tento postup nelze spolehlivě zajistit prostým obrácením uvedených pravidel. Natrénovaný model byl poté použit k automatické konverzi původních dialektologických anotací nahrávek z Databáze nářečních promluv pro odbornou veřejnost do normalizované folklorní transkripce při zachování jejich vazby na příslušné zvukové úseky. Výsledný soubor tvořený nahrávkami a takto odvozenými cílovými přepisy představoval vlastní trénovací data pro adaptaci české větve rozpoznávače řeči; celková stopáž využitého zvukového materiálu činila přibližně 29 hodin.

Odůvodnění naplnění definice software s prvky VaV (odlišení od negativní části definice software)

Výsledek nenaplňuje negativní vymezení software (např. běžné „přidání uživatelských funkcí do stávajících aplikačních programů (včetně funkčnosti základních vstupních dat)“) z následujících důvodů:
1. Absence existujícího řešení: Na trhu ani ve výzkumné sféře neexistuje volně dostupný ani komerční software schopný přímé automatické konverze nářečního akustického signálu do folklorní transkripce.
2. Řešení vědecko-technické nejistoty: Převod nářeční mluvené řeči nelze vyřešit pouhým rozšířením slovníku nebo přidáním funkcí do stávajících ASR programů (např. Whisper), neboť běžné ASR systémy dialektismy ignorují nebo degradují na nejbližší spisovná slova. Bylo nutné vyvinout nový postup učení s učitelem s využitím pravidlové konverze paralelních korpusů pro přípravu trénovacích dat a trénovat hlubokou neuronovou síť.
3. Nový algoritmus/metodika přípravy anotací: Vývoj vyžadoval vytvoření unikátního převodního můstku mezi normalizovanou dialektologickou transkripcí a folklorní transkripcí pomocí regulárních výrazů a reverzního strojového učení, což představuje nový metodologický a softwarový přínos, nikoli rutinní programování.
Vzhledem k uvedenému je vyvinutý software zcela jedinečný a první na světě.

Popis demonstrátoru

Automatický přepis nářečí je implementován jako nástroj napsaný v jazyce Python a ovládaný z příkazové řádky. Daná forma je tudíž primárně vhodná pro použití softwaru v rámci vědeckovýzkumné činnosti, nikoliv pro testování méně technicky zdatnými uživateli. Z toho důvodu bylo implementováno i webové rozhraní a „cloudové“ zapouzdření rozpoznávače. Zjednodušené schéma demonstrátoru zachycuje obrázek 4.
Demonstrátor se skládá z následujících komponent:

- Minio = datové úložiště
- RabbitMQ = obsluha fronty zpracovávaných úloh
- Celery = distribuovaný systém zpracování úloh
- Django = backend pro Celery; webový frontend

Výše uvedené komponenty běží jako služby zapouzdřené do Docker kontejnerů na dedikovaném serveru na Fakultě informačních technologií Vysokého učení technického

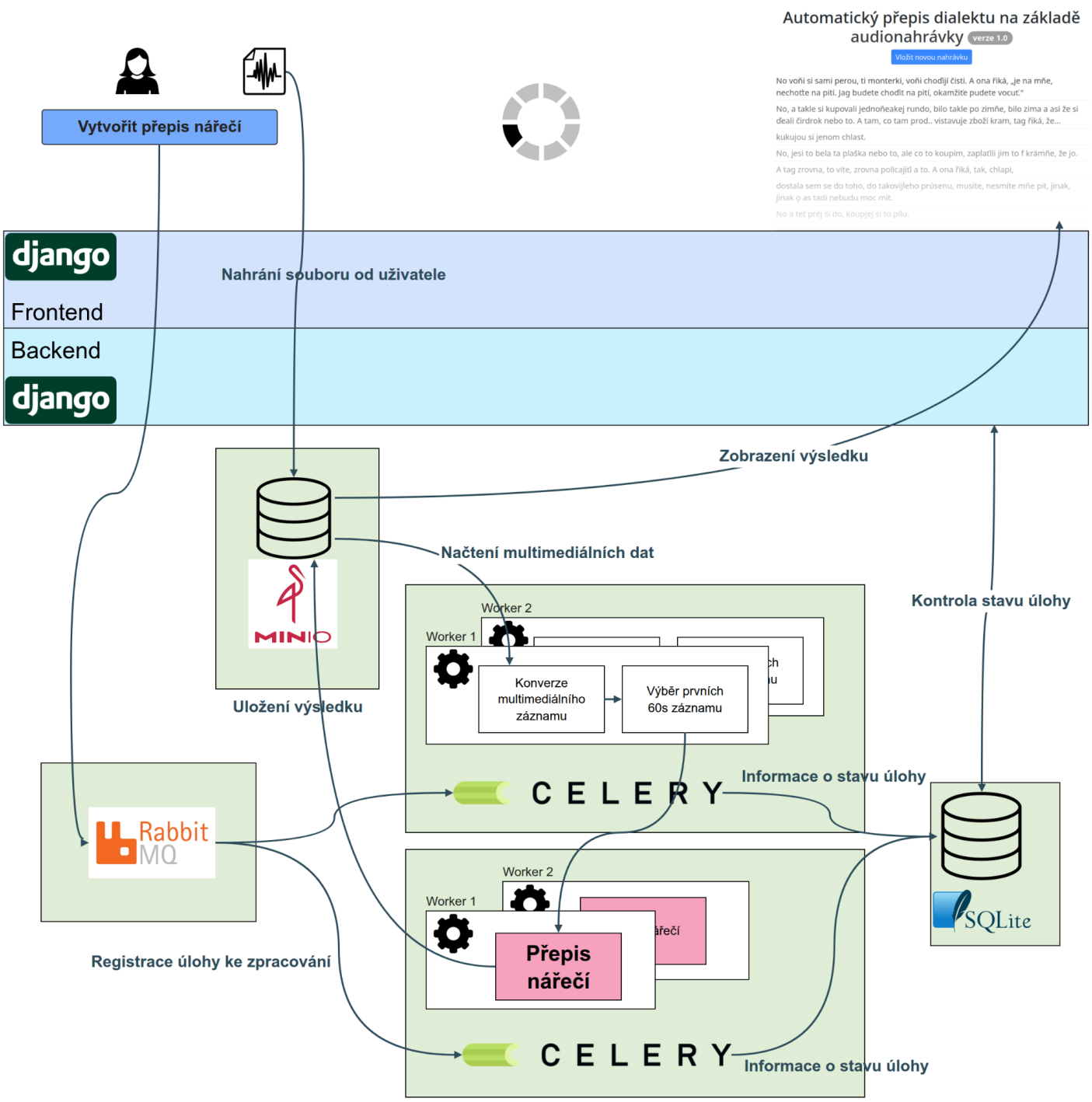
v Brně. Demonstrátor je přístupný z adresy https://folkprepis.jamap.cz/. Po vložení nahrávky uživatelem a stisku tlačítka Vytvořit přepis nářečí (viz obr. 1) je tato nahrávka odeslána a uložena do Minio úložiště. Dále je webovým frontendem vytvořena úloha Přepis nářečí z dané nahrávky. Uživateli se zobrazí rotující ikona, která znázorňuje, že se úloha zpracovává. Frontend se průběžně dotazuje backendu na výsledek zpracování.
Úloha je přes frontu RabbitMQ poslána ke zpracování. Pokud jsou k dispozici volné prostředky pro zpracování (worker), tak Celery přijme úlohu a je zahájeno zpracování. Demonstrátor zpracovává v jednu chvíli jednu nahrávku. Pokud uživatelé vloží v krátkém sledu více úloh, musí počkat, než se dostanou na řadu. Vzhledem k tomu, že je úloha automatického přepisu náročná na výpočetní zdroje, byla zapouzdřena do samostatného kontejneru. Tím lze tuto úlohu pouštět na fyzicky jiném hardwaru.
Úloha je rozdělena do 3 po sobě jdoucích kroků:

- Konverze vstupních dat na požadovaný formát PCM mono s16le 16kHz.
- Extrakce prvních 60 vteřin z nahrávky. K tomuto kroku se přistupuje z důvodu omezení paměťových nároků demonstrátoru rozpoznávače a také zkrácení čekací doby na výstup.
- Automatický přepis nářečí z 60vteřinové nahrávky a uložení výsledku do databáze backendu včetně indikace úspěšného dokončení úlohy.

V momentě, kdy frontend detekuje úspěšné dokončení úlohy, zobrazí výsledky uživateli (viz obr. 2).
Upozornění: úloha automatického přepisu nářečí vždy před zpracováním zvukové nahrávky načítá modely. Tento krok trvá cca 15 vteřin. Velmi krátká nahrávka se tedy zpracovává cca 20 vteřin, minutová pak cca 40 vteřin (pokud je prázdná fronta).

Implementace a konkrétní využití programovacích jazyků

Navrhovaný software využívá kombinaci programovacích jazyků a frameworků, přičemž novost a unikátnost spočívá v jejich konkrétním zapojení do výpočetního a konverzního řetězce:
- Python (Inference, strojové učení a backend): Python je využit jako primární jazyk pro běh neuronového modelu (s využitím PyTorch a NVIDIA NeMo frameworku), distribuované zpracování úloh (Celery) a backend webové aplikace (Django). V Pythonu byly implementovány vlastní skripty pro dynamickou transformaci a konverzi audiosignálu na PCM mono 16kHz s paměťově optimalizovaným streamingem pro 60s úseky, který zabraňuje přetížení GPU RAM. Byla vyvinuta vlastní obslužná rutina v Pythonu sbírající načítání váhových matic adaptovaného Parakeet-TDT modelu, která provádí přímou dekódovací konverzi výstupu neuronové sítě na textové řetězce ve folklorní transkripci.
- JavaScript / HTML5 (Webový frontend a audiovizualizace): JavaScript (ES6+) na frontendu zajišťuje asynchronní komunikaci s REST API backendu (AJAX/Fetch polling), řízení uživatelského rozhraní a přehrávání zvuku. Frontendové skripty v JS zajišťují reálné propojení vygenerovaného textu ve folklorní transkripci s časovými úseky přehrávače audio záznamu, což umožňuje uživateli okamžitou vizuální a akustickou kontrolu specifických dialektických znaků (např. *dęž, bęlo*).
- Shell/Dockerfile/YAML (Orchestrace a kontejnery): Příkazové skripty a konfigurace pro izolovaný chod služeb (Minio, Celery, RabbitMQ, Django). Izolované zapouzdření náročného AI modelu do samostatného GPU worker kontejneru umožňuje škálovatelné nasazení v cloudovém prostředí akademické infrastruktury bez nutnosti modifikace klientských stanic.



Obr. 4: Zjednodušené schéma demonstrátoru přepisu nářečí. Komponenta automatického přepisu nářečí je zvýrazněna červeně. Zelené boxy indikují použití Docker kontejneru

Literatura

ŠIMEČKOVÁ, M. – STUPŇÁNEK, B. – KARAFIÁT, M. – VOŽENÍLEK, V. – VONDRÁKOVÁ, A. – NÉTEK, R. (2025). *Metodika pro převod strukturovaných znalostí z oboru dialektologie do strojového učení*. Brno – Olomouc: Ústav pro jazyk český Akademie věd České republiky – Vysoké učení technické v Brně – Univerzita Palackého v Olomouci, 2025. Dostupné z: https://www.jamap.cz/metodika.

ŠIMEČKOVÁ, M. (2024). *Archiv zvukových záznamů nářečních promluv*. Praha: Academia. Též online. Dostupné z: http://www.vedakolemnas.cz/sys/galerie-download/VKN-132_.pdf.

NVIDIA 2025. *NVIDIA Parakeet-TDT 0.6b-v3* [počítačový program]. Hugging Face, 2025 [cit. 2026-10-14]. Dostupné z: https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3

REKESH, D. et al. (2023). FastConformer: Faster Conformer Architecture for Automatic Speech Recognition. arXiv preprint arXiv:2305.05084.

XU, Y. et al. (2023). Token-and-Duration Transducer for Speech Recognition. arXiv preprint arXiv:2304.06836.