METODIKA PRO PŘEVOD STRUKTUROVANÝCH ZNALOSTÍ Z OBORU DIALEKTOLOGIE DO STROJOVÉHO UČENÍ

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

Program na podporu aplikovaného výzkumu
v oblasti národní a kulturní identity 2023–2027

JaMap

Ústav pro jazyk český
Akademie věd České republiky

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

KATEDRA GEOINFORMATIKY
Univerzita Palackého v Olomouci

METODIKA PRO PŘEVOD STRUKTUROVANÝCH ZNALOSTÍ Z OBORU DIALEKTOLOGIE DO STROJOVÉHO UČENÍ

Marta Šimečková (ed.)

Ústav pro jazyk český Akademie věd České republiky, v. v. i.
Vysoké učení technické v Brně
Univerzita Palackého v Olomouci

Brno 2025

Metodika pro převod strukturovaných znalostí z oboru dialektologie do strojového učení

Editorka: Marta Šimečková

Autorský kolektiv: Mgr. Marta Šimečková, Ph.D.[1]
Mgr. Bronislav Stupňánek, Ph.D.[1]
Ing. Martin Karafiát, Ph.D.[2]
prof. RNDr. Vít Voženílek, CSc.[3]
RNDr. Alena Vondráková, Ph.D., LL.M.[3]
RNDr. Rostislav Nétek, Ph.D.[3]

Redakčně zpracovaly: Mgr. Marta Šimečková, Ph.D.[1], RNDr. Alena Vondráková, Ph.D., LL.M.[3]

Pracoviště: [1] Dialektologické oddělení, Ústav pro jazyk český Akademie věd České republiky, v. v. i.
[2] Ústav počítačové grafiky a multimédií, Fakulta informačních technologií, Vysoké učení technické v Brně
[3] Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Odborní konzultanti: prof. Dr. Ing. Jan Černocký
Mgr. Filip Kubeček
Ing. Oldřich Plchot, Ph.D.
Ing. Igor Szőke, Ph.D.
PhDr. Milena Šipková, CSc.
PhDr. Veronika Štěpánová, Ph.D.

Odborní oponenti metodiky: Mgr. Eva Chodějovská, Ph.D., Moravská zemská knihovna
PhDr. Marie Kopřivová, Ph.D., Ústav bohemistiky pro cizince a komunikace neslyšících, Filozofická fakulta, Univerzita Karlova

Korektura: Mgr. Petra Přadková

Technická redakce a design: RNDr. Alena Vondráková, Ph.D., LL.M.

Metodika je výstupem projektu č. DH23P03OVV010 *Jazyková paměť regionů České republiky. Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka*; poskytovatel podpory Ministerstvo kultury, Program NAKI III.

www.jamap.cz

OBSAH

1 Úvod / 8

2 Cíle metodiky / 11

3 Audiální data: sběr, archivace, katalogizace a příprava pro strojové učení / Marta Šimečková / 15

3.0 **Úvod** / 16
3.1 **Dialektologické minimum: pojmy, termíny, struktury** / 17
3.2 **Cesty ke zdrojům audiálních dat** / 21
3.2.1 Terénní výzkum a metoda rozhovoru / 21
3.2.1.1 Typologie rozhovorů/interview a role jejich aktérů / 21
3.2.1.2 Strategie pro přepnutí jazykového kódu mluvčího směrem k běžněmluvenostnímu vyjadřování / 23
3.2.1.3 Fáze rozhovoru a strategie pro řešení různých situací / 33
3.2.2 Zprostředkovaný sběr audiálních dat cestou citizen science / 42
3.2.3 Zapojení již existujících sbírek do výzkumu / 43
3.2.4 Metody a techniky nahrávání / 44
3.2.5 Právní a etický rámec pořizování, archivace a zveřejňování zvukových nahrávek, uchovávání osobních dat a problematika anonymizace / 45
3.2.5.1 Pořizování, archivace a zveřejňování nahrávek z hlediska legislativy a informovaný souhlas / 45
3.2.5.2 Osobní a citlivé údaje a jejich ochrana / 48
3.2.5.3 Pořizování, archivace a zveřejňování nahrávek z hlediska etických principů / 51
3.3 **Budování zvukového archivu** / 53
3.4 **Transkripty audiálních dat: pravidla transkripcí a příprava dat pro strojové učení** / 56
3.4.1 Typy transkripčních soustav / 56
3.4.2 Dialektologická transkripce: principy a základní poučení / 58
3.5 **Shrnutí** / 62

4 Textová data: výběr, digitalizace, normalizace, analýza a převod textů a jejich příprava pro strojové učení / Bronislav Stupňánek / 64

4.0 **Úvod** / 64
4.1 **Výběr textů** / 67
4.1.0 Úvod / 67
4.1.1 Základní účely existujících písemných záznamů nářečí / 67

4.1.2 Typologie nářečních textů / 68
4.1.2.1 Podle transkripce / 68
4.1.2.2 Podle média / 71
4.1.2.3 Podle původu autora / 72
4.1.2.4 Podle odbornosti autora / 72
4.1.2.5 Podle žánru / 73
4.1.3 Shrnutí / 78
4.2 **Digitalizace (OCR) nářečních textů** / 78
4.2.0 Úvod / 78
4.2.1 Software pro OCR / 79
4.2.2 Snímání digitálních obrazů stránek (skenování, focení) / 79
4.2.2.1 Vhodné vlastnosti obrazů stránek pro OCR / 79
4.2.2.2 Skenery / 80
4.2.2.2.1 Plošné (flatbed) skenery / 80
4.2.2.2.2 Skenery se snímáním shora / 81
4.2.2.2.3 Ruční skenery / 81
4.2.2.2.4 Průtažné skenery / 81
4.2.2.3 Fotoaparáty / 82
4.2.2.4 Shrnutí / 82
4.2.3 OCR nářečního textu / 82
4.2.3.1 OCR bez nářečního slovníku / 82
4.2.3.2 Před samotným OCR / 83
4.2.3.3 Zásady OCR nářečního textu / 83
4.3 **Čištění a formální sjednocení textu** / 85
4.3.1 Čištění textu / 85
4.3.1.1 Identifikátory při čištění textu / 85
4.3.1.2 Odstraňované prvky textu / 86
4.3.1.3 Extrahované prvky textu / 87
4.3.2 Formální sjednocení textu / 87
4.4 **Normalizace folklorního textu** / 92
4.4.0 Úvod / 92
4.4.1 Problém sjednocení folklorních textů / 92
4.4.2 Obojetnosti / 92
4.4.3 Normalizovaný folklorní text / 93
4.4.4 Normalizace folklorního textu podle nářečních podskupin / 94
4.5 **Normalizace dialektologického přepisu** / 108
4.5.0 Úvod / 108
4.5.1 Problém sjednocení dialektologických textů / 108
4.5.2 Normalizované náhrady za znaky dialektologického přepisu / 108
4.5.3 Přehled nářečních znaků normalizované dialektologické transkripce / 114
4.5.4 Přehled znaků normalizované dialektologické transkripce podle nářečí / 116
4.6 **Převod folklorního přepisu na dialektologický přepis** / 123
4.6.0 Úvod / 123
4.6.1 Číslice / 124
4.6.2 Zkratky / 125

4.6.2.1 Typy zkratek / 125
4.6.2.2 Zpracování vokalizovaných zkratek majuskulemi / 127
4.6.2.3 Odstranění nezkratkových textů psaných verzálkami / 129
4.6.2.4 Převod zkratek / 130
4.6.3 Slova cizího původu / 133
4.6.3.1 Citáty z jiných jazyků / 133
4.6.3.2 Citátové výrazy / 133
4.6.3.3 Výrazy graficky neadaptované / 134
4.6.3.4 Německá příjmení / 135
4.6.3.5 Cizí slova částečně graficky adaptovaná / 144
4.6.3.6 Izolované znaky / 149
4.6.4 „Di“, „ti“, „ni“, „dě“, „tě“, „ně“ > *ďi*, *ťi*, *ňi*, *ďe*, *ťe*, *ňe* / 149
4.6.5 Znaky „x“ a „ch” / 150
4.6.6 Příprava před asimilacemi u jedinečných souhlásek, *v* a *h* a před vokály / 151
4.6.7 Mezislovní asimilace u jedinečných souhlásek, *v* a *h* a před vokály / 154
4.6.7.1 Mezislovní asimilace před jedinečnými / 154
4.6.7.2 Mezislovní asimilace před *v* / 154
4.6.7.3 Mezislovní asimilace u *h* / 155
4.6.7.4 Mezislovní (a předložkové) asimilace před vokálem / 156
4.6.8 Předložkové asimilace / 156
4.6.8.1 Předložkové asimilace před jedinečnými / 156
4.6.8.2 Předložkové asimilace před *v* / 157
4.6.8.3 Předložkové asimilace u *h* / 157
4.6.8.4 Předložkové asimilace před vokálem / 158
4.6.8.5 Odstranění zbytků po fixaci předložkových asimilací / 159
4.6.9 Příprava na asimilace znělosti před znělými a neznělými / 159
4.6.9.1 Znělé souhlásky před pauzou / 159
4.6.9.2 Neznělá před *h* v rámci slova / 160
4.6.9.3 Umožnění asimilace přes závorku obojetností / 160
4.6.10 Asimilace znělosti před znělými a neznělými souhláskami / 161
4.6.10.1 První sada regulárních výrazů / 162
4.6.10.2 Druhá až osmá sada regulárních výrazů / 163
4.6.10.3 Oprava nesoustavností v obojetnostech / 163
4.6.11 Nářeční změny / 164
4.6.11.1 Měkké retnice, jotace a *ň* po retnicích / 164
4.6.11.2 Geminace souhlásek / 165
4.6.11.3 Dvojí „i“, „y” / 166
4.6.11.4 Skupiny *čy*, *žy*, *šy*, *řy*, *cy*, *zy*, *sy* / 166
4.6.11.5 Progresivní asimilace *v* ve skupinách *kf*, *tf*, *sf*, *šf*, *chf* / 167
4.6.11.6 Typ „se sestrou” / 168
4.6.11.7 Měkkosti u sykavek a polosykavek / 168
4.6.11.8 Měkkosti u velár *k*, *g* / 169
4.6.12 Závěrečné změny / 169
4.6.12.1 Výslovnost skupiny *t* + neznělá sykavka, polosykavka / 169
4.6.12.2 Hiátové *j* / 170

4.6.12.3 Kroužkované „ů" / 170
4.6.12.4 Opětné zavedení digrafu „ch" / 171
4.7 **Shrnutí** / 171

5 Konverze textových dat pomocí strojového učení / Martin Karafiát / 172

5.0 **Úvod** / 173
5.1 **Attention mechanismus** / 173
5.1.1 Rozšíření a vlastnosti attention mechanismu / 175
5.2 **Self-attention mechanismus a transformer modely** / 177
5.3 **Předtrénování transformer modelů** / 178
5.4 **Translingvální modely (XLM)** / 179
5.5 **Trénování cross-lingual jazykového modelu na dialektologických datech** / 180
5.5.1 Předtrénování / 180
5.5.2 Doladění modelu / 181
5.6 **Shrnutí** / 183

6 Strojový přepis mluvené řeči do textu / Martin Karafiát / 184

6.0 **Úvod** / 185
6.1 **Tradiční přístup** / 185
6.1.1 Využití neuronových sítí v hybridním přístupu / 186
6.1.2 Výhody hybridního přístupu / 186
6.2 **End-to-End přístup** / 187
6.3 **Příprava dialektologických dat pro trénování přepisu řeči** / 188
6.4 **Shrnutí** / 189

7 Zpracování prostorových dat / Vít Voženílek, Alena Vondráková, Rostislav Nétek / 190

7.0 **Úvod** / 191
7.1 **Geolokalizace audiálních dat** / 191
7.1.1 Geolokace existujících nahrávek na území České republiky / 192
7.1.2 Geolokace nových nahrávek na území České republiky / 193
7.2 **Příprava pro interaktivní geovizualizaci** / 195
7.2.1 Datové formáty prostorových dat / 195
7.2.2 Prostorová dialektologická data / 199
7.2.3 Metody vizualizace prostorových dat / 202
7.2.4 Multimediální interaktivní vizualizace nářečních nahrávek / 204
7.2.5 Požadavky na tvorbu multimediální interaktivní nářeční mapy / 206
7.3 **Shrnutí** / 211

8 Srovnání novosti postupů / 212

9 Uplatnění metodiky v praxi / 216

10 Pro koho je metodika určena / 220

11 Seznam použité literatury / 221

12 Seznam publikací, které předcházely metodice a byly publikovány / 235

PŘÍLOHY / 237

Příloha 1
Struktura obsahových metadat v *Databázi nářečních promluv pro odbornou veřejnost* / 238

Příloha 2
Leták s výzvou k nářečnímu výzkumu na jihovýchodní Moravě / 249

Příloha 3
Informovaný souhlas k pořízení zvukového záznamu, jeho archivaci a nakládání a ke zpracování a zpřístupnění osobních údajů / 250

Úvod

1

ÚVOD

V rámci lingvistiky je jako samostatná disciplína vydělována **dialektologie**, kterou lze chápat jako **systematické studium teritoriálních dialektů určitého národního jazyka**, též studium jejich transformací do interdialektů, hyperdialektů a dalších variant v důsledku rozrušování starších nářečních kontinuí. V České republice se soustavným výzkumem této jazykové vrstvy (resp. vrstev) zabývají odborníci z dialektologického oddělení Ústavu pro jazyk český AV ČR, v. v. i., a to od 50. let 20. století. Jejich snahou je zejména **dokumentace dialektů cestou pořizování audiálních i textových záznamů**. Součástí sbírek jmenovaného pracoviště je mj. **Archiv zvukových záznamů nářečních promluv**, který představuje vůbec největší fond nahrávek souvislých projevů prezentujících různá nářečí českého jazyka z různých období. Dané nahrávky mají v dialektologii různé využití, slouží mj. k excerpci slovníkových exemplifikací pro vznikající celoúzemní nářeční lexikon. Práce s tímto ojedinělým zdrojem je však komplikována skutečností, že předpokladem pro **využití audionahrávek** je pořízení jejich textového přepisu, a to navíc dialektologickou (někdy i folklorní) transkripcí. Jde o záznam zohledňující fonetickou stránku řeči, přičemž se v něm užívá řada specifických znaků a postupů, které jsou tradičnímu pravopisnému systému cizí. Pořizování takovýchto přepisů je časově náročnou činností, navíc ne všechny transkripty, kterými oddělení již disponuje, vyhovují novým normám transkripčních zásad, a tak jsou nutné jejich revize, konverze i opravy. Je zřejmé, že cestou pro optimalizaci přístupu a zpracování těchto dat jsou **algoritmy strojového učení** (angl. machine learning, ML) spadající pod populární termín umělá inteligence (angl. artificial intelligence, AI).

V návaznosti na současný rozvoj technologií v oblasti automatického rozpoznávání jazyka byla nastolena otázka, zda by nebylo možné vyvinout **nástroje umožňující dialektologům efektivnější práci s nahrávkami**. Na vývoj systémů pro přepis řeči se dlouhodobě specializují odborníci z Vysokého učení technického v Brně, kteří zaštítili projekt *Jazyková paměť regionů České republiky. Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka* (č. DH23P03OVV010), řešený v rámci Programu na podporu aplikovaného výzkumu v oblasti národní a kulturní identity na léta 2023 až 2027 (NAKI III). Ve spolupráci s dialektology z Ústavu pro jazyk český AV ČR, v. v. i., a s geoinformatiky z Univerzity Palackého v Olomouci nyní vyvíjejí dva softwary, a to *Rozpoznávač řeči adaptovaný pro generování dialektologické transkripce z audionahrávek* a *Multilingvální rozpoznávač západoslovanských jazyků pro generování folklorní transkripce z audionahrávek*. Tyto nástroje jsou založeny na budované *Databázi nářečních promluv pro odbornou veřejnost* a *Databázi nářečních textů*, které jsou zdrojem dat audiálních a čistě textových.

Samotnému trénování a testování rozpoznávačů však musela nutně předcházet konsolidace existujících dat, jejich strukturace, standardizace a vůbec sestavení pravidel pro práci s nimi při využití metod strojového učení. Způsob, jakým toho bylo v rámci zmíněného projektu dosaženo, jakož i výsledky algoritmů strojového učení aplikované na dialektologii, jsou shrnuty v předložené metodice, která je i jedním z projektových výstupů. Metodika představuje **manuál k přípravě specifického jazykového, v tomto případě nářečního materiálu, audiálního i textového, která je nutná pro pozdější využití dat při trénování modelů strojového učení**. Metodika obsahuje jednak teoretické části, které přibližují možnosti ve sběru a zpracování nářečního materiálu a současné technologie strojového učení, jednak části aplikační, v nichž jsou blíže popsány jednotlivé kroky pracovního postupu a nabídnuty způsoby technického řešení. Předestřená doporučení, která jsou nyní aplikována při tvorbě výše jmenovaných rozpoznávačů, se přitom

mohou stát obecným návodem i pro další výzkumné týmy a inspirací pro české i zahraniční odborníky z řad specialistů na strojové učení a dialektologii, potažmo i z řad lingvistů pracujících se souvislou přirozenou řečí. Metodika rovněž předestírá způsoby **vizualizace dialektologického materiálu metodami tematické kartografie**, které uživatele metodiky a dialektologických dat dovedou k vytvoření tematických map, případně multimediálních interaktivních map či webových nářečních atlasů. Kromě praktických cílů spočívajících v transferu informací si metodika klade za cíl též poukázat na potřebu zapojit do humanitních věd interdisciplinární přístup, vedoucí jednak k inženýrskému rozvoji metod umělé inteligence, jednak k vytvoření nástrojů pro zefektivnění práce na straně věd humanitních.

Cíle metodiky

2

CÍLE METODIKY

Pokroky v oblasti umělé inteligence (angl. artificial intelligence, dále AI) přinášejí možnosti rozvoje nejen v přírodních vědách, ale i ve vědách humanitních. Delší dobu jsou techniky AI užívány v oblasti aplikované lingvistiky, třeba v rámci počítačového zpracování přirozeného jazyka jsou její specifické postupy uplatňovány pro analýzu či generování mluvených i psaných textů, jsou vyvíjeny softwary pro automatické překlady a korektury textů, chatboty určené k mluvené interakci s uživateli ap. Podoblastí AI je strojové učení, které pomocí sad algoritmů určených k trénování umělých neuronových sítí umožňuje **automatické rozpoznávání řeči** (angl. automatic speech recognition, též speech to text translation, ASR/STT), čímž se rozumí nejen detekce konkrétního jazyka, ale i jeho převod z mluvené podoby do podoby psané.

V této oblasti jsou nejpokročilejší rozpoznávače optimalizované na jazyky s velkým množstvím trénovacích dat (akustických i textových), a tedy i s velkou variabilitou mluvčích. Jedná se zejména o angličtinu, ale výborných výsledků dosahují třeba i rozpoznávače mandarínštiny a moderní standardní arabštiny. Oproti tomu u menších jazyků je dat méně, k tomu přistupují některé další problémy spojené s jejich specifiky – v případě češtiny je to zejména její flektivní charakter (lze předpokládat, že v mentálním lexikonu mluvčích jsou slovní jednotky uloženy v abstraktní formě společně s pravidly o jejich hláskové či tvarové variaci, např. slovo *kniha* nabývá v singuláru tvarů *knihy*, *knize*, *knihu*, *kniho*, *knihou*, což je nutné v rozpoznávačích zohlednit). V českém prostředí jsou automatické rozpoznávače řeči vyvíjeny v Brně, Praze, Plzni a Liberci. Jazykové modely byly donedávna optimalizovány na češtinu spisovnou, popř. na její interdialektickou obecněčeskou variantu. **K územním variantám se dosud nepřihlíželo**, a to ze zjevných důvodů – jednak jde o data obecně nedostatková, jednak značně rozrůzněná lexikálně, foneticky, morfologicky i syntakticky, mimo to i graficky v případě textových dat, což může být bariérou pro jejich zapojení do strojového učení.

Cílem metodiky je předložit podrobný návod k **vytvoření a přípravě nářečních dat** pro strojové učení a také ke geoinformatickému zpracování, též **nastínit možnosti jejich konkrétní aplikace** s využitím metod zmíněných přírodovědných oborů (viz obrázek 2.1), a to ve čtyřech krocích:

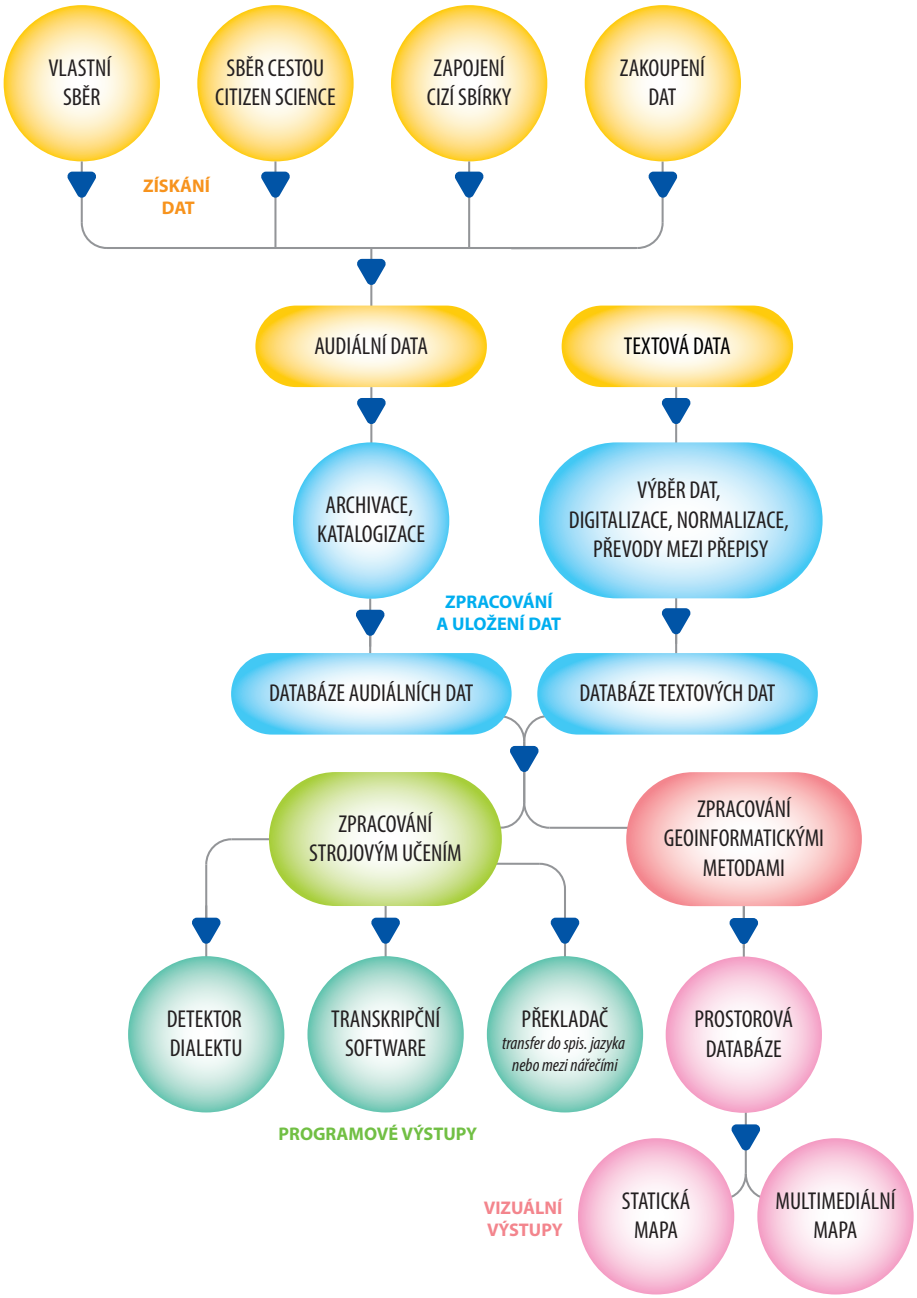
1. **krok**: stanovení zdroje nářečních dat a vybudování datové základny (kapitoly 3, 4);
2. **krok**: zpracování dat jednotným způsobem (kapitoly 3, 4);
3. **krok**: konverze textových dat pomocí strojového učení a strojový přepis mluvené řeči do textu (kapitoly 5, 6);
4. **krok**: geolokalizace dat a jejich interaktivní geovizualizace (kapitola 7).

Stěžejním bodem je **stanovení zdroje dat**, což lze řešit:

- vyhledáním již existujícího zdroje, v němž by byla obsažena data audiálního, textového, či obojího charakteru;
- sběrem zcela nových dat, a to cestou exploračních i excerpčních metod;
- kombinací obou výše zmíněných metod, kdy vstupní data již existují a jsou postupně doplňována daty novými.

Jazyková data je přitom nutno **zpracovat jednotným způsobem**. To obnáší:

- představení postupů pro převod dat z existujících zdrojů a z nových sběrů na formy optimální prostřednictvím sekvence opakovatelných kroků;
- stanovení optimální formy a formátu dat (audiálních, textových), mj. sestavení soustav znaků pro přepis diferenčních nářečních jevů a způsobu jejich aplikace v závislosti na původní formě dat i na cíli jejich dalšího zpracování, stanovení způsobu jednoznačné identifikace dat (nahrávek, transkriptů, textů vůbec) v rámci sady dat;
- revizi existujících dat vedoucí k jejich formální unifikaci, včetně sestavení jejich databáze umožňující jednodušší transfer ke strojovému/geoinformatickému zpracování;
- uplatnění týchž pravidel na data nově přibývající.



Obrázek 2.1 Transfer nářečních dat (audiálních, textových) a jejich zpracování metodami strojového učení a geoinformatiky

CÍLE METODIKY

Metodika nabízí **vodítko pro jednotlivé body, a to názorným a srozumitelným způsobem** – s vědomím, že předložená doporučení mohou být uplatňována vědci z různých vědních disciplín, přírodních, humanitních, formálních i aplikovaných. Z toho důvodu je její součástí vysvětlení základních pojmů, výklad rovněž zahrnuje množství praktických příkladů, modelových situací i sekvencí vzorových algoritmizovaných kroků, které lze na data přímo uplatnit. Představeny jsou i konkrétní výsledky testování metodiky na připravovaných softwarech vyvíjených pro generování dialektologické a folklorní transkripce z audionahrávek a rozpoznávání dialektů.

Metodiku je možno číst jednak jako souvislý text, jednak po jednotlivých kapitolách (a sekcích), které nabízí množství inovativních postupů pro různé obory. **Zájemci o dialektologii** z řad odborníků, studentů i laiků se budou moci například dozvědět, jak postupovat při:

- sběru nářečních nahrávek v terénu (včetně etických aspektů výzkumu);
- katalogizaci nahrávek a budování elektronického archivu/katalogu;
- zapojení veřejnosti do rozšiřování archivů cestou citizen science;
- pořizování přepisů nahrávek formou dialektologické/folklorní transkripce;
- výběru a prioritizaci nářečních textů pro jejich počítačové zpracování s ohledem na jeho rychlost a kvalitu výsledku;
- digitalizaci a optickém rozpoznání nářečních textů, aby byly co nejpřesnější a současně efektivní;
- čištění a formálním sjednocení digitálních textů;
- sjednocování různorodých zápisů ve folklorní i dialektologické transkripci;
- algoritmickém převodu folklorní transkripce do transkripce dialektologické.

A naopak zájemci o **strojové učení** naleznou odpovědi na to, jak využít dostupných dat v maximální míře, konkrétně při tvorbě jazykového modelu z textových dat (převážně folklorní transkripce) pro přepisovač řeči v dialektologické formě. V metodice jsou nejen definována komplexní pravidla pro převod z folklorní transkripce do transkripce dialektologické, ale jsou tu popsány i algoritmy pro reverzní konverzi. Shrnuje a porovnává dva hlavní přístupy:

- statistický přístup: na základě vstupních trénovacích dat jsou spočítány pravděpodobnostní statistiky pro kombinaci „vstupní/výstupní grafém : kontext“. Ty jsou pak aplikovány na neznámých datech a vybrána je nejpravděpodobnější sekvence;
- neurální přístup: využívá neurální modely pro strojový překlad. Ty umožňují načíst celou sekvenci znaků v jednom jazyce/formátu a generují sekvenci v jazyce/formátu jiném.

Z **geoinformatického pohledu** je cílem metodiky poskytnout jednak návod ke geolokalizaci audiálních dat, jednak návod pro výběr vhodného formátu dialektologických dat k jejich možné interaktivní geovizualizaci. Obsahem těchto návodů jsou postupy:

- přiřazení prostorové složky nářečním nahrávkám (tj. vytvoření prostorových dialektologických dat) pomocí číselníku částí obcí s jednoznačnými identifikátory;
- výběru vhodného datového formátu prostorových dialektologických dat, včetně jeho struktury, s doporučením pro vizualizaci nářečních nahrávek v multimediálních interaktivních mapách.

Některá doporučení jsou vázána na existenci specifického materiálu či zařízení, většina z nich je však univerzální a v případě přenosu předložených znalostí lze dojít k obdobným výsledkům.

3

Audiální data: sběr, archivace, katalogizace a příprava pro strojové učení

AUDIÁLNÍ DATA: SBĚR, ARCHIVACE, KATALOGIZACE A PŘÍPRAVA PRO STROJOVÉ UČENÍ

3.0 Úvod

Technologie automatického rozpoznávání řeči z nahrávek, včetně automatického převodu audiálních dat do podoby textové, se neobejde bez příslušných řečových dat. V ideálním případě jsou modely ASR trénovány na co **nejobjemnějších datových sadách**. Badatel v oboru ASR z toho důvodu stojí na začátku výzkumu před zásadním rozhodnutím, a to jakým způsobem potřebná data získat. Jeho volba je samozřejmě závislá na charakteru dat, zejména na tom, který jazyk konkrétně a která jeho varieta budou ASR podrobovány. Zatímco u větších jazyků (např. u angličtiny, čínštiny, hindštiny) existuje řada potenciálních zdrojů, u jazyků menších (mj. u češtiny) je cesta k datům problematičtější. Audiální data (i textová) jsou navíc často **omezena na spisovnou formu jazyka**, nikoliv na jeho další strukturní útvary a poloútvary.

Soustředíme-li se na **českojazyčná data, nadto rozrůzněná teritoriálně**, přece jen několik možností nabídnout můžeme. Lze k nim teoreticky dojít:

- nákupem;[1]
- vlastním sběrem;
- zprostředkovaným sběrem, třeba cestou citizen science;
- spoluprací s institucemi nebo jednotlivci vlastnícími již existující sbírky.

První bod odpadá – na trhu v současnosti neevidujeme žádnou sbírku českojazyčných regionálně rozrůzněných dat. Výzkumníkovi tak nezbude než buď spojit síly s vlastníky nějakých již existujících archivů (třeba z jiné instituce, a to i nad rámec přírodovědné oblasti), nebo si vytvořit sbírku vlastní. A právě **cesty k audiálním datům** jsou předmětem této kapitoly, která je určena nejen odborníkům na strojové učení, ale také dalším badatelům, kteří prostřednictvím techniky rozhovoru (interview) získávají a následně i zpracovávají zvuková data vyznačující se spontaneitou jazykového projevu.

V první podkapitole (3.1) jsou objasněny základní dialektologické pojmy, s nimiž je v kapitole a také dále v metodice pracováno. Dále je vysvětleno dělení nářečí českého jazyka, v případě ASR nutné pro geografické uchopení dat. V následující podkapitole (3.2) jsou již popsány způsoby nabytí dat, a to vlastním terénním výzkumem (3.2.1), zapojením veřejnosti (3.2.2) a také použitím již vybudovaných datasetů (3.2.3). U jednotlivých metod jsou uvedeny klady i zápory, přičemž největší pozornost je věnována samosběru. V souvislosti s ním jsou zmíněny různé typy rozhovorů, a to včetně strategií, jak dospět k audiálním datům jednak souvislým, jednak po stránce jazykového projevu přirozeným, nejlépe regionálně zabarveným, nebo přímo nářečním. Představeny jsou základní techniky nahrávání (3.2.4), např. optimální nastavení rekordéru, umístění mikrofonu ap. Výklad o pořizování audiálních dat (a také o jejich archivaci a případném zveřejnění) je doplněn právním a etickým rámcem (3.2.5), v současnosti hojně diskutovaným. Zájemce v této části nalezne základní informace o tajném/veřejném nahrávání, informovaném souhlasu, ochraně osobních a citlivých údajů, anonymizaci nebo mikroetice. Nastíněny jsou možnosti vytvoření digitálního zvukového

1 Dostupné jsou sbírky spisovných projevů, v omezené míře též sbírky spontánní řeči, avšak v nich obsažená data nejsou tříděna regionálně/nářečně.

archivu (3.3), a to včetně názorné ukázky z budované *Databáze nářečních projevů pro odbornou veřejnost*.[2] V poslední podkapitole jsou popsány vybrané systémy pro transkripci spontánních jazykových promluv (3.4), se zvláštním zřetelem k aktuální dialektologické transkripci, sestavené pro účely strojového učení. Audiální data je totiž vhodné (ne-li nutné) podpořit i daty textovými, z toho důvodu jsou k nahrávkám pořizovány transkripty, případně lze vycházet též z textových dat nemajících oporu v audiu, jak je podrobně popsáno v kapitole 4.

3.1 Dialektologické minimum: pojmy, termíny, struktury

Předmětem metodiky je zpracování českojazyčných dat (zvukových i textových) strojovým učením. Primárně jde o data nářeční povahy. Jako **teritoriální nářečí/dialekt** (synonymně též topolekt, lokolekt, geolekt, regionalekt či regiolekt[3]) se označuje soubor jazykových prostředků, které jsou vymezeny regionálně (oblastně) a které se užívají v mluvené komunikaci, zvláště v běžném, každodenním styku (podrobnější vymezení viz Kloferová, 2017). Teritoriálním dialektům je pak nadřazen pojem **běžná mluva** (**běžně mluvený jazyk**), kterou lze zjednodušeně charakterizovat jako „spontánní způsob projevu, který určitá osoba zaujímá v každodenních, neformálních situacích, kdy ke svému užívání jazyka neupírá přílišnou pozornost" (Chromý, 2021, s. 46; srov. též Krčmová, 1997; Krčmová a Chloupek, 2017 aj.).

Nářečí českého jazyka se hierarchicky dělí na:

- nářeční skupiny;
- nářeční podskupiny;
- nářeční úseky;
- nářeční typy.

Stratifikace českých dialektů v širším smyslu vychází z konce 19. století, kdy jako základní dělítko posloužily hranice vývojových střídnic za staročeské vokály *ý* (popř. *í*) a *ú* (srov. středočeské podoby *dobrej*, *mouka* × středomoravské *dobré*, *móka* × východomoravské *dobrý/dobrí*, *múka* × slezské *dobry*, *muka*). Při dělení hierarchicky nižších kategorií (podskupin, úseků, typů) bylo přihlédnuto i k dalším hláskovým, tvaroslovným, lexikálním, případně i syntaktickým jevům. Dělení českojazyčného území na jednotlivé nářeční skupiny a podskupiny zobrazuje uvedená mapa (viz obrázek 3.1), která byla sestavena na základě dat převážně z 50.–60. let 20. století.[4]

Nářeční hranice jsou v současnosti narušovány (a někdy dokonce již silně porušeny) působením velkého množství faktorů, vnějších i vnitřních (viz Šimečková, 2022, s. 115–116, též zde uvedená literatura). Dochází jednak ke zmenšování původních nářečních oblastí směrem od okrajů,[5] jednak k narušování kompaktnosti areálů vůbec. Z toho důvodu je nutné při sběru nových dat rozlišovat, zda je zachycená mluva nářeční (ve smyslu relativně uchovaného tradičního dialektu, jak je popsáno ve starších dialektologických pracích,

2 Databáze je jedním z výstupů projektu NAKI *Jazyková paměť regionů České republiky*, s plánovaným datem uplatnění výsledku v r. 2026.

3 V tomto případě nejde vždy o synonymní výraz. Např. v německém prostředí splývá termín *regiolekt* (tedy *Regiolekt*, též *Regionalsprache*, *regional Ungangssprache*) s interdialektem (takto např. v *Norddeutscher Sprachatlas*; viz Elmentaler, Rosenberg a kol., 2015), totéž platí třeba o polské dialektologické terminologii (např. Wyderka, 2014, s. 1).

4 Data sesbíral, utřídil a k mapování připravil B. Stupňánek. Jde o enormní objem dat, textových i zvukových, ukazujících oproti starším mapám hranice areálů mnohem detailněji, mj. s přihlédnutím k četným německojazyčným ostrovům ve vnitrozemí.

5 Např. ústup hranic v oblasti centrálně středomoravské popsal již v polovině 20. století F. Kopečný (1957, výklopná mapa). Ten porovnal rozsah areálu stanoveného F. Bartošem v poslední čtvrtině 19. století a stav zjištěný vlastním výzkumem. Na základě této komparace sestavil mapu, na níž jsou vyznačeny oblasti s již neužívanými širokými vokály, jejichž rozsahem je právě oblast centrálně středomoravských dialektů tradičně definována.

např. Bělič, 1972; Balhar a kol., 1992–2011), nebo zda již nejde „pouze" o **regionálně zabarvenou mluvu**, v níž je využíváno jen některých nářečních prostředků, navíc nesystematicky.

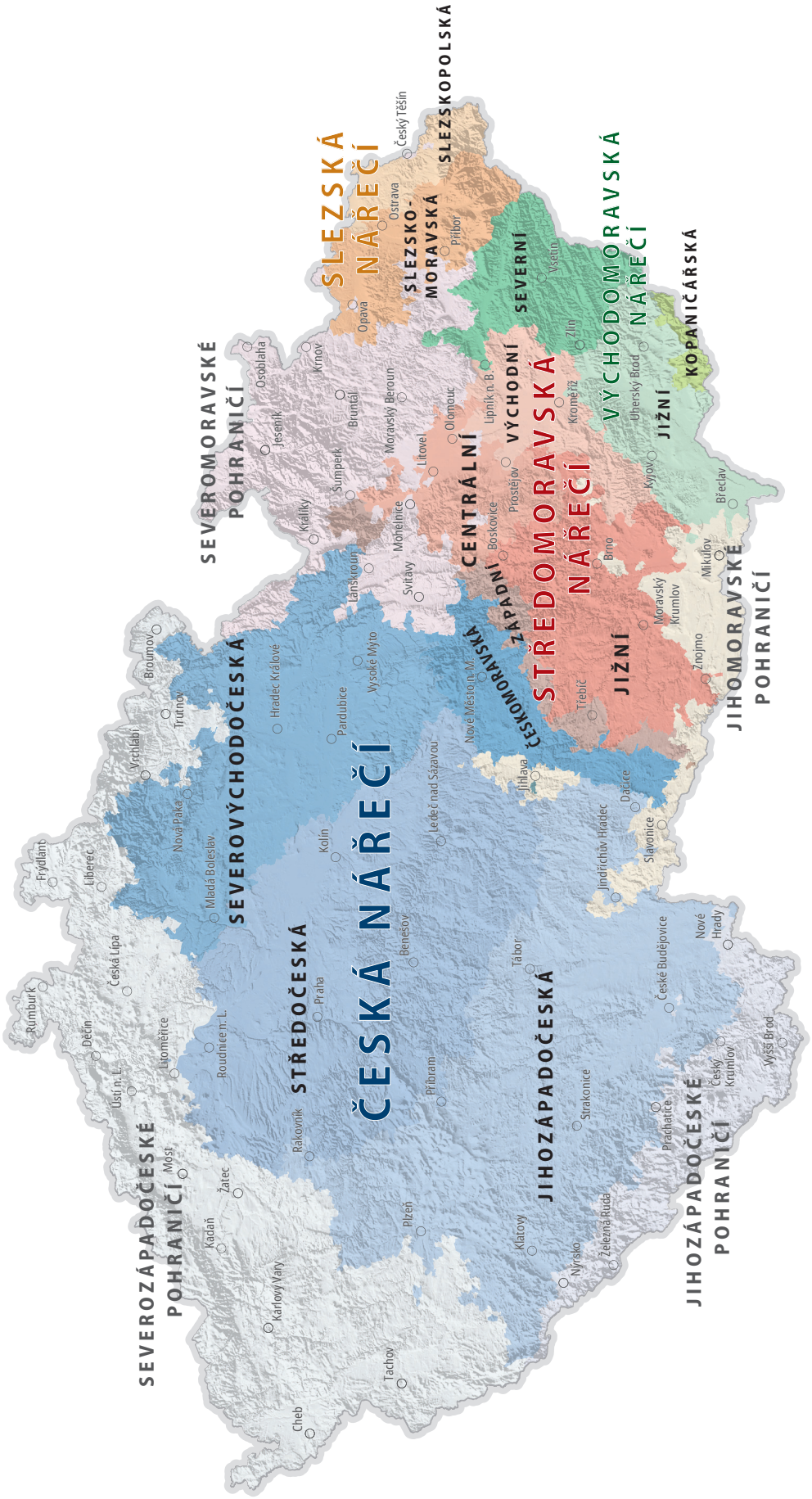
V tomto ohledu lze vydělit ještě interdialekt, hyperdialekt a kulturní dialekt. Zatímco **interdialekt** je chápán jako nadnářeční útvar, vzniklý ústupem některých výraznějších nářečních jevů původního dialektu (tento proces je pak nazýván jako nivelizace) a postupným sbližováním více původně samostatných dialektů, u **hyperdialektu** jsou naopak nářeční prostředky posíleny stylizací, a to nad úroveň jejich výskytu v přirozeném projevu (Šimečková, 2024a, s. 13). Označení **kulturní dialekt** se pak vztahuje na jazyk užívaný v nářeční beletrii, který je v porovnání s přirozenými projevy zdeformovaný a v němž se odráží autorský styl spolu se silným vlivem spisovného jazyka.

Ve všech případech (tj. u tradičních dialektů, též u interdialektů, hyperdialektů, kulturních dialektů i regionálně zabarvené mluvy) je zásadní výskyt regionálně specifických jazykových prvků, s jejichž pomocí lze určit, ze které oblasti mluvčí (popř. pisatel) pravděpodobně pochází, případně která oblast silně ovlivnila jeho projev třeba při dlouhodobějším studijním pobytu, při výkonu práce atd.

Vzhledem ke zmíněné územní distribuci jazykových prostředků lze při dialektologické analýze, též při dolování dat pro strojové učení využívat jak data starší (tj. dokumentující tradiční nářeční vrstvu), tak data nová, která jsou buď tradičně nářeční, nebo jen nářečně zabarvená. Může jít přitom o data audiální, audiovizuální či textová. Naopak nezbytné je **vyloučit ta data, která regionální povahy nejsou** (tedy projevy nenářeční, tj. spisovné, polospisovné nebo smíšené), popř. data, která jsou jiným způsobem nevyhovující (např. pocházející z nářečně nepůvodních oblastí /= z pohraničí a německých jazykových ostrovů uvnitř českojazyčného území/ nebo získaná v zahraničních enklávách češtiny).

Data, která výzkumník identifikuje jako nevyhovující, není nutné likvidovat; naopak mohou posloužit budoucímu výzkumu. Je však výhodné všechna data třídit podle jednotného klíče, zaručujícího snadnou eliminaci nevyhovujících dat i případné přeskupování dat v rámci jednotlivých kategorií.

Příkladem takového **klíče pro audiální data**, na něž se v této kapitole budeme soustředit, je způsob třídění zvukových nahrávek v připravované *Databázi nářečních promluv pro odbornou veřejnost* (viz tabulka 3.1). V této databázi je všem nahrávkám přidělován perzistentní kód označující jejich příslušnost (resp. příslušnost hlavního mluvčího) k určité nářeční oblasti (např. kód 2-4-3 = slavkovsko-bučovický typ), popř. k oblasti nářečně nepůvodní či zahraniční (speciální kategorii pak tvoří projevy nenářeční a smíšené). Při dolování dat strojovým učením může být využito **méně detailní členění** nářečních či regionálně zabarvených promluv, než které je uvedeno v klíči – z důvodu malého zastoupení dat z některých nářečních typů a úseků je někdy přistupováno ke zpracování dat na úrovni nářečních podskupin, zatímco k nižším úrovním se nepřihlíží (takto např. Šimečková, Karafiát a Plchot, v tisku).



Obrázek 3.1 Nářečí českého jazyka, tradiční stav.
Nové vymezení hranic nářečních oblastí dosud nebylo publikováno.

Tabulka 3.1 Klíč pro třídění nahrávek na základě jejich teritoriální příslušnosti v Databázi nářečních promluv pro odbornou veřejnost

1 ČESKÁ NÁŘEČNÍ SKUPINA
- **1-1 severovýchodočeská nářeční podskupina**
  - 1-1-1 podještědský a podkrkonošský typ
  - 1-1-2 východolitomyšlský typ
  - 1-1-3 náchodský typ
- **1-2 středočeská nářeční podskupina**
  - 1-2-1 lounsko-litoměřický typ
- **1-3 jihozápadočeská nářeční podskupina**
  - 1-3-1 západočeský úsek
    - 1-3-1-1 domažlický typ
    - 1-3-1-2 manětínský typ
    - 1-3-1-3 stříbrský typ
  - 1-3-2 jihočeský úsek
    - 1-3-2-1 doudlebský typ
    - 1-3-2-2 prachatický typ
- **1-4 českomoravská nářeční podskupina**
  - 1-4-1 žďársko-bystřický typ
  - 1-4-2 jemnický typ

2 STŘEDOMORAVSKÁ NÁŘEČNÍ SKUPINA
- **2-1 centrální středomoravská nářeční podskupina**
- **2-2 jižní středomoravská nářeční podskupina**
  - 2-2-1 znojemský typ
  - 2-2-2 tišnovský typ
  - 2-2-3 židlochovický typ
- **2-3 západní středomoravský okrajový úsek**
  - 2-3-1 zábřežský typ
  - 2-3-2 kunštátsko-budějovický typ
- **2-4 východní středomoravský okrajový úsek**
  - 2-4-1 týnecko-tištínský typ
  - 2-4-2 kojetínsko-přerovský typ
  - 2-4-3 slavkovsko-bučovický typ

3 VÝCHODOMORAVSKÁ NÁŘEČNÍ SKUPINA
- **3-1 jižní východomoravská nářeční podskupina**
  - 3-1-1 kyjovský typ
- **3-2 severní východomoravská nářeční podskupina**
  - 3-2-1 hranický typ
  - 3-2-2 kelečský typ
  - 3-2-3 spálovský typ
  - 3-2-4 starojičínský typ
- **3-3 kopaničářská nářeční podskupina**

4 SLEZSKÁ NÁŘEČNÍ SKUPINA
- **4-1 slezskomoravská nářeční podskupina**
  - 4-1-1 západní slezskomoravský úsek
    - 4-1-1-1 západoopavský typ
    - 4-1-1-2 branický typ
    - 4-1-1-3 baborowský typ
  - 4-1-2 východní slezskomoravský úsek
    - 4-1-2-1 hornoostravický typ
  - 4-1-3 jižní slezskomoravský úsek
- **4-2 slezskopolská nářeční podskupina**
  - 4-2-1 karvinský typ
  - 4-2-2 bohumínský typ
  - 4-2-3 těšínský typ
  - 4-2-4 jablunkovský typ

5 NÁŘEČNĚ NEPŮVODNÍ LOKALITY
- **5-1 severozápadočeské pohraničí**
- **5-2 jihozápadočeské pohraničí**
- **5-3 severomoravské pohraničí**
- **5-4 jihomoravské pohraničí**

6 ZAHRANIČNÍ LOKALITY
- **6-1 Polsko**
- **6-2 Slovensko**
- **6-3 Chorvatsko**
- **6-4 Bosna a Hercegovina**
- **6-5 Srbsko**
- **6-6 Rumunsko**

7 NENÁŘEČNÍ NEBO SMÍŠENÉ PROJEVY

Určení příslušnosti audiálních (a někdy i textových) dat je mnohdy nadmíru těžké, a to zvláště u dat získaných novějšími sběry, v nichž dochází k míšení a prolínání různých typů projevů, včetně jazykových prostředků z různých nářečních oblastí. Z toho důvodu je nutné pečlivě dbát na výběr dat, která mají být podrobena analýze či experimentu, a dat, která k těmto účelům naopak vhodná nejsou. Obzvlášť pozorně by se mělo přihlížet k **přirozenosti projevů**, přičemž pro sběr takovýchto audiálních dat existuje řada strategií, jak je popsáno dále.

3.2 Cesty ke zdrojům audiálních dat

V lingvisticky orientovaných kvalitativních výzkumech se užívá čtyř základních způsobů ke sběru dat:

- dotazníkový výzkum;
- rozhovor/interview;
- zúčastněné pozorování;
- explorace dat dostupných v korpusech (např. excerpce dat z korpusu DIALEKT, zabudovaného v Českém národním korpusu).

Tyto metody jsou dobře propracovány zvláště v zahraničních sociolingvistických pracích (některé citujeme níže), v českém prostředí je shrnul zejména J. Chromý (2014, 2021). Pro strojové učení je zásadní pracovat s co největším množstvím souvislých dat, z toho důvodu bude v této kapitole **podrobně představena metoda rozhovoru**, umožňující vytvoření vlastního, datově bohatého archivu s audiálními daty.

3.2.1 Terénní výzkum a metoda rozhovoru

3.2.1.1 Typologie rozhovorů/interview a role jejich aktérů

Rozhovor neboli interview lze v lingvistickém výzkumu definovat jako sběr jazykových dat rozmlouváním s lidmi. Rozhovory lze dělit na tři základní typy:

- **strukturovaný rozhovor**, který sestává z pokládání série předem daných otázek, přičemž u každého účastníka rozhovoru je použita stejná sada dotazů. Jde vlastně o mluvenou formu dotazníkového výzkumu (podrobně Fontana a Frey, 1994, s. 363–364);
- **nestrukturovaný rozhovor**, spočívající v pokládání nepřipravených otázek a oproti strukturovanému interview mající nedirektivní charakter (Fontana a Frey, 2004, s. 366);
- **polostrukturovaný/semistrukturovaný rozhovor**, jehož průběh je sice značně improvizovaný, co se týče obsahu dotazování, avšak počítající předem s určitým tematickým okruhem otázek. V české dialektologii jde o nejčastější výzkumnou metodu.

Při sběru mluvených dat lze vydělit **čtyři formy rozhovorů**, které probíhají:

- osobně;
- telefonicky;
- přes videorozhovor;
- kombinací osobního rozhovoru / videorozhovoru a psané komunikace.[6]

Tyto rozhovory lze realizovat třemi způsoby:

- přímo (rozhovor je veden ze strany výzkumníka);
- polopřímo (rozhovor je veden zprostředkovanou osobou podle instrukcí výzkumníka);
- nepřímo (rozhovor je veden za pomoci umělé inteligence /virtuální asistent nebo chatbot/ podle zadání výzkumníka).

V této metodice se primárně věnujeme **přímému osobnímu rozhovoru**, který má několik pozitiv – nahraná data jsou mnohem kvalitnější po zvukové stránce (třeba v porovnání s telefonickými rozhovory) a také jde, zvláště u příslušníků starších generací zapojených do výzkumu v roli mluvčích, o vítanější formu interview. Navíc u některých starších lidí lze předpokládat nižší počítačovou gramotnost, což ztěžuje možnost využít metody online rozhovoru (obecně k metodě internetového rozhovoru viz James a Busher, 2012, též zde uvedená literatura). Osobní setkání nadto umožňuje rozvést různé strategie pro uvolnění mluvčího a snazší přechod k běžnému vyjadřování – jde totiž o **nápodobu přirozeného, přátelského rozhovoru**.

Do rozhovoru tedy vstupují dva hlavní aktéři:

- **explorátor**, tedy výzkumník, popř. výzkumný tým, který sleduje výzkumné cíle a těmto cílům podřizuje průběh rozhovoru. Jde o nahrávající osobu, která by měla osobu nahrávanou vést k běžnému, každodennímu, popř. nářečnímu vyjadřování a pobízet ideálně k souvislým promluvám prostřednictvím vhodně kladených otázek;
- **informátor**[7], **informant** neboli **respondent**, tedy mluvčí, který odpovídá na dotazy kladené explorátorem a optimálně k tomu užívá běžné, nespisovné, popř. nářeční jazykové prostředky.

V dialektologických výzkumech býval dosud za **prototypického informanta** pokládán starousedlík, který žil v jedné obci od narození, měl základní vzdělání a pracoval v zemědělství nebo byl (v případě ženských informátorek) v domácnosti. Šlo přitom majoritně o příslušníky starších generací (k tzv. ideálnímu mluvčímu viz 3.2.1.3, bod 1). Takovému profilu dnes vyhovuje minimum mluvčích, z toho důvodu se upouští od nutnosti nejnižšího stupně vzdělání a od zaměření na určitou pracovní vrstvu obyvatelstva. Do výzkumů jsou nově dokonce zapojováni lidé, kteří změnili své bydliště; podmínkou však zůstává aktivní udržování nářečí rodiště (ve smyslu lokality, v níž žil mluvčí dlouhodobě v dětství). Do budoucna bude nezbytné přihlédnout i k mluvčím, u nichž se prolínají jazykové kódy více lokalit, šlo by však o výzkum se specifickými cíli.

---

6 Zvláštní formu představuje interview vedené se sluchově znevýhodněnými informanty. Vzhledem k zaměření metodiky na výzkumy každodenních, regionálně rozrůzněných mluvených projevů lze pracovat s mluvčími, kteří ztratili sluch (částečně, či úplně) v pozdějším věku. Komunikace s nimi pak sice může probíhat osobně, avšak prostřednictvím pomůcek sloužících ke kladení otázek ze strany badatele v písemné formě (klasické psací potřeby, tablet, obrazovka notebooku; lze též využít automatický přepis mluveného slova do textu). Otázky lze případně pokládat i znakovou řečí, pokud ji ovládají oba účastníci komunikace (s tím, že odpovědi jsou pak vyžadovány ve verbální formě). Neslyšící, kteří jsou znevýhodněni od narození nebo u nichž došlo ke ztrátě sluchu v raném dětství, do výzkumu mluvených projevů zařazováni z pochopitelných důvodů nebývají, totéž platí o lidech postižených němotou. Překážkou pro zařazení osoby do výzkumu běžné a nářeční mluvy mohou být také další tělesné vady, úrazy a postižení (včetně těch mentálních), např. defekty chrupu, orofaciální rozštěp, prodělaná cévní mozková příhoda vedoucí ke zbavení schopnosti mluvit aj.

7 Termín je někdy považován za nevyhovující, neboť se k němu vážou negativní konotace (informátor coby donašeč; informaci čerpáme z ústního sdělení několika českých lingvistů). Z toho důvodu je vhodnější užívat méně zatíženého pojmu informant nebo neutrálních označení jako respondent nebo mluvčí. V orální historii se pro informanta (avšak bez ohledu na používané jazykové prostředky, které v daném oboru nejsou předmětem výzkumu) vžilo označení narátor (např. Nosková, 2014).

Ideálně je informant jeden; případné zaznamenávání mluvy více lidí hovořících zároveň vede k těžkostem při následném vyhodnocování dat strojovým učením, též při pořizování přepisu (ručního i /polo/automatického). **Skupinovému rozhovoru**, kterého se účastní více informantů coby nositelů téhož jazykového kódu (nářečí) nebo kódů blízkých, však nelze upřít tu výhodu, že navozuje přirozenější komunikační situaci, a tak je spontánnější i nahraný projev. Je tak na zvážení výzkumníka, kterým směrem se ve svém výzkumu vydá.

V případě explorátora je ideální sestava dvou osob, z nichž jedna se věnuje mluvčímu, tj. pokládá otázky a tím řídí chod rozhovoru, druhá osoba[8] pak může pořizovat poznámky nebo kontrolovat a případně přenastavovat nahrávací zařízení v procesu výzkumu (k tomu viz 3.2.4). Větší počet explorátorů není vhodný, neboť by se tím jednak informant mohl cítit diskomfortně, a to i z důvodu někdy vnímané hierarchicky nadřazené pozice explorátora obecně, jednak by rozhovor ztratil ráz důvěrnosti.

V některých případech si tito aktéři (explorátor, informant) role vymění, a tak se z informanta stává osoba kladoucí otázky explorátorovi; tato situace však není žádoucí, neboť při získávání dat je cílem nahrání co nejdelší promluvy informantů. Mezi explorátorem a informantem totiž není rovnocenné partnerství, a vlastně ani nejde o skutečný (přirozený) rozhovor, nýbrž o situaci, během níž se explorátor snaží přimět informanta k co nejobsáhlejšímu mluvnímu projevu pomocí různých strategií.

PŘÍKLAD Z PRAXE

V roce 2022 byl F. Kubečkem a M. Šimečkovou uskutečněn výzkum v Protivanově na Prostějovsku. Zúčastnilo se ho 7 mluvčích, které nebylo přes veškerou snahu možné rozdělit do menších skupin, nebo dokonce s nimi nahrávat rozhovory jednotlivě. Přes původní domluvu, že si mluvčí nebudou zasahovat do projevu a že se budou „hlásit" o slovo, docházelo k častým překryvům, kvůli nimž není nahrávka vhodná třeba pro účely strojového učení. Úsek pořízené nahrávky si lze poslechnout prostřednictvím QR kódu.

3.2.1.2 Strategie pro přepnutí jazykového kódu mluvčího směrem k běžněmluvenostnímu vyjadřování

Pro zvýšení spontánnosti projevu a zmírnění napětí na straně informanta existuje několik technik. Doporučené taktiky lze během výzkumu uplatnit veskrze, anebo výběrově, vždy v závislosti na sledovaných cílech. Z těch nejdůležitějších jmenujme:

1. tajné nahrávání a zatajení účelu výzkumu (tento bod je však sporný, viz níže);
2. nedirektivní formát rozhovoru;
3. navázání důvěry mezi explorátorem „cizincem" a informantem;
4. přítomnost blízké osoby v roli druhého informanta;
5. přítomnost blízké osoby v roli explorátora;
6. vyjadřování podpory mluvčího ze strany explorátora;
7. akomodace ze strany explorátora;
8. výběr témat rozhovoru;
9. způsob pokládání otázek.

8 Může jít též o začínajícího výzkumníka, který společným nahráváním, vedeným hlavním explorátorem, teprve získává zkušenosti, jak vést rozhovor. Zaučení v terénu je neocenitelné, stejně tak vrozené nadání pro interview – tyto propriety nemohou být vynahrazeny žádnými teoretickými radami.

Ad 1: Tajné nahrávání a zatajení účelu výzkumu

Podle současné české legislativy lze v rámci vědeckého výzkumu, byť s jistými omezeními, **pořizovat nahrávky osob bez jejich vědomí** (podrobně viz 3.2.5), tato technika je ale někdy vnímána jako „pirátská", tedy neetická (např. Zíková a Křivan, 2014, s. 80; Chromý, 2021, s. 48[9]). Sporné ze strany etiky je rovněž případné zveřejnění takto získaných nahrávek. Nevýhodou tajného nahrávání je horší kvalita zvuku, neboť tato metoda neumožňuje umístit k mluvčímu externí mikrofon a i samotný diktafon musí zůstat skryt. Zjevným kladem je naopak větší spontánnost projevu a menší nervozita mluvčího, pro kterého může očitá přítomnost nahrávacího zařízení vyvolat stres a napětí.

Obecně by se měl výzkumník snažit najít jiné postupy a metody, jak data získat, přičemž tajné nahrávání by mělo být až poslední variantou (podrobně se nad tajným nahráváním zamýšlí na poli nelingvistického výzkumu Miovský a kol., 2004). Je totiž možné, že v budoucnu dojde ke změně legislativy, a starší data, pořízená pomocí klamné strategie, nebude možné nadále používat k výzkumným účelům, stanou se tedy bezcennými.

Lze též užít strategii **zatajení účelu výzkumu** – na otevřené lingvistické bádání, zvláště zaštítěné institucí, jako je prestižní univerzitní pracoviště nebo Ústav pro jazyk český, totiž někteří informanti reagují buď spisovným vyjadřováním, nebo přepnutím do přehnaně nářečního kódu, tj. do hyperdialektu. Z toho důvodu se i v českém prostředí dialektologický výzkum zastřešoval zvláště v minulosti výzkumem etnografickým či historickým, konkrétně výzkumem místního zvykosloví nebo historie obce. I toto částečné zatajení je však sporné.

PŘÍKLAD Z PRAXE

Vzpomínka dialektologa J. Balhara na praxi tajného nahrávání ve druhé polovině 20. století: „V seminářích jsme byli staršími pracovníky nabádáni, abychom nejprve získali důvěru informátorů a abychom se s nimi a s jejich problémy nejprve blíže seznámili, pak nepozorovaně vytáhli magnetofon a po delší době jej spustili. Informátor neměl vůbec tušit, že je nahráván. Sám jsem se tohoto doporučení zuby nehty držel." (Goláňová, 2009, s. 22) Od taktiky tajného nahrávání bylo později členy dialektologického oddělení ÚJČ AV ČR upuštěno.

Při tvorbě Pražského mluveného korpusu, vedeného F. Čermákem, byly využity nahrávky pořízené v letech 1988–1992 a 1994–1996. Nahrávaným osobám tehdy bylo sděleno, že je záznam pořizován nikoliv za účelem lingvistického zkoumání, nýbrž že objektem výzkumu jsou názory na aktuální společenské problémy (Čermák a kol., 2007, s. 12).

Sociolingvistka L. Milroyová svůj výzkum sociálních sítí v severoirském Belfastu zdůvodňovala jako „zkoumání společenských změn nebo života v daném společenství" (Milroyová a Gordon, 2012, s. 88). Obdobně. A. Tagliamonteová původně užívala jako zástěrku výzkum historie a kultury místních komunit, ovšem později byla nucena participantům nových výzkumů specifikovat výzkum jako lingvistický (resp. sociolingvistický), a dokonce popsat proměnné, které byly předmětem jejího zájmu (Tagliamonte, 2006, s. 33).

Explorátoři zapojení do výzkumu protetického *v*-, vedeného J. Chromým, měli doporučeno zdůvodňovat své počínání nahráváním neformálních rozhovorů primárně se týkajících života v daném městě. O skutečném, tj. sociolingvistickém záměru, byli zpraveni až po završení sběru v dané lokalitě (Chromý, 2017, s. 144).

Originální zastírací manévr byl využit při výzkumu výslovnostního úzu realizovaném týmem českých fonetiků v r. 2014. Ve snaze o zatajení lingvistického charakteru bádání byli mluvčí informováni, že je výzkum zaměřen na čtení a zapamatování krátkých vět, z čehož mohlo být vyvozeno, že jde o výzkum krátkodobé paměti (Duběda a kol., 2014, s. 135–136). Není známo, zda později došlo ze strany explorátorů k odhalení výzkumného cíle, či nikoliv.

9 Chromý ve starší práci dokonce takto získaná data zcela zavrhl: „V některých případech nastává problém, že některé jevy nemůžeme zachytit jinak než potají. V tomto ohledu by se nabízelo skryté nahrávání. To je považováno za zcela neetické a výzkumy na jeho základě by neměly být realizovány, natožpak publikovány." (Chromý, 2014, s. 52)

Ad 2: Nedirektivní formát rozhovoru

Pokud je cílem výzkumu dokumentace souvislých promluv bez nutnosti získat ode všech mluvčích data stejného charakteru, pak je výhodnější jít cestou **nestrukturovaného nebo polostrukturovaného rozhovoru** (viz 3.2.1.1). U strukturovaného rozhovoru hrozí, že bude mluvčí příliš svázán tématy, o nichž navíc nemusí být vždy ochoten hovořit, a tím by se mohl uzavřít do sebe. Navíc ve světle vynucování odpovědí na konkrétní otázky se explorátor může jevit jako hierarchicky nadřazená osoba, což mohou mluvčí pociťovat negativně a má to i dopad na jejich slovní projev.

Ad 3: Navázání důvěry mezi explorátorem „cizincem" a mluvčím

Pokud je explorátor pro informanta cizí osobou, může být pro mluvčího obtížné přepnout do běžného vyjadřovacího modu. V takovém případě informant často hovoří spisovně (popř. využívá jen některé nespisovné prvky, zvláště interdialektické), nadto může být zdrženlivý při rozvíjení jednotlivých témat, a tak na otázky odpovídá zkratkovitě. Z toho důvodu je výhodné, pokud mezi explorátorem a informantem již v době výzkumu existuje nějaký vztah (příbuzenský či jiný); v opačném případně je doporučováno postupné **vyvinutí důvěry opakovanými návštěvami**.

Pro posílení důvěry je navíc možné k tomu využít nějakou zprostředkovatelskou osobu, která buď explorátora informantovi doporučí, nebo je dokonce přítomna při nahrávání (viz též bod 4). Ve druhém případě je však nutné danou osobu instruovat ohledně nezasahování do průběhu rozhovoru, popř. se s ní domluvit na určitých pravidlech chování ve snaze o omezení řečových překryvů.

Ad 4: Přítomnost blízké osoby v roli druhého informanta

Jako ideální taktika se jeví přítomnost dalšího informanta (nebo i více informantů), který je v nějakém vztahu (příbuzenském, přátelském) s hlavním informantem. Jde o metodu **skupinových rozhovorů**, která je doporučována napříč sociolingvistickými pracemi (např. Labov, 1972, s. 210; Codó, 2008, s. 163). Zjevnou nevýhodou této metody je – kromě nižší kvality takto pořízených nahrávek – také obtížná identifikace mluvčích a nerovnováha mezi zastoupením promluv jednotlivých účastníků výzkumu.

Jednotlivé typy skupinových interview včetně jejich ne/výhod popsali sociologové A. Fontana a J. H. Frey (2004, s. 364–365).

Ad 5: Přítomnost blízké osoby v roli explorátora

Pokud není možné vybudovat si s informantem vztah a nechceme přistoupit ke skupinovému rozhovoru, lze jít cestou tzv. **zprostředkovaného explorátora**, který na základě instrukcí získaných od výzkumníka uskuteční výzkum za něj a který pro informanta představuje právě osobu blízkou. Jde o metodu, která vede informanta k větší otevřenosti a uvolněnosti ve vyjadřování, co se týče obsahu i jazyka. Oproti přímému explorátorovi však hrozí menší metodologická přesnost. V lingvistice jde o metodu zatím minimálně užívanou (v českém prostředí třeba Tkáč, 2022).[10]

Jinou strategií je využití zprostředkovaného explorátora coby pomocníka, který sice neřídí průběh rozhovoru, ale napomáhá jeho hladšímu průběhu. Jeho úkolem je pak buď přeformulovávání dotazů do podoby uchopitelnější pro mluvčího (třeba i s převodem do místního dialektu), nebo vymýšlení doplňujících otázek opřených o znalost mluvčího. V případě využití této metody je nezbytné vymezit pomocnému explorátorovi pole jeho působnosti, aby se tak předcházelo přehnanému aktivismu, který by mohl mít negativní dopad na výzkum.

---

10 Částečně roli zprostředkovaného explorátora sehrávají studenti, kteří na základě instrukcí sbírají data pro účely sepsání studentské práce (seminární, diplomové, disertační). Například pracovníci dialektologického oddělení ÚJČ AV ČR spolupracují se studenty dlouhodobě, zvláště z Masarykovy univerzity v Brně, přičemž získaná data se následně stávají součástí zvukového archivu, deponovaného v tomto oddělení. Informanty zapojenými do těchto rozhovorů přitom bývají často příbuzní nebo známí studentů.

PŘÍKLAD Z PRAXE

V roce 2023 sbíral data ve vybraných jihovýchodomoravských obcích externista T. Macalík, a to pro dialektologické oddělení ÚJČ AV ČR. Nápomocna mu při tom byla babička, která mu nejen zprostředkovávala kontakty na místní nářeční mluvčí, ale také s ním byla přítomna u rozhovorů, do nichž často vstupovala v roli druhého (zprostředkovaného) explorátora. Výzkumník tuto spolupráci zhodnotil slovy: „Vztah mluvčích k explorátorově babičce implikuje vřelý a nadstandardní vztah i k explorátorovi samému, tím více, pokud byla u rozhovoru také ona."

Pomocný explorátor se ukázal být velkou výhodou v oblasti česko-polského smíšeného pruhu. V r. 2023 zde proběhla intenzivní dokumentace nářečí výzkumníky ÚJČ AV ČR, komunikace s mluvčími však byla ztížena neznalostí jejich specifického dialektu na straně explorátorů. V Oldřichovicích, části Třince, z toho důvodu bylo využito nabídky rodáka z blízkých Kojkovic H. Szlaura, který byl přítomen u rozhovoru se čtyřmi mluvčími a ochotně překládal explorátorovy dotazy do místního nářečí.

Jinou strategií je částečné převzetí role explorátora účastníky výzkumu, jedná-li se o skupinový rozhovor; explorátor pak záměrně nechává prostor volnému vyprávění, jen částečně řídí jeho průběh. Pro minimalizaci rušivého faktoru „cizinectví" může explorátor z místnosti odejít, a nechat tak mluvčím volné pole, přičemž nahrávací zařízení je po odchodu explorátora ponecháno zapnuté (Braber a Davies, 2016, s. 100). Lze se tak vyhnout tzv. **paradoxu pozorovatele**, který výstižně popsal Labov: „Cílem lingvistického výzkumu v komunitě musí být zjistit, jak lidé mluví, když nejsou systematicky pozorováni; avšak tato data můžeme získat pouze systematickým pozorováním." (Labov, 1972, s. 209, vlastní překlad)[11]

Ad 6: Vyjadřování podpory mluvčího ze strany explorátora

Explorátor informanta podporuje v mluveném projevu explicitně (slovním vybídnutím), nebo implicitně (např. úsměvem, přikývnutím, souhlasným gestem ruky). K základní **verbální technice** patří kladení otázek, které má explorátor buď připravené, nebo jsou vytvářeny až během rozhovoru. Otázky jsou pokládány jasně a stručně (podle Labova by měla být délka otázky kladené v angličtině do 5 sekund, viz Labov, 1984, s. 34), samozřejmostí je jejich zdvořilostní forma. Naopak zcela nevhodný je přílišně ležérní, familiární styl (nepojí-li explorátora s informantem bližší vztah), popř. afektovanost vůbec.

Explorátor může informanta výslovně vybízet také k určitému stylu vyjadřování, tj. k vyjadřování běžnému, každodennímu, popř. nářečnímu. Žádoucí je to zvláště v případě, pokud mluvčí užívá výhradně spisovný jazyk, ovšem explorátorovým cílem je záznam tradičního dialektu a přitom ví, že daný informant ho ovládá.

PŘÍKLAD Z PRAXE

V roce 2022 byl dialektology F. Kubečkem a M. Šimečkovou uskutečněn nářeční výzkum v Hluchově na Prostějovsku. Asi po hodinovém zaznamenávání promluvy, vyznačující se jen některými středomoravskými rysy, informantka použila v replice osoby, o níž vyprávěla, široké vokály *ę*, *ǫ*, tedy tradiční prvky centrálně středomoravské. Na dotaz ze strany explorátorky, zda už tyto formy ona sama v mluvě neužívá, informantka překvapeně reagovala: „Já s vama jen mluvim tak česky, ne hanackę." (Pozn.: *česky* = spisovně.) Poté rozhovor pokračoval asi další hodinu, tentokrát již v centrálně středomoravském dialektu.

11 Další cestou, jak obejít daný paradox, je uplatnění tajného nahrávání obecně, viz bod 1.

Diskutabilní je jazyková stránka dotazování, tj. zda má explorátor užívat výhradně spisovných jazykových prostředků, nebo „přepnout" do každodenního vyjadřovacího kódu, popř. nářečí (k tomu viz bod 7).

Do verbální komunikace se řadí i užívání hezitačních zvuků a kontaktových formulí (*aha*, *no vidíte* ap.), jimiž explorátor vyjadřuje v průběhu výzkumu přitakání, a tak pobízí informanta k pokračování ve vyprávění. Explorátor by měl mít však na zřeteli, že by svými otázkami neměl přílišně narušovat promluvu informanta, totéž platí o hezitačních zvucích a kontaktových obratech, jejichž nadužívání může znehodnotit jinak kvalitní nahrávku (nevhodnou pak třeba pro detailní fonetickou analýzu).

Existuje také řada **neverbálních technik**, které lze aplikovat i na lingvistický výzkum, např. R. L. Gorden (1980, s. 335) rozlišuje čtyři základní typy neverbálních komunikací:

- proxemická komunikace = využití prostoru (resp. horizontální vzdálenosti) mezi aktéry rozhovoru k vyjádření postojů;
- chronemická komunikace = využití rytmu mluvy a délky ticha v konverzaci;
- kinezická komunikace = zahrnuje pohyby nebo postoje těla;
- paralingvální komunikace = zahrnuje všechny variace hlasitosti, tónu a barvy hlasu.

Jindy se v rámci neverbální komunikace hovoří o prostředcích paralingvistických a extralingvistických (např. Gavora, 2005, s. 100), přičemž do paralingvistických prostředků bývá řazena zvučnost hlasu, jeho barva, rychlost řeči, kladení a délka pauz, slovní důraz; do extralingvistických prostředků pak gestikulace, mimika, pohled (tj. zaměření zraku), haptika, posturika, proxemika, a dokonce i vzhled (tamtéž; podrobně k jednotlivým prostředkům viz s. 99–110). V našem případě je oním vzhledem myšlen zevnějšek explorátora, a to nejen ve smyslu úpravy vlasů, nehtů nebo výběru oblečení, ale také známek určitého fyzického stavu (např. únavy v podobě kruhů pod očima), které mohou být některými informanty hodnoceny nelibě, popř. znaků určitého věku a pohlaví (někteří informanti by nemuseli být tolik ochotni hovořit s příslušníky opačného pohlaví nebo s explorátory mladšího/staršího věku). Podrobněji ke vzhledu explorátora viz bod 7.

Explorátor by si měl osvojit základy body language; při budování vztahu s informantem jsou nápomocné základní techniky, jako je oční kontakt (udržován by měl být po většinu rozhovoru), přátelský úsměv, jemná přitakávací gestikulace a rovné držení těla. Nežádoucí jsou naopak zlozvyky, jako je prokřupávání kloubů v prstech, okusování nehtů, poklepávání tužkou apod.

Ad 7: Akomodace ze strany explorátora

Explorátor se může informantovi přizpůsobit:

- jazykovým projevem;
- regionem původu/bydliště či životní zkušeností;
- svými názory;
- svým zevnějškem.

Přizpůsobením jazykového projevu máme na mysli jak volbu jazykových prostředků (ne/spisovných, ne/nářečních), tak tón hlasu, tempo řeči či pauzování. Explorátor modifikuje svůj jazykový projev směrem k projevu informanta, přičemž **jazyková akomodace** je přirozeným procesem (podrobně Trudgill, 1986; Chromý, 2012; Wilson, 2017 a zde uvedená literatura). Možné je též vědomé přizpůsobení jazyka, zde je však zapotřebí dbát na přirozenost vlastního projevu (Tagliamonte, 2006, s. 41). Případné nadužívání nespisovných prvků explorátorem může u mluvčích vyvolat nelibost, totéž platí pro užívání nářečních prostředků, které jsou buď odlišné od prostředků užívaných informantem (protože jsou součástí jiného dialektu, jehož nositelem je explorátor), nebo dokonce chybné (explorátor předstírá příslušnost k témuž dialektu jako informant, a to ve snaze o podpoření nářečního vyjadřování na straně participanta výzkumu, avšak je přistižen při neznalosti).

PŘÍKLAD Z PRAXE

Se selháním strategie jazykové akomodace má zkušenost autorka této kapitoly, a to z výzkumu realizovaného ve Staré Břeclavi v r. 2023. Coby explorátorka někdy při dokumentaci nářečí v moravské oblasti užívá nářeční prostředky typické pro její rodiště, tj. východní Prostějovsko. Tato taktika obvykle vede k uvolnění ze strany informantů a k přechodu do vlastního nářečního modu. Ve jmenované lokalitě ale na vyslovení tvaru (v) *kroju* (namísto *kroji*) v rámci dotazu *Chodíte tady ještě v kroju?* informantka reagovala podrážděně; jde totiž o lokalitu, pro kterou je v daném tvaru charakteristické zakončení *-i*, nikoliv *-u*. Rozhořčení informantky pramenilo z přesvědčení, že jí explorátorka vnucuje jiný tvar, popř. že neví, co je pro danou oblast typické. Také jde prý o tvar, na který jsou místní lidé obzvlášť citliví. V takovém případě je vhodnější přejít na spisovné vyjadřování, které v informantovi nevyvolá nežádoucí reakci.

Příkladem dobré praxe je strategie J. Chromého, kterou v letech 2013–2015 uplatnil při výzkumu protetického *v-* v pěti městech, konkrétně v Praze, Plzni, Českých Budějovicích, Hradci Králové a Brně. Sběrem dat metodou neformálních rozhovorů pověřil externí spolupracovníky (vysokoškolské studenty) pocházející z výzkumných lokalit, čímž se vyhnul možnému zkreslení dat vlivem jazykové akomodace (Chromý, 2017, s. 143).

Jazyková akomodace může nechtěně vést k přílišnému posílení sebevědomí ze strany informanta (Braber a Davies, 2016, s. 100) a k nadužívání nářečních prostředků (ve snaze „udělat explorátorovi radost“, popř. ukázat, nakolik je informant dobrým nářečním mluvčím). Projevy hyperdialektu jsou však nežádoucí, nejde totiž o součást přirozeného projevu. Jindy může užívání jiného dialektu, než je dialekt informanta, vést k nechtěnému míšení různých nářečních prvků. Z toho důvodu je zapotřebí při využití strategie jazykové akomodace jednat uváženě.

Výhodnou strategií je **ztotožnění se s informantem** regionálně – pokud explorátor pochází z téže oblasti, lze na to upozornit, čímž se zprostí role „cizince“. S informantem lze rovněž najít další body, s nimiž se lze vzájemně identifikovat, jako je navštěvování téže školy, společné koníčky, společní známí, znalost nějakého tématu (např. hospodaření, krojové odívání, vaření) ap.

PŘÍKLAD Z PRAXE

Pracovníci dialektologického oddělení ÚJČ AV ČR při výzkumech na Moravě užívají strategické umístění pracoviště – to se nachází v Brně, nikoliv v Praze. Poukázáním na tuto skutečnost je u informantů někdy viditelná úleva, že nemluví s osobou pocházející z hlavního města, nýbrž s někým, kdo patří do šíře pojatého moravského regionu. Obdobně někteří dialektologové poukazují na úzce lokální původ explicitním určením místa svého rodiště/bydliště, navíc s využitím specifických jazykových prostředků (např.: *já su taky z Prostějovska* namísto spisovného *já jsem také z Prostějovska*).

Rozhovory probíhají v klidné atmosféře, tomuto požadavku lze vyhovět mj. **přizpůsobením názorů** ze strany explorátora. A. Fontana a J. H. Frey (1994, s. 371) přímo uvádějí: „The researcher must adapt to the world of the individuals studied and try to share their concerns and outlooks.“[12] Neznamená to, že by musel explorátor bezmyšlenkovitě přebírat veškeré názory informanta – jak píše orální historička J. Nosková (2014, s. 49), výzkumník by si měl zachovat důstojnost, obzvlášť pokud jsou probírána některá problematizovaná témata dotýkající se politického názoru, etnicity, náboženského vyznání, sexuality ap. Doporučení se týká spíše skutečnosti, že není žádoucí vyvolávat s informantem hádku. Informantovy názory lze zaznamenat s tím, že je explorátor nekomentuje ani nevyvrací, případně se snaží rozhovor nenápadně otočit jiným směrem.

12 „Výzkumník se musí přizpůsobit světu zkoumaných osob a pokusit se sdílet jejich obavy a pohledy.“ Překlad M. Š.

Explorátor může snazšímu průběhu rozhovoru napomoci **úpravou svého zevnějšku**. Informant, obzvlášť pokud je příslušníkem starší generace, zpravidla očekává výzkumníka upraveného vzhledu, neboť jde v očích veřejnosti o představitele určité institucionální pozice (zejména pokud je explorátor pracovníkem nějaké univerzity nebo Akademie věd). Explorátor s avantgardním účesem, výrazným tetováním a piercingem, navíc oblečený ve vyzývavém či nedůstojném oblečení, by mohl u mluvčího vyvolat negativní reakci (Fontana, 1977; Codó, 2008, s. 163, s. 169). Totéž platí v případě nadmíru slavnostního oblečení, které by mohlo navodit příliš formální, a tedy nežádoucí komunikační situaci. Pro některé mluvčí však mohou být rozhodující též faktory, které explorátor nemůže ovlivnit, a to pohlaví, věk, etnicita, popř. sociální třída a úroveň vzdělání (takto např. Braber a Davies, 2016, s. 100; o vlivu pohlaví explorátora na průběh rozhovoru v orálně historickém výzkumu viz třeba Gluck a Patai, 1991).

PŘÍKLAD Z PRAXE

Při dotazníkovém šetření *Metoda rozhovoru (interview) ve společenských a humanitních vědách* (dále MetRoz; viz Šimečková, 2024c)[13] se věk jako omezení při výzkumu neprokázal (100 % respondentů uvedlo, že se s omezením nesetkali). Jeden respondent poukázal na výhodu mladšího věku explorátora oproti informantovi, citujeme výňatek odpovědi: „Starší generace si více dovolí (nezdráhá se tykat), více vysvětluje, jelikož tomu všemu mladší člověk přece nemůže tolik rozumět, cítí se bezpečněji, nota bene mladý výzkumník zastoupí vnoučata, která třeba zanedbávají návštěvy, atd.“

S nabídkou tykání se setkali i někteří mladší výzkumníci dialektologického oddělení ÚJČ AV ČR; tykání lze brát v tomto případě jako výhodu, neboť je tak s mluvčím navázán bližší vztah, což napomáhá k jeho uvolnění, tudíž i ke každodennímu vyjadřování.

Naopak v r. 2023 se při výzkumu obalovaného *l* v jihovýchodomoravském areálu setkal jeden člen týmu, vedeného M. Šimečkovou, s negativní reakcí na své mládí. Informant, jakmile spatřil výzkumníka, vyjádřil svou nelibost nad věkem badatele – nepokládal ho za rovného partnera pro konverzaci. Tento případ je však ojedinělý. Doporučit pak lze zkusit i přes úvodní komplikace zapříst rozhovor, popř. v případě neúspěchu ho ukončit a přenechat práci s daným informantem jinému, staršímu výzkumníkovi.

Ad 8: Výběr témat rozhovoru

Obsah vyprávění je buď dopředu dán, pokud explorátor sleduje konkrétní výzkumné cíle (např. sběr autentických vyprávění o krizových situacích, viz Šimečková, 2024b). Je-li obsah rozhovoru volný, pak by měl explorátor vybírat taková témata, která jsou informantovi blízká, tj. taková, která se odvíjejí od jeho profese, zálib nebo prožitků. Je vhodné dopředu získat informace o mluvčích, popř. je zjistit při úvodním dotazování na osobní údaje. Explorátor by se měl vyhýbat citlivým otázkám, ať už jde o témata tabuizovaná (neštěstí, smrt, sexualita), nebo obecně intimní (finance, podnikání, politika). Jsou však mluvčí, kteří dokáží bezprostředně vyprávět i o těchto věcech (buď jsou obecně sdílnější, nebo je to dáno tím, že zprostředkovávají zážitky dnes už časově vzdálené, např. vzpomínky na násilné činy na ženách při osvobozování v r. 1945).

Pro vypnutí spisovného jazykového modu je možné využít **strategii překvapení**. Znamená to, že explorátor položí informantovi nezvyklý dotaz, ten si ve stavu překvapení zapomene dávat pozor na způsob projevu, a tak přejde do každodenního způsobu vyjadřování. Doporučováno je též ptát se na témata, která v informantovi vyvolají **silné emoce**. Ty jsou spouštěčem spontánní mluvy, neboť právě kvůli emočnímu vypětí nebude informant s to věnovat pozornost způsobu projevu.

13 Elektronický dotazník byl sdílen mezi sociálními a humanitními vědci, kteří využívají (nebo v minulosti využívali) metodu rozhovoru pro sběr dat. Dotazník byl šířen v období od 15. 6. do 30. 7. 2024, a to prostřednictvím služby GoogleForms. Zapojilo se do něj 70 vědců z Česka, Slovenska, Polska, Rakouska, Německa, Slovinska a Maďarska.

PŘÍKLAD Z PRAXE

Strategie překvapení byla uplatněna při výzkumu německých nářečí ve státě Missouri – explorátoři informanty vybízeli, ve snaze o uvolněnější vyjadřování, ke sdílení písní, receptů, ale také vtipů či anekdot, což je velice netradiční přístup (Johnson, 2023, s. 1).

V zahraničních sociolingvistických výzkumech se uplatňují otázky týkající se obecných témat, které informanta nutí k hlubšímu zamyšlení. S. A. Tagliamonteová (2006, s. 40) v rámci otázek, které jsou prý aplikovatelné napříč všemi komunitami a které jsou inspirovány Labovovými moduly, uvádí:

Did you ever have a dream that really scared you?[14]

Were you ever in a situation where you thought, 'This is it'?[15]

Obdobným příkladem z českého prostředí jsou otázky, které kladli výzkumníci při sběru jazykových dat pro *Pražský mluvený korpus* (Čermák a kol., 2017, s. 11–12). Výběrově citujeme:

Dneska chodí do práce asi 98 procent žen. Je podle vás postavení žen (u vás na pracovišti) rovnoprávný s muži?

Jak budeme vychovávat děti, taková bude jednou společnost. Otcové našich otců bývali často přísný a brali na děti pásek. Myslíte si, že dnešní rodiče trestaj děti málo nebo hodně?

V Praze nikdy nebyla nouze o cizince, ale teď je jich tady všude plno. Co si myslíš/-te o cizincích a cestovním ruchu?

Při některých dialektologických (popř. vůbec lingvistických) výzkumech v minulosti byla témata včetně konkrétních otázek pevně stanovena (Labov, 1984, s. 39; Tagliamonte, 2005; táž, 2006, s. 37–49; Vašíček, 2020, s. 402–407 aj.), v české dialektologii byly teprve nedávno rozpracovávány **soubory návodných otázek**, a to v souvislosti se zapojením veřejnosti do sběru dat (viz 3.2.2). Na půdě dialektologického oddělení ÚJČ AV ČR byl rovněž vypracován seznam obsahových metadat, kterými jsou značkovány jednotlivé nahrávky (a jejich úseky) v interním zvukovém archivu, ty vypovídají o tématech opakujících se napříč rozhovory. Soubor vypracovali B. Stupňánek a M. Šimečková, témata jsou dělena do pěti konverzačních modulů (viz obrázek 3.2), dále členěných do sekcí, podsekcí atd. Celkový počet obsahových metadat je 263, s novými sběry je však pravděpodobné, že se počet bude zvyšovat. Struktura metadat viz příloha 1.

S volbou témat souvisí **metoda reminiscence**. V takovém případě je informant vybídnut k autobiografickému vyprávění – hovoří o svém životě, rodině a práci, a to nejen o současném stavu, ale také o minulosti. Doporučit lze zvláště otázky zaměřené na dětství a mládí, díky kterým může být způsob promluvy přirozenější, běžnější (Ellis, 1974, s. 37). Je to dáno jednak tím, že je pozornost mluvčího více zaměstnána vzpomínáním, a tak nemá dostatek prostoru pro kontrolu jazykových prostředků, ale také vybavováním si prostředí, ve kterém vyrostl a k němuž se váže konkrétní místní dialekt.

Obrázek 3.2

Základní dělení obsahových metadat v Databázi nářečních promluv

14 *Měli jste někdy sen, který vás opravdu vyděsil?* Překlad M. Š.

15 *Byli jste někdy v situaci, kdy jste si pomysleli: To je konec?* Překlad M. Š.

Ad 9: Způsob pokládání otázek

V bodě 6 bylo poznamenáno, že otázky, s jejichž pomocí explorátor zjišťuje konkrétní odpovědi nebo díky kterým udržuje informanta v procesu vyprávění, by měly být pokládány jasně, srozumitelně, stručně a zdvořile.

Jednou z technik, jak „rozpovídat" mluvčího, je pak **předstíraná nevědomost**; v praxi to znamená, že pokud se informant ujišťuje, zda explorátor zná určitou reálii nebo slovo, měl by explorátor svou případnou znalost zapřít a požádat informanta o vysvětlení. Na druhou stranu je toto doporučení v rozporu se skutečností, že někteří informanti předpokládají připraveného explorátora, který bude mít alespoň základní přehled v místních poměrech (v orální historii viz třeba Nosková, 2014, s. 43). To však neznamená, že by měl explorátor informanta poučovat. Je tak nutné vhodně vyvážit obě strategie a odhadnout, která z nich bude u konkrétního informanta lépe fungovat. Výzkumník přitom nesmí zapomínat na to, že by celý rozhovor měl působit co nejpřirozeněji; pokud by tedy přílišně hrál jednu z rolí („neználka", či naopak „vševěda"), bylo by to kontraproduktivní, protože by byl zanedlouho odhalen.

PŘÍKLAD Z PRAXE

S rozporem mezi strategií nevědomosti a strategií obeznámenosti se členové dialektologického oddělení ÚJČ AV ČR setkávají poměrně často. Informanty jsou pokládáni, vzhledem ke své profesi, za znalce v oboru, mluvčí tak někdy vyjadřují zklamání, pokud se výzkumník přizná k neznalosti (opravdové, či strojené) určitého nářečního výrazu nebo tvaru, popř. místní reálie. Z vlastní zkušenosti lze výzkumníkovi doporučit alespoň minimální přípravu před návštěvou informanta, která by se týkala jak místního nářečí, tak oněch reálií. V případě znalosti určité reálie (pokud je na ni výzkumník explicitně dotazován), pak může připustit, že zná odpověď, avšak zároveň požádat informanta o podrobnější vysvětlení (*Ano, to znám, ale mohl/a byste mi to prosím ještě popsat? / Já bych to rád/a slyšel/a od vás.*).

Spíše výjimečně dojde k situaci, kdy se za „všeználka" pokládá informant (což bezesporu většinou je – žádný dialektolog nemůže dopodrobna znát veškerá nářečí konkrétního jazyka natolik detailně jako jejich nositelé) a své role zneužívá v tom smyslu, že výzkumníka začne zkoušet. Tento výslech by měl výzkumník co nejdříve přerušit, protože mu jde primárně o záznam souvislé řeči. Takovou zkušenost má za sebou M. Šimečková z výzkumu na Hodonínsku v r. 2023, do něhož byla zapojena coby informantka místní folklorně činná osoba; přes snahu přenést konverzaci do spontánního vyprávění a po nekončícím sledu otázek (včetně předčítání dlouhého textu z knihy) bylo nakonec lepší rozhovor ukončit.

Patrně nejobtížnějším úkolem je správná **technika kladení otázek**. Explorátor by se měl vyhnout zjišťovacím otázkám, na které by informant mohl odpovídat jednoslovně *ano*, či *ne* (např. *Chodí se tu v kroji?*); naopak je žádoucí užívat otázky doplňovací (*V jakém kroji se tady chodilo nebo chodí? Jak vypadaly a jak se jmenovaly jednotlivé části tohoto oděvu?*), popř. kombinace obou dotazů (*Chodí se tu v kroji? A jak vypadá?*). Výzkumník by neměl pokládat více otázek najednou, neboť by se v nich pak informant ztratil (většinou pak začne odpovídat pouze na poslední položenou otázku).

Obecně platí, že je lepší postupovat od obecného, nespecifického k tématům konkrétnějším, osobním (Tagliamonte, 2006, s. 38; obdobně Nosková, 2014, s. 51). Právě osobní otázky, zvláště pokud jsou vztaženy k minulosti (= využití reminiscence) a/nebo pokud jsou emočně vypjaté (= využití překvapení), produkují v mysli mluvčích „živé vzpomínky"; informant se díky tomu přestane soustředit na jazykovou stránku svého projevu, což je v lingvistickém výzkumu žádoucí (Labov, 1984, s. 34). Nosnost jednotlivých tematických modulů by se měla nejdříve ověřit průzkumnými dotazy; pokud zjistíme, že je informant schopný o tématu

mluvit, můžeme klást dotazy podrobnější. Jednotlivé moduly maximálně rozvíjíme (mj. pomocí pobídek typu *To je zajímavé! O tom mi ještě něco řekněte…*), rychlé skoky mezi moduly by naopak mohly narušit tok vyprávění. Modul věnovaný jazyku, pokud je plánován, by měl být zařazen až na konec rozhovoru (Tagliamonte, 2006, s. 40) – je to pochopitelné, neboť uvědomování si zvláštností vlastního jazykového projevu (nebo projevů jiných lidí, např. z nejbližšího nářečního okolí) může narušit každodenní ráz mluvy.

Otázky a také způsob jejich formulace se odvíjí od věku, pohlaví, zkušeností a zálib informanta (a také explorátora). Explorátor by měl jít „do terénu" vybaven sadou otázek, a to i v případě neřízeného rozhovoru. Tagliamonteová (2006, s. 39–40) doporučuje zejména otázky zaměřené na význam historických událostí (světových, národních, místních) a otázky demografické. Lze sem zařadit i dotazování na kulturu místní komunity, jejíž členové jsou participanty výzkumu – explorátor tak prokáže svůj zájem, navíc takové otázky mohou u některých informantů vyvolat delší mluvní projevy (např. v některých moravských regionech se nabízí dotazování na jízdu králů, královničky, stavění máje, krojové odívání ap.).

Některá témata vyžadují citlivý přístup – pokud například potřebujeme zjistit věk mluvčího, je vhodnější zeptat se ho na rok narození (a věk následně dopočítat). Totéž platí třeba v případě dotazu na nejvyšší dosažené vzdělání – pro některé mluvčí může jít o citlivou otázku, neboť v určitých obdobích a v rodinách určitých sociálních poměrů nebylo možné vzdělávat se. Dotaz lze formulovat třeba ve stylu: *Měli jste možnost chodit do školy?* nebo *Co jste dělal/a po základní škole?* (k opatrnému formulování dotazů viz Tagliamonte, 2006, s. 41). **Otázky jsou kladeny zdvořilým způsobem**, přičemž explorátor by měl být pozorný a přirozený – to jsou podle Tagliamonteové (2006, s. 45) hlavní pravidla úspěchu. K tomu lze připojit ještě citlivost, taktnost a obecně sociální zručnost.

Dodejme, že dotazy by měly být pronášeny zpaměti (nikoliv čteny z papíru) – navodí se tak ona požadovaná **přirozenost rozhovoru**. Výzkumník si může dělat **psané poznámky** (např. o tématech, která by bylo žádoucí dále rozvinout, popř. si vypisovat slova, obraty nebo tvary, které by stály za vysvětlení). V tom případě by měl participantovi výzkumu vysvětlit, k čemu tyto poznámky slouží, v opačném případě by mohlo takové počínání vyvolat negativní reakci, k tomu srov. poznámku od respondenta zapojeného do dotazníkového šetření MetRoz: „Když si něco napíšete v přítomnosti mluvčího, může ho to vystresovat a může to vypadat jako situace ve škole nebo na výslechu." Pořizování zápisků by mohlo být také mylně interpretováno jako projev nezájmu ze strany badatele, který si raději něco zapisuje, než aby pozorně poslouchal.

Méně využívanou strategií v lingvistice (avšak o to více užívanou v orální historii) je **práce s egodokumenty** (Buchner-Fuhs, 1977; Hurtworth, 2003). Jako egodokument je označován pramen soukromé povahy, např. fotografie z rodinného alba, dopisy, osobní doklady, deníky, diáře atd. Zvláště **fotointerview** (Nosková, 2014, s. 50) může u méně komunikativních mluvčích elicitovat souvislý projev při komentování výjevů na snímcích a následném vybavování si vzpomínek. Metoda má však dvě nevýhody: během probírání se fotografiemi mnohdy zaznívají citlivé údaje (jména osob, data narození ap.), které je v případě zveřejnění nahrávky/přepisu nutné anonymizovat. Druhým negativem je někdy převládající popis pomocí ostenze na úkor slovního popisu (např. *A jak vypadal rukávec u kroje? – No takhle, to vidíte tady. Takový byl.*). V takovém případě musí výzkumník informanta poučit o nutnosti podrobnější slovní deskripce. Egodokumenty v podobě písemností (vč. psaných poznámek, např. vypsaných nářečních výrazů) pak ohrožují výzkum přirozené řeči v tom smyslu, že mluvčí mohou mít tendenci z nich předčítat, čímž přepínají do jazyka psané komunikace, tj. většinově do jazyka spisovného.

Dílčí strategie, které jsme nabídli ve výše uvedených devíti bodech, mohou vést k podnícení každodenního, běžného jazykového kódu na straně informanta a k jeho celkovému uvolnění. Jednotlivé postupy lze navíc kombinovat, čímž lze docílit lepšího výsledku. Je však namístě upozornit na skutečnost, že **žádná z těchto technik není stoprocentně úspěšná**; je vždy zapotřebí přihlížet jak k potřebám mluvčích, tak k záměrům výzkumu.

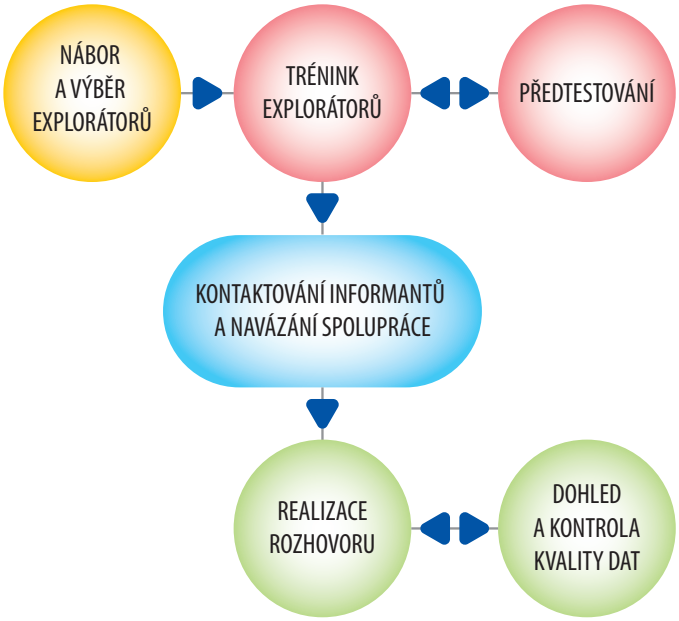
3.2.1.3 Fáze rozhovoru a strategie pro řešení různých situací

V rámci výzkumné metody rozhovoru lze vydělit pět základních částí. Jsou to:

1. předpřípravná fáze = vytipování informantů a prvotní kontakt s nimi; organizace výzkumu (domluva místa, termínu); kontrola nahrávacího zařízení;
2. přípravná fáze = úprava výzkumného místa; informovaný souhlas (vč. navázání důvěry); zahájení dokumentace;
3. vlastní výzkum = ostrý sběr zvukových dat;
4. závěrečná fáze = rozloučení s informantem; ukončení dokumentace;
5. pozávěrečná fáze = archivace a katalogizace nahrávky; uplatnění principu reciprocity.

Délka jednotlivých fází je individuální. Popsaný proces je nutno chápat jako základní, je samozřejmě možné vydělit další kroky, např. u vlastního výzkumu vyčlenit ostrou fázi sběru dat a fázi neostrou, která je potřebná pro uvolnění mluvčího.

V případě zapojení **zprostředkovaného explorátora** je ještě nutno počítat s jeho výběrem a tréninkem (včetně testovacího sběru dat; viz obrázek 3.3). Úkolem takového spolupracovníka je pak nejen samotná realizace výzkumu, ale také sjednání informantů v místě, zajištění prostor pro nahrávání atd. Rolí výzkumníka je naopak zajištění školení, kontrola techniky při sběru testovacích dat a také kontrola dat získaných při ostrém sběru. Takto rozfázovaný výzkum je podrobně popsán v pracích obecně věnovaných metodě rozhovorů (např. Singleton a Straits, 2012).



Obrázek 3.3 Fáze výzkumu při zapojení externího explorátora (čerpáno z: Singleton a Straits, 2012)

Ad 1: Předpřípravná fáze

Před zahájením předpřípravné fáze se očekává, že výzkumník má již stanoven konkrétní výzkumný záměr, tedy že má představu i o **výzkumném vzorku**. Výzkumným vzorkem se rozumí soubor informantů, kteří jsou vybráni na základě určitých kritérií (nepracuje-li badatel s náhodným výběrem). Takovými kritérii bývají v lingvistickém výzkumu faktory ovlivňující užívání jazykových prostředků, jako jsou: pohlaví/gender, věk, místo původu či bydliště, vzdělání, profese, zájmy.

V dialektologických výzkumech se tradičně přihlíží zejména k věku[16] a místu původu mluvčího (k autochtonnosti mluvčích viz 4.1.2.3). Vzhledem k tomu, že je naší intencí získat nahrávky tradičních nářečních promluv, doporučuje se k výzkumu přizvat příslušníky starších generací, přičemž platí, že čím starší mluvčí, tím lépe (ovšem neplatí to stoprocentně). V některých zahraničních výzkumech se pro **ideální nářeční mluvčí** vžila zkratka **NORMs**, což je označení pro usedlé starší venkovské muže (jde o akronym z angl. *nonmobile, older, rural males*; Chambers a Trudgill, 2004, s. 29).[17] Podobná charakteristika výzkumného vzorku se uplatňovala i v českém prostředí (srov. definici hlavních nositelů nářečí, zapojených do celoúzemního výzkumu pro *Český jazykový atlas*, během nějž byly také pořizovány zvukové záznamy mluvy – šlo o „příslušníky starousedlé vrstvy zemědělského obyvatelstva"; Balhar a kol., 1992, s. 19), kde se však za prototyp nářečního mluvčího pokládaly spíše ženy,[18] které lze analogicky označit za **NORFs**. Participantkami výzkumů se stávaly nejen díky přirozeně větší výřečnosti, ale také vyššímu věku dožití, jehož dosahují v porovnání s příslušníky opačného pohlaví. Jindy jsou ideální mluvčí charakterizováni jako „starší muži s dobrými zuby" (*The Survey of English Dialects /SED/*, 2024); je pravda, že mluvčí by neměli trpět žádnou deformací chrupu nebo poruchou řeči, takovéto záznamy promluv by mohly vést ke zkreslení vzorku.

Stanovit věkové hranice pro jednotlivé generace je problematické. Zůstaneme-li v **kategorii starší/staré generace** (nebo lépe starší dospělosti), tak při zmíněném výzkumu pro *Český jazykový atlas* byl minimální věk participantů 65 let (uvádí se, že věk mluvčích byl „65 až 75 let a výše"; Balhar a kol., 1992, s. 19), při výzkumech zahájených v r. 2023 v rámci projektu Jazyková paměť regionů České republiky (dále JAMAP) byl minimální věk ponížen na 60 let (mj. z důvodu velké úmrtnosti starších občanů zapříčiněné pandemií covidu-19). Jako výhodnější se však jeví i v rámci starší dospělosti vydělovat další kategorie; například při výzkumu vzpomínkových vyprávění starších generací byly vydělovány dvě skupiny mluvčích, a to ve věku 60 (popř. 65) až 75 let a ve věku nad 75 let (Hoffmannová a Zeman, 2007, s. 9).

Pokud by měly být součástí vzorku **nezletilé osoby** (případně osoby nesvéprávné), je nutné zajistit souhlas s jejich účastí od zákonných zástupců nebo opatrovníků.

Metod, jak **kontaktovat informanty** a přimět je k zapojení do výzkumu, je několik. Lze je oslovit:

- přímo, pokud víme o osobách vyhovujících kritériím vzorku;
- nepřímo, a to prostřednictvím námi pověřených osob, kterými mohou být jednotlivci (příbuzní, známí) nebo zástupci určité instituce (např. pracovník obecního úřadu, místní knihovny, školy, folklorního souboru, domova důchodců);
- nepřímo, a to veřejnou výzvou šířenou sociálními sítěmi, médii ap.

Zatímco u prvních dvou metod lze předpokládat, že domluvení mluvčí budou vyhovovat na ně kladeným požadavkům, v případě veřejné výzvy je nutné nejdříve zjistit základní informace o osobě, která se na výzvu ozvala, až poté domluvit realizaci výzkumu s ní; motivací těchto osob totiž může být nejen snaha napomoci v jejich očích smysluplnému výzkumu (mj. z důvodu pociťovaného závazku vůči společnosti, např. pokud se považují za jedny z posledních znalců/nositelů konkrétního nářečí), ale také touha po zviditelnění sama sebe nebo své práce (může jít např. o autory regionálních nářečních slovníčků či nářeční beletrie, lidové vypravěče ap., kteří jsou však mnohdy nositeli hyperdialektu, nikoliv běžného nářečí).

---

16 Při sběru nahrávek pro účely strojového učení je ideální získat co největší objem dat od mluvčích všech generací. Jestliže jsou k dispozici výhradně data od určité věkové skupiny (většinově od té nejstarší), pak se vyvíjené softwary (rozpoznávače dialektu, automatické transkriptory) optimalizují pro danou část populace, přičemž v případě aplikace na nahrávky s projevy mladších informantů dochází k chybovosti.

17 Termín NORMs byl zaveden už ve vydání z r. 1980.

18 Explicitně na tuto skutečnost upozorňuje třeba Kellner (1939, s. VI): „… ženy, které zůstávají doma, neodcházejí za prací mimo obec, zachovávají rázovitost v řeči a nepodléhají tak vlivu spis. jazyka." Viz též Vážný, 1955, s. 170.

PŘÍKLAD Z PRAXE

V projektu JAMAP je kladen důraz na popularizaci nářečních výzkumů. Nápomocna jsou v tomto ohledu média; o probíhající dokumentaci dialektů bylo natočeno několik reportáží Českou televizí a regionálními televizními stanicemi, také vzniklo několik rozhovorů pro různé stanice Českého rozhlasu, o akci informovaly i tištěné/online časopisy, jako je iDNES, Novinky.cz nebo Brněnský deník (podrobně viz www.jamap.cz/media). Po každé zprávě v médiích zaznamenal projektový tým zájem o jejich činnost ze strany veřejnosti, též několik pozvánek do obcí nebo přímo k jednotlivým mluvčím. Následně byl ve většině případů výzkum realizován.

Participant výzkumu by měl být už v předpřípravné fázi, kdy je s ním navazován kontakt, obeznámen s tím, jak bude rozhovor probíhat a co se od něj bude očekávat. K tomuto účelu může výzkumník užít tiskoviny (např. letáky) obsahující název projektu, v jehož rámci je sběr dat prováděn, dále cíle projektu, odkaz na web projektu nebo jméno výzkumníka (popř. vedoucího projektu), a to včetně kontaktních údajů. Příkladem této praxe jsou terénní výzkumy realizované ve výše jmenovaném projektu, leták pro jihovýchodomoravskou oblast viz příloha 2.

Po zajištění výzkumného vzorku je s informantem (popř. zprostředkující osobou) smluven **termín** výzkumu a také místo jeho uskutečnění. V obou případech platí, že se výzkumník snaží informantovi co nejvíce vyjít vstříc. Je s ním domluveno datum a čas příjezdu explorátora (včetně odjezdu/odchodu; mluvčí by měl být dopředu informován, jak dlouhý časový interval si má vyhradit pro účast ve výzkumu). Obecně platí, že není zcela vhodné zkoumat o víkendech, státních svátcích, během prázdnin (jde o období dovolených) nebo v době místních folklorních a kulturních akcí (např. hodů), kterých se informanti chtějí zúčastnit. Další obecná doporučení týkající se času:

- příslušníci starších generací v důchodovém věku: ideální jsou dopolední nebo odpolední hodiny; naopak nevhodný je obědový čas (včetně poobědové siesty) a večerní doba (kvůli únavě);
- příslušníci středních generací (nebo starších generací, pokud ještě patří mezi pracující): ideální je pozdější odpoledne (po návratu z práce);
- příslušníci mladších generací: ideální je odpoledne (po návratu ze školy).

Délka rozhovoru se odvíjí od výzkumného záměru, též od schopnosti informanta hovořit pokud možno souvisle a od jeho ochoty. Při výzkumech realizovaných pracovníky dialektologického oddělení ÚJČ AV ČR byly v minulosti pořizovány nahrávky zpravidla o délce kolem 20 minut, důvodem však bylo šetření nahrávacím materiálem (konkrétně cívkovými kotouči, později páskovými kazetami; viz Šimečková, 2024a). V současnosti mívá rozhovor s jedním mluvčím kolem 1,5 hodiny délky, podle J. Noskové (2014, s. 50) by neměl přesáhnout délku 2 hodin. Platí ale, že pokud by došlo k fyzickému nebo psychickému vyčerpání informanta, je nutné rozhovor ukončit předčasně a případně si domluvit jiný termín schůzky.

Co se týče **místa**, v minulosti se nahrávalo převážně v příbytku mluvčího. **Domácí prostředí** v informantovi evokovalo větší pocit bezpečí, a tak u něj bylo snazší přejít do každodenního jazykového kódu. V současné době však mají lidé obavu vpouštět cizí osoby do svých domovů, a tak se výzkum mnohdy přesouvá do prostor obecních úřadů, knihoven, škol ap. V domácnostech nyní probíhají výzkumy zpravidla na žádost mluvčího, např. kvůli jeho snížené pohyblivosti.

Specializované výzkumy, zvláště počítající s využitím nahrávek pro přístrojové fonetické analýzy, se mohou odehrávat ve zvukotěsných místnostech, které jsou součástí **lingvistických laboratoří**. Hrozí však nebezpečí, že mluvčí nebudou s to v takovýchto podmínkách mluvit přirozeným způsobem; tehdy je nutno využít různé strategie, zejména ty zmíněné v souvislosti s navozením pocitu důvěry mezi explorátorem a infor-

mantem. Jiným důvodem pro odmítnutí participace na takovémto výzkumu je nemožnost mluvčího na takovéto místo se dopravit; tento problém odpadá v případě zapojení **pojízdné laboratoře** (srov. výzkum nizozemštiny pomocí mobilního karavanu v rámci projektu SPRAAKLAB, viz Wieling a kol., 2023).

PŘÍKLAD Z PRAXE

Při výzkumu výslovnosti hovorové nizozemštiny, který v letech 1995 a 1996 vedla psycholingvistka M. Ernestusová, byly rozhovory nahrávány ve zvukotěsné místnosti ve Fonetické laboratoři Amsterdamské univerzity. Pro uvolnění mluvčích bylo využito zajímavé strategie, a to předmětu zútulňujícího nahrávací prostor: „Next to each microphone there was a potted plant, whose function was to take the speakers' minds off the microphones, and to make the room somewhat less austere, which was necessary, as the poor lighting, the foam rubber on the walls, and the grid on the floor gave the room a gloomy atmosphere."[19] (Ernestus, 2000, s. 98)

Posledním krokem přípravné fáze je **kontrola nahrávacího zařízení**. Doporučuje se připravit si zařízení dvě, jedno pro ostré nahrávání, druhé jako záložní (pro případ defektní nahrávky pořízené z prvního zařízení, popř. vůbec poruchy hlavního přístroje znemožňující snímání zvuku). Jestliže jde o zařízení nové, měl by se s ním explorátor důkladně obeznámit – nastudovat si manuál a také vyzkoušet zařízení nanečisto se simulací různých situací (Clemente, 2008, s. 183).

Je nezbytné zkontrolovat stav baterie a také kapacitu volné paměti. Pokud je diktafon (nebo některé z jeho externích příslušenství) napájen vyměnitelnými bateriemi, měly by se v případě vybití vyměnit a také by se měly připravit baterie náhradní (pro případ nouze by měly být vždy po ruce). Pokud diktafon umožňuje nabíjení z elektrické sítě, pak by si měl výzkumník s sebou přibalit adaptér, popř. notebook, plánuje-li dobíjení z něj (v tomto případě pozor na hlučné chlazení procesoru, které by mohlo kvalitu nahrávky znehodnotit). Obdobnou kontrolu vyžaduje paměť nahrávacího zařízení, obzvlášť pokud jsou pořizované soubory v nekomprimovaném formátu WAV. V případě nedostatku paměti je nezbytný přesun uložených nahrávek do jiného úložiště a následný výmaz z diktafonu, popř. vložení nové paměťové karty. Podrobně k výběru a nastavení nahrávacího zařízení viz 3.2.4.

Ad 2: Přípravná fáze

Přípravná fáze se týká té části výzkumného procesu, kdy je explorátor již připraven na místě, v němž má výzkum proběhnout (předpokládá se samozřejmě, že výzkumník přijde včas – nechat informanta čekat by bylo neslušné). Část přípravných úkonů se přitom odehrává bez přítomnosti informanta (i proto je vhodné být na místě s předstihem) a část po jeho příchodu. Jde konkrétně o tyto nezbytné činnosti:

- před příchodem informanta: příprava nahrávacího stanoviště a nahrávacího zařízení;
- po příchodu informanta: uvítání se s mluvčím a vzájemné představení se (pokud nejde o osobu blízkou), představení výzkumných cílů (pokud není přistoupeno k úmyslnému zatajení záměru, např. při výzkumu specifických hláskových či tvaroslovných jevů) a popis průběhu výzkumu; zajištění informovaného souhlasu.

Příprava prostoru, v němž proběhne nahrávání, je prerekvizitou pro pořízení kvalitní nahrávky. Ideální podmínky z hlediska nároků na prostor lze zajistit pouze v případě, že nahráváme ve zvukotěsné místnosti nebo jinak speciálně upravené laboratoři (třeba i pojízdné). Pokud takový prostor k dispozici nemáme

19 „Vedle každého mikrofonu byla rostlina v květináči, jejímž úkolem bylo odvést pozornost účastníků od mikrofonů a učinit místnost o něco méně strohou, což bylo nutné, protože špatné osvětlení, pěnová guma na stěnách a mřížka na podlaze dodávaly místnosti ponurou atmosféru." Překlad M. Š.

(nebo nechceme mít ve snaze o přirozenější průběh rozhovoru), je vhodnější upřednostnit interiér před exteriérem (např. nahrávání na zahradě může být rušeno zpěvem ptáků, štěkáním psa, hlukem sousedovy sekačky ap.). V případě interiéru se dává přednost soukromým prostorům (např. u mluvčího doma nebo ve speciálně vyhrazené místnosti na obecním úřadě, knihovně ap.) před veřejnými prostory (jako je třeba kavárna, restaurační zařízení). Pokud je možnost zvolit konkrétní místnost, tak se doporučuje menší prostor bez ozvěny; tomuto požadavku zpravidla nevyhovují velké obřadní síně v obecních úřadech, při výzkumu doma je naopak optimální obývací pokoj, a to z důvodu členitého povrchu a menšího počtu zařízení produkujících hluk (Tagliamonte, 2006, s. 45).

PŘÍKLAD Z PRAXE

Výzkumník by měl mít na paměti, že zatímco v reálu může být rozhovor probíhající na hlučném pozadí snadno uchopitelný, při pozdějším zpracování nahrávky (zvláště při pořizování jejího přepisu) představují tyto rušivé zvuky někdy až nepřekonatelný problém. Takovým příkladem je nahrávka pořízená středoškolskou studentkou v rámci soutěže Staň se superdialektologem v r. 2024, přičemž jako místo pro nahrávání byl zvolen domov důchodců. Bohužel nebyl zajištěn privátní pokoj, a tak vstupy různých osob do rozhovoru a hluk v pozadí znehodnotily celý audiozáznam, jak se lze přesvědčit z nahrávky dostupné přes QR kód.

Pokud můžeme zasahovat do prostoru, v němž probíhá zvuková dokumentace, pak se snažíme:

1. eliminovat hluk (přicházející zvenku i zevnitř);
2. eliminovat hladké povrchy;
3. optimalizovat polohu nahrávacího zařízení vůči usazení informanta.

Ad 1: Eliminace hluku

Minimalizace hluku je samozřejmostí. Jak bylo zmíněno, nahrávání uvnitř je vhodnější než venku. Ovšem i ve vnitřních prostorech „číhá“ řada hlučných objektů, jako je např. vrzající židle či podlaha, jejichž rušivého zvuku si možná explorátor zpočátku nevšimne, avšak při zpracovávání nahrávky je evidentní. U některých spotřebičů lze informanta požádat o vypnutí (rádio, televize, počítač s hlučným chladičem, vyzvánění na mobilním telefonu, větrák, klimatizace aj.), u jiných to možné není, a tak je vhodné být od nich v co největší vzdálenosti (lednička, nástěnné hodiny, akvárium). Nemělo by se zapomenout na zavření oken a dveří, aby nebylo nahrávání narušováno děním venku.

Ad 2: Eliminace hladkých povrchů

Hladké povrchy způsobují nechtěnou rezonanci produkovaného zvuku – ten se od nich odráží jako od zrcadla. Z toho důvodu se doporučuje volit místnosti s minimem hladkých povrchů (Zíková a Křivan, 2014, s. 74) – třeba zmíněný obývací pokoj je ideální, obzvlášť pokud je v něm koberec a dostatek nábytku (např. velká knihovna). Pokud takovou místnost nemáme k dispozici, posadíme mluvčího alespoň dále od okna nebo holé stěny, popř. hladkou stěnu snadno upravíme – pověsíme před ni deku. K této úpravě lze samozřejmě přistoupit pouze v případě, že zkoumáme třeba na obecním úřadě nebo v soukromí u informanta, kterého dobře známe a který nám danou úpravu dovolí; stěží budeme přeskládávat nábytek nebo věšet deky u cizí osoby doma či v restauraci.

M. Zíková a J. Křivan (2014, s. 73–74) zdůrazňují též zvláštní péči o povrch, na němž je položeno nahrávací zařízení – ani tento povrch (zpravidla stůl) by neměl být hladký, protože by se od něj přílišně odrážely doprovodné zvuky, jako je poklepávání prstů nebo pokládání šálků na stůl. Doporučují tak podložení rekordéru šátkem, což případné rezonance zmírní, nebo použití stativu.

Ad 3: Optimalizování usazení informanta vůči poloze nahrávacího zařízení

Pokud užíváme stolní mikrofon nebo mikrofon zabudovaný do nahrávacího zařízení, měl by být umístěn co nejblíže k informantovi, ideálně ve vzdálenosti 50–70 cm (Zíková a Křivan, 2014, s. 73). Kromě vzdálenosti je nutné zkontrolovat také orientaci mikrofonu – měl by směřovat směrem k nahrávané osobě (viz 3.2.4). Pro zajištění kvalitního zvuku lze využít též mikrofon klopový nebo headsetový, u nichž odpadá starost s kontrolou vzdálenosti. Diktafon by měl být připraven na nahrávací pozici od začátku výzkumu, aby měl informant dostatek času se s ním seznámit – postupně si na jeho přítomnost zvykne, mnohdy na něj informanti dokonce zapomenou (důkazem je skutečnost, že po skončení rozhovoru mnohokrát odcházejí i s klopovým mikrofonem přichyceným na jejich oděvu).

Je-li prostor připraven, zbývá už nahrávací zařízení (včetně příslušenství) nastavit na nahrávání a otestovat jeho funkčnost. Nahrávání zahajujeme ideálně ihned po příchodu informanta. V tento moment začíná fáze, kterou lze označit za neostrou část výzkumného procesu – s mluvčím je totiž navázána úvodní komunikace, avšak zapotřebí je jednak získat od něj informovaný souhlas k nahrávání a následnému využití nahrávky k výzkumným účelům, jednak ho zbavit nervozity, a tím i uvolnit, co se týče stylu promluvy.

Setkání s informantem probíhá v této posloupnosti:

- formální uvítání se s příchozím, popř. jeho doprovodem (ideální je doprovázející osobu poté přimět k odchodu, neboť by mohla narušovat průběh rozhovoru);
- představení výzkumníka informantovi = uvedení jména, profesního zařazení, instituce;
- představení výzkumného účelu (není-li záměrně zatajován) = objekt zájmu, způsob nakládání s osobními údaji a nahrávkami (vč. uložení), druhy plánovaných výstupů materiálově čerpajících z rozhovoru;
- vysvětlení procesu udělení souhlasů a anonymizace;
- upozornění na eliminaci rušivého chování (např. poklepávání prsty na stole nebo hůlkou, listování knihou);
- upozornění na styl promluvy = vybídnutí ke každodenní, popř. nářeční mluvě (během rozhovoru lze vybízet opakovaně, avšak jen přiměřeně); vyzvání k nepřipravenému, nikoliv připravenému (či dokonce čtenému) projevu;
- obeznámení informanta s nahrávacím zařízením, popř. jeho příslušenstvím (vč. svolení k uchycení externího mikrofonu, pokud je užíván);
- vyjádření informovaného souhlasu, popř. podepsání příslušného dokumentu (podrobně viz 3.2.5);
- zjištění základních sociolingvistických údajů = nejčastěji rok narození, rodiště/bydliště, rodiště/bydliště rodičů (popř. prarodičů nebo osob, se kterými informant vyrůstal), nejvyšší dosažené vzdělání, zaměstnání (vč. případných předchozích zaměstnání), zájmy;
- krátký zkušební rozhovor.

Pokud je u informanta zjištěna výrazná nervozita, lze využít některé strategie, které již byly popsány (viz 3.2.1.2). Mnohdy pomáhá ujišťování, že probíhající komunikace není žádným zkoušením, nýbrž rozpravou. Stimulem pro uvolnění je také **povzbuzování a chválení** ve stylu: *No vidíte, jak to všechno pěkně vyprávíte! Vy si toho tolik pamatujete! To je úžasné, jak si ještě pamatujete nářečí.*

PŘÍKLAD Z PRAXE

Je otázkou, jakým způsobem se má výzkumník představit. Pokud jde třeba o vědeckého pracovníka dialektologického oddělení ÚJČ AV ČR, nabízí se možnosti představit se jako lingvista/jazykovědec, dialektolog, pracovník konkrétní instituce (včetně jejího plného oficiálního názvu), akademik, vědec, výzkumník, popř. zájemce o nářečí. Konkrétní forma vyplývá z konkrétní situace – zatímco méně vzdělanému informantovi některá označení nic neřeknou (např. *dialektolog*, *lingvista*), u vzdělanějšího mluvčího je užít můžeme, popř. je můžeme uvést s vysvětlením (*jsem dialektolog, to znamená, že se zabývám tím, jak se kde mluví*). Výzkumník může někdy svou pracovní pozici nebo příslušnost k instituci záměrně tajit, ovšem v tom případě musí vymyslet nějakou zástěrku (např. zájemce o nářečí), která však musí být důvěryhodná; v opačném případně by mohlo vyvstat podezření z nekalého úmyslu. Dnes již historickým příkladem je vzpomínka dialektologa J. Balhara na výzkum z počátku kolektivizace: „Terénní výzkum (...) nesl v různých dobách různé problémy. Když jsem shromažďoval materiál pro diplomovou práci, začala se u nás provádět násilná kolektivizace. Lidé na vesnici byli tehdy nedůvěřiví a ani starší lidé neměli moc času vysedávat a sloužit jako informátoři." (Goláňová, 2009, s. 26)

Ad 3: Vlastní výzkum

Jakmile jsou vyřízeny formality (obeznámení informanta s průběhem výzkumu, informovaný souhlas, popř. krátká zkušební rozprava), může se přejít k ostrému výzkumu. Již bylo zmíněno, že jde o nápodobu přirozeného rozhovoru – explorátor pokládá otázky, na které informant odpovídá, přičemž je žádoucí, aby byly odpovědi jednak co nejdelší (jde nám o sběr souvislých promluv), jednak aby byly užívány přirozené, každodenní jazykové prostředky (nikoliv prostředky spisovné, popř. hyperdialektické). **Úkolem explorátora** tak je:

- klást otázky, též promýšlet jejich obsah a návaznost, formulovat je vhodným způsobem a za pomoci různých taktik (např. strategie emočně vypjatých vyprávění, strategie překvapení ap.) se snažit rozvíjet samostatnou promluvu informanta;
- soustředit se nejen na jazyk, ale i na obsah sdělení (explorátor by měl mít přehled, o čem mluvčí momentálně vypráví, a při tom promýšlet rozvíjení tématu; výzkumník, který je roztržitý a nedává pozor, na informanta nepůsobí dobrým dojmem, srov. nevhodnost otázek typu *Á, kde jsme to vlastně přestali? O čem jsme se to bavili?*);
- vybízet informanta ke každodennímu jazykovému modu; povzbuzovat ho ve vyprávění a chválit ho;
- minimalizovat vlastní zvukové projevy, a to včetně hezitačních zvuků – kladené otázky by měly být krátké, informantovi by se nemělo vpadat do řeči (není-li to nezbytně nutné)[20] a namísto verbálních prostředků vyjadřujících přitakání obsahu či postoj obecně (*ano*, *no ba*, *hm*, *skutečně?*, *no vidíte*, *to je pravda*) je vhodnější užívat gesta nebo mimiku (ovšem přirozeně, nikoliv strojeně);
- eliminovat rušivé zvuky a projevy;
- kontrolovat nahrávací zařízení, tj. opakovaně ověřovat, zda běží nahrávání, zda se baterie nevybíjí nebo zda jsou hodnoty zvuku (v případě jejich vykreslování na displeji rekordéru) ve správné míře.

20 Někdy je však těžké odhadnout, zda informant svou výpověď již dokončil, či nikoliv. Z toho důvodu lze doporučit respektování přirozené pauzy za replikou informanta, a to pro případ, že by chtěl k řečenému ještě něco sám doplnit bez pobízení.

PŘÍKLAD Z PRAXE

Uvádíme dva příklady, z nichž první ilustruje chybně vedený rozhovor, v jehož průběhu výzkumník neustále přerušuje promluvu informanta. Jde o výzkum realizovaný v r. 1969 v Otaslavicích na Prostějovsku, explorátorem tu byl F. Matějek. Druhá nahrávka, prezentující správný postup při kladení otázek, byla získána v r. 2024 v Šumvaldě na Olomoucku. Dialektoložka M. Šimečková nechává informantovi dostatek prostoru pro vyjádření se a rozvíjí téma na sebe navazujícími doplňkovými otázkami. Nahrávky jsou dostupné přes QR kód vlevo a QR kód vpravo.

Ad 4: Závěrečná fáze

Závěr výzkumu spočívá v poděkování informantovi za zapojení se do výzkumu, zodpovězení případných dotazů ohledně zpracování dat a výstupů výzkumu a rozloučení. Totéž se vztahuje na osoby, které výzkum zprostředkovaly. Lze se případně domluvit na opakované návštěvě, popř. spolupráci s obcí (viz bod 5).

Nahrávání lze ukončit buď na konci vlastního výzkumu, nebo až v samotném závěru (o neostrosti začátku a konce výzkumu viz Lindbloom, 2004, s. 91). Z praxe lze doporučit spíše druhou variantu, neboť mluvčí někdy teprve na konci (třeba po rozloučení) začne rozvíjet některá témata, na která si dříve nemohl vzpomenout, případně o nich nebyl ochoten hovořit, neboť zpočátku neměl důvěru k explorátorovi. Někdy je otevřenost způsobena přesvědčením, že explorátor již nahrávání ukončil; je tak nutné znovu upozornit na probíhající dokumentaci. Informant někdy projeví přání tyto pasáže, značně neformální jazykově i obsahově, vypustit z nahrávky, popř. omezit jejich využití (např. zakázat případné zveřejnění); v tom případě se očekává, že mu bude vyhověno.

Rozhovor lze také ukončit předčasně, a to ze strany výzkumníka i informanta. Důvodem může být fyzické či psychické vyčerpání, naléhavá událost, vzájemné neporozumění mezi účastníky výzkumu, neochota či nekomunikativnost informanta, případně nemožnost přepnutí jazykového kódu směrem k běžnému, každodennímu vyjadřování. Pokud impulz vzejde od explorátora, nemusí na skutečnost, že jde o ukončení předčasné, upozorňovat.

Ad 5: Pozávěrečná fáze

Pozávěrečnou fázi lze vést ve dvou směrech, a to:

- směrem ke zpracování nahrávek;
- směrem k participantům výzkumu.

První bod znamená, že je po získání nahrávek nezbytné tyto audiální dokumenty (dnes již výhradně v digitalizovaném formátu) popsat a uložit, archivovat a katalogizovat, též v celé délce nebo pouze vybrané úseky transkribovat. Tyto kroky podrobně popisujeme v podkapitolách 3.3 a 3.4.

Druhý bod, na který se v této části zaměříme, spočívá v **odpovědnosti vůči jazykovému společenství** (např. Wolfram, 1993;[21] Milroyová a Gordon, 2012, s. 91–94; Braber a Davies, 2013, s. 104). Znamená to, že explorátor, který zkoumal v určité komunitě (u obyvatel konkrétní obce nebo její části, členů folklorního spolku ap.), by měl nějakým způsobem tomuto společenství oplatit ochotu zapojit se do výzkumu nebo výzkumu vůbec pomoci (organizačně, zapůjčením prostor k nahrávání ap.). Způsobů návratnosti je několik, výběrově:

- věnování nahraných rozhovorů (zaslaných virtuálně, uložených na USB flash disku nebo vypálených na CD) informantovi, popř. komunitě, projeví-li o ně zájem (věnování komunitě je možné pouze za podmínky, že k tomu svolí participanti výzkumu);
- publikační výstup (studie, regionální slovník, videodokument, popularizační brožura) shrnující výsledky výzkumu a jeho postoupení společenství (= věnování určitého množství výtisků informantům, obecnímu úřadu, místní knihovně);
- publikování zprávy o výzkumu v regionálním periodiku (např. v regionálním zpravodaji vydávaném obcí);
- uspořádání workshopu nebo přednášky (pro žáky/studenty místní školy nebo veřejnost) se shrnutím výsledků výzkumu a s edukačními cíli;
- zapůjčení nebo uspořádání výstavy v místě, popř. vytvoření expozice věnované jazyku místní komunity pro místní muzeum;
- pozvání informantů a jiných pomocných osob na exkurzi do výzkumného pracoviště s ukázkou zpracování získaných dat.

Některé z těchto aktivit mohou mj. **přispět k posílení jazykové tolerance a k akceptaci jazykové diverzity**, a to zejména pokud je navázána spolupráce s místními školami (Wolfram, 1993).

Sociologové M. Hammersley a P. Atkinson (2007, s. 218) v souvislosti s etnologickým výzkumem výslovně píší o reciprocitě ze strany explorátorů ve formě služeb nebo **finanční odměny**. U druhé možnosti je otázka, nakolik by taková odměna měla být vysoká. V případě vyplácení odměn je zapotřebí mít na to naplánované peníze v rozpočtu projektu či instituce; hrozí však, že si mluvčí obecně navyknou na tento druh odměny, a tak u případných budoucích výzkumů bez finanční podpory budou odmítat účast (k důsledkům pro budoucí výzkum viz 3.2.5). Vyplácení odměny také může vést k nežádoucímu posílení hierarchicky nadřazené pozice explorátora (Havlíková, 2004) nebo k mylné představě na straně informanta ohledně výzkumníkova očekávání, vedoucí k hyperdialektickému projevu. Finanční benefit může nadto přilákat participanty, kteří sice nesplňují základní požadavky na ně kladené, avšak kvůli vidině zisku některé skutečnosti zapřou, což může vést k chybovosti výzkumného vzorku, tedy i takto získaných dat.

Opomenout nelze skutečnost, že pro informanta může mít nemalý přínos už rozhovor samotný – zvláště osamělí jedinci tak mají možnost svěřit se se svými problémy nebo životním příběhem naslouchající osobě, proud řeči tak má mnohdy přímo **katarzní účinek**, což může být také pokládáno za formu reciprocity.

21 Wolfram pro tuto strategii používá pojmenování *the principle of linguistic gratuity* (Wolfram, 1993, s. 227).

PŘÍKLAD Z PRAXE

J. Chromý se ve své knize o protetickém *v-* zmiňuje, že peníze pro nahrávané osoby v projektové žádosti neplánoval, což prý „nebylo soudné“ (Chromý, 2017, s. 145). Následně se však daný druh odměny pokusil uplatnit, patrně s financováním z institucionálního rozpočtu nebo z vlastních zdrojů, ovšem ani tato motivace k účasti na výzkumu nebyla úspěšná: „Z mojí zkušenosti však vyplývá, že ani finanční pobídka nezaručuje souhlas s účastí ve výzkumu.“ (tamtéž, s. 136)

V letech 2022 a 2023 proběhl v obci Prušánky na Hodonínsku výzkum obalovaného *l* pod vedením M. Šimečkové. Výsledkem reciprocity byly dva příspěvky v obecním zpravodaji, dvě přednášky pro místní školu, příslib zapůjčení putovní výstavy do prostor obecního úřadu a zapojení místních obyvatel do natáčení reportáže pro Českou televizi. Do budoucna je plánována další spolupráce s Muzejním spolkem Prušánky.

3.2.2 Zprostředkovaný sběr audiálních dat cestou citizen science

Vlastní sběr nahrávek je sice optimální způsob, jak získat kvalitní data, jde ale také o velice časově, finančně a personálně (mj. psychicky a fyzicky) náročnou činnost, obzvlášť pokud je zapotřebí shromáždit velký objem zvukových záznamů. Z toho důvodu se lze zamyslet nad využitím veřejnosti, a to cestou **citizen science neboli občanské vědy**.

V současnosti jde o oblíbenou formu spolupráce mezi vědeckou komunitou a laiky, kteří se nadchnou pro výzkumný projekt a aktivně se na něm podílejí. Základní principy občanskovědních projektů vymezila asociace ECSA (European Citizen Science Association) v dokumentu nazvaném *Ten Principles of Citizen Science* (2015; v čes. překladu *Deset principů občanské vědy*, 2016).

V souvislosti s daty nasbíranými zapojenou veřejností je závazný bod 7: „Výzkumná data a metadata občanské vědy jsou veřejně přístupná a výsledky jsou, pokud je to možné, publikovány s otevřeným přístupem.“ Výzkumník, který by chtěl využít veřejnost k nabytí dat, tak musí mít na paměti, že by měl tato data také **veřejně sdílet**, třeba v podobě otevřeného zvukového archivu nebo zvukové mapy.

S tím však souvisí nutnost tato data, jsou-li zveřejněna, ohlídat z hlediska **etických principů** (viz bod 10 v téže úmluvě: „Vedoucí projektů občanské vědy dbají na právní a etické aspekty týkající se autorského práva, ochrany práv duševního vlastnictví, smluv o sdílení dat, důvěrnosti dat, uvádění autorství a dopadů aktivit na životní prostředí.“). Veřejnost, která zprostředkuje a nahraje rozhovor, by měla být výzkumníkem dobře informována nejen o výzkumném procesu (tedy co se zkoumá a jak na to), ale také o získání informovaného souhlasu ze strany nahrávané osoby (viz 3.2.5.1). Součástí tohoto souhlasu by mělo být bezpodmínečně i svolení se zpřístupněním zvukové nahrávky (nebo jejích úseků), dále by měl být informantovi vysvětlen způsob ochrany osobních údajů, anonymizace atd. Veřejnost by měla mít k dispozici podrobný dokument, v němž by se mohla (i zpětně) obeznámit s veškerými náležitostmi výzkumu, a to včetně oněch etických pravidel.

PŘÍKLAD Z PRAXE

Příkladem fungujícího sociálněvědního projektu je aktivita pro školáky *Staň se superdialektologem*, kterou od r. 2023 vyhlašuje pod vedením M. Šimečkové dialektologické oddělení ÚJČ AV ČR ve spolupráci s Vysokým učením technickým v Brně a Univerzitou Palackého v Olomouci. Cílem soutěže je zapojení veřejnosti do dokumentace nářečí a běžné mluvy na území České republiky, a to cestou pořizování zvukových záznamů. V prvním roce se soutěže zúčastnilo 64 školáků z 21 škol, do soutěže bylo přijato 114 nahrávek s celkovou stopáží přes 27 hodin. Část dat (149 jazykově a obsahově reprezentativních úseků, sestříhaných z došlých záznamů o celkové délce přes 8 hodin) byla zveřejněna na projektovém webu jamap.cz v podobě audiomapy (Šimečková a kol., 2024). S dalšími ročníky bude tato mapa obohacována novými nahrávkami.

Dalším příkladem z českého prostředí je nářeční fonotéka, což je hlavní výstup projektu JAMAP, jehož zveřejnění je naplánováno na r. 2027. Počítá se zde nejen s otevřením původního zvukového archivu dialektologického oddělení ÚJČ AV ČR, ale také s doplněním o nahrávky nové, a to cestou komunitního plnění. Veřejnost tak bude moci obsah zvukové mapy rozšiřovat, a zaplnit tak mj. některá „bílá“ místa na mapě. U přijatých dat bude samozřejmě nutná kontrola z hlediska jejich obsahu (a to včetně osobních a citlivých údajů).

3.2.3 Zapojení již existujících sbírek do výzkumu

Možnou cestou k audiálním datům jsou již existující sbírky, které jsou ve vlastnictví institucí nebo jednotlivců. V úvahu připadají (výběrově) tyto možnosti:

- zvukové archivy (popř. jednotlivé nahrávky) budované napříč různými vědními obory – kromě lingvistiky jsou rozhovory využívány v historii (v rámci orálněhistorických výzkumů), etnologii/etnografii, antropologii, sociologii, psychologii, lékařství aj.;
- zvukové archivy (popř. jednotlivé nahrávky) jsoucí ve vlastnictví ne/veřejnoprávních médií – např. televizní/rozhlasový fond;
- zvukové nahrávky shromažďované jinými institucemi – např. školami (pro účel výuky např. v návaznosti na regionální ráz edukace), obcemi (jako audiální součást obecních kronik), folklorními spolky aj.;
- zvukové nahrávky ve vlastnictví jednotlivců (zejm. laických zájemců o jazyk a jeho regionální podobu).

Některé z uvedených zdrojů jsou veřejně přístupné (formou korpusů, archivů, doprovodného materiálu k publikačním výstupům různého typu), jiné nikoliv. V obou případech je možné požádat vlastníka dat o jejich poskytnutí k vědeckovýzkumným účelům, popř. o zařazení do vlastního archivu (s přesným stanovením nově nabytých práv k datům a ke způsobu jejich využití, popř. dalšího zveřejňování). Tajné využití dat, nedovoluje-li to přímo typ zdroje (např. korpus otevřený pro využití dat k lingvistickému bádání, ovšem s příslušnou citací čerpání), je nepřípustné. Některé zdroje je možné použít za úplatu (např. archiv České televize).

Nevýhodou takto čerpaných dat (a též dat získaných cestou citizen science, viz 3.2.2) je jejich **rozdílná kvalita**. Může jít o data nevyhovující jazykově, obsahově či po technické stránce, mnohdy u nich přitom **schází důležité údaje** týkající se informantů (Braber a Davies, 2016). Také využití dat k jinému účelu, než byla sesbírána, představuje **etický, někdy i legislativní problém** (Bornat, 2003).

PŘÍKLAD Z PRAXE

Dialektologické oddělení ÚJČ AV ČR ve snaze o rozšíření zvukového archivu od r. 2022 intenzivně šíří výzvu směrem k veřejnosti týkající se darování nářečních záznamů. Může jít jak o digitalizované soubory, tak o nahrávky na starých zvukových nosičích, u nichž je pak zajištěna profesionální digitalizace (mnohdy jsou tak doslova zachráněny cenné záznamy, které by jinak byly zničeny kvůli opotřebování nahrávacího materiálu). Tyto nahrávky se stávají pevnou součástí dialektologického fondu, a v budoucnu tak mohou být (svolí-li k tomu dárce) zpřístupněny v nářeční fonotéce. Nahrávky jsou darovány jak ze strany laické veřejnosti (např. nedávno byl fond rozšířen o 12 kazet věnovaných aktivním folkloristou z Topolné na Uherskohradišťsku), tak vysokoškolskými institucemi (např. katedrou českého jazyka Filozofické fakulty Ostravské univerzity).

3.2.4 Metody a techniky nahrávání

Zásady a techniky nahrávání podrobně popsali M. Zíková a J. Křivan (2014), z jejich textu většinově vychází tato část metodiky. Jsou zde uvedena základní ponaučení, pro podrobný návod např. ohledně nastavení zvukového zařízení odkazujeme právě na tuto studii.

V současnosti se ke zvukové dokumentaci užívají výhradně **digitální rekordéry**, přičemž pro kvalitu nahrávky po technické stránce je stěžejní jednak mikrofon, jednak **nastavení samotného přístroje**, které se liší v závislosti na okolnostech. Doporučuje se nahrávání při 48 kHz (Clemente, 2008, s. 182; Zíková a Křivan, 2014, s. 68) a hloubce 24 bitů (Zíková a Křivan, tamtéž), přičemž ve starších pracích je uváděna i hodnota nižší, totiž 16 bitů (např. Clemente, 2008, tamtéž). Zásadní je nastavení hlasitosti vstupu (= inputu), které je nutné v průběhu nahrávání kontrolovat, aby nedošlo k „přepálení" nahrávky (tj. k bodu, kdy se input dostane do červených hodnot a zvuková vlna je nenávratně zdeformována). Mluvčí totiž mají ve zvyku postupně zesilovat svůj mluvený projev, jakmile z nich opadne prvotní nervozita. Podle Clemente (2008, s. 184) by měl mít výzkumník v terénu po ruce sluchátka pro občasné monitorování kvality nahraného zvuku; v optimálním případě by tak byli na místě dva explorátoři, přičemž jeden by se věnoval informantovi, druhý nahrávacímu zařízení.

Mikrofon může být buď interní (tj. zabudovaný v nahrávacím zařízení), nebo připojený externě. V případě interního mikrofonu je možné zvolit mikrofon všesměrový (snímající více mluvčích najednou), nebo směrový (určený pro jednoho mluvčího, snímající pouze část prostoru). Výhodou směrového mikrofonu je to, že nezachycuje hluk z okolí. K umístění interního mikrofonu viz 3.2.1.3.

V případě použití externího mikrofonu (nebo mikrofonů) se rozlišuje mikrofon kondenzátorový, disponující fantomovým napájením vhodnějším pro záznam mluvy, a mikrofon dynamický, který byl vyvinut zvláště pro zachycení zpěvu a hudby. I u externích mikrofonů hraje velkou roli, zda jde o přístroj směrový, či všesměrový. Rozlišují se tři základní typy **externích mikrofonů**:

- ruční = explorátor ho během rozhovoru drží v ruce, což může být nepohodlné v případě delších rozhovorů. Doporučuje se používat ho společně s molitanovou ochranou, která mikrofon chrání nejen před nárazy větru, ale třeba i při výraznější výslovnosti exploziv (Zíková a Křivan, 2014, s. 78);
- klopový = přichycený na oděv informanta. Výhodou je větší komfort pro explorátora i informanta (často zapomene, že má mikrofon připojen) a kvalita takto získaného zvuku. Je skvělým pomocníkem zvláště u mluvčích, kteří mluví potichu, „polykají" slova, mumlají nebo si při mluvení nevědomky zakrývají ústa (Ritchie, 2015, s. 46). Nevýhodou je naopak zachycení možných rušivých zvuků (např. dotyk ruky, tření textilu), případně shození nebo posunutí (např. při úpravě šátku). Clemente (2008, s. 183) též upozorňuje na rušení jiných elektronických zařízení a magnetických polí, též na ztrátu přenosu kvůli přílišné vzdálenosti mezi vysílacím a přijímacím zařízením v případě bezdrátového klopového mikrofonu;

- headsetový = připojený ke sluchátkům nebo samostatný, nasazený za uši (užívá se též označení náhlavní mikrofon); umožňuje vysokou kvalitu zvuku, podle Zíkové a Křivana (2014, s. 78) je „ideální pro detailní zvukovou analýzu". Nespornou nevýhodou je však to, že si je jeho přítomnosti informant více vědom, a tak jeho projev nemusí být zcela přirozený.

Ideálně se záznam pořizuje v **nekomprimovaném formátu (WAV)**. Dříve se hojně užívaly formáty MP3, MP4, které jsou však ztrátové. Nevýhodou formátu WAV je velký objem souboru, s čímž je nutno počítat při ukládání a zálohování dat.

K nahrávání se užívají také **mobilní telefony** nebo **tablety**, ve kterých je zabudován záznamník zvuku; i v tomto případě platí, že by mělo jít o wavovou nahrávku a že by měl být ideálně použit kvalitní externí mikrofon.

PŘÍKLAD Z PRAXE

V minulosti se při dialektologických výzkumech užívaly cívkové magnetofony, dnes již legendární je magnetofon Tesla Sonet Duo. Zvuk byl snímán prostřednictvím externího ručního mikrofonu (k historii sběru nářečních nahrávek viz Šimečková, 2024a). Kvalita těchto historických nahrávek, které byly později digitalizovány, je různá v závislosti na různých okolnostech, u většiny z nich je však slyšitelná horší kvalita zvuku. Pro porovnání si lze poslechnout jednu historickou nahrávku, pořízenou r. 1968 při výzkumu v Zápech v okrese Praha-východ (viz QR kód vlevo), a jednu novou nahrávku, pořízenou již moderním digitálním rekordérem v r. 2024 v Radíkově, dnešní části Olomouce (viz QR kód vpravo). Rozdílná kvalita je tu evidentní.

3.2.5 Právní a etický rámec pořizování, archivace a zveřejňování zvukových nahrávek, uchovávání osobních dat a problematika anonymizace

3.2.5.1 Pořizování, archivace a zveřejňování nahrávek z hlediska legislativy a informovaný souhlas

V souvislosti s budováním archivu nahrávek je namístě otázka, zda je lze pořizovat potají, či nikoliv. Předmětem zájmu je totiž běžná, každodenní mluva, nikoliv strojený projev, do něhož mohou mluvčí přepínat, pokud tuší, či dokonce ví, že jsou nahráváni. **Legislativní stránka tajného nahrávání** je řešena v jednotlivých státech různě, vždy je nutné předem se seznámit s aktuálními předpisy a zákony. V České republice je pro účely vědeckého výzkumu takovéto pořizování zvukových záznamů možné, viz Nový občanský zákoník č. 89/2012 Sb., znění předpisu k datu 1. 1. 2024: „Podobizna nebo zvukový či obrazový záznam se mohou bez svolení člověka také pořídit nebo použít přiměřeným způsobem též k vědeckému nebo uměleckému účelu a pro tiskové, rozhlasové, televizní nebo obdobné zpravodajství." Týž předpis lze vztáhnout na zveřejnění nahrávek,[22] aplikujeme-li na ně postulát veřejného zájmu, ovšem v takovém případě je nezbytné mít na zřeteli ochranu osobních a citlivých údajů (viz níže). Podle J. Noskové (2014, s. 64–65) je řešením **anonymizace osobních a citlivých údajů**, a to ve smyslu nejen uvedení informanta pod pseudonymem (namísto plného jména), ale úplné anonymizace i dalších údajů, na jejichž základě by mohlo dojít k jeho

22 Zpřístupňování archivů, včetně těch zvukových, je jedním z hlavních cílů veřejných vědeckovýzkumných institucí. Nelze souhlasit s názorem, že by měl být přístup k nahrávkám omezen na členy výzkumného týmu, srov. prohlášení lingvistů Milroyové a Gordona: „Volně dostupné by nahrávky neměly být za žádných okolností" (2012, s. 89). Je však pochopitelné, že archivy mohou být zpřístupněny až ve fázi, kdy jsou pro to podmínky, a to nejen po technické a finanční stránce, ale také po zajištění ochrany osobních údajů, a to u popisků nahrávek včetně jejich obsahu a transkriptů, jak je popsáno dále.

identifikaci. V malých místních komunitách je to přitom úkol téměř nemožný (v případě nahrávky je mluvčí identifikovatelný už jen podle barvy hlasu).

Kromě národní legislativy je nezbytné obeznámit se také s **etickým kodexem výzkumné instituce**, pod kterou jsou nahrávky pořizovány, archivovány a případně zveřejňovány (k činnosti těchto komisí a kolizi mezi procedurální etikou[23] a etikou v praxi viz Guillemin a Gillam, 2004; též Hejnal a Lupták, 2013; Novotná, 2014 a zde uvedená literatura). V některých zemích je výzkum založený na záznamech pořízených tajně právně napadnutelný, a tak je institucionálními komisemi pro etiku výzkumu omezený, nebo přímo zakázaný (Murray a Murrayová, 1992). Tajné nahrávání naráží také na obecné etické principy. Možným řešením je dodatečné uvědomění informanta o proběhlém nahrávání a **dodatečné udělení souhlasu** se zvukovou dokumentací ze strany nahrávaného (dodatečný souhlas byl většinově užíván třeba při sběru dat pro korpus ORAL2013, viz Benešová a kol., 2015, s. 48–49). Hrozí však nebezpečí, že se výzkumník setká se zamítnutím (v takovém případě by měla být veškerá data od daného informanta smazána). Dodatečný souhlas je využíván zejména v případě výzkumu založeného na skrytém účastnickém pozorování, kdy by informování účastníka již na začátku výzkumu mohlo vést k falešným výsledkům. V ideálním případě by měl být dodatečný souhlas získán bezprostředně po realizaci výzkumu (u informanta, popř. v místní komunitě); v případě prodlevy mezi realizací výzkumu a řešením souhlasu může dojít k situacím, kdy by mohla být snaha výzkumníka komplikována (např. změnou bydliště informanta), nebo dokonce znemožněna (např. úmrtím).

Na udělování **informovaného souhlasu** (a jeho formální náležitosti) mají výzkumníci různý názor, netradiční pohled nabízí třeba J. Lindbloomová (2004), podle které podepsaný informovaný souhlas zaručuje pouze bezpečnou práci s daty pro výzkumníka, avšak pro informanta nepředstavuje žádné výhody, spíše naopak. Z toho důvodu Lindbloomová informovaný souhlas jako takový spíše odmítá; neznamená to však, že by neměl být participant výzkumu informován o výzkumných cílech, práci s daty atd. (mělo by to však být věcí důvěry, nikoliv smluvního závazku).

V současné dialektologické (potažmo lingvistické) praxi převládá vědomé nahrávání s tím, že informant ještě před zahájením nahrávání nebo na jeho začátku udělí k této činnosti souhlas. **Vědomý souhlas** by měl zejména obsahovat svolení s těmito základními body:

- nahrávání rozhovoru;
- zařazení nahrávky do konkrétního archivu, fondu, databáze;
- použití nahrávky ke konkrétním (vědeckovýzkumným) účelům;
- případné zveřejnění nahrávky.

Informovaný souhlas by měl být udělen explicitně, a to ústně, nebo písemně. Konkludentní souhlas, tj. vyjádřený implicitně kývnutím hlavy, pokynem ruky, popř. tacitně, je nedostatečný.

Přestože je v současnosti zdůrazňována potřeba **písemné formy informovaného souhlasu** (např. Zíková a Křivan, 2014, s. 80), nelze upřít některé klady jeho ústní podobě (ústní souhlas ostatně připouští i Milroyová a Gordon, 2012, s. 87). Hlavním pozitivem je menší formálnost celého procesu a také menší obava z možného zneužití podpisu na dokumentu (v návaznosti na různé podvody, jimž byli v minulosti vystaveni zvláště příslušníci starších generací; k nedůvěřivosti z praxe sociologického výzkumu viz Lindbloom, 2004, s. 89). Během udílení **ústního souhlasu** je navíc možné jednotlivé body podrobně vysvětlit, a předejít tak možným nepochopením ze strany informanta. Veškeré informace podávané explorátorem informanto-

23 Procedurální etikou máme na mysli získání souhlasu příslušných etických komisí k provedení výzkumu, jehož participanty jsou lidé (Guillemin a Gillam, 2004, s. 263).

vi (v mluvené i psané podobě) by měly být formulovány jednoduchým, srozumitelným jazykem.[24] Ústní souhlas by měl být ideálně zaznamenaný na nahrávce, aby se předešlo případným právním komplikacím. I v tom je výhoda ústního souhlasu – celý proces jeho udílení je zdokumentován, zatímco v případě pouhého psaného dokumentu je doložen pouze výsledek, totiž samotné udělení (Guillemin a Gillam, 2004, s. 272).

Písemné udělení informovaného souhlasu znamená podepsání písemného prohlášení informantem, v případě nezletilé nebo nesvéprávné osoby jejím zákonným zástupcem či opatrovníkem. **Povinnou součástí takového prohlášení** by mělo být:

1. jméno odpovědného výzkumníka (popř. výzkumníků, včetně výzkumné instituce) a jeho kontaktní údaje;
2. předmět výzkumu a v něm uplatněné výzkumné metody dotýkající se nahrávaného subjektu;
3. konkretizace výstupů výzkumu (zvukový archiv, zvuková mapa, tištěná/elektronická publikace ap.);
4. způsob nakládání s nahrávkami a jejich transkripty (např. zda budou uloženy v interním/externím archivu, kdo bude jejich data stewardem, v jaké míře mohou být zveřejněny, jací uživatelé budou mít k nahrávkám/transkriptům přístup);
5. nakládání s osobními údaji během výzkumného projektu a po něm (vč. určení správce těchto dat);
6. informace o možnosti úprav vybraných bodů prohlášení (např. výmaz vybraných částí nahrávky včetně transkriptu, zákaz zveřejnění nahrávky /avšak ponechání možnosti nahrávku zpracovat pro účely výzkumu/), popř. podmínky odstoupení od dohody v celém znění;
7. jméno, příjmení a podpis nahrávaného subjektu (popř. jeho zákonného zástupce či opatrovníka), datum a místo podpisu.

V závislosti na sledovaných cílech je možné do informovaného souhlasu začlenit další body, které lze oproti výše vyjmenovaným chápat jako fakultativní, např. délku nahrávání, potenciální rizika a benefity vyplývající pro nahrávané subjekty, způsob kompenzace v případě placené spolupráce aj. Informant by měl mít možnost, aby mu ze strany výzkumníka byly méně jasné body osvětleny, a to podobně jako v případě ústního souhlasu. Písemný souhlas se pořizuje vždy ve dvou vyhotoveních, z nichž jeden exemplář získá podepsaná osoba (informant), druhý je archivován ve výzkumné instituci. Příklad informovaného souhlasu viz příloha 3, další příklady obdobných prohlášení viz třeba Johnstoneová (2000, s. 44–47).

U informovaného souhlasu vyjádřeného ústně je na úvaze výzkumníka, zda informanta seznámí se všemi body uvedenými v souvislosti s písemným souhlasem v maximálním rozsahu, nebo zda poskytne pouze **omezené informace**. Jak upozornili M. Hammersley a P. Atkinson (2007, s. 211), explorátor v době sběru dat nemusí znát veškeré detaily výzkumu (např. konkrétní typy výstupů) a ne všichni informanti se o výzkum zajímají do té míry, že by chtěli znát veškeré podrobnosti, a tak by jednak výklad mohl působit rušivě, jednak by mohl ovlivnit způsob mluvy druhé strany, a tak zneplatnit získaná jazyková data.

Je také otázkou, zda by neměl být mluvčí **o nahrávání informován opakovaně** v jeho průběhu; participanti výzkumu často na skutečnost, že je jejich promluva zaznamenávána, zapomenou, a to z důvodu postupného budování vztahu s výzkumníkem (Hammersley a Atkinson, 2007, s. 210). Řešení tohoto etického problému závisí na sledovaném cíli; při sběru dialektologických (potažmo lingvistických) dat je opakovaný souhlas bezpředmětný – výzkumník by měl naopak naplno využívat strategii aktivního budování vztahu s informantem (viz 3.2.1).

V ideálním případě by měl být informovaný **souhlas pořízen od všech osob** zapojených do zvukové dokumentace. Platí to také pro osoby, které se do nahrávání zapojily neplánovaně v jeho průběhu (např. náhod-

24 S ústním souhlasem, nahraným na začátku nebo konci rozhovoru, pracuje též kolektiv Slovenského hovoreného korpusu. Respondentovi je vždy položena otázka ve znění: „Ak súhlasíte s tým, aby bola táto nahrávka prepísaná, zaradená do databázy Slovenského hovoreného korpusu a slúžila na vedeckovýskumné ciele, povedzte, prosím, áno, súhlasím.“ (Gajdošová a Šimková, 2014, s. 67).

ně příchozí návštěva); v takovém případě je žádoucí alespoň dodatečné vyjádření souhlasu. Ne vždy má ale explorátor možnost zajistit informovaný souhlas od všech zúčastněných; badatel totiž nemá vždy kontrolu nad výzkumným procesem, neboť výzkum probíhá v přirozeném prostředí, jako je příbytek mluvčího, volně přístupné prostory v obecním úřadě, v knihovně ap. (Hammersley a Atkinson, 2007, s. 211).

PŘÍKLAD Z PRAXE

Způsob přirozeného vysvětlení výzkumného cíle ilustruje nahrávka, kterou si lze poslechnout přes přiložený QR kód. Jde o úvodní část rozhovoru dialektoložky M. Ireinové s mluvčím ze Skvrňan (části Plzně), nahraného v r. 2024.

Zvláštní případ představují data, která byla získána v minulosti, tj. v době, kdy informovaný souhlas nebyl součástí výzkumné praxe, navíc se mnohdy nahrávalo potají. Modelovým příkladem jsou historické nahrávky z 50. až 70. let 20. století, které tvoří jádro Archivu zvukových záznamů nářečních promluv, jenž je ve vlastnictví dialektologického oddělení ÚJČ AV ČR. Pokud bychom na tento problém nahlíželi dnešní optikou, nebylo by možné takovéto zdroje využívat, neboť nesplňují podmínku informovaného souhlasu. Tím bychom ale přišli o cenná (někdy opravdu jedinečná) data, nadto lze takovýto přístup označit za ahistorický. Data byla získána v té době standardním postupem, navíc od mluvčích už nelze dodatečný souhlas získat, neboť již nejsou mezi živými.

3.2.5.2 Osobní a citlivé údaje a jejich ochrana

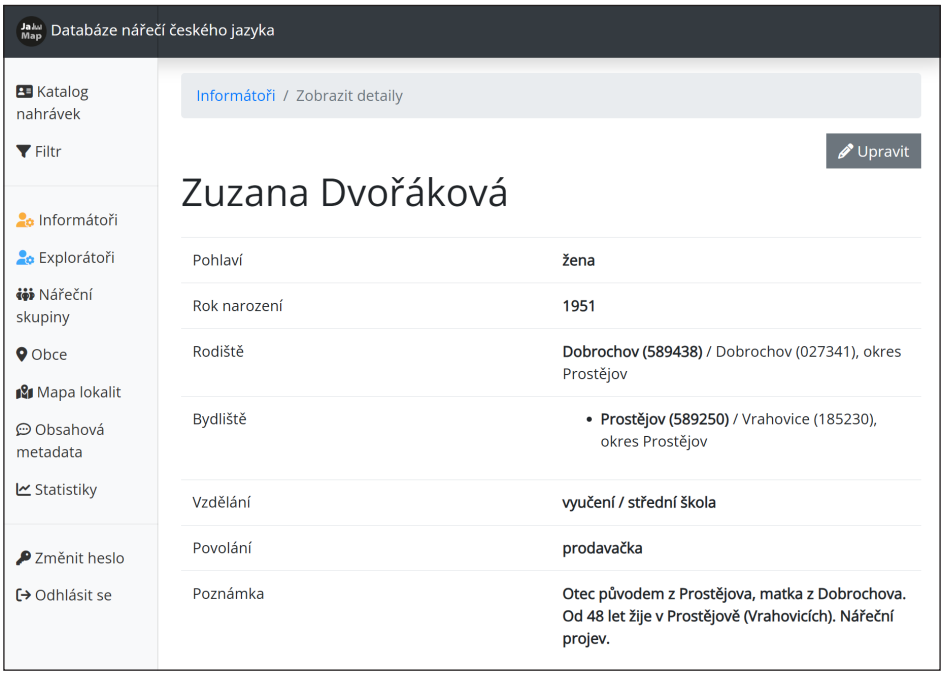
Prerekvizitou správného vyhodnocení jazykové stránky nahrávek je získání co nejvíce informací o mluvčích (informantech, popř. i explorátorech, kteří rovněž do promluvy zasahují a jejichž jazyk může být také podroben zkoumání). Sbírají se **informace relevantní pro (socio)dialektologický výzkum**, které jsou nutností mj. pro správné zařazení promluv do určité nářeční oblasti nebo vůbec regionu s určitými typickými jazykovými prostředky. Pro tyto účely jsou zjišťovány následující údaje:

- pohlaví mluvčího;
- rok narození;
- místo jeho původu;
- místo původu matky;
- místo původu otce;
- přesuny bydliště (vč. přechodných bydlišť, např. v době studia, během povinné vojenské služby, během pobytu v zahraničí ap.);
- zaměstnání (vč. těch dřívějších);
- nejvyšší stupeň dosaženého vzdělání.

Vytčené údaje představují nezbytné minimum, které lze v závislosti na charakteru výzkumu dále rozšiřovat. Zvláště v sociolingvistických výzkumech je např. někdy doporučováno zjišťovat také rozsah jazykových kompetencí mluvčího (Zíková a Křivan, 2014, s. 80), a to na úrovni rodného nářečí, mateřského jazyka i nářečí a jazyků cizích; obdobně tomu bylo při celoúzemním výzkumu pro *Český jazykový atlas*, během nějž byli mluvčí dotazováni mj. na zálibu ve čtení. Informace by měly být uváděny ve strukturované podobě pro snazší práci s nimi, jako je tomu např. v elektronické *Databázi nářečních promluv pro odbornou veřejnost* (viz obrázek 3.4); takovýto způsob vedení informací umožňuje badateli mj. filtrování na základě jednotlivých kritérií.

V interních záznamech je vhodné u jednotlivých mluvčích uvádět jméno informanta (rodné jméno, příjmení, popř. přezdívka v obci) a kontaktní údaje (adresa, e-mail, telefonní číslo), popř. kontaktní údaje osoby,

která zprostředkovala kontakt s mluvčím. Žádoucí je to zejména tehdy, pokud je plánován opakovaný výzkum. Vzhledem k etickým zásadám a právním normám[25] jsou tyto údaje soukromého charakteru, a jsou tak určeny výhradně výzkumníkovi nebo jeho výzkumnému týmu. Případně lze zvolit správce těchto dat, který zajišťuje jejich zabezpečení. Osobní údaje by měly být uloženy ve zvláštních adresářích, ideálně na jiných úložištích, než na kterých se nacházejí nahrávky. Platí tu obzvláštní opatrnost, a to i kvůli stále se zvyšujícímu počtu kybernetických útoků vedoucích k neoprávněnému zveřejnění osobních a citlivých údajů.



Obrázek 3.4 Ukázka karty s osobními údaji informanta, zabudované do Databáze nářečních promluv pro odbornou veřejnost *(fiktivní příklad)*

V případě **zveřejnění promluv v audiální či textové podobě** nesmí dojít k úniku osobních informací, které nejsou určeny veřejnosti. Nahrávky jsou z toho důvodu podrobovány **anonymizaci**, která je buď povrchová (informanta není možné na základě zveřejněných informací okamžitě identifikovat; vypustí se např. jméno, kontaktní adresa, e-mail), nebo hloubková (identifikace je zcela nemožná; Chromý, 2014, s. 12). V lingvistickém výzkumu se majoritně pracuje s povrchovou anonymizací. V praxi to znamená, že se pro označení mluvčích namísto osobních jmen užívají zástupná označení (např. *mluvčí 1*, *mluvčí 2* atd., *M1*, *M2* atd., *A*, *B* atd., případně zkratka složená z prvního písmene rodného jména a příjmení; viz příklady starší praxe níže). Utajeny zůstávají též zmíněné kontaktní údaje, přesné datum narození a další citlivé údaje (např. název firmy, ve které informant pracuje; výše platu nebo důchodu; rodinné vazby, např. výčet vnoučat a jejich jména). Z hlediska etiky by měly být anonymizovány veškeré údaje, o které mluvčí požádá, a to i zpětně.

Anonymizaci podléhají také samotné nahrávky, které jsou prověřovány z hlediska **citlivosti obsahu**; znamená to, že pokud na nahrávce vystupují osobní údaje mluvčího nebo jiných aktérů, o nichž je v promluvě zmínka, jsou tyto údaje v případě zveřejnění nahrávky smazány (dané místo je zvukově naznačeno uměle

25 Právní stránku ochrany osobních údajů v České republice ošetřuje Zákon č. 110/2019 Sb., o zpracování osobních údajů, § 16 pracování osobních údajů za účelem vědeckého nebo historického výzkumu nebo pro statistické účely. Viz též Nařízení Evropského parlamentu a Rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů).

vygenerovaným tónem). V přepisech se pak buď užije znak pro výpustku (trojtečka), nebo náhradní výraz (např. zástupné rodné jméno a/nebo příjmení).

Jindy je doporučováno u nahrávky změnit veškeré popisy, které by mohly vést k **jednoznačné identifikaci nahrávaného** (Johnstone, 2000, s. 43), v některých sociolingvistických pracích z toho důvodu byly dokonce anonymizovány názvy obcí, v nichž byly záznamy pořízeny, a to uvedením přibližného údaje o místě, ve kterém mluvčí žili nebo strávili většinu života, nikoliv přesného názvu obce (Milroyová a Gordon, 2012; v české praxi viz třeba Hoffmannová a Müllerová, 2007). Je to pochopitelné, neboť v obcích s menším počtem obyvatel může být mluvčí identifikován na základě uvedeného roku narození a hlasu; na druhou stranu je právě místo původu či pobytu pro dialektologické výzkumy stěžejní.

PŘÍKLAD Z PRAXE

Příklady starší praxe:
osobní údaje mluvčích a různé způsoby (ne)anonymizace v českých lingvistických pracích

Hoffmannová a Müllerová, 2007:
pohlaví (*pan/paní*), zkratka jména, věk, přibližné místo původu/bydliště
Pan K., 77 let, Prostějovsko
Paní R., 81 let, Jindřichův Hradec, Praha

Balhar a kol., 2011:
pohlaví, rok narození, místo původu/bydliště
NÁVOJNÁ
žena, nar. 1904

Kloferová, 2000:
rok narození, místo původu/bydliště
LOMNICE
(Nar. 1974)

Bachmann, 2001:
zkratka rodného jména, příjmení, rok narození, povolání, původ (ne/rodák), místo původu/bydliště
SRBCE, M. Stratílek, nar. 1889, bývalý krejčí, rodák

Lamprecht a kol., 1976:
rodné jméno a příjmení, rok narození, místo původu/bydliště
Boskovice, okr. Blansko. Mluvila Marie Kruťová, nar. 1908

Utěšený, 1972:
pohlaví, rok narození, místo původu/bydliště
Z Jezeřan, okr. Znojmo (tři muži, 1904, 1905, 1899)

Kopečný, 1957:
nejednotnost v údajích
84letý stařeček
A – stařenka, B – její 21letý vnuk
A – 26letý hlídač švestek, překupník z U., B – já, C – 23letý vysokoškolák, D – jeho 55letý otec

3.2.5.3 Pořizování, archivace a zveřejňování nahrávek z hlediska etických principů

Některé obecné etické zásady již byly zmíněny v souvislosti s ne/vědomým zapojením informantů do výzkumu. Existují však další **etické problémy**, k nimž by měl výzkumník zaujmout stanovisko. Na poli etnografie bylo vyděleno pět základních bodů (Hammersley a Atkinson, 2007, s. 209–229), které lze zobecnit na veškeré vědní oblasti, a to:

1. vědomý souhlas;
2. soukromí;
3. újma;
4. vykořisťování;
5. důsledky pro budoucí výzkum.

Jednotlivé body rozvedeme v návaznosti na dialektologický výzkum.

Ad 1: Vědomý souhlas

První bod se týká otázky uděleného, či neuděleného souhlasu, a v případě souhlasu uděleného pak jeho formy. Podrobný výklad viz 3.2.5.1.

Ad 2: Soukromí

Týká se informací získaných od mluvčího, které mohou být veřejného nebo soukromého charakteru. V případě zveřejnění nahrávek musí výzkumník předjímat, které úseky je ne/vhodné zpřístupnit, musí též **zamezit úniku soukromých údajů**, jak je popsáno v podkapitole 3.2.5.1.

Otázka soukromí je rovněž namístě při pokládání otázek, z nichž některým by se měl výzkumník vyhnout; záleží však na domluvě s informantem a na jeho ochotě některá témata otevřít (např. při popisu konce války v r. 1945 někteří mluvčí odmítají hovořit o dění v obci, jiní naopak velmi otevřeně a detailně popisují třeba znásilňování žen a dívek, kterým byli zasaženi mj. rodinní příslušníci).

Ad 3: Újma

Újma (hmotná, fyzická, duševní, sociální aj.) může vzniknout během výzkumného procesu i po zveřejnění výsledků na straně informanta. Mluvčí může pociťovat stres, nebo dokonce úzkost. Při výzkumu se může **cítit diskomfortně** z důvodu nevhodného prostředí (někteří mluvčí se cítí „nesví“ v prostorách obecního úřadu, jiní neradi pouští „cizince“ do svých příbytků). Jindy se mluvčí svěřují se **strachem ze samotného nahrávání**, negativně reagují na diktafony, zapojení klopového mikrofonu ap. V tomto případě je zapotřebí ze strany výzkumníka osvětlit potřebu těchto zařízení, případně najít s mluvčím nějaký kompromis (např. nahrávání na zabudovaný, nikoliv externí mikrofon). Mluvčí také někdy vyjadřují **obavu ze zneužití nahrávky** nebo **posměchu** v případě zveřejnění, a to ze strany příbuzných nebo místní komunity. I z toho důvodu je nezbytné mluvčím vše řádně osvětlit, zejména způsob nakládání s nahrávkami, výběr zveřejněných úseků a anonymizaci.

Újma se může týkat i osob, o nichž je zmínka v nahrávce, a to zvláště pokud se o nich hovoří v negativním kontextu. Z toho důvodu je nutné dbát na anonymizaci obsahu zveřejněných nahrávek, případně na důkladný výběr zveřejňovaných úseků po obsahové stránce. **Obavu z poškození** výjimečně vyjadřují příbuzní informanta, obrací se pak na výzkumníky s žádostí o výmaz nahrávky a osobních dat mluvčího; při komunikaci s těmito osobami by měl mít výzkumník na zřeteli nejen výzkumné cíle, ale také případnou újmu, kterou by mohl utrpět informant. Je vhodné nejdříve se zkusit s rodinou domluvit, případně zvážit jiná ře-

šení, např. hloubkovou anonymizaci (např. změnou místa sběru nahrávky) nebo zákaz zveřejnění nahrávky (avšak se souhlasem s jejím uložením v archivu pro vědeckovýzkumné zpracování).

V souvislosti s citovou újmou je vhodné dbát na **mikroetiku**, tedy na obecně etická pravidla vycházející z reflexivity výzkumníka a aplikovaná přímo v praxi (Guillemin a Gillam, 2004). Základním principem mikroetiky je požadavek, aby výzkumník jednal s informantem lidsky a ohleduplně, bez manipulací a zneužívání (tamtéž, s. 264). Dotýká se to jak obsahu rozhovoru, který výzkumník vede s informantem, tak způsobu pokládání otázek, popř. i odpovědí a otázek ze strany informanta, které výzkumník nechává bez povšimnutí.

PŘÍKLAD Z PRAXE

Neexistují žádná komplexní doporučení, jak by měl výzkumník reagovat v konkrétních situacích během výzkumu (nebo i po jeho ukončení). Příkladem je výzkum uskutečněný M. Šimečkovou v r. 2022 v jedné obci na Břeclavsku (kvůli anonymizaci neuvádíme konkrétní obec), během něhož bylo zaznamenáno vyprávění starší ženy plné bolestných rodinných vzpomínek (týkaly se např. alkoholismu v rodině, odnětí dětí příbuzné a následná péče o ně informantkou v pokročilejším věku, úmrtí milované vnučky aj.). Vyprávění nezůstalo bez reakce, výzkumnice své emoce vyjádřila slzami a následným uznáním, jak mohla informantka překonat tolik nesnází a zachovat si optimistickou mysl. Zda šlo o správnou reakci, či nikoliv, nelze kriticky zhodnotit. Ovšem pokud by vyprávění bylo výzkumnicí pouze zaznamenáno bez jakékoliv odezvy, působilo by to dozajista nelidsky, tedy i neeticky, pokud bychom na jednání nahlíželi z perspektivy mikroetiky. Doplňme, že s informantkou bylo domluveno, že nahrávka nebude zveřejněna.

Jinou situaci zažil dialektolog F. Kubeček v r. 2024 při dokumentaci nářečí na Vyškovsku (obec z téhož důvodu jako výše neupřesňujeme). Informant, starší muž, se ke konci rozhovoru svěřil, že ho po úmrtí manželky opustila pozitivní energie a trápily ho myšlenky na sebevraždu; popsal dokonce přesný postup, kterým by ji provedl. Výzkumník se informanta snažil povzbudit, a to připomenutím neobyčejného životního optimismu informantova otce, o němž byla dříve v rozhovoru řeč, a vyjádřením obavy z ochuzení se o hezké chvíle, které by ještě mohl zažít. Další reakce v tomto případě nebylo zapotřebí, neboť informant navázal ve vyprávění tím, že právě naděje na světlejší zítřky společně s jeho dětmi a vírou v Boha mu nakonec pomohly sebevražedné myšlenky překonat. V případě takto netypického sdělení týkajícího se sebevraždy lze přitom reagovat více způsoby: 1) nereagovat a přejít raději k veselejšímu tématu; 2) téma rozvíjet a přitom se snažit informanta navést jiným směrem (což by však mohlo být hodnoceno jako zásah do soukromí, totiž jako ovlivňování informanta, k němuž výzkumník nemá právo); 3) svěřit se s obavou o informantův život osobě, která je mu blízká (např. někomu z rodiny nebo místní komunity), což by ale mohlo být vysvětlováno jako narušení důvěry.

Ad 4: Vykořisťování

Myšlenka vykořisťování informantů zapojených do rozhovorů spočívá v tom, že mluvčí poskytují výzkumníkovi informace (v našem případě jazyková data), které následně výzkumník využije ve svém výzkumu, avšak **druhá strana tím nic nezískává** (Hammersley a Atkinson, 2007, s. 213). Tento pohled je však diskutabilní, neboť ze strany mluvčích jde o dobrovolné zapojení. Ze zkušeností s terénním výzkumem v českém prostředí lze poukázat na skutečnost, že mluvčí jsou naopak participací ve výzkumu potěšeni, dokonce mnohdy vyzývají výzkumníka k opětovné návštěvě. Je to dáno tím, že výzkum dosud převážně probíhal mezi příslušníky starších generací, kteří se cítí být vyloučeni, proto vítají možnost mezigeneračního dialogu. V mnoha oblastech je navíc nářečí vnímáno jako součást regionální identity (Šimečková, 2022), a tak se mluvčí rádi dělí o své jazykové znalosti (a také o životní zkušenosti). K reciprocitě výzkumu v zahraničí s ukázkou dobré praxe v českém prostředí viz 3.2.1.3.

S tématem vykořisťování je spojena otázka **vlastnických práv**. Sběr dat je časově a finančně náročný, výzkumník je placen z institucionálních nebo projektových zdrojů. Je tak přirozené, že takto získaná data jsou ve vlastnictví instituce, nikoliv informanta. Mluvčí, jejich rodiny nebo obce (popř. knihovny, folklorní spolky) se někdy obrací s žádostí o poskytnutí nahrávky; v tomto případě by se mělo žádosti vyjít vstříc pro zachování dobrých vztahů a kvůli odpovědnosti vůči jazykovému společenství (viz 3.2.1), avšak s upozorněním na způsob zacházení s nahrávkou (např. nemožnost jejího zveřejnění a dalšího šíření), též s předcházejícím udělením souhlasu s vydáním nahrávky ze strany informanta.

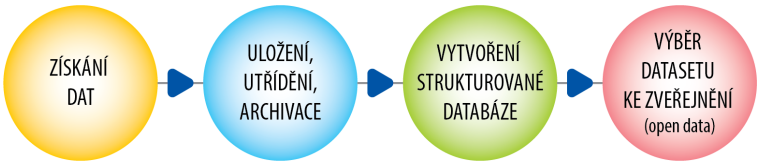
Ad 5: Důsledky pro budoucí výzkum

Badatel by měl **vést výzkum zodpovědně** vůči informantovi, též vůči kolegům, a to současným i budoucím. Pokud by informant nabyl s výzkumem špatnou zkušenost, vedlo by to k zamítnutí účasti na dalším výzkumu. Z toho důvodu by měl explorátor během výzkumu sledovat informantovy reakce a řídit se jimi, popř. by se měl po dokončení rozhovoru ujistit, zda byla naplněna informantova představa o celém procesu, či nikoliv.[26] V českém prostředí byla negativní reakce participantů výzkumu zaznamenána zřídka, a pokud ano, tak převážně u mluvčích zapojených do celoúzemního výzkumu pro *Český jazykový atlas* – tehdejší sběr, při němž byly zjišťovány odpovědi na 2 649 položek, představoval pro starší lidi velkou fyzickou i mentální zátěž, z toho důvodu někdy nebyli ochotni zapojit se do opakovacích výzkumů.

Problematika etických principů v exploračním výzkumu je samozřejmě mnohem komplikovanější, výčet bodů by bylo možné rozšířit, avšak to by bylo již nad rámec této metodiky.

3.3 Budování zvukového archivu

Po návratu z výzkumu by mělo dojít k uložení nahrávky, její archivaci a katalogizaci, též případně ke zveřejnění, je-li plánováno (viz obrázek 3.5). Stejně tak je nezbytné archivovat a katalogizovat nahrávky získané cestou citizen science nebo z jiného zdroje. Výzkumník by měl mít promyšleno, jakým způsobem bude postupovat – na jaké úložiště bude data ukládat, jakým způsobem je bude zálohovat a co bude obnášet jejich popis.



Obrázek 3.5 Proces zpracování audiálních dat od sběru po zpřístupnění

Základní poučka zní: nikdy při procesu zpracování nepoužívat originální nahrávky. Mohlo by totiž dojít k jejich nechtěnému výmazu (jedné nahrávky, nebo všech nahrávek uložených v rekordéru), čímž by veškerá práce v terénu přišla vniveč (Clemente, 2008, s. 183). Je lepší soubory nejdříve zkopírovat (na pevný/externí disk, server pro ukládání většího množství dat) a vytvořit také zálohy (v případě placených úložišť lze počítat s několikerým zrcadlením, tzn. při ztrátě dat by mělo být možné jejich obnovení). Několikeré uložení, včetně verzování dat, je opravdu nezbytné; i v případě uložení na více externích discích se doporučuje obzvláštní opatrnost, např. není vhodné mít tyto disky ve stejné místnosti nebo budově – v případě požáru nebo jiného neštěstí či katastrofy by došlo k jejich nenávratné ztrátě.

26 Někdy se dokonce hovoří o „nesmiřitelném zájmu“ vědy a těch, kteří jsou studováni (Becker, 1964, s. 276). Nepřátelskou reakci ze strany informantů předpokládal třeba H. S. Becker (tamtéž), ovšem to je tvrzení značně přehnané.

Jakmile jsou data uložena, mělo by dojít k jejich utřídění. Ideální je vytvořit si strukturovaný digitální archiv, v němž by se data řadila podle konkrétního klíče (popř. podle více kritérií, třeba podle data pořízení, způsobu promluvy, nářeční oblasti ap.). Utříděná data by měla být podrobně popsána.

Popisná metadata by měla obsahovat tyto informace:

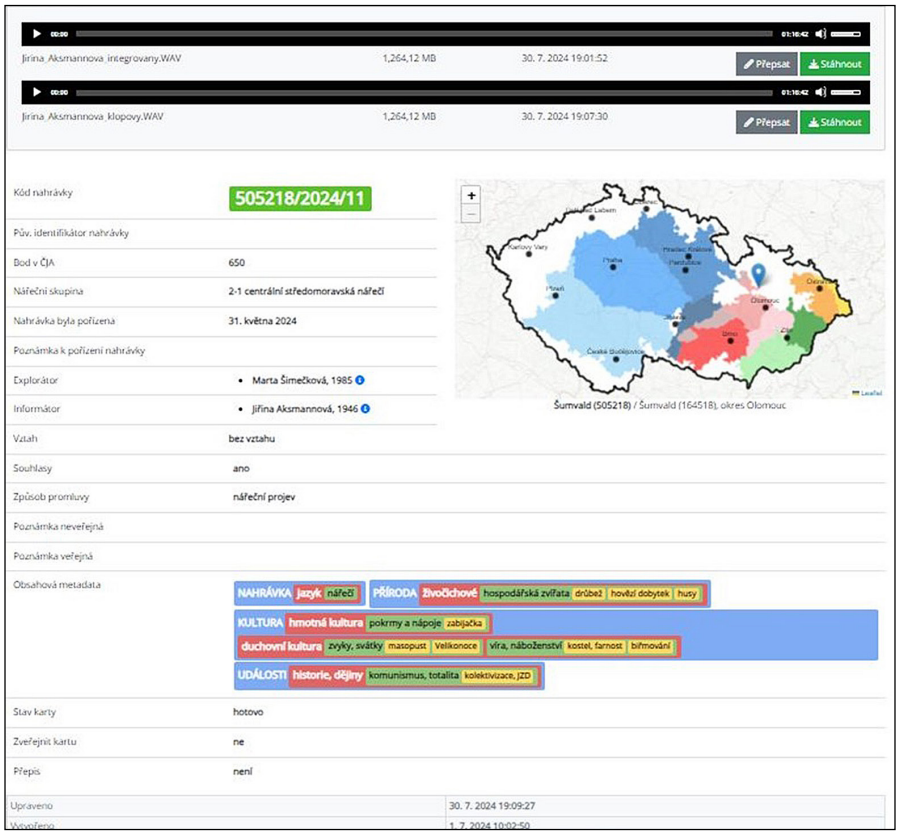
- jednoznačný identifikátor nahrávky (nutný mj. pro případné uložení v repozitářích; je vhodné uvažovat o jednotném kódování podle doporučení Národní fonotéky – metodika je v přípravě);
- místo sběru (ideálně název obce, její část, okres, popř. konkrétní lokalizátor, např. GPS souřadnice nebo kód obce podle číselníku CISOB; ke geolokaci nářečních dat viz 7.1.1);
- datum sběru (den, měsíc, rok);
- jméno explorátora (rodné jméno, příjmení);
- jméno informanta (rodné jméno, příjmení, popř. přezdívka v obci), jeho adresa a kontaktní údaje (při zveřejnění anonymizováno); dále rok narození, pohlaví, místo původu a další bydliště (název obce, část, okres, lokalizátor), místo původu rodičů, nejvyšší dosažené vzdělání, zaměstnání, záliby, případně doplňující informace;
- jméno kontaktní osoby (rodné jméno, příjmení), kontaktní údaje (e-mail, telefonní číslo);
- vztah mezi explorátorem a informantem (základní kategorie: bez vztahu / rodinný příslušník / známý);
- způsob zachycené mluvy (základní kategorie: nářeční projev / polospisovný projev / spisovný projev / střídání kódů / hypernářečí);
- v případě nářečního projevu zařazení do konkrétního nářečí (třeba pomocí kódu, viz 3.1);
- informace o informovaném souhlasu;
- informace o přepisu (zda je přepis zadán k vyhotovení, popř. zda jsou přepisy již hotové nebo zrevidované);
- obsahová metadata (viz 3.2.1);
- případně další poznámky (např. v rámci jakého projektu byla data získána).

Součástí zápisu mohou být také další materiály vložené formou samostatných souborů, např. fotografie, rukopisné soupisy slov nebo souvislá vyprávění ap.

PŘÍKLAD Z PRAXE

Příkladem popisných metadat zabudovaných do zvukového archivu je připravovaná *Databáze nářečních promluv pro odbornou veřejnost* – k jednotlivým nahrávkám jsou pořizovány popisné karty,[27] jejichž struktura je vidět na obrázku 3.6.

27 Jde vlastně o elektronický protokol (takto Nosková, 2014, s. 59), jindy nazývaný též jako záznam o rozhovoru (Vaněk a kol., 2007, s. 110).



Obrázek 3.6 Ukázka karty nahrávky z Databáze nářečních promluv pro odbornou veřejnost *(fiktivní příklad)*

Na tomto místě je ještě vhodné upozornit, že **ne všechny pořízené zvukové záznamy musí být nutně použitelné** pro sledovaný výzkumný záměr. Některé z nich mohou být vyhodnoceny jako nedostačující, a to obsahem, způsobem mluvy nebo i kvalitou záznamu po technické stránce. Jindy je důvodem nevhodně zvolený výzkumný vzorek – od některých informantů můžeme teprve v průběhu natáčení rozhovoru (tedy ex post) obdržet informace, které je vyřadí z užšího výběru. Při sběru nahrávek tradičních nářečních promluv je to zejména značné ovlivnění projevu mluvčího způsobené třeba:

- odlišným místem původu u jednoho rodiče (nebo obou rodičů) oproti rodišti/bydlišti informanta;
- delším pobytem na jiném, nářečně odlišném místě (a to i v zahraničí);
- manželstvím s osobou, která je nositelem jiného dialektu.

Tato audiální data mohou být vyloučena z archivu, popř. se vyloučení může týkat pouze jazykové analýzy (včetně zpracování dat strojovým učením), neboť by mohla generovat chybné závěry.

PŘÍKLAD Z PRAXE

J. Chromý ve starších pracích o protetickém *v-* problematické mluvčí z výzkumného vzorku vylučoval (2015a, 2015b), ovšem později své stanovisko změnil (Chromý, 2017, s. 147). Ke změně názoru přispěl dozajista malý objem dat k ověřování stavu protetického *v-* v současné češtině. Tento přístup však nelze pokládat za správný. Chromý svůj krok „obhajuje" zpřístupněním dat (pouze excerpcí, nikoliv nahrávek), z nichž mohou být problémové doklady jiným výzkumníkem vypuštěny, a tím mohou být sestaveny i nové statistiky a vyhotoveny nové propočty vyjadřující vliv faktorů na ne/užití sledované proměnné.

3.4 Transkripty audiálních dat: pravidla transkripcí a příprava dat pro strojové učení

3.4.1 Typy transkripčních soustav

Zvuková data jsou v rámci lingvistického výzkumu zpravidla převáděna do psané podoby – data se tak dají snáze uchopit. Tvorba přepisů (transkriptů) je nezbytná i pro využití těchto dat ve strojovém učení – při vývoji některých softwarů může jít o data primární, anebo sekundární (podpůrná, a to vedle primárních dat akustických).

Přepisy mohou být pořizovány standardním souborem znaků užívaných v daném jazyce i pro zápisy spisovné, nebo pomocí speciálních transkripčních pravidel. Při zpracování nářečních nebo vůbec regionálně rozrůzněných dat se tradičně využívá **dialektologická transkripce**, což je „záznam skutečného znění mluvy bez ohledu na pravopis" (Krčmová, 2017). Jde o druh fonetického zápisu, v němž je navíc využito speciálních prostředků pro diferenční hlásky, tedy takové, které jsou spisovnému jazyku cizí (např. pro široké centrálně středomoravské vokály se užívají složené znaky *ę*, *ǫ*, pro různé realizace *l*-ové hlásky znaky *l*, *u̯*, *ł*, *ĺ*, *ł́*, *l´* aj.; podrobně k dialektologické transkripci viz kapitola 4).

V minulosti byla sestavena *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských* (Hála, Vážný a kol., 1944; rozšířená verze 1951), která se dočkala zjednodušení v nářeční čítance *České nářeční texty* (Lamprecht a kol., 1976, s. 7–9). Těmito zjednodušenými doporučeními se řídil také autorský kolektiv *Českého jazykového atlasu*, a to u přepisů publikovaných v dodatkovém svazku (Balhar a kol., 2011).

Přestože existují jistá doporučení, jak transkribovat nářeční promluvy, přepisy nebývají jednotné;[28] např. v korpusu DIALEKT byl v dialektologické rovině přepisu oproti výše zmíněným zjednodušeným doporučením opět zaveden znak *ɣ* pro znělé *ch* v návaznosti na stará *Pravidla* (*Transkripce v korpusu DIALEKT*, 2018). Důležitým aspektem tu je výzkumný cíl; pokud je objektem zájmu fonetická stránka řeči, pak výzkumník mnohdy potřebuje podrobnější transkripci, zohledňující i drobné nuance (např. znaky pro úzké, popř. ultraúzké vokály jako *ę*, *ǫ*). Z toho důvodu by měl výzkumník vždy podrobně vysvětlit, jaká pravidla jsou na přepisy aplikována.

Lze využít i fonetickou **transkripci IPA** (angl. International Phonetic Alphabet), která disponuje větším počtem znaků a také umožňuje zachycovat slovní přízvuk nebo výdech. Pro počítačové zpracování jazyka byl vyvinut **transkripční systém SAMPA** (angl. Speech Assessment Methods Phonetic Alphabet), který je využíván i v českém prostředí ve strojovém učení.

Nářeční přepisy lze dohledat i v textech mimo lingvistickou oblast – jde nejčastěji o díla etnografického/etnologického, historického či muzikologického charakteru, jako jsou zápisy pohádek, lidových vyprávění, lidových písní atd. V takovýchto textech je využívána značně jednoduchá transkripce, bez pevně stanovených pravidel. Jde o přepis využívající znaky spisovného jazyka a přihlížející jen k výraznějším rysům nářečí, pro které jsou užívány speciální znaky (ne nutně shodné s dialektologickou transkripcí; např. obalované *l* je zapisováno jako *u*, *u̯*, nikoliv *u̯*; široké vokály jako *ê*, *ô*, nikoliv *ę*, *ǫ* ap.). Vzhledem k tomu, že takovýto zjednodušený princip využívali zvláště folkloristé, vžil se pro tento způsob zápisu název **folklorní transkripce** (Bachmannová, 2008; podrobně k této transkripci viz kapitola 4).

28 Další způsoby přepisů v lingvistických pracích viz třeba Hoffmannová, 1992; Kaderka a Svobodová, 2006. K transkripčním systémům, které našly využití napříč humanitními vědami, viz třeba Leix, 2003; Nosková, 2014, s. 85–88.

Ukázka přepisu ze zvukového archivu Veslovech.cz (dialektologická transkripce)

Lokalita: Lanžhot (okres Břeclav)

Rok pořízení nahrávky: 2022

Mluvčí: žena narozená r. 1932

No a tak potom uš to tak uťichu̯o, tak uš sme potom, že, náš taťka, že: „Púdeme dom, púdeme sa poďivat." Já sem biu̯a také ďefčisko, dvanást roku, já sem sa nebáu̯a. To biu̯o sa, fšadi samá su̯ama, koňe, šecko na zemi, P*, aš s teho Pastiřska aš sem, aš na Lučňí. Tak sme šli dom. Došli sme dom, to šecko, okna porozbíjané, fšecičko, su̯áma a fšecko. A spješ, diž biu̯o mau̯é ďecko, nebiu̯a postelka, ešťe biu̯a taká kolépka, to sa to ďecko f tem tak kolébau̯o, to bilo také dřevjené. A koňe s* tam, pu̯oti ot súsedú rozbité, na dvoře koňe a fšecko. No a ti koňe žrali s tej, s tej kolépki su̯á*, to seno. A oňi ch*, enom mňeli mjechi a oňi enom ďeu̯ali mňechi plné, co viďeli, šecko brali a do tich mňechú a zašivali a posiu̯ali dom, posiu̯ali. Já to řeknu, jak to je, protože to pra*, posiu̯ali to do Ruska. A mi sme tam mňeli tři vlňáki. Toto, co je, to sú, temu sa říká vlňák. A ten jeden uš ten jeden bráu̯ a uš to žák* a dáu̯ do teho mjecha. A taťka mu řiká, že: „Co to bereš?" Že, že: „To je mój." Že to je jeho, že on z Rusije dovézeu̯. A já sem, já, že: „Ti z Rusije?" Já, ja takové ďecko blbé, ja, že: „Tis to dovezeu̯ z Ru*, z řiťi, a ne z Rusije." A on to třeba nevi, to je jedno. Ale oňi nás osvoboďili, ale ot fšeckého.*

Ukázka přepisu z korpusu DIALEKT (modifikovaná dialektologická transkripce + ortografická transkripce)

Nářeční ukázka z Petrovic u Karviné (Dolní Marklovice)

Rok: 2019

Mluvčí: žena (80 let, nar. 1939)

Dialektologický přepis (rovina dial):

(a pjeku̯i se tak, jag ďišo sum ti jidaše, nebo jakśi, [tuš co se pjekło]?) [ňi, to se pjeku̯a] šołdra. to śe šołdra @ pjeku̯a. to było dicki polym... pirše se mjenso uvařiło a, a vindzune a špirka. á to se zarobiu̯o tum polifkum. jedyňe se mlika dało teľa, co ti drožďe @ vyšu̯y. a tak se to tum polifkum zarobjało. no a to se dało na, na pľech, to ćasto, na to se ukłodało špirk'i i bočku v*, tego vindzunego a to vindzune mjenso, zavinyu̯o. no i upjekło a to se nazyvała šołdra.*

Ortografický přepis (rovina ort):

(a pekli se tak . jak dišo sum ty jidáše nebo jaksi .. [tož co se peklo]?) [ne . to se pekla] šoldra .. to se šoldra @ pekla .. to bylo vždycky polym . pirše se maso uvařilo a . a . vindzuné a špirka .. a .. to se zarobilo tum polívkum . jedině se mlíka dalo tela co ty droždě @ vyšly .. a tak se to tum polívkum zarobjalo . no a to se dalo na . na plech to těsto .. na to se ukládalo .. špirky i bočku v* . teho vindzuného a to vindzuné maso zavinulo .. no i upeklo a to se nazývala šoldra*

Ukázka přepisu z knihy *Všelijaký poudání*, 2020 (folklorní transkripce)

Jo, vono se pořád poudá a piše v novinách vo těch Kérkonoších, jaký tam bejvaj v zejmě foukanice a všelijaký mélhy, že není ani na krok viděť. Ale vo Pasekách, Polomným a Přichovicích to žádnej v novinách nepiše, a co se tam vodjakživa stalo všelijakejch přítřefuňku. To vám taky něco povím, co tuhle poudal Franta Potockej.

Prej se pořád chtěl poďvať ke švagrovej na Přichovice. Ale jen tak pěšky se mu nechtělo, je to vod nich přec lán cesty, a k tomu ten kopec. To už na starý nohy neni.

3.4.2 Dialektologická transkripce: principy a základní poučení

V této části metodiky se soustředíme na základní popis fungování dialektologické transkripce, která byla stanovena v projektu JAMAP. Platí následující **obecná pravidla**:

- zaznamenává se vše, jak je řečeno, včetně asimilací, splynulin, artikulačních chyb, přeřeků, opakování; v přepisu musí figurovat všechna slova, slabiky, hlásky;
- projev se nijak neupravuje směrem ke spisovnému jazyku, ani neidealizuje směrem k nářečnímu vyjadřování;
- číslice a zkratky jsou rozepisovány podle vysloveného (např. *osndesáté deváté rok*, nikoliv *89. rok*; *jé zé dé*, nikoliv *JZD*);
- suprasegmentální jevy (tj. intonace, přízvuk, frázování, pauzování, rytmus řeči) se nezachycují (není-li to součástí výzkumného plánu);
- neverbální projevy (např. pláč, smích) se neevidují (není-li to součástí výzkumného plánu);
- interpunkce se řídí tradičními pravidly pravopisu (užívání čárek je přitom mnohdy komplikované vzhledem k syntaktické struktuře mluvených projevů značně odlišné od projevů psaných);
- výraznější hezitační zvuky (*hm*, *hmm* ap.) se značí speciálním znakem (křížkem):[29] *no, tag mňe # dali pokoj*;
- nedořečená slova se značí hvězdičkou: *oňi bu*, budó*;
- úseky pronesené zároveň (překryvy, angl. overlaps) se nijak nevyznačují;[30] tato místa však mají vliv na segmentaci projevu do odstavců;
- nedokončené věty se značí trojtečkou: *Čekali zaťim uš… On bil na cesťe*. Trojtečka se nenadužívá, vkládá se zpravidla na konec věty (nikoliv dovnitř vět);
- splynulé hlásky se značí obloučkem, např. *pořá‿ce* (= *pořát se*; splynutí souhlásky *t*, *s* > *c*), *pořá‿tam* (= splynutí souhlásky *t*);
- nejasné úseky, u kterých si pořizovatel přepisu není jistý, lze označit hranatými závorkami, např. *šla aj[i] ona*; *jak to pravili, ajn grós [empfank]*. Pokud nelze slovo nebo jeho část vůbec identifikovat, lze pouze naznačit počet nesrozumitelných slov, např.: *a ten [2] zaklét*;
- transkriptor člení nahrávku na promluvy respondenta (R1, R2, atd.) a explorátora (E1, E2, atd.); případně je možné využít automatické segmentace, nutná je však manuální kontrola.

Pořizování přepisů je **mravenčí práce**, hodinový záznam řeči je v průměru přepisován 5 až několik desítek hodin (v závislosti na zkušenosti přepisovatele, též na nářeční oblasti, tím pádem i na počtu diferenčních jevů oproti spisovnému vyjadřování nebo na nářečí pořizovatele přepisu, tempu řeči, počtu překryvů, srozumitelnosti nahrávky atd.). Pracuje-li transkriptor v týmu, je vhodné zavést víceúrovňovou kontrolu s tím, že revizor se nejdříve soustředí na vyznačené nesrozumitelné úseky, posléze provede revizi celého zápisu.

29 Např. v korpusu DIALEKT je užíván znak zavináče: @. Viz *Transkripce v korpusu DIALEKT*, 2018.

30 V řadě transkripčních systémů je však k tomuto častému jevu přihlíženo, většinou se k jeho zápisu používají hranaté závorky (např. v korpusu DIALEKT). Takto též návrh od Leix, 2003.

PŘÍKLAD Z PRAXE

Členové dialektologického oddělení ÚJČ AV ČR byli požádáni o zodpovězení zdánlivě jednoduché otázky: Jak dlouho vám trvá pořízení přepisu hodinové nahrávky v dialektologické transkripci? Odpovědi se různily – jednak z důvodu rozdílné zkušenosti dotazovaných s danou činností, jednak proto, že nelze odpovědět jednoznačným časovým údajem (popř. intervalem) kvůli řadě faktorů, které mohou postup na přepisu zpomalit. Zatímco někteří odpověděli 4–6 hodin, 5 hodin ap., v jednom případě zněla odpověď dokonce 50–60 hodin (jednalo se totiž o přepis velice důkladný, s víceúrovňovou autokontrolou ze strany transkriptora, v té době méně zkušeného).

Složitost celého procesu si lze představit na základě odpovědi jedné z dotazovaných: „Nahrávku vždy poslouchám v autentické rychlosti a píšu první záznam; poté poslouchám zpomaleně, abych zajistila, že jsou zapsána všechna slova; poté poslouchám ještě jednou zpomaleně, abych doladila dialektologickou transkripci; poté poslouchám v autentické rychlosti, abych upravila syntaktické členění apod. Myslím, že pod 7 hodin se nedostanu. Toto by byl příklad přepisu, který je pro mě „čistý“, srozumitelný a bezproblémový, což znamená, že by byl z Čech a nemusela bych se soustředit na žádné nestandardní hlásky. Pokud bych přepisovala nahrávku, kde bych se musela soustředit ještě například na tvrdé *l*, široké vokály apod., trvalo by to ještě déle (řekněme o 2 hodiny).“

Níže uvádíme **přehled znaků užívaných v dialektologické transkripci**, a to v několika soustavách: ve standardní dialektologické transkripci, IPA (standardní i zjednodušené verzi) a SAMPA (viz tabulka 3.2).[31]

Tabulka 3.2 Znaky pro české hlásky (včetně nářečních) v různých transkripčních soustavách

dial. tran.	IPA (stand.)	IPA (zjedn.)	FIT X-SAMPA	vysvětlivky diferenčních jevů
a	a	A	a	
á	aː	aː	aː	
ą̊		aː	aː	dlouhé labializované *a* (*dobrą̊* = dobrá; *hlą̊pé* = hloupý)
ą		A	a	krátké široké *a* (*tą̊m* = tam)
ą́		aː	aː	dlouhé široké *a* (*bohą́č* = boháč)
b	b	B	b	
b′		B	b	měkké *b* (*hříb′atom* = hříbatům)
c	ʦ	ʦ	c	
c′		C	c	palatalizované *c* (*c′esto* = těsto)
ć		ʧ	C	polské palatální *č* (*ćesto* = těsto)
č	ʧ	ʧ	C	
d	d	D	d	
d′	ɟ	ɟ	D	
e	ɛ	ɛ	e	
é	ɛː	ɛː	eː	
ę		ɛ	e	krátké široké *e* (*ręb̥ę* = ryby)
ę́		ɛː	eː	dlouhé široké *e* (*bę́k* = býk)
ẹ		ɛ	e	krátké zavřené *e* (*tẹn*)

31 Za pomoc při sestavení tabulky děkuji fonetičce PhDr. Veronice Štěpánové, Ph.D.

dial. tran.	IPA (stand.)	IPA (zjedn.)	FIT X-SAMPA	vysvětlivky diferenčních jevů
ę̣		ɛ:	e:	dlouhé zavřené *e* (*dobrę̣*)
f	f	F	f	
f′		F	f	měkké *f* (*šf′ec* = švec)
g	g	g	g	
g′		g	g	měkké *g* (*švag′er* = švagr)
h	ɦ	ɦ	h	
ch	x	X	x	pozn.: nerozlišuje se neznělé *ch* a znělé *ɣ*
i	ɪ	ɪ	i	
í	i:	i:	i:	
y		Y	i	krátké tvrdé *i*
ý		y:	i:	dlouhé tvrdé *i*
j	j	J	j	
k	k	K	k	
k′		K	k	měkké *k* (*kočk′i* = kočky)
l	l	L	l	pozn.: nerozlišuje se slabikotvorné a neslabikotvorné *l*
l′		L	l	měkké *l* (*al′e* = ale)
ĺ		l:	l	dlouhé *l* (*slp* = sloup)
ł		L	l	krátké tvrdé *l* (*vúł* = vůl)
ł́		L	l	dlouhé tvrdé *ł* (*vł́ča* = vlče)
m	m	M	m	pozn.: nerozlišuje se *m* a *ɱ*
m′		M	m	měkké *m* (*m′esto* = město, *m′ano* = jméno)
n	n	N	n	pozn.: nerozlišuje se *n* a *ŋ*
ň	ɲ	ɲ	N	
o	o	O	o	
ó	o:	o:	o:	
ǫ		O	o	krátké široké *o* (*kǫs* = kus)
ǫ́		o:	o:	dlouhé široké *o* (*ďǫ́ravę́* = děravý)
ọ		O	o	krátké zavřené *o* (*stọdọla* = stodola)
ọ́		o:	o:	dlouhé zavřené *ó* (*mọ́ka* = mouka)
p	p	P	p	
p′		P	p	měkké *p* (*p′es* = pes, *p′ata*, *p′ynta* = pata)
r	r	R	r	pozn.: nerozlišuje se slabikotvorné a neslabikotvorné *r*
ŕ		r:	r	dlouhé *r* (*tŕňí* = trní)
ř	r̝	r̝	R	pozn.: nerozlišuje se znělé a neznělé *ř*
s	s	S	s	
s′		S	s	palatalizované *s* (*s′eno* = seno)
ś		ʃ	S	polské palatální *š* (*śeno* = seno)
š	ʃ	ʃ	S	
t	t	T	t	
ť	c	C	T	

dial. tran.	IPA (stand.)	IPA (zjedn.)	FIT X-SAMPA	vysvětlivky diferenčních jevů
u	u	U	u	
ú	uː	uː	uː	
v	v	V	v	
v′		V	v	měkké *v* (*v′ečir* = večer)
w	w	W	v	bilabiální *v* (*bejwal* = býval, *wečir* = večer)
z	z	Z	z	
z′		Z	z	palatalizované *z* (*z′ima* = zima)
ź		ʒ	Z	polské palatální *ž* (*źima* = zima)
ž	ʒ	ʒ	Z	
ə		ɛ	e	redukce (*ňat′* = nit; *zedňak* = zedník; *protəže* = protože); průvodní střídnice u slabikotvorného *r* (*kərk* = krk)
u̯		U	u	obalované *l* (*biu̯* = byl); *u*-ová realizace protetického *v*- (*u̯okno* = okno); *u*-ová realizace *v* (*prau̯da* = pravda)

V rámci prvního sloupce jsou modře vyznačeny ty znaky, které **nejsou z důvodu přílišné detailnosti užívány při využití dat strojovým učením** (důvodem tohoto rozhodnutí je malé množství dat, na nichž by mohlo být užívání těchto specifických znaků trénováno, viz 4.6.0). Jak je vidět, některé znaky pro nářeční hlásky nelze v některých soustavách dohledat, případně obsahují stejný znak pro více různých nářečních hlásek; řešením by bylo vytvoření znaků zcela nových.

K představené soustavě znaků lze ještě doplnit další symboly, vyznačující jevy typické pro mluvený jazyk, popř. napomáhající práci při pořizování transkripce – např. pro vyjádření nejistých míst nebo anonymizaci (viz tabulka 3.3).

Tabulka 3.3 Symboly pro značení dalších specifik mluvených projevů

znak	vysvětlení
*	nedokončené slovo
#	hezitační zvuk
‿	splynulina (*poř‿ce* = pořád se)
…	nedokončená věta (*Čekali zatím uš… On bil na cesťe*)
[]	nesrozumitelné slovo; nejistá interpretace: *[empfank]*, částečně nejistá interpretace: *[empf]ank*, zcela nejistá interpretace → odhad počtu vyslovených slov: *[2]*
{}	anonymizace, využití zástupného jména v odpovídajícím tvaru: *{Eva Nováková}, s {Nováčkó}*

Přepisy byly v české dialektologii dosud pořizovány v textových procesorech, přičemž nahrávka byla poslouchána ze separátního přehrávače. V současnosti je doporučován zvláště multiplatformní editor digitálního zvuku Audacity, který je volně dostupný a který umožňuje mj. zpomalení zvuku (rychlost by neměla být menší než 85 %) nebo označení konkrétního úseku k opakovanému přehrávání. Existují však i softwary umožňující více funkcí naráz; je jím třeba **transkripční program** ELAN, který slouží k poslechu nahrávky, její

segmentaci a přepisu.[32] Výhodou je to, že přepis (na úrovni určitého segmentu, tj. úseku nahrávky, který lze definovat jako větu, více vět o určitém počtu slov nebo nepřerušenou repliku jednoho mluvčího) je přímo spárován s příslušnou zvukovou stopou.

V souvislosti s projektem JAMAP je vyvíjen unikátní transkripční systém optimalizovaný na dialektologické přepisy, umožňující mj. značkování úseků podle jejich obsahové náplně,[33] systematické zachycování výrazných nonverbálních projevů nebo mimojazykové reality (např. narušení rozhovoru příchodem nějaké osoby, pozastavení nahrávání).

Podrobněji je o problematice přepisu nářečních projevů pojednáno v kapitole 4, a to včetně specifik spojených s konkrétními nářečními oblastmi.

3.5 Shrnutí

Cílem této kapitoly bylo představit možnosti sběru, archivace a katalogizace audiálních nářečních dat, též jejich přípravu pro strojové učení zahrnující tvorbu zvukové databáze a pořizování transkriptů. Zájemce o danou problematiku tak má před sebou manuál, díky kterému bude s to získat vlastní kvalitní audiální data, též je smysluplně utřídit, přepsat a případně i zveřejnit. Pojednáno bylo též o obecnějších vědecko-výzkumných otázkách, jako jsou etické principy při shromažďování dat nebo zapojení veřejnosti do sběrné činnosti cestou citizen science.

V jednotlivých podkapitolách byly podrobně popsány různé strategie, které lze v jednotlivých krocích uplatnit, a to včetně problematických případů, opřených o četné ukázky z praxe. Čerpáno přitom bylo nejen z poznatků z oboru dialektologie, ale i sociolingvistiky, konverzační analýzy, sociologie nebo etnologie. Cílem bylo popsat jednotlivé strategie srozumitelně, podrobně a v co možná nejširší škále (včetně strategií užívaných okrajově, popř. již vyšlých z úzu). Při aplikaci předložených metod v praxi je nutno mít na paměti, že žádná strategie není absolutně zakázána, avšak některé mohou s sebou nést jistá úskalí týkající se jejich proveditelnosti nebo obtížnější obhajitelnosti (srov. např. současný negativní názor na tajné pořizování nahrávek).

32 Daný program je využíván při přepisu nahrávek Českým národním korpusem (viz např. Komrsková a kol., 2017, s. 221–224).

33 Zatím je obsahová anotace zadávána manuálně, později by měl být proces automatizován.

4

Textová data: výběr, digitalizace, čištění, normalizace a převod textů a jejich příprava pro strojové učení

TEXTOVÁ DATA: VÝBĚR, DIGITALIZACE, ČIŠTĚNÍ, NORMALIZACE A PŘEVOD TEXTŮ A JEJICH PŘÍPRAVA PRO STROJOVÉ UČENÍ

4.0 Úvod

Jako textová data jsou v rámci této metodiky pojímána taková **nářeční data v textové podobě, která nelze využít jako data audiální** (k audiálním datům viz kapitola 3). To znamená, že jsou to specificky taková textová data, která nemají svou audiální podobu, a to z několika důvodů:

- zvukový záznam existoval nebo existuje, ale není nám dostupný (zanikl, ztratil se, je poškozen, je na nosiči, který dostupná technika neumí přehrát, je v soukromém držení apod.);
- text je přímým zápisem promluv, které nebyly zvukově zaznamenávány (byly zaznamenány těsnopisem, písmem);
- text není přímým zápisem nářečních promluv, ale:
  - vychází z autorovy jazykové kompetence (většinou kompetence rodilého mluvčího);
  - vychází z paměťové stopy recepce nářeční promluvy v minulosti.

Cílem této části je předestřít způsob takového zpracování těchto nářečních dat tak, aby se z nich stala funkční, jednotná, rozsáhlá a kvalitní **data pro strojové učení**.

Textová data, která připravíme, by měla sloužit ke dvěma účelům:

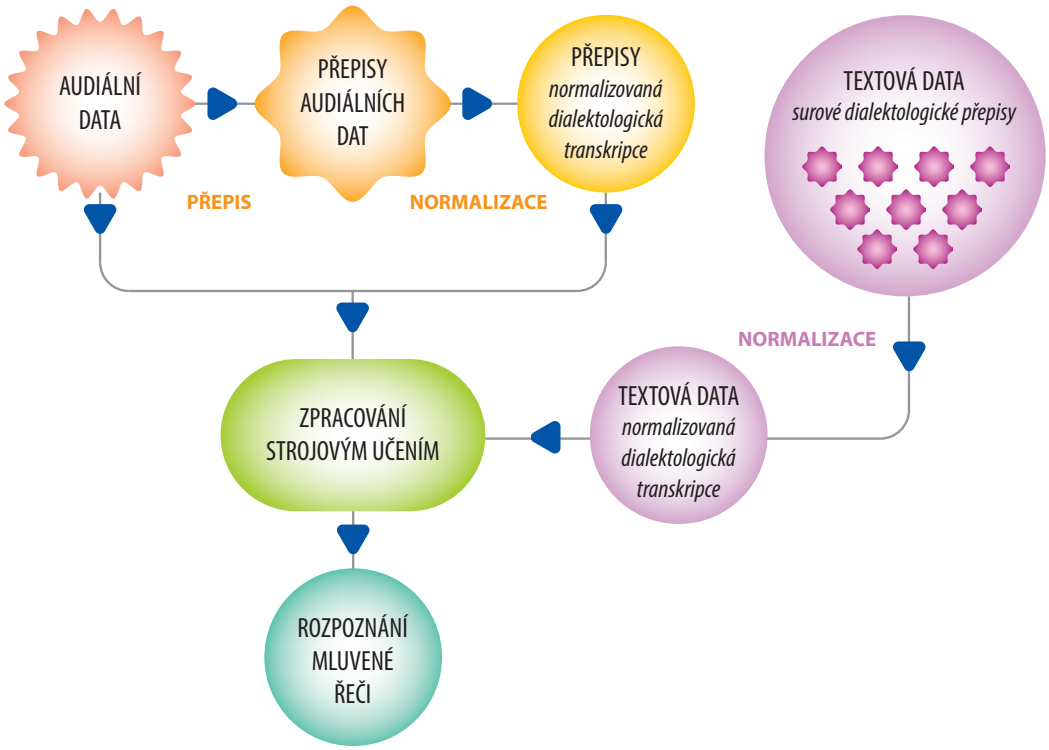
1. jako **podpora rozpoznání audiálních dat** (data z přepisů audiálních dat už ze své podstaty nejsou rozsáhlá,[34] proto je vhodné opřít rozpoznání i o strojové učení textové podoby dialektu s jeho běžnými výrazy a syntagmatickými vzorci);
2. jako **zdroj trénovacích dat pro automatický převod mezi dialektologickým a folklorním přepisem** (podrobněji k oběma přepisům viz 3.4.1, 4.1.2.1, 4.4 a 4.5; tento převod by měl sloužit také k tomu, aby audiální data mohla být automaticky rozpoznávána nejen v primárním a zvukově přesnějším dialektologickém přepisu, ale také v rozšířenějším a obecně srozumitelnějším přepisu folklorním, a to formou automatického převodu z přepisu dialektologického).

K těmto dvěma účelům je tedy zapotřebí:

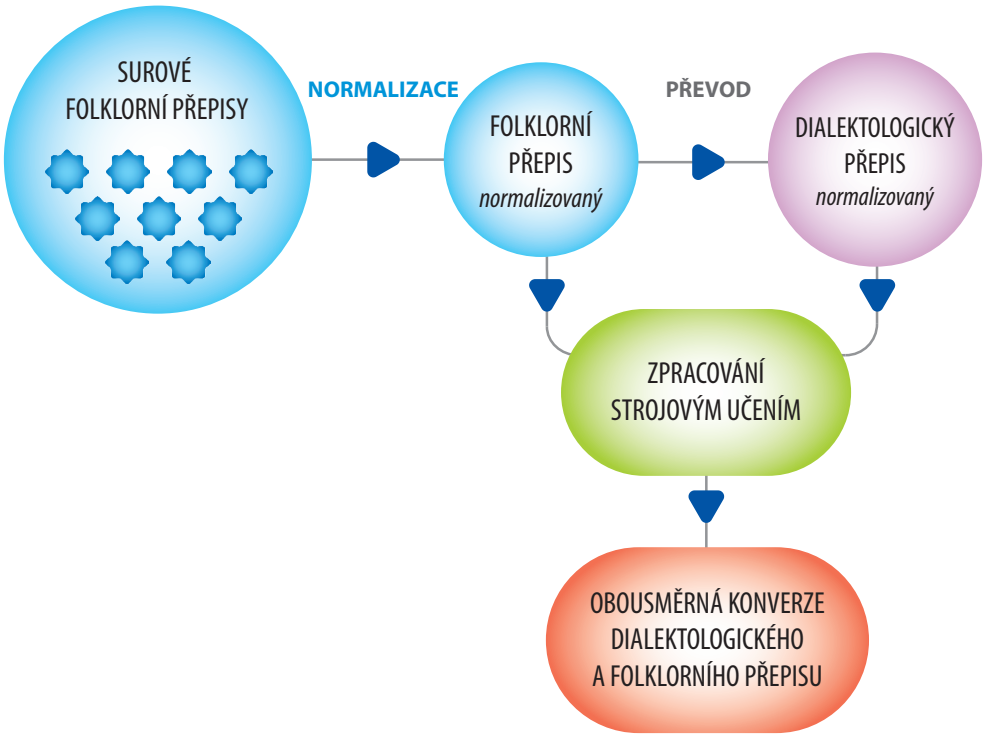
1. Získat textová data s dialektologickým přepisem a zbavit je všech nejednotností.
2. Získat přepisy týchž textů v jedné i druhé transkripci, které jsou zbavené všech nejednotností. Vzhledem k tomu, že taková data neexistují a bez nich nelze strojové učení vytrénovat, je nutné je vytvořit. K jejich tvorbě je potřeba algoritmický, na pravidlech založený převod mezi přepisy. To je možné pouze jedním směrem, a to z přepisu folklorního do přepisu dialektologického. Strojové učení by mělo být schopno na základě těchto dat převádět směrem opačným.

---

34 Brání tomu obecná nákladnost jejich pořizování (časová, personální i finanční), od samotného zajištění terénních výzkumů a kvalitních zvukových záznamů až po jejich přepisy, které jsou po všech stránkách vůbec nejnáročnější (srov. k tomu kapitola 3.4.2).



Obrázek 4.1 Schéma uplatnění textových dat s dialektologickým přepisem při rozpoznání mluvené řeči



Obrázek 4.2 Schéma uplatnění převodu folklorního přepisu na přepis dialektologický

Cílem zpracování nářečních textových dat je tedy získat **normalizovaný přepis folklorní** a **normalizovaný přepis dialektologický** a **folklorní převádět na dialektologický**.

TEXTOVÁ DATA: VÝBĚR, DIGITALIZACE, ČIŠTĚNÍ, NORMALIZACE A PŘEVOD TEXTŮ A JEJICH PŘÍPRAVA PRO STROJOVÉ UČENÍ

Tato kapitola nás provede celou cestou, kterou musí textová data urazit od samotného výběru surových dat přes jejich zpracování až po jejich konečnou normalizaci a vzájemný převod.

4.1 Výběr textů

Tato část představí, jaké typy nářečních textů jsou v českém prostředí k dispozici a jaké jsou jejich klíčové vlastnosti pro tvorbu trénovacích dat. Určuje, které texty upřednostnit, ať už po stránce kvality, nebo po stránce vydatnosti a množství dat i snadnosti a rychlosti jejich zpracování.

4.2 Digitalizace (OCR) nářečních textů

Tato podkapitola ukazuje, jak vybrané texty zpracovat po stránce digitalizace a optického rozpoznání znaků. Zabývá se jak hardwarovou, tak softwarovou stránkou digitalizace. Ukazuje postupy vyvinuté speciálně pro rozpoznání nářečního textu, který nemá vlastnosti textu spisovného a kterému běžné OCR (angl. optical character recognition, optické rozeznání znaků) není uzpůsobeno.

4.3 Čištění a formální sjednocení textu

Po získání elektronické podoby textu je obvykle potřeba text pročistit od všech jeho částí, které nejsou nářeční. Tato podkapitola uvádí zásady a postupy, jak toho dosáhnout a získat čistý nářeční text. Vlivem OCR i různosti textových zdrojů panuje také v získaných elektronických textech základní formální a typografická nejednotnost, která by zabraňovala dalšímu efektivnímu zpracování. Proto je zde představena série regulárních výrazů, které text uvedou do formální jednoty a vytvoří relativně homogenní prostředí pro další postup normalizace.

4.4 Normalizace folklorního přepisu

Podkapitola představuje standard folklorního přepisu, tak aby byl vhodný k dalšímu zpracování. Uvádí postupy, jak jednotlivé variabilní formy přepisu na normalizovaný přepis převést. Pro každou nářeční podskupinu vypočítává přehled jejích znaků v normalizované podobě, který bude mít značné využití při praktické aplikaci kapitoly 4.6.

4.5 Normalizace dialektologického přepisu

Zde je představen standard dialektologického přepisu a postupy, jak ho dosáhnout. Standard je vytvořen a zdůvodněn s ohledem na další zpracování textu jako trénovacích dat. Je uvedena tabulka převodů jednotlivých speciálních dialektologických znaků používaných v praxi na znaky normalizované. Pro každou nářeční podskupinu je pak sestaven přehled užívaných znaků.

4.6 Převod folklorního přepisu na dialektologický přepis

Podkapitola zevrubně popisuje převod z folklorní do dialektologické transkripce prostřednictvím regulárních výrazů. Krok po kroku postupuje soustavou regulárních výrazů, představuje jejich vzory, smysl, jazykové pozadí a praktickou aplikaci, začíná u širokého spektra výjimek a nepravidelností a pokračuje k pravidelným a rozsáhlým změnám, které přebudovávají folklorní zápis na dialektologický.

V celé této kapitole používáme pro **zápis textu ve folklorním přepisu „uvozovky“**, pro **zápis textu v dialektologickém přepisu *kurzívu***. Uvozovky používáme též pro texty, které jsou na půli cesty mezi folklorním a dialektologickým přepisem (např. jde o texty ve fázi nedokončeného převodu z folklorního na dialektologický přepis), a také pro uváděné regulární výrazy uvnitř věty. Kurzívu pak používáme pro vyjádření zvukové a hláskoslovné stránky textu, uvozovky ke zdůraznění její znakové stránky. V naprosté většině však jde o manifestaci protipólu folklorního a dialektologického přepisu.

4.1 Výběr textů

4.1.0 Úvod

Prvním krokem při vytváření textových trénovacích dat pro strojové učení je výběr textů. Tento krok je na první pohled méně důležitý a zvládnutelný i intuitivně, přesto dalece ovlivňuje výsledek práce, a nelze ho tudíž podceňovat. Výběr textů nám totiž určí jednak rychlost zpracování textů (a tedy množství zpracovaného materiálu), jednak jeho kvalitu. Množství a kvalita materiálu jsou nejpodstatnějšími faktory podmiňujícími úspěšnost strojového učení, takže správná selekce textů je naopak klíčová a rozhoduje o celém výsledku další práce.

Nejde přitom o to, vybírat pouze texty ideální. V oblasti nářečí se prakticky vždy potýkáme s nedostatkem materiálu, a pokud bychom tento měli ještě redukovat na materiál zcela neproblematický (jednotný, bezchybný, odborně revidovaný atd.), bylo by jeho množství zcela nevyhovující. Je proto třeba vybírat texty pomocí následujících obecných kritérií:

kvalita materiálu:

- autentičnost – trénovací data by měla pocházet od mluvčích, pro něž je nářečí mateřským a prvním jazykem, nezanedbatelná je i osoba výzkumníka či zaznamenavatele textu, který tento autentický projev dovede adekvátně percipovat a zapsat;
- reprezentativnost – vybrané texty by jako trénovací data měly být co nejpodobnější datům, na nichž výsledek strojového učení bude uplatněn; reprezentativnost je tu tedy míněna v mnohostranném smyslu: výběr textů by měl budoucímu účelu co nejlépe odpovídat transkripčně, teritoriálně, časově, generačně, sociálně, tematicky atd.;
- rozsah textového vzorku – souvislé texty jsou jako trénovací data vhodnější než krátké úseky textu (jednotlivá slova, slovní spojení, úryvky vět a souvětí);

množství materiálu:

- minimální nutnost zpracování – materiál, který je už před zpracováním v řadě ohledů blízký stanovené normě pro trénovací data, bude rychleji narůstat; může jít o jeho hotové technické zpracování, typografickou shodu s normou, blízkost normalizované transkripci apod.;
- snadná zpracovatelnost – materiál nemusí být shodný se stanovenou normou, ale může na ni být snáze, nebo hůře převoditelný; rozdílům, které jsou snadno odstranitelné, je třeba dávat přednost před těmi, které snadno odstranitelné nejsou: je tak možné zpracovat nejen větší množství dat, ale také obvykle ve vyšší kvalitě;
- rozsah jednotlivého textu – rozsáhlejší text zaznamenaný jedním způsobem se zpracovává rychleji než řada různých textů v úhrnu o shodném rozsahu.

Tato obecná kritéria budeme dále konkretizovat prostřednictvím typologie nářečních textů, ale už v této obecné podobě jsou prakticky vyčerpávající. Výběr z dostupného materiálu by měl postupovat na základě těchto kritérií od nejlepších textů k horším. Měl by tedy začít u nejkvalitnějšího a nejsnáze zpracovatelného materiálu a pokračovat k méně kvalitním zdrojům, zpracovatelným obtížněji. Bude tak i dobře patrné, zda trénovací data v dané fázi ještě nějak podstatně vylepšujeme, nebo zda je dokonce nepoškozujeme.

4.1.1 Základní účely existujících písemných záznamů nářečí

Textový nářeční materiál se obvykle objevuje v několika ustálených formách daných kulturním ovzduším, zvyky a trendy minulé i současné (nejen) české společnosti. Teritoriální nářečí, útvar národního jazyka, jehož

doménou je soukromá mluvená komunikace (podrobněji viz 3.1), vstupovalo do psané podoby na základě několika tendencí.

1. Jednou ze základních tendencí, viditelnou už od dob národního obrození, byla snaha zaznamenat a zmapovat tradiční českou kulturu, jejíž neoddělitelnou, ale nikoli jedinou součástí bylo nářečí. Výsledkem těchto laických i odborných národopisných a folkloristických snah byly sběry lidových písní, lidových vyprávění, pohádek a pověstí, ale i výzkum zvyků a hmotné kultury spolu s lexikem, které se k ní vázalo, případně i se záznamy promluv osob udržujících lidové tradice, lidových řemeslníků a umělců, pamětníků, účastníků lidových obyčejů atd. (srov. Jeřábek, 1997; Markl, 1987; Klímová a Otčenášek, 2012). Tyto záznamy můžeme najít ve sbírkách lidové slovesnosti, ve folklorních studiích a monografiích i v archivech jako dosud nepublikovaný materiál.
2. Druhou tendencí, která se z první vyvinula o něco později, byla snaha zaznamenat přímo a výhradně jazyk, tedy teritoriální nářečí. Opět šlo a stále jde o odborné i laické snahy, jejichž produktem jsou nářeční monografie, slovníky, slovníčky, více nebo méně vážně míněné učebnice nářečí, soubory ukázek nářečních promluv nebo nářečních vyprávění (srov. Kloferová, 2007, s. 338–367; Stupňánek a Ireinová, 2020).
3. Třetí tendencí je tendence umělecká, která používá nářečí jako charakteristiku postav nebo jako ozvláštňující (často komický) jazykový prostředek prózy, poezie či dramatu, ale i např. reportáže a jiných publicistických útvarů (viz např. Slavík, 1940; týž, 1947). Může se pohybovat na široké škále od ojedinělých dialektismů po výhradní užívání nářečí, vesměs bývá úzce tematicky spjata s tradiční lidovou kulturou a způsobem života a často mívá již zmíněné humorné ladění (je-li nářečí zastoupeno ve vysoké míře). Výsledkem bývají literární a publicistické útvary všeho druhu, mnohdy jsou blízké lidové slovesnosti.

Je zřejmé, že tyto tři tendence se mohou vzájemně prolínat a nelze mezi nimi stanovit ostré hranice. I když pro sběr trénovacích dat jsou obvykle nejvhodnější texty druhého typu, ve všech třech typech zacílení textů se může objevovat materiál velmi kvalitní i velmi nekvalitní. K jeho kvalitativnímu určení dopomůže podrobnější typologie nářečních textů.

4.1.2 Typologie nářečních textů

Nářeční texty můžeme typologizovat na základě řady kritérií. Ta nám napomáhají určovat jak kvalitu textu, tak další způsob a efektivitu jeho zpracování.

4.1.2.1 Podle transkripce

Podle transkripce dělíme texty na:

- folklorní (přepis je založený v zásadě na spisovném úzu doplněném případně o speciální nářeční znaky, transkripce je víceméně fonologická, tedy nezohledňuje znělostní asimilace ani řadu měkkostních rozdílů, zachovává typické historické prvky pravopisu, jako např. rozlišení *i-y*, *ú-ů*, i když zvukově je nářečí nereflektuje; v případě kopaničářských nářečí někdy přejímá i prvky slovenského pravopisného úzu, omezeně se to děje i u dialektů slezskopolských a pravopisu polského);
- dialektologické (přepis je založený na fonetickém zápisu, většinou pro dialektologické účely zjednodušeném, jsou reflektovány znělostní asimilace a obecně je mnohem přesněji zápisem zaznamenávána zvuková stránka nářečí, důkladněji je rozlišována kvalita hlásek, jejich měkkost, mnohdy i kvantita nebo jiné charakteristiky; součástí dialektologické transkripce mohou být i grafické indikátory pauz, intonací, rázů a jiných zvukových charakteristik, mnohdy bývá přepis doplněn i metadaty s dalšími detaily ohledně zvukové realizace dané promluvy, viz 3.4).

Transkripční pravidla folklorního a dialektologického přepisu zdaleka nejsou jednotná. Dialektologický přepis si zpravidla přizpůsobují odborníci podle jevů, které studují, a které naopak zanedbávají. Přestože existují *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských* (Hála, Vážný a kol., 1944; rozšířená verze 1951), pravidla uplatňovaná v různých dialektologických přepisech jsou velmi variabilní.

Jednó, dъž se procházeł po mněstě, potkał svyho bratra, kerě tam přъjeł po své prácъ. I pozvał jé k sobje a vъpravoval mu dukładně všecъ přihodъ svy, jak viděł na jeho polu ščesti kłáskъ sbirat, jak a hde se své Bidъ zbavil a jiny věcъ. Častovał

Obrázek 4.3 Dialektologický přepis z Chromče v Bartošově Dialektologii moravské II *(Bartoš, 1895, s. 144)*

do.ma || Pag mňe po.wdā ho.spo.dār || Viš | ve.zne.š si vo.la | ·a jedne.ho. ko.ňe || Tak sem jel | ·e.šče. s tē.m ko.ňem | ·a jak jedu | jeja | jeja | napro.ťi mňe | z Le.zňice. | de. pro.ce.stvi || ·A f tē.m | muzəkaňťi spusťilə | marijānskō. p'i.se.ň | ·a ne.vim | co. se. vo.-lo.vi stalo. | že. se dal do. tance. | ·aš se kuň svaləl | ·a že.le.zni radło. | se mu do.stało. | aš k ple.ce.ťi || Mālo. chəbjeło. | bəl bə mňel nohu

Obrázek 4.4 Dialektologický přepis z Chromče ve Výslovnosti na Zábřežsku *V. Mazlové (Mazlová, 1949, s. 168)*

Dəbəch jā bəla | vo. de.se.t le.t mlači | jā ne.vim | co. bəch ďāla || Dəť sme. v no.vi re.publice. || Prf | dəš smə bələ v rajchu | se.m řikā-vala || dəbə Pāmbiček dal | ·a mə mňelə našu zlatō. republiku | jā bəch tancuvala po. səlnicə || ·A taki se.m si zatancuvala z jednē.m Ruse.m | ·a ə se.m jedne.ho. po.lō.bla | dəš přəšlə || To. bəla hubička | vo.d babə ||

Obrázek 4.5 Dialektologický přepis z Jedlí v téže publikaci (Mazlová, 1949, s. 168)

No a to sme vəďeáli — přet tó prvňí válkó — tak prúmjerňe patnást, névéž dvacet korun za štrnázd dňí. Teda za štrnázd dňí sme brávali véplatu tak patnást, névidz dvacet, nehdə³ takə dvanást, dəž nebəlo moc práce, i dvanást korun sme brali nehdə. No a to bəlo celkovje málo, iní továrňe, hde bəla podobná taková továrna, tak si tam vəďelali dvakrát

Obrázek 4.6 Dialektologický přepis z Jedlí v Českých nářečních textech *(Lamprecht a kol., 1976, s. 196)*

Za prvňi vojny choj'yli z hetmanstfi a vojsko a zbirali sedlakum obile. Tak se uš' kaž'dy ňeco skovał. My sme mjeli fš'ecko pochraněne pod humnem. Na humňe było pylno š'eč'ky, na dňe był oves pro koňe. Jak přyš'li, to sme uš' strachem trnuli, jak teš' to enem dopadňe. U suš'eda to doš'č' lechko proš'ło a u nas teš' to jaksyk š'ło. Ale na druhy j'eň rano sme stanuli, naraz u suš'eda zas pylny dvur vojaku a nanovo rekvizyce. Gdosyk

Obrázek 4.7 Dialektologický přepis z Březové v Balharově studii
K charakteristice lašského okrajového nářečí na jihozápadním Opavsku *(Balhar, 1957, s. 114)*

No v ňedželu už było zle. Už hazali, byli ve Skřypovje a uš to z našy zahrady było opčaz vidžeč, takove, takove střylaňi slyšedž hrozne. No uš sme tak, było takove, takovy strach jakysyk a stařyčeg naž byli tam vedla, kuseg u přyzňe jedne a oňi tam byli v nedželu k połedňu a naras přyšli s takym strachem. U Balhara tam uhodžyło, granad lebo co. Přylečeli s takym popłachem. Toš potem zas to

Obrázek 4.8 Dialektologický přepis téže informátorky z Březové v Balharově Skladbě lašských nářečí *(Balhar, 1974, s. 176)*

U textů ve folklorním přepisu nacházíme podobnou nejednotnost, zde však není ani tolik dána různými odbornými záměry autorů, jako spíše jejich obecnou různorodostí. Folklorní zápis je forma zápisu naprosté většiny nářečních textů 19., 20. i 21. století, ať už je zapisují laikové, nebo odborníci. U laických autorů většinou platí, že jsou více ovlivňováni spisovnou normou než inspirováni jinými texty ve folklorním přepisu, nářeční zvláštnosti pak často vyřeší zcela po svém. Ani u odborných textů, u nichž jsou vzájemné intertextuální vztahy mnohem silnější, nelze spoléhat na jednotný způsob zápisu. Na obrázcích 4.9–4.11 vidíme např. různý přístup k zápisu diftongů s první klouzavou *i*-ovou složkou („ia", „ie" vs. „ja", „já", „je", „jé") u kopaničářských dialektů:

Tak zebral tam jakési štýlko od motyky a tak čerta fliakal: po ušoch, po hlave, po chrbce, no čert prosil:
„Nie, nie, ty musíš dostac, ked sem chodzíš, abys sem viac razí něšól."

Obrázek 4.9 Folklorní přepis z Vápenic z Lidových povídek ze Slovácka *(Frolec a Holý, 1967, s. 128)*

nazpátek: „Urežce si obá porjádné šípové prútky, vyroscené v jedném roku bez konárkóv. Na ohnišče postavce škarúpky od vajec a nalejce do ních vody. Dzecisko staně a pójdze sa dzívac, čo to je, čo sa robí? Aj rečnovac si budze ňjéčo pre sebja. Vbehnice do izby a len ho rúbajce a rúbajce, něch kričí kolko chce. Prídze dzivoženka, doněse vašého a svójho podhodzenca si zebere."

Obrázek 4.10 Folklorní přepis z Vápenic z Rychlíkových
Pověstí, pohádek a vyprávění z moravských Kopanic *(Rychlík, 2001, s. 26)*

Strýko Marcin umrel. Němal ešče moc roków, ale jako sa povedá, smrc si něvybjérá. A tak sa zešla rodzina a susedzja na svjéceňje – to bolo modleňje za něboščíka. Jako tak všecci klačali okolo poscele, svjéčky svjécily, ludzja sa modlili a čerstvá vdova Marína jačala a narjékala:" Nale preboha živého, čos mi to Marcínku urobil. Vedla cetky Maríny klačala souseda Rozára. Kedz vidzela a čula, jako

Obrázek 4.11 Folklorní přepis z Vápenic z publikace O kopanickej reči *(Křižková a kol., 2010, s. 58)*

Ačkoli mezi folklorní a dialektologickou transkripcí existuje jakési širší přechodové pásmo (srov. např. Fojtík, 2011; Kašík, 1908), je třeba je přísně rozlišovat, shromažďovat je, zpracovávat a trénovat odděleně a texty, které se ocitají na půli cesty, vyloučit z dalšího zpracování. Vyšší přesnost u folklorního přepisu nevadí, pokud zůstane přepisem v zásadě fonologickým. Takový přepis je obvykle snadno převoditelný na normalizovaný folklorní přepis a povětšinou se i snáze a ve vyšší kvalitě převádí na přepis dialektologický. Pokud však autor ve výraznější míře začne zohledňovat asimilace znělosti, tedy upustí od fonologického principu, ale není v tom důsledný, nelze už takový text použít jako folklorní ani ho bez dalšího na takový

text převést. Je možné se ho pokusit převést na text dialektologický (prostřednictvím postupů představených v podkapitole 4.6), ale i zde jsou rizika, že nejrůznější výjimky, vlivem odlišného zápisu, nebudou správně detekovány, a nepřevedou se tedy korektně.

V celé této kapitole pojednávající o textových datech budeme používat termíny **folklorní text** a **dialektologický text** ve shodném smyslu jako **text ve folklorním přepisu** a **text v dialektologickém přepisu**. Lze říci, že texty v dialektologickém přepisu opravdu spadají téměř výlučně do dialektologické literatury, tedy jsou zaměřeny na jazyk a věrný záznam jeho zvuku prostřednictvím textu. To, co na straně druhé nazýváme jako folklorní texty, má mnohem širší spektrum účelů, od textových ukázek v odborných etnografických a folkloristických studiích po nářeční beletrii a lidové záznamy nářečí. I tyto texty jsou téměř vždy nějakým způsobem spjaty s folklorem, případně folklorismem, a to nám tedy dovoluje toto mírné terminologické zjednodušení.

Při sběru materiálu je nutné texty v těchto dvou přepisech oddělovat.

4.1.2.2 Podle média

Dle média můžeme rozlišovat nářeční texty:

- elektronické (databáze, webové stránky, elektronické publikace nebo elektronické verze tištěných publikací; nezahrnujeme sem skeny obsahující jen obrazová, nikoli textová data), takovéto texty můžeme dále rozlišit na:
  - texty pořizované, editované člověkem;
  - texty vzniklé automatickým OCR;
- tištěné (jakékoli tištěné texty dostupné v papírové formě, zahrnujeme sem i nepublikované strojopisy nebo strojopisy studentských prací, domácí tisky laických autorů nebo opět prací studentů apod.);
- rukopisné (jedná se vesměs o archivní materiály v různém stadiu zpracování).

U elektronických textů je extrakce dat nejsnazší, z elektronických zdrojů lze v krátkém čase získat textová data ve značném množství. Podstatně obtížnější je získání textu u zdrojů tištěných a nejobtížnější u zdrojů rukopisných, v obou případech nejlépe prostřednictvím automatického OCR (ruční přepis je proti tomu extrémně časově náročný a efektivitou nedokáže konkurovat ostatním způsobům získávání dat). Proto je vhodné dát nejprve přednost zdrojům elektronickým, a to specificky těm, které byly pořizovány nebo důkladně editovány člověkem. Takováto textová data totiž bývají zkontrolovaná a mohou dosahovat (a zpravidla dosahují) nejvyšší kvality. V tom jsou závislá ještě na dalších faktorech, které je třeba zhodnotit (autor textu, žánr textu). Naproti tomu elektronické texty vzniklé prostřednictvím OCR je třeba předběžně spíše vyřadit, resp. zařadit až za texty tištěné. U zdrojů tištěných se totiž rovněž spoléháme na automatický převod OCR, který má vždy určitou chybovost, ale pokud je proveden podle návodu uvedeného v podkapitole 4.2, bývá zpravidla podstatně kvalitnější než neodborný převod OCR nebo převod homologovaný na spisovný jazyk. Zpravidla se lze setkat právě s takovýmito digitalizáty, jejichž kvalita bývá znatelně nižší než kvalita digitalizátů, které si můžeme pořídit sami. Pokud jsou tedy u elektronických nářečních textů k dispozici i obrazová data tištěného textu v dostatečné kvalitě, je většinou lepší provést OCR znovu. I kvalita těchto obrazových dat může být velmi snadno negativně ovlivněna (viz oddíl 4.2.3), proto je nutné přistupovat k elektronickým textům vzniklým automatickým OCR velmi obezřetně.

Tištěné texty bývají stále velmi dobrým zdrojem dat, pokud provedeme vlastní OCR. Platí to však především pro texty folklorní, texty dialektologické bývají bohužel často vysázeny kurzívou a méně typickými písmy, takže OCR u nich chybuje mnohem více.

Extrakce textových dat z rukopisů je potom vůbec nejnáročnější. I když už v současnosti existují softwary OCR, které fungují při rozpoznání rukopisů relativně velmi spolehlivě (např. projekt PERO-OCR, dostupný z: https://github.com/DCGM/pero-ocr.git), přesto je zde chybovost stále dosti vysoká a časové náklady na trénování konkrétního rukopisu nezanedbatelné. Krom toho rukopisné texty mohou často být neuspořádané, opatřené vpisky, dodatky a škrty, což opět znesnadňuje extrakci dat. OCR rukopisu tedy lze zatím považovat za nepreferovaný pramen textových dat, použitelný jen v okrajových případech, kdy je text rozsáhlý, čitelný a psaný vypsanou rukou, konzistentním písmem.

4.1.2.3 Podle původu autora

Text můžeme typologizovat také na základě původu jeho autora a z toho plynoucí autorovy jazykové kompetence. Autor může být dle původu:

- autochtonní (nářečí je autorovou mateřštinou a současně jeho jazyk nebyl výrazně ovlivněn vzděláním a pobytem v jiných nářečních oblastech nebo je schopen toto ovlivnění vědomě reflektovat a odfiltrovat);
- částečně autochtonní (je přechodným typem mezi autorem autochtonním a neautochtonním; mezi autory tohoto typu mohou být nemalé rozdíly, přesto by měli mít k nářečí, kterým nebo o kterém píší, velmi silnou a ideálně i ranou vazbu; může jít např. o autory, kteří se přistěhovali do dané nářeční oblasti v dětství nebo se z ní naopak brzy odstěhovali, kteří se v průběhu života stěhovali v rámci jedné nářeční podskupiny, v krajním případně sem lze zahrnout i autory, kteří se do dané nářeční lokality přistěhovali až v dospělosti, dlouho v ní žijí a místní nářečí dlouhodobě studují);
- neautochtonní (autor nepochází z dané oblasti);
- neznámého původu (informace o autorovi nemusíme být schopni dohledat, případně text nemusí poskytovat jednoznačnou identifikaci autora).

Původ autora je velmi důležité sledovat. Nejde jen o místo jeho narození (neboť v nejnovější době už není běžným zvykem rodit doma), ale především o místo pobytu v jeho dětství a také o původ jeho rodičů, místo jeho studia i profesního života. Tyto informace jsou podobně důležité jako u informátorů při sběru zvukových dat (viz kapitola 3.2.1.3). Ideálem je autor autochtonní, nejlépe takový, který jen minimálně opouštěl danou oblast. Autochtonie je velkou výhodou u výzkumníka, který nářečí studuje a zaznamenává, a téměř nutnou podmínkou u autora, který jím píše. Neautochtonní autor naproti tomu může být stále výborným výzkumníkem, ale jen stěží dokáže produkovat po všech stránkách autentické nářeční texty. Obtíž je to koneckonců i u autorů autochtonních, kteří jsou ovlivněni vzděláním, četbou, masmédii, migrací, dojížděním za prací apod. U autorů částečně autochtonních je častým problémem, že jsou schopni dobře rozlišit, která část jejich jazykové výbavy neodpovídá spisovné normě, ale nejsou už schopni dobře rozpoznat, odkud daný jazykový prostředek pochází, a směšují tak nechtěně různé jazykové kódy (srov. např. Kadoch, 2008; o něm pak Balhar, 2010; Popelka, 2016; týž, 2016–2018).

V současnosti se však zejména vlivem masových médií a moderních způsobů komunikace prakticky všichni autochtoni posunují do role autochtonů částečných.

4.1.2.4 Podle odbornosti autora

Na základě odbornosti lze autory rozlišit zhruba v této škále:

- odborník;
- student;
- laik.

Máme zde na mysli formální vzdělání a odbornost jazykovědnou, dialektologickou, případně i etnografickou, ovšem nelze opomíjet fakt, že značné míry odbornosti lze dosáhnout i bez formálního vzdělání, pouze soustavným a metodickým zájmem a studiem. Takže ačkoli je vhodné dát přednost autorům s jazykovědným vzděláním a dialektologickou zkušeností, je třeba posuzovat původce textů individuálně a nevylučovat a priori ani laiky nebo studenty. I etnografická odbornost autora je do velké míry garancí kvality textového materiálu. Studentské práce mají svá specifika a jsou kvalitativně velmi rozkolísané. Je zde zapotřebí značné obezřetnosti, neboť i na první pohled vzorná práce může obsahovat nečekané chyby a opomenutí. U textů pocházejících od laiků je třeba kriticky posoudit a odhalit jejich omezení a zvážit, zda s nimi lze při dalším zpracování textu konstruktivně pracovat.

4.1.2.5 Podle žánru

Nářeční texty se zpravidla objevují v několika žánrech literatury, které mají opakující se formy. Tyto formy pak podmiňují způsob a obtížnost extrakce nářečního textu i vhodnost jeho akvizice do trénovacích dat. Tato typologie na základě žánrů literatury, v nichž se nářeční text nachází, není vytvořena zcela konzistentně, protože zahrnuje nejen žánry celých textů, ale i žánry a typy úseků textů, pasáží literatury. Často totiž dochází k tomu, že nářeční je pouze určitý oddíl textu (např. typicky ukázky souvislejších nářečních promluv v dialektologické literatuře nebo nářeční exemplifikace), který se v této ustálené formě vyskytuje napříč různými typy literatury (např. nářeční exemplifikace ve slovnících i v nářečních monografiích). Sledujeme zde však primárně nikoli klasifikační, nýbrž praktické hledisko, jak získat z dostupné literatury co nejkvalitnější a nejhojnější materiál, kterému tuto klasifikaci podřizujeme.

Dostupné texty můžeme tedy dle žánru členit takto:

- **odborná literatura:**
  - dialektologická (jazykovědná):
    - nářeční monografie, studie, studentské práce (viz např. Bělič, 1954; Michálková, 1971; Kolařík, 1998; Steuer, 1932; Wodarz, 1955; Šipková, 1992; Studnička, 1953; Fic, 1971; Valihrachová, 1971; odborné dialektologické texty tohoto rázu bývají zpravidla doprovázeny exemplifikačním materiálem; ten bývá vesměs zapsán v dialektologické transkripci a většinou je vysázen kurzívou, případně i jinak zvýrazněn, což obvykle umožňuje jeho extrakci); daný nářeční materiál lze lišit na základě toho, zda je umístěn uvnitř odborného textu jako krátký a názorný doklad jednoho pojednávaného jevu, nebo zda je umístěn relativně samostatně a slouží jako rozsáhlejší ilustrace komplexního stavu a podoby daného nářečí; podle toho lze v těchto textech rozlišovat:
      - exemplifikace (lze použít pouze takové zdroje, kde exemplifikace spadají do jedné lokality, případně do jedné nářeční podskupiny, anebo jsou jejich lokality zapisovány formalizovaně a konzistentně; v pasážích a studiích fonetických, fonologických, morfologických a slovotvorných může být tento exemplifikační materiál v rozsahu pouhých hlásek nebo morfémů, je třeba zajistit, aby měl extrahovaný exemplifikační text minimálně délku celého slova; v syntaktických výkladech bývají exemplifikace zpravidla nejrozsáhlejší, proto jsou nejvhodnější);
      - ukázky souvislých nářečních promluv (jsou tradiční součástí nářečních monografií i studentských prací; rozsah jedné ukázky bohužel většinou nepřesáhne délku odstavce; pokud jsou nářeční ukázky z různých lokalit, bývá jejich zpracování poměrně pomalé; s rostoucím rozsahem jedné ukázky nebo souboru ukázek z jedné lokality roste i efektivita zpracování nářečních textových dat);
      - materiálové studie (představují obvykle systematický výběr nářečního materiálu dle určitého tématu; materiál většinou nezahrnuje delší části textu, bývá spíše segmentován na slova

nebo velmi krátké textové úseky, proto nepředstavují nijak vydatný zdroj nářečních textů); s jistou mírou zjednodušení je možné je dělit na základě toho, zda je materiál:

  - z jedné oblasti (takovéto studie, zvláště ve své nyní již vymizelé formě soupisu různých lexikálních jednotek nebo frazémů, bývají obvykle vhodné k další extrakci materiálu, viz např. Jirsák, 1932–1934);
  - z různých oblastí napříč celým/širším územím (tyto studie bývají většinou k extrakci materiálu nevhodné, pokud nevykazují značnou míru formalizace a konzistence při lokalizaci jednotlivých prvků materiálu, srov. Kloferová a Šipková, 2018);

- dialektologické slovníky (nebývají obvykle nejlepším zdrojem trénovacích dat, protože nenabízejí nářeční texty o větším rozsahu, tyto texty mohou však být velmi kvalitní, viz SNČJ, 2011–2024; tamtéž viz také dialektologickou slovníkovou literaturu), mohou poskytovat textová data tří druhů:
  - nářeční lemmata (jde většinou o jednoslovné texty v dialektologické transkripci, zpravidla vysázené jiným typem nebo řezem písma);
  - nářeční tvary (bývají uváděny ve formě koncovek i celých tvarů, někdy je možné je extrahovat na základě typu nebo řezu písma, případně pravidelně sestavit z lemmatu a koncovky, obvykle však toto druhotné sestavení tvarů způsobuje řadu problémů a přináší materiálově nevelké zisky);
  - nářeční exemplifikace (nebývají přítomny ve všech dialektologických slovnících ani soustavně ve všech heslech jednoho slovníku, jde typicky o slovní spojení, krátké věty a souvětí, které jsou jako zdroj dat vhodnější než slovníková lemmata; často bývají vysázeny na rozdíl od ostatního textu kurzívou);
- dialektologické atlasy (nebývají vhodným zdrojem textového materiálu, srov. Balhar a kol., 1992–2011; Čižmárová, 2000; totéž platí o atlasových pasážích dialektologických monografií, srov. např. Jančáková, 1987, s. 129–173);
- dialektologické kartotéky (dialektologické kartotéky nejsou odbornou literaturou vydanou, ale bývají přípravou pro ni, často nerealizovanou; obsahují obvykle stejný typ údajů jako nářeční slovníky; v případě, že jsou strojopisné nebo tištěné, je možné z nich dobře extrahovat data např. na základě jejich polohy na kartotéčním lístku, ovšem výtěžnost, obdobná jako u slovníků, bývá snižována nepravidelnostmi, vpisky, opravami, nedůslednou úpravou, srov. ALJ, 1952–2024; kartotéky s relevantními dialektologickými daty mohou vznikat i v rámci jiných oborů, zejména etnografie);
- dialektologické databáze (bývají elektronickou obdobou dialektologických kartoték a na rozdíl od nich mohou být i přímo publikovány, srov. Stupňánek a kol., 2022; Nétek, Štrubl a Stupňánek, 2022; je obvykle možné z nich extrahovat stejné tři druhy informací, nářeční lemmata, tvary a exemplifikace, jejich elektronická forma však zpravidla zajišťuje systematické zpracování a snadnou extrakci dat);

- etnografická:
  - národopisné monografie a studie (jazykový materiál zde bývá většinou vedlejším produktem zkoumání některého aspektu lidové kultury, i tak ale může být velmi bohatý, srov. např. Zíbrt, 1909–1911; nářeční text se zde objevuje ve folklorním přepisu, a to v následujících podobách:
    - izolované nářeční lexikum (může být v odborném textu formálně odlišeno, např. uvozovkami, což ho umožňuje sesbírat, daný formální prostředek však musí být použit pouze pro vyznačení nářečního lexika, případně i ukázek promluv; další podmínkou je též možnost hromadné nebo automatické atribuce lokality u extrahovaných textů);

    - ukázky, citace nářečních promluv (popisy řemeslných postupů, zvyků, tradic, vzpomínková vyprávění…; nářeční promluvy bývají v textu graficky odlišeny, je většinou možné je extrahovat; je třeba zvážit účinnost takovéhoto postupu, pokud jsou tyto ukázky lokalizovány neformálně, ve volném textu, a není tak možná hromadná nebo automatická atribuce lokalizace);
    - ukázky lidové slovesnosti (ukázky bývají graficky odlišeny od zbytku textu, jejich využitelnost je závislá zejména na druhu lidové slovesnosti; platí o nich totéž, co o lidové slovesnosti ve specializovaných sbírkách; o nich bude pojednáno dále);
  - etnografické kartotéky (kartotéky s relevantními dialektologickými daty existují i v oboru etnografie, platí o nich totéž, co o kartotékách dialektologických, srov. Spilka, 1956–1975);
- vlastivědná (vlastivědné publikace týkající se regionů nebo obcí často obsahují kapitolu o nářečí, která může obsahovat ukázky souvislých nářečních promluv, ukázky lexika a jeho exemplifikace, případně i exemplifikace gramatických jevů; obvykle se jedná o relativně stručné pasáže, z nichž nelze vytěžit velké množství materiálu, také často citují nářeční ukázky z jiných, povětšinou jazykovědných publikací, takže může hrozit nežádoucí duplikace téhož textu v materiálu; součástí vlastivědných publikací bývají i zpravidla o poznání rozsáhlejší pasáže etnografické, které mohou nabízet rovněž určitý dialektologický materiál, srov. Stolařík a Štika, 1997–2001; Tausch, 2006; vlastivědné publikace tedy při dostatečné obezřetnosti mohou sloužit jako nevydatný zdroj obohacení trénovacích dat);
- jiná (nářeční jazykový materiál v omezené míře můžeme nalézt i v odborné literatuře z jiných oborů, jako např. orální historie, sociálněvědné, psychologické případové studie, literárněvědné práce o autorech píšících v nářečí apod.; jde však převážně o zdroje méně vhodné k získávání kvalitních a rozsáhlých dat);

- **odborné a poloodborné sbírky lidové slovesnosti, souvislých nářečních mluvených projevů** (obecně zde platí, že čím sevřenější a ustálenější útvar, tím více je ohrožena jeho nářeční autenticita, neboť ustálenější formy lidové slovesnosti částečně tendují k přijímání spisovných prvků a současně se šíří mimo místo svého vzniku i spolu s nářečními prvky z původní lokality):
  - neformální vyprávění, přepisy rozhovorů (představují nejcennější zdroj nářečních textů, takto zaznamenaný jazyk je obvykle nejautentičtější, současně mívají rozsah dostatečný k tomu, aby extrakce textů byla velmi efektivní):
    - nářeční čítanky (zdroje nářečních vyprávění, za jejich distinktivní rys, odlišující je od druhého typu, můžeme považovat dialektologický přepis těchto projevů, viz Lamprecht a kol., 1976);
    - sbírky vyprávění, nářečních rozhovorů a besed (jde opět o nářeční vyprávění či rozhovory, někdy i soubor autentických popisů zvyků a tradic, jsou zapisovány folklorním přepisem, viz např. Pospíšilová, 2016; Bachmannová, 2008);
    - ustálená a tradovaná vyprávění (útvary lidového vyprávění a prozaické epické lidové slovesnosti mohou mít různou míru ustálenosti a reprodukovat je může lidový vypravěč i běžný informátor, přednost mají spontánnější, nepřipravené podoby textů):
    - vyprávění lidových vypravěčů (většinou jde o povídky, anekdoty „ze života“, ale může jít i o pohádky a pověsti, často aktualizované, obohacené o humornou složku; jde o připravená nebo částečně připravená vystoupení, což do jisté míry může narušovat autenticitu zejména lexikálního a syntaktického jazykového plánu; svým charakterem se – i v tomto ohledu – často blíží krásné literatuře v nářečí, srov. Palátová, 1958; Satke, 1958);
    - pohádky, pověsti (např. Kubín, 1964–1971; Jech, 1959; sběry se liší v míře připravenosti, často se objevují i sbírky, které mají v nářečí jen několik textů, a ve zbytku případů jde o spisovné

parafráze; důsledně je třeba vylučovat texty, které představují směs nářečí a spisovnosti, srov. Kulda, 1874–1894; vyskytují se menšinově též případy, kdy pohádka má v nářečí jen přímé řeči; obecně je tedy třeba obezřetnosti, zda a kde konkrétně v takovéto sbírce těžit; ve sbírkách pohádek nebo pověstí pro děti, které nemají alespoň částečně odborné zaměření, se nářeční texty objevují jen zřídka, nebo jde o nežádoucí kompromis mezi nářečím a spisovným jazykem);

- přísloví, pranostiky, lidová moudra, frazeologie (viz např. Rybnikář, 2021; Stará, 2008; Zaorálek, 1996; existují v řadě forem, v mnoha z nich autentické nářečí nenacházíme nebo nacházíme silně porušené);
- písně, říkadla (představují okrajový a nepreferovaný zdroj pro trénovací data, nechceme-li trénovat přímo veršované nářeční texty; sbírky písní jsou velmi četné a mohou představovat velmi rozsáhlý zdroj textů, srov. např. Sušil, 1998; Poláček, 2010–2011; Bartoš, 2006; ale písňové i veršované texty v sobě zahrnují ve velké míře nenářeční, spisovné jevy, různé deformace slov, slovosledu, syntaxe ve prospěch rytmu, slova a zkomoleniny mimo standardní nářeční slovník, cizorodé prvky z jiných nářečí, viz k tomu např. Hošek, 1897; Horálková, 1962; Nejedlý, 2021; nadto většinou neužívají příliš širokou slovní zásobu, opakují se v nich ustálená témata, a tím se i statisticky neúměrně zvyšuje přítomnost některých slov atd.; úhrnem představují nejposlednější zdroj pro trénovací data, která mají být aplikována na běžné nářeční promluvy);

- **laická literatura o nářečí** (představuje laickou obdobu odborných dialektologických publikací, které sestavují a mnohdy i vydávají nebo ineditně šíří zájemci o nářečí a lokální patrioti):
  - slovníky, slovníčky (lidové slovníky a slovníčky jen zřídka mívají exemplifikace, mohou však v některých případech dosáhnout značné bohatosti slovní zásoby, např. Kazmíř, 2012; Borocký, 2003; extrakce dat z nich může být obtížnější, pokud si autoři nestanoví jasné formální zásady, neodlišují různé typy informací pomocí různých typů písem apod.; většinou je však možné relevantní data z lidových slovníčků získat);
  - gramatiky, učebnice (v laickém prostředí nabývá útvar nářečních učebnic a gramatik na popularitě, zvláště v posledních dekádách, srov. k tomu Stupňánek a Ireinová, 2020; některé učebnice se snaží mít seriózní, jiné však až parodický charakter, obvykle používají různé typy písma a formálních prostředků, takže přinejmenším některé druhy nářečních jazykových dat lze izolovat; samotná kvalita textů bývá spíše nižší, často se totiž nářečí objevuje v pro něj nezvyklých komunikačních situacích, ale to už velmi záleží na provedení konkrétní publikace; některé nářeční gramatiky a učebnice jsou celé sepsány v nářečí, což obvykle spíše zvýrazní jejich negativní vlastnosti stran reprezentace autentického dialektu; v úhrnu je tento typ zdroje spíše nevhodný pro akvizici trénovacích dat);

- **literatura v nářečí:**
  - krásná literatura:
    - próza (je jediným typem krásné literatury, který lze v některých případech doporučit k těžbě textových dat); nářečí se v prozaické literatuře může objevovat v různých podobách:
      - kompletní text v nářečí (prózy tohoto typu často představují rozsáhlý nářeční komunikát, který může být bohatým a pestrým zdrojem nářečních dat, srov. např. Přikryl, 1925; týž, 1928; týž, 1929; Křemela, 1940; Baruch, 1934–1937; týž 1948; Kynčl, 1928; takovéto texty mají však po stránce nářeční autenticity svá omezení: často napodobují spisovné literární vzory, kombinují spisovnou a nářeční syntax, zahrnují nebo nářečně adaptují spisovné lexikum, objevují se aktualizace výrazu, okazionalismy, autorská přirovnání a nově vytvořené frazémy, lze si také povšimnout, že z takovéto literatury lze obvykle získat mnohem více slovesných prefixací a vůbec slovotvorných odvozenin, než se podaří zachytit prostřednictvím terénních výzkumů; může to být dáno tím, že krásná literatura a sám stylový konstituent psanosti

umožňuje využívat i periferie slovní zásoby, ale zčásti může jít i o autorské novotvary; spolehlivosti nářečního literárního jazyka nenapomáhá ani častá nadsázka a humorné ladění literatury v nářečí, která autory může svádět k velmi uvolněnému a svévolnému nakládání s jazykem a obecně vede autory k výraznému upřednostňování expresivního lexika);

- nářeční přímé řeči (častým typem prózy je i spisovný text, v němž postavy hovoří v přímých nebo i nepřímých řečech v nářečí, srov. např. Zábranský, 1919; Břeněk, 1921; Javořická, 2004; Martínek, 1977; dialektické přímé řeči byly častým prostředkem realistické a naturalistické prózy i takzvaného socialistického realismu; vypravěčův text bývá spisovný, i když může být také protkaný dialektismy; přímé řeči jsou snadno identifikovatelné prostřednictvím uvozovek, je však nutná předběžná analýza textu, neboť ne všechny postavy musí mluvit nářečím, postavy venkovanů, případně městských nářečních mluvčích se mohou střetat se světem úřadů, soudů, s velkoměstským, oficiálním prostředím nebo také s mluvčími jiných nářečí a jazyků; jazyková charakteristika postav také vždy není přesvědčivá, i když text píše autochton, může dělat ústupky srozumitelnosti a kontaminovat nářečí spisovnými výrazy a spojeními);
- záměrné dialektismy (jde o dialektismy, které jsou součástí autorské intence, vědomého záměru nářeční výrazy v textu používat, ať už v pásmu vypravěče, nebo v přímých řečech; je možné je extrahovat, pokud jsou vyznačeny uvozovkami, protože jsou však obklopeny spisovným textem, je nutné se ujistit, že jejich hláskosloví a tvarosloví nebylo upraveno nebo zcela převedeno do spisovné podoby, srov. např. Herben, 1946);
- nezáměrné dialektismy (najdeme je u velkého množství autorů, jde o dialektismy většinou neuvědomované, které jsou součástí jejich přirozené slovní zásoby a neovlivněné vzděláním a četbou; tyto dialektismy nebývají v textu nijak formálně vyznačeny, dají se zčásti zvýraznit např. prostřednictvím kontroly pravopisu, jde však o velmi hubený a nespolehlivý zdroj dat, neboť daný výraz si mohl autor osvojit četbou nebo komunikací/poslechem mluvčích jiného nářečí v cílené snaze rozšířit své literární výrazové prostředky; bývá většinou hláskoslovně a tvaroslovně zformován do spisovné podoby, a vyžadoval by tedy srovnávací studium k ověření autenticity a zpětný převod do nářečí);
- nářeční překlady (jde o svébytný žánr, který se v české literatuře objevuje poměrně zřídka; mohli bychom ho ještě dále dělit na překlady, jejichž součástí je nářečí, a na překlady čistě nářeční, na překlady prózy, překlady dramat a překlady poezie, také na překlady seriózní a překlady humoristické, ale o všech typech platí z hlediska nářečních trénovacích dat totéž, totiž že obvykle obsahují příliš mnoho neautentických, případně i okazionálních prvků, a bývají tak pro naše využití zcela nevhodné; většinou bývá v překladech užíváno nářečí k charakteristice postav, které jsou zvláštním dialektem charakterizovány i ve výchozím textu; překladatel tu obvykle přizpůsobuje nářečí prostředí, jež pro ně není autentické, známý je např. Procházkův překlad Steinbeckových *Hroznů hněvu* /Steinbeck, 1941/, který užívá východočeské nářečí jako jazykovou charakteristiku lidí z Oklahomy; existují i seriózní překlady kompletních děl v nářečí češtiny, např. Homolův překlad *Dopisů poslance bavorského zemského sněmu* /Thoma, 1966/ v nářečí centrálně středomoravském nebo Krušinův překlad Hauptmannovy divadelní hry *Tkalci* v podkrkonošském nářečí /Hauptmann, 1898/; v novější době se objevuje žánr humoristických překladů, které jsou mnohdy spíše humoristickou adaptací původního díla, typickým příkladem jsou v současnosti různé překlady Saint-Exupéryho díla *Malý princ*, které testují možnosti dialektů /Eliáš a Saint-Exupéry, 2020; Odehnal a Saint-Exupéry, 2021; Hoffmanová, Bachmannová a Saint-Exupéry, 2023; Škarnétka, Kubíkova a Saint-Exupéry, 2024/, překlad je zde pojat obvykle s velkou nadsázkou);

    - *slovníček* (bývá součástí próz v nářečí nebo s nářečními prvky; obvykle jde o málo užitný zdroj textových dat);
  - poezie, umělé písňové texty (nejsou vhodným zdrojem nářečních textů, jsou deformovány ohledy na rytmus a rým, mohou se objevovat deformace jazyka, básnické novotvary i kontaminace jiným jazykovým kódem; současně jde o množstevně nepočetný zdroj ve srovnání s lidovými písněmi, srov. např. Přikryl, 1943, k písňovým textům viz např. Stupňánek a Ireinová, 2020);
  - drama (může být výjimečně použitelným zdrojem, je však třeba kritické posouzení autenticity a dialektologické kvality textu, srov. Preissová, 1910; pokud jsou nářečím charakterizovány jen některé postavy, je v tomto případě extrakce textů vybraných postav snadná);
- laické texty v nářečí:
  - sbírky nářečních vyprávění (jsou obdobou sbírek odborného charakteru, často však nejsou výsledkem terénního výzkumu, nýbrž autorské činnosti kolektivu nářečních mluvčích, případně je terénní výzkum pouze dílčím zdrojem prezentovaných textů, srov. Křižková a kol., 2010; mohou být relativně dobrým zdrojem);
  - synkretické texty v nářečí (jde o texty, u nichž není účel zcela vyhraněný, respektive směšují více různých účelů (humoristických, publicistických, vzdělávacích, memoárových apod.; řadí se sem např. nářeční články v obecních zpravodajích, nářeční zpravodajství a publicistika, fejetony, různé prozaické útvary s poučením o lokálním nářečí a zmizelém způsobu života apod.).

4.1.3 Shrnutí

Uvedená kritéria je nutné vzájemně kombinovat. Nejlepšími zdroji obvykle bývají sbírky neformálních, nepřipravených nářečních vyprávění a rozhovorů a nářeční čítanky, dále jakékoli odborně zpracované databáze s nářečním texty nebo lokalizovanými exemplifikacemi, ukázky souvislých nářečních promluv v nářečních monografiích, studentských pracích i etnografických textech, dále pak nářeční texty autochtonů bez nápadné tendence k aktualizaci výrazu. Přitom postupujeme od většího rozsahu k menšímu a později můžeme přidávat další zdroje podle jejich spolehlivosti, dostupnosti a objemu získaných textových dat.

4.2 Digitalizace (OCR) nářečních textů

4.2.0 Úvod

Ačkoli nejvýhodnější formou textu je forma elektronická (pořízená nebo editovaná člověkem), nejhojnějším a nejlépe dostupným zdrojem nářečního materiálu jsou v současnosti tištěné publikace. Abychom z nich získali textová data, je zapotřebí je digitalizovat, přesněji získat prostřednictvím skenování nebo digitální fotografie obrazy jejich stránek a na nich provést optické rozpoznání znaků (angl. optical character recognition, OCR) pomocí specializovaného softwaru.

V současnosti existuje řada nářečních publikací, které jsou k dispozici v elektronické podobě, jež byla získána prostřednictvím OCR. Jsou takto dostupné zejména v databázích veřejných knihoven,[35] ale i v jiných veřejně přístupných databázích digitalizátů.[36] Existuje také řada dobře dostupných aplikací, které OCR bez větší námahy zajistí, i firem, které digitalizaci včetně OCR nabízejí. Problémem však je, že takováto OCR nejsou pro nářeční texty vhodná, jsou totiž upravena a nastavena pro spisovné texty.

OCR software se opírá o dva základní pilíře. Jedním z nich je samotné rozpoznávání obrazové podoby

35 Digitální knihovna Kramerius, Městská knihovna v Praze aj.

36 Např. Internet Archive, Google Books.

jednotlivých tištěných znaků a druhým je potom korekce tohoto rozpoznání pomocí databáze slov daného jazyka. Předpřipravené databáze slov a tvarů v těchto softwarech jsou však pouze pro spisovné jazyky, nářeční databáze slov a jejich tvarů nejsou k dispozici (alespoň pro češtinu). OCR programy tak mají tendenci přizpůsobovat rozpoznaná slova spisovnému jazyku, čímž nářeční texty nevratně deformují. Takto chybně rozpoznaný, deformovaný text už může opravit pouze člověk, s vynaložením značných časových nákladů. Mnohem rychlejší je provést OCR znovu a správnými postupy. Pokud tedy zjistíme, že v OCR jsou nahrazena některá nářeční slova za spisovná, je nutným postupem zajistit si OCR nářečního textu svépomocí, a to nejlépe na obrazovém materiálu, který si sami vhodně připravíme. Existují pochopitelně i texty, u nichž automatické OCR ještě kontroloval a editoval člověk. V takovém případě ho můžeme využít, pokud neobsahuje jiné závažné chyby. Ve většině případů však není k dispozici žádný digitalizát a digitalizace a OCR dané publikace je jedinou cestou, jak z ní textová data vytěžit.

4.2.1 Software pro OCR

Prvním krokem jakéhokoli OCR je příprava obrazového materiálu (zpravidla skenováním), avšak tato příprava už by měla probíhat v souladu se softwarem, který má být použit. Proto nejprve pojednejme o něm. Softwaru pro OCR obrazů textu je nepřeberné množství. Pro snadnou extrakci textu a pro uživatelská nastavení, která umožní co nejlépe rozpoznávat nářeční text, je však nutné mít software, který je vybaven maximem možných funkcí. Především je důležité mít v daném softwaru následující možnosti:

- vypnout vestavěné (spisovné) slovníky;
- nastavit sestavu znaků použitých v textu, včetně speciálních znaků, které se užívají v dialektologických publikacích;
- posílit rozpoznání písma a zvláštních znaků prostřednictvím učení uživatelem;
- automaticky rozpoznávat styly, typy a řezy písem;
- automaticky rozpoznávat záhlaví a zápatí;
- automaticky rozpoznávat obrázky a ornamenty;
- automaticky rozpoznávat tabulky;
- mít možnost dodatečně editovat nebo znovu rozpoznat vybrané pasáže textu;
- exportovat text do různých formátů za nastavitelných podmínek (např. zachovat styly písma, odstranit záhlaví a zápatí, obrázky atd.).

Tyto možnosti jsou nesmírně důležité při čištění textu i při vyjímání nářečního textu z textu spisovného nebo naopak odstraňování spisovných textů z textu nářečního. V současnosti tyto možnosti nabízí jediný software, a to ABBYY Finereader (současná verze je 16, ale všechny tyto funkcionality mělo i několik předchozích verzí). Většinu těchto možností nabízí v současnosti i Adobe Acrobat Pro DC (verze 2024), problematické jsou zde však první tři uvedené body, které jsou pro OCR nářečí nejpodstatnější, tedy především vypnutí vestavěného slovníku, nastavení přesné sestavy použitých znaků a možnost uživatelského učení speciálních a nezvykle vysázených znaků. Proto je v současnosti možné pro OCR nářečí doporučit pouze program ABBYY Finereader, pokud však jakýkoli software bude mít tyto možnosti, nebo je bude jakýmkoli efektivním způsobem nahrazovat či obcházet, není nutné u Finereaderu nadále setrvávat.

4.2.2 Snímání digitálních obrazů stránek (skenování, focení)

4.2.2.1 Vhodné vlastnosti obrazů stránek pro OCR

Skeny nebo fotografie stránek je možné pořizovat v různých rozlišeních a v různých typech barevného snímání. Ačkoli samotný manuál pro software Finereader (ABBYY, 2019, s. 253) preferuje rozlišení 300 dpi, nám se při uživatelském ladění programu pro rozpoznávání nářečí osvědčilo používat rozlišení 600 dpi, ze-

jména při využití uživatelského učení speciálních (i běžných) znaků. I bez výuky znaků jsou však výsledky při rozlišení 600 dpi o něco lepší než při 300 dpi a výrobce produktu k tomuto doporučení vede spíše ohled na rychlost rozpoznávání než na jeho přesnost. Při použití vestavěných slovníků jsou rozdíly v přesnosti OCR zanedbatelné, při jejich vypnutí však patrné jsou.

Výrobce softwaru rovněž doporučuje, aby skeny byly ve stupních šedé (nikoli pouze černobílé ani barevné). S tímto doporučením lze souhlasit. Skeny ve stupních šedé jsou podstatně lépe rozpoznávány než skeny pouze černobílé (bez šedých přechodů). U barevných skenů nebo fotografií je však výsledek zcela srovnatelný se stupni šedé, a není tak nutné barevné snímky dále zpracovávat na stupně šedé nebo používat na fotoaparátu odpovídající digitální filtr. Barevné obrazy jsou pouze datově větší, což opět může do určité míry zpomalovat proces OCR (a v případě méně výkonných PC vést i k nestabilitě), barevné skeny jsou také časově náročnější na pořízení, ovšem při samotném procesu OCR nepředstavují barevné obrazy nevýhodu, pokud jde o konečnou přesnost výsledku.

Při pořizování obrazů je rovněž nutné nastavit jas, kontrast, gammu u skenerů a expozici, clonu a ISO u fotoaparátu tak, aby písmo bylo co nejzřetelnější, tedy ani přesvětlené, ani podexponované, příliš tmavé a slévající se, konkrétnější doporučení by však záviselo na použitém přístroji. U řady modernějších přístrojů lze spolehnout na automatiku.

Nejlépe se osvědčilo pořizovat skeny ve formě obrázků, nikoli PDF (nejčastěji skenery nabízejí formáty JPG a TIFF). Nevyplácí se ani skenovat přímo do projektu OCR, což program ABBYY Finereader umožňuje. Řada změn v projektu OCR je nevratná, čemuž lze předcházet neustálým zálohováním projektu OCR, ovšem vzhledem k velké datové náročnosti takových projektů je to značně nepraktické. Proto je vhodné mít externě uložené skeny jednotlivých stránek, které je možné hromadně i jednotlivě do projektu doplnit v případě, že v procesu dojde k porušení obrazu některé z nich.

4.2.2.2 Skenery

K pořízení digitálních snímků stránek je možné použít více technických prostředků.

4.2.2.2.1 Plošné (flatbed) skenery

Plošné skenery, které bývají obvykle nejsnáze dostupné, jsou většinou pro digitalizaci dokumentů nejvhodnější a nejuniverzálnější, a to i v případě, že většinu těchto dokumentů tvoří knihy. Výhodou oproti ostatním typům skenerů je možnost přitlačit knihu stránkami dolů na sklo skeneru, což většinou způsobí jejich narovnání a srovnání řádků na skenu. U silnějších knih a méně poddajných stránek však může docházet kolem vazby ke vzniku obloukovitého zakřivení stránek, a tím k deformaci (zúžení) písmen, případně i zakřivení řádku. U některých starších knih s méně dokonalou vazbou může také dojít ke zvlnění stránky, které se přitlačením knihy ke sklu ještě zvýrazní a opět deformuje vzhled písmen i rovinu řádků. I tyto problémy lze však v rámci kategorie flatbed skenerů dobře vyřešit. Ideálním řešením je mít více skenerů pro různé případy, aby digitalizace probíhala co nejhladčeji.

V kategorii flatbed skenerů existují různé typy odlišující se zejména dle typu snímače:

- CIS snímače nabízejí nejvyšší rychlost, ale nízkou hloubku ostrosti. Pro většinu knih jsou zcela postačující a skenování na 600 dpi s nimi bývá rychlejší nebo srovnatelně rychlé jako skenování na 300 dpi u jiných typů skenerů (s výjimkou CIS skenerů průtahových). Nízká hloubka ostrosti způsobuje, že při zvlnění nebo ohnutí stránky skener už nezaostří písmo, které není dostatečně blízko ke skleněné podložce, potažmo snímací hlavě (problém může vznikat už při nízkých jednotkách milimetrů). Tento typ skeneru má také potíže s průhlednými a lesklými stránkami.

- CCD snímače naproti tomu všechny tyto nevýhody nemají. Jsou tudíž vhodné pro skenování problematičtějších dokumentů. Vyrábějí se i specializované knižní flatbed skenery, které nemají po jedné straně skleněné plochy lištu, a umožňují tak skenovat knihu až k vnitřku vazby a současně ji nerozevírat více než na 90 stupňů, takže se stránka při vazbě příliš neohýbá. Všechny tyto dobré vlastnosti jsou však vykoupeny menší rychlostí skenování a vyšší pořizovací cenou CCD skenerů, také jejich většími rozměry a váhou. Není-li CCD skener k dispozici, je možné pro problematičtější dokumenty použít fotoaparát v dobrých světelných podmínkách, jeho využití má však některé nevýhody (viz 4.2.2.3).
- CMOS (CMOS CIS) snímače jsou u plošných skenerů vzácnější. V zásadě nedokážou kompenzovat problémy, které řeší CCD snímač, i když kvalitou obrazu stojí mezi CIS a CCD snímači. Ve vyšším rozlišení jsou o něco pomalejší než jednoduché CIS skenery, takže jejich použití představuje nedokonalý kompromis a za současného stavu je lepší se spíše přiklonit ke dvojici CIS a CCD skeneru. Tato technologie se však postupně vylepšuje a možná brzy dokáže integrovat výhody obou.

4.2.2.2.2 Skenery se snímáním shora

Jde o typické knižní skenery, na rozdíl od flatbed skenerů jsou knihy při skenování otočené hřbetem dolů a stránkami vzhůru, knihy lze přitom umístit na podložku nastavitelnou do tvaru V, takže kniha nemusí být při skenování plně rozevřena a stránky se díky tomu tolik neohýbají a nevlní jako při otevření knihy na 180 stupňů. Lze rozlišovat dva druhy těchto skenerů:

- Skenery s horní kamerou (zpravidla CMOS snímačem) mívají tu nevýhodu, že i při mírném rozevření knihy se stránky mnohdy ohýbají, otáčejí, takže je často nutné je fixovat buď prsty, nebo svorkami, jinak jsou řádky zakřivené nebo se sken vůbec nezdaří. Prsty i svorky se pak objevují v obrazech stránek. Použití svorek výrazněji obrazu nevadí, pokud nepřekrývá písmo, je však velmi časově náročné. Použití prstů v nemálo případech vede ke špatnému zaostření kamery a vadnému skenu. To vše vyúsťuje v pomalejší průběh skenování a často i v méně uspokojivý výsledek. Jde současně o dražší typ skeneru, nelze ho tedy bez dalšího doporučit.
- Planetární skenery jsou nejdražším typem skeneru a pořizují si je pouze specializovaná digitalizační pracoviště. Obraz snímá shora více snímačů (CCD nebo CMOS), takovéto skenery mohou být vybaveny automatickým otáčením stránek nebo dvojicí průhledných skel ve tvaru V, spouštěných na knihu shora, které zajišťují ideální polohu stránek bez nutnosti jejich jiné (improvizované) fixace. Nevýhody skeneru s horní kamerou mohou být tímto překonány, ale vysoká pořizovací cena i prostorová a váhová nadměrnost tohoto typu skeneru z něj činí výlučné a za běžných okolností nedostupné zařízení.

4.2.2.2.3 Ruční skenery

Tyto skenery pracují rovněž s dokumenty otočenými lícem vzhůru, a to ručním tažením snímacího tělesa (CIS) po dokumentu. Výsledky takovéhoto typu skenování jsou však nedokonalé jak po stránce obrazového rozlišení u tohoto typu skenerů, tak po stránce mechanického provedení skenování. Nelze je proto doporučit.

4.2.2.2.4 Průtažné skenery

Mohou pracovat pouze s volnými listy, které se umístí do podavače a které skener automaticky odebírá a skenuje. Tento typ skenování bývá v průměru nejrychlejší (využívá rovněž CIS snímač), není však vhodný pro svázané knihy, pokud nepřistoupíme k jejich rozřezání na volné listy. Za běžných okolností je tedy průtažný skener pro skenování nářečních textů málo využitelný, najde své využití pouze při skenování kartoték, archivních dokumentů nebo dokumentů vytištěných na osobních tiskárnách, nesvázaných strojopisů apod.

4.2.2.3 Fotoaparáty

Fotoaparát sice může zabezpečovat velmi rychlou digitalizaci některých dokumentů, překonávající rychlost skenerů, ale má také řadu nevýhod. K jeho použití je vhodné sáhnout, pokud nemáme k dispozici CCD skener a u CIS skeneru dochází k odleskům nebo problémům s nedostatečnou hloubkou ostrosti. Použití fotoaparátu však není jednoduché. K tomu, aby fotoaparát ideálně fungoval pro naše účely, měly by jeho vlastnosti a způsob použití splňovat řadu podmínek:

- měl by mít rozlišení alespoň 35 megapixelů (stránka A4 v maximální velikosti tak bude mít kolem 600 dpi, u menších stránek nebo při vyšších rozlišeních fotoaparátu je možné rozlišení na 600 dpi dodatečně softwarově snížit);
- je třeba zajistit jeho polohu kolmou k podložce (většinou desce stolu), nejlépe pomocí sklopného stativu (stativu s horizontálním ramenem) vybaveného vodováhami, jehož pomocí ho lze v této poloze přesně zafixovat;
- během focení by se vzdálenost objektivu od dokumentu ani zoom neměly měnit (vedlo by to k odlišné bodové velikosti písmen na každé straně a výrazně snížené funkčnosti výuky znaků v OCR softwaru);
- aby se předcházelo odleskům a ostrým stínům, je třeba zajistit rozptýlené externí světlo, focení musí probíhat bez blesku (naproti tomu neostrý stín vrhaný fotoaparátem, stativem nebo rukou tisknoucí spoušť obvykle úspěšnosti OCR nijak nevadí);
- velmi potřebná je stabilizace obrazu, neboť při stisku spouště může dojít k pohybu nebo vibraci fotoaparátu na stativu.

Takovéto použití fotoaparátu stále může trpět podobnými nedostatky jako u skeneru s horní kamerou, u fotoaparátu lze však mnohem lépe určit bod, podle kterého se automaticky zaostřuje, a tak je použití prstů k fixaci otevřených stran mnohem spolehlivější.

4.2.2.4 Shrnutí

Nejvhodnější výbavou ke skenování je kombinace plošných skenerů CIS a CCD. Případně lze CCD skener nahradit fotoaparátem se stativem s otočnou středovou tyčí. Skeny je nejvhodnější dělat v rozlišení 600 dpi, které lze na fotoaparátu napodobit rozlišením 35 Mpix u stránky A4. U vyššího rozlišení nebo menší stránky lze dpi později uměle softwarově snížit. Skenovat i fotografovat je možné v odstínech šedé nebo barevně.

4.2.3 OCR nářečního textu

4.2.3.1 OCR bez nářečního slovníku

Je potřeba předeslat, že následující zásady OCR nářečního textu jsou vytvořeny pro situaci, kdy nemáme k dispozici nářeční slovník, přesněji řečeno databázi slovníkových jednotek daného nářečí a jejich tvarů. Takový uživatelský nářeční slovník není technicky nemožné vytvořit a v programu Finereader ho použít, ale znesnadňuje to řada okolností. Existují dostatečně bohaté slovníky některých nářečí, ale chybí nástroj pro generování jejich tvarů u ohebných slovních druhů, respektive soustavná data pro takovéto generování. Současně je problém jednotlivá nářečí vyhraničit. Zásadní hranice jsou vedeny prostřednictvím izoglos markantních hláskoslovných jevů (viz 3.1), ale uvnitř každé takto vymezené oblasti vede řada dalších zásadních i méně zásadních izoglos a v mnoha oblastech by bylo možné sestavit unikátní slovník tvarů pro každou jednotlivou obec. Data z *Českého jazykového atlasu* (Balhar a kol., 1992–2011) ukazují, že izoglosy řady hláskoslovných a tvaroslovných jevů se liší u každého jednotlivého slova, tedy týž gramatický jev má různý územní rozsah v závislosti na lexikální jednotce, u které se vyskytuje. Tyto jevy se také proměňují v čase, do

hry vstupuje dubletnost, ale také variabilita způsobů zápisu nářečí, neustálená i v rámci dvou existujících transkripcí, dialektologické a folklorní. Kombinatorika všech těchto variujících faktorů je neúprosná, existuje totiž jen omezené množství dostupných nářečních textů, fakticky jsme tedy v situaci, kdy máme k dispozici malý vzorek s vysokou variabilitou. Pro strojové učení můžeme tuto variabilitu snížit procesem normalizace textu, a počet textů naopak zvýšit převodem z jedné normalizované transkripce do druhé. Ale pro OCR surových textů to není možné, protože OCR všem těmto úpravám předchází jako předpoklad a první krok. Tudíž sestavit funkční slovníky pro OCR surových textů, které by pokrývaly celé území, časový vývoj, různé varianty zápisu folklorního i dialektologického, lokální dublety a tvaroslovná paradigmata ohebných slov, není úkol, jehož řešení by bylo nasnadě. Proto v našem doporučovaném postupu počítáme s faktem, že uživatelský nářeční slovník není k dispozici.

4.2.3.2 Před samotným OCR

Předtím, než přistoupíme k samotnému OCR, je vhodné přehlédnout obrazy stránek a případně je vytvořit znovu, pokud jakýmkoli způsobem nevyhovují (chybí na nich část textu, jsou místy rozostřené, písmo je porušené odlesky, případně obrazy pocházejí z externího zdroje a nemají dostatečné rozlišení). Pořizování obrazových digitalizátů je většinou nejméně časově náročnou částí přípravy textových dat, přitom na ní závisí úspěšnost všech dalších kroků, proto se vyplatí této přípravě věnovat čas navíc.

Program ABBYY Finereader nabízí automatickou úpravu obrazů stránek, tuto automatiku je však nejlepší kompletně vypnout a provést pomocí programu úpravy ručně. Jde zejména o správnou orientaci (otočení) stránek a jejich případné rozdělení, pokud jsou naskenovány dvoustrany. Vzhledem k tomu, že při úpravách více stránek najednou neumožňuje program krok zpět, je vhodné mít obrazy stránek uložené externě, nikoli pouze v projektu OCR, aby bylo možné se k jejich původní podobě vrátit. Zvláště automatické rozdělování dvoustran způsobuje nepříjemné chyby, proto je vhodné nesprávně rozdělené dvoustrany vymazat, do projektu znovu nahrát a rozdělit ručně.

Jedinou další vhodnou úpravou je korekce zešikmení. Vyrovnání řádků mnohdy obraz stránky zcela zdeformuje, automatická oprava rozlišení obrazu funguje spíše chaoticky a v nesouladu s dalšími požadavky programu. Rozlišení je možné v programu snížit u všech stran naráz ručně a systematicky (ideálně na 600 dpi). U fotografických obrázků je možné zkusit nechat opravit lichoběžníkové zkreslení, pokud nebyly fotografie pořízeny kolmo ke knize. Takto jsou strany připraveny k OCR.

4.2.3.3 Zásady OCR nářečního textu

Všechny zásadní kroky OCR nářečního textu už zde byly naznačeny v kapitole 4.2.1.

1. **Vypnout vestavěný slovník češtiny.** V jazykových nastaveních je nutné vypnout všechny jazyky, vytvořit vlastní uživatelský jazyk založený na češtině a v něm vypnout vestavěný slovník. Vypnutý slovník zajistí, že slova podobající se slovům spisovným nebudou mylně identifikována a změněna na slova spisovná.
2. **Nastavit sestavu používaných znaků.** Nářeční publikace používají variabilní sady znaků, včetně různých znaků speciálních (viz podkapitoly 4.4, 4.5). Je zapotřebí před OCR textu nastavit sadu rozpoznávaných znaků přesně ve shodě se sadou znaků v publikaci. Nastavení znaků by mělo být jedna ku jedné, pro každý znak v dokumentu bychom měli mít přesně jeden odpovídající znak ve znakové sadě projektu OCR. V následujícím bodě si ukážeme, že takovýto postup není jedinou možností, je však dobré tuto zásadu dodržet, protože tím zlepšíme efektivitu rozpoznávání. Vypnutím vestavěného slovníku jsme oslabili efektivitu OCR a je nyní potřeba ji všemi prostředky posílit na úrovni rozpoznání znaků. Pokud přesně identifikujeme sadu znaků, bude program používat přesně ty naučené vestavěné vzory, které odpovídají znakové sadě dokumentu. Může se ovšem stát, že některé znaky

z digitalizovaného dokumentu v široké nabídce sady znaků Finereaderu nenajdeme. V takovém případě přistoupíme k výuce daného znaku, což je proces uplatnitelný i v dalších situacích a popisuje ho následující bod. Finereader se po několika pokusech naučí znak rozpoznávat a přidělovat mu znakovou hodnotu, kterou určíme.

3. **Výuka znaků (uživatelských vzorů).** Jde o proces, kdy na konkrétních příkladech z rozpoznávaného dokumentu učíme program rozpoznávat lépe jednotlivé znaky. Tento postup při správném provedení vylepšuje efektivitu OCR, což je potřeba vzhledem k nedostupnosti slovníku jako druhého pilíře OCR. K výuce znaků můžeme sáhnout kdykoli jako k prostředku zvýšení efektivity OCR, zvláště je užitečná u méně obvyklých typů a řezů písma. Platí to zejména u kurzívy, v níž je formátována většina publikovaných odborných dialektologických přepisů, výuka je ale velmi užitečná rovněž u strojopisů a u dalších specifických písem. Jejím prostřednictvím je také nutné řešit speciální znaky v dokumentu, které nelze nalézt v nabídce znaků Finereaderu, např. pokud autor daný znak do dokumentu dokresloval ručně (ve studentských a strojopisných pracích) a v podobných speciálních případech. I když předem víme, že budeme daný znak při normalizaci zjednodušovat a přepisovat shodně jako jiný, standardní znak, přesto je nejlepší v této fázi dodržovat zásadu jedna ku jedné a naučit program rozpoznávat tento speciální znak jako jakýkoli znak, který dosud nebyl použit, nejlépe však znak vzhledově co nejpodobnější. Teprve po exportu textových dat tento speciální znak soustavně nahradíme. Při výuce uživatelských vzorů je vhodné k dosažení co nejvyšší efektivity dodržovat následující doporučení:
   a. Výuku provádět jen na dobře vykreslených znacích. Chybně vykreslené znaky (nedotištěné, přeškrtnuté či podtržené tužkou, ovlivněné nečistotami papíru apod.) přeskočit.
   b. Pokud nám jde specificky o výuku jednoho konkrétního znaku nebo jeho doučení, protože v něm program soustavně chybuje, můžeme textovým polem označit právě jen daný znak (nebo slovo s ním) a výuku aplikovat na něm (na dalších výskytech v dokumentu pak opakovat).
   c. Je velmi důležité se při výuce uživatelských vzorů vyvarovat chyb. Naučené vzory se sice v programu propisují do editoru vzorů, kde je možné dělat opravy nebo jednotlivé vzory mazat, přesto se však tato chyba často dále replikuje a k jejímu potlačení je mnohdy potřeba podniknout celou výuku znovu.
4. **Nastavení OCR.** Pokud proběhla výuka znaků, je třeba nastavit použití uživatelských i vestavěných vzorů. Pokud nebyla výuka znaků provedena, pak zůstaneme u nastavení použití pouze vestavěných vzorů. Také zvolíme důkladné rozpoznávání (oproti rychlému), a pokud dokument v sobě obsahuje strukturální prvky jako záhlaví a zápatí, poznámky pod čarou nebo číslované seznamy, odsouhlasíme jejich rozpoznání.
5. **Datové typy.** Po provedení rozpoznání je vhodné zkontrolovat, zda odpovídají rozpoznané datové typy (texty textům, obrázky obrázkům, tabulky tabulkám, záhlaví záhlavím atd.), a případně udělat ruční zásahy a nechat rozpoznat chybnou stránku znovu.
6. **Možná kompenzace chyb další výukou znaků.** Při zaznamenání nějaké chyby v rozpoznání znaků je možné se ji pokusit kompenzovat prostřednictvím výuky znaků.
7. **Export dat.** Příprava pro čištění textu. Dalším krokem, který bude následovat po OCR, je čištění dat, pro které musíme data připravit. Čištění každého dokumentu probíhá jinak a v závislosti na způsobu zpracování, pro který se rozhodneme, budeme data také exportovat do patřičného formátu. Pokud jsme schopni důležitá data rozpoznat na základě stylů a řezů písma, exportujeme do formátu DOCX, případně do HTML, laické slovníčky uspořádané do protilehlých sloupců nebo tabulek můžeme exportovat jako XLSX, abychom mohli extrahovat pouze jeden sloupec; v dalších případech můžeme exportovat rovnou do TXT. Některé kroky čištění dat můžeme realizovat ještě před exportem přímo

v editoru OCR, např. odstranění popisků obrázků, které je někdy nejlepší provést ručně a orientovat se právě obrázky. Při exportu dat je obvykle vhodné všechny strukturální prvky, v nichž se nenachází nářečí (záhlaví, zápatí, poznámky pod čarou, obrázky aj.) rovnou z exportu vyloučit, což Finereader umožňuje.

4.3 Čištění a formální sjednocení textu

4.3.1 Čištění textu

Jako čištění textu označujeme proces, při němž se z dialektologických nebo folklorních textů odstraňují všechny prvky mimo čistý nářeční text. K této fázi zpracování přistupujeme v případě, že máme k dispozici elektronický text (ať už náš vlastní digitalizát, nebo elektronický text z jiného zdroje), ale ten je smíšený z nářečního textu, který chceme zachovat, a ze zbylého obsahu, který chceme beze zbytku odstranit. Tento zbylý obsah je povětšinou textový (nadpisy, poznámky pod čarou, tiráž), ale může být i obrazový (obrázky, fotografie, ornamenty, dekorace) nebo smíšený (titulní strana).

4.3.1.1 Identifikátory při čištění textu

Čištění textu nelze dobře algoritmizovat, nelze univerzálně předepisovat jeho konkrétní provedení, protože vyplývá z poměrů v každém jednotlivém digitalizátu. Na druhou stranu se vždy řídí určitými obecnými zásadami. Vždy se snažíme najít nějaký identifikátor žádoucího nebo nežádoucího prvku textu, na jeho základě tento prvek uchopit a ze zbylého textu ho vyjmout. Např. v odborných dialektologických publikacích můžeme většinou identifikovat nářeční ukázky v dialektologickém přepisu na základě toho, že jsou všechny psány kurzívou. Očistíme-li tedy text od všech ostatních typů písma, můžeme získat čistý nářeční text.

Různé identifikátory se dají uchopit v různých prostředích a programech, proto na základě přítomných identifikátorů volíme export/import dat nebo převod formátů, jak už bylo naznačeno v závěru předchozí kapitoly. Nyní podrobněji rozdělme typy identifikátorů:

- **formátování textu** – styl, velikost, řez nebo druh písma bývá klíčovým identifikátorem u dobře formálně vyvedených textů a jejich zdařilých digitalizátů; bývají to typicky odborné texty, ale není to podmínkou; uchopit část textu za tento identifikátor je možné zejména v prostředí Microsoft Word, kde existuje možnost hledat a nahrazovat na základě řady vlastností písma i odstavce, prostředí umožňuje i vyhledávání pomocí wildcards, tedy prostřednictvím zjednodušených regulárních výrazů, dané výrazy neumožňují sice vše, ale ve vzájemné kombinaci s identifikací formátování jde o velmi silný nástroj. Běžné regulární výrazy můžeme použít při vyhledávání formátování v kódu webových stránek, export do HTML umožňuje i Finereader, obvykle je však snazší pracovat s exportem do DOCX a Wordem;
- **formální vlastnosti textu** – pokud je text do určité míry formalizovaný, může být identifikátorem právě vzorec či šablona, podle kterých je text vystavěn. Například v rámci souvislé nářeční promluvy může autor zápisu dávat do závorek poznámky o gestech a jiných reakcích mluvčího, závorky jsou tedy identifikátorem a potom stačí prostřednictvím regulárního výrazu závorky a jejich obsah uchopit a odstranit. Identifikátory tohoto typu bývají velmi variabilní, mohou to být nejen různé kombinace znaků, ale třeba i klíčová slova, která se používají jen v daném kontextu, a dokáží tak povahu celého textu (např. odstavce) identifikovat. Mohou to být i specifické hlásky nebo hláskové kombinace (buď nářeční, nebo naopak spisovné). Např. přítomnost dlouhých vokálů bývá neklamným znakem spisovného jazyka v textu, který pojednává o slezských dialektech;
- **sloupec (atribut) tabulky** – možnost vytáhnout si z tabulky jeden nebo více sloupců je velmi rychlým způsobem extrakce nářečních dat, ať už jde o databázovou tabulku, tabulku na webových stránkách nebo např. o laický nářeční slovník vytvořený (nebo i digitalizovaný a exportovaný) jako excelovská tabulka;

- **vizuální a common sense identifikátory** – tyto identifikátory jsou typické při ruční extrakci textu, která může být v některých případech tou nejrychlejší cestou. Když chceme například z knihy odstranit její obsah nebo první stránky před začátkem vlastního textu knihy, použijeme vizuální identifikátory, případně naše přirozené porozumění uspořádání knihy.

Identifikátory je možné různě kombinovat, čištění textu je činnost obvykle vyžadující všímavost a vynalézavost a téhož nebo podobného efektu je možné dosáhnout mnoha různými cestami. Často je nutné, aby text prošel několika programy a uplatnily se na něm důmyslné kombinace hledání identifikátorů, než se podaří ho řádně vyčistit. Někdy je také při čištění nevyhnutelné část nářečního textu obětovat, aby se do dat nezanesly cizorodé prvky, mnohdy se v textu nedaří zjevné a funkční identifikátory najít, jde tedy o disciplínu značně kreativní a proměnlivou.

4.3.1.2 Odstraňované prvky textu

Přestože je nemožné podat konečný přehled prostředků, cest a postupů čištění textu, můžeme dobře vypočítat hlavní prvky textu, které při čištění odstraňujeme. Jsou to:

- **Začátek a konec digitalizované knihy (titulní list, úvod, doslov, vysvětlivky, rejstříky, tiráž, …):** Tyto části odstraňujeme proto, abychom eliminovali všechny spisovné metatexty, které se vztahují k předmětnému nářečnímu textu. Zpravidla je tento proces zapotřebí udělat ručně u každého jednotlivého digitalizátu.
- **Textové okolí článku v periodiku, kapitoly v knize:** Jde prakticky o týž případ jako předchozí bod, pouze textovým okolím nejsou metatexty, ale jiné texty, které nejsou předmětem zájmu.
- **Nadpisy:** Nadpisy obvykle obsahují spisovné texty. Ale i když se v nich objevuje nářečí (což jsou vesměs případy textů ve folklorním přepisu), je vhodné tyto nadpisy odstranit. I při užití nářečí jde totiž namnoze o nepřirozený, stylizovaný nářeční text, v němž se mohou objevovat nepatřičná slova jako „aneb“ nebo jiné neautentické prvky. Eliminace nadpisů je však vhodná i pro další zpracování nářečních textů, alespoň v případě, že nadpisy jsou psány majuskulemi. Zohledňovat totiž text majuskulemi v regulárních výrazech při převodu folklorního přepisu na dialektologický je mimořádně náročné a je mnohem výhodnější se takovýchto krátkých sekvencí nářečních textů zbavit.
- **Jména mluvčích, údaje o nich:** Jde o texty, které se obvykle objevují před nebo za souvislým nářečním textem. Bývají formalizované, proto i snadno uchopitelné.
- **Označení mluvčích v rozhovoru (jmény, iniciálou apod.):** V zápisech rozhovorů je obvykle text členěn prostřednictvím označení mluvčích před jejich replikou, ať už prostřednictvím iniciály/iniciál jména, plným jménem, jejich kombinací nebo jiným způsobem (např. „[Mluvčí 1]“; k označení mluvčích viz 3.2.5.2). Tyto textové úseky nejsou nářečními texty a jejich opakování v rozhovoru pouze znehodnocuje trénovací data, proto je potřeba je odstranit, obvykle skrze formální vlastnosti textu. Velmi podobně bychom postupovali i v případě dramat či scénářů v nářečí, pokud by poskytovaly autentický jazykový materiál.
- **Popisy neverbální komunikace, komunikační situace:** Jde o popisy reakcí a gest („směje se“, „přikyvuje“, „informátor ukazuje asi do výše svého pasu“), které bývají formálně odděleny od zbylého textu. Na základě zvolené formy je možné je identifikovat.
- **Okolnosti záznamu, zápisu promluvy:** Tvoří zpravidla samostatné pasáže, někdy kapitoly, jindy odstavce, oddělené od nářečních ukázek, mohou se však vyskytovat i mezi nimi. Pokud nejdříve odstraníme nadpisy a pasáže o informátorech, nemusejí být tyto části už formálně snadno identifikovatelné.

- **Vysvětlivky, poznámky pod čarou, vnitrotextové vysvětlivky v nářečních promluvách:** Obvykle se týkají významu nářečních výrazů nebo frazémů, bývají dostatečně formálně odlišeny od zbytku textu. Ve vnitrotextové podobě nemusejí být na první pohled patrné.
- **Záhlaví a zápatí stránek (jména autorů, publikace, kapitoly, čísla stránek apod.):** Při tvorbě vlastních digitalizátů je možné exportovat data bez dat ze záhlaví a zápatí, ne vždy je však záhlaví a zápatí rozeznáno korektně, můžeme mít také zdroj, který záhlaví a zápatí má, ale digitalizát je nerozlišuje; většinou jde však o ustálené texty, u kterých není obtížné nalézt identifikátor, čísla stránek se v některých případech odstraňují hůř, je však třeba tato data odstraňovat důsledně např. i proto, abychom později mohli spojit slova rozdělená na hranici stránek.
- **Členění textu číslovanými, písmennými seznamy, odrážkami:** Mohou se objevovat např. u variant pohádek, příběhů, formálně je lze dobře hromadně uchopit.
- **Obrázky, fotografie, ornamenty, dekorace stránek:** Je možné je odstranit hromadně převodem textu do formátu, který obrázky nepodporuje (např. TXT), předtím je však nutné udělat všechny kroky vyžadující zachované formátování, případně i přítomnost právě obrázků.
- **Popisky obrázků:** Pokud nemají výraznou formální vlastnost (např. specifické číslování, v dokumentu ojedinělou kurzívu), je nutné je odstranit ručně na základě vizuálního identifikátoru – obrázku.

Řada prvků obsažených v tomto seznamu přitom nemusí být v textu nápadná. Je vhodné po nich cíleně pátrat, např. prozkoumat, zda a jak jsou v nářečním textu používány různé typy závorek, číslic apod.

4.3.1.3 Extrahované prvky textu

Nepostupujeme pouze tím způsobem, že nářeční text očišťujeme od cizorodých částí, ale někdy také nářeční text (tvořící jen malou, formálně dobře uchopitelnou část textu celkového) izolujeme od zbytku elektronického textu na základě nějakého identifikátoru. Nemusí jít nutně o vynětí textu, můžeme někdy odstranit vše, co není definováno identifikátorem. Na základě identifikátorů můžeme tedy očistit především tyto prvky textu:

- **Nářeční exemplifikace v odborných textech:** Zpravidla je můžeme uchopit díky kurzívě nebo uvozovkám.
- **Ukázky nářečních promluv v odborných textech:** Mívají podobné identifikátory jako exemplifikace.
- **Nářeční přímé řeči:** Bývají většinou definovány uvozovkami.
- **Slovníková lemmata:** Mohou být identifikována sloupcem tabulky, specifickým druhem písma nebo určitým znakem či sekvencí znaků po slovníkovém lemmatu.

Vždy je třeba v datech zkontrolovat, zda daný identifikátor neuchopí i jiné než žádoucí texty. V případě odstraňování prvků textu to není tak nebezpečné, protože takovýmto způsobem data nekontaminujeme, pouze okleštíme, zde však může dojít ke kontaminaci, a proto je o to více nutná kontrola.

4.3.2 Formální sjednocení textu

Při akvizici nářečního materiálu získáme řadu textů, které budou různě formátované, a to i po jejich převedení do jednoduchého TXT formátu. Budou se řídit různými typografickými pravidly a také v nich bude řada chyb vzniklých při jejich digitalizaci. Vzhledem k tomu, že našimi dalšími cíli bude texty normalizovat a převádět z jedné normalizace na druhou, musíme nejprve zajistit jejich základní formální shody, a to shody v typografických pravidlech. Zatímco v jednom typu textů budeme mít pevné mezery (např. proto, že byly vytvořeny ve Wordu, který je vytváří automaticky), v jiném pevné mezery nebudou. Zatímco jeden text bude zapisovat měkkosti pomocí znaku „’“, jiný pomocí znaku „ˈ“. Protože máme v plánu texty sjedno-

covat transkripčně nebo v nich systematicky vyhledávat regulárními výrazy, nejprve si musíme maximálně sjednotit textové prostředí a odstranit typické chyby a nežádoucí typografické jevy (dvojí mezery, mezery na začátku nebo konci odstavce, dělení slov apod.).

- **Pevné a nezlomitelné mezery:** (Nahradíme je jednoduchým regulárním výrazem.)

  ```
  Search: ( | )
  Replace: " "
  ```

  (Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

- **Sjednocení tří teček a případných chyb jejich OCR:**

  ```
  Search: [\.,•··]{3}
  Replace: …
  ```

- **Chybějící mezera po interpunkci** (typ „dost.Byl“)**:**

  ```
  Search: ([\.:;\?!…])(\b)
  Replace: \1 \2
  ```

- **Tabulátory:**

  ```
  Search: \t
  Replace: " "
  ```

  (Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

- **Znaky měkkosti (apostrofy):** (Znaky měkkosti, objevující se zpravidla jen v dialektologické transkripci, chceme sjednocovat na znak „ʼ“, bývají však rozpoznány nebo používány i znaky „,“ a „´“. Ve folklorní transkripci je třeba všechny apostrofy odstranit.)

  U dialektologických textů můžeme postupovat buď jednoduchým nahrazením:

  ```
  Search: [ʼ'´]
  Replace: ʼ
  ```

  nebo se můžeme pojistit a nahrazovat znak měkkosti jen v přesných souvislostech. Potom nahrazujeme v jednotlivých textech na základě jejich nářečí či nářeční skupiny/podskupiny:

  U severní východomoravské nářeční podskupiny (3-2) a vzácně i u jižní východomoravské nářeční podskupiny (3-1) na Břeclavsku a u jihozápadočeské nářeční podskupiny (1-3), konkrétně v okolí Soběnova, nahrazujeme takto:

  ```
  Search: ([bpvfm])( ?)[ʼ'´]
  Replace: \1ʼ
  ```

  Ve slezskomoravské nářeční podskupině (4-1) nahrazujeme tímto regulárním výrazem:

  ```
  Search: ([szc])( ?)[ʼ'´]
  Replace: \1ʼ
  ```

  Ve slezskopolské nářeční podskupině (4-2) pak tímto:

  ```
  Search: ([bpvfmkg])( ?)[ʼ'´]
  Replace: \1ʼ
  ```

V případě, že transkripce značí přízvuky „ˈ“ nebo rázy „ˀ“, je třeba postupovat opatrněji a zajistit jejich odstranění dříve, než přistoupíme k tomuto typografickému sjednocení.

Ve folklorních textech všechny apostrofy (v kterékoli jejich grafické verzi) odstraňujeme. Mohou se v textech objevit obvykle ve třech funkcích:

- měkkost u konsonantů – děje se velmi vzácně, ale zvláště někteří starší autoři s méně vyhraněnými pravidly přepisu zapisovali měkkosti i v textech odpovídajících folklorní transkripci;
- příčestí minulé typu „vlez'“, „pad'“ – jev zániku koncového *-l* se objevuje v nářečích českých (1) a slezských (4) a v severní části severní východomoravské nářeční podskupiny (3-2); apostrof značí nepřítomnost tohoto koncového formantu;
- příklonné *-s* druhé osoby singuláru préterita, zvláště po slovních druzích mimo slovesa, např. „hned's mu měl jednu vrazit“, „kde's byl?“, „to's neměl“ – jev se může objevit prakticky po celém území, není však vlastní nářečím slezským, kde se vyskytuje jen jako novější, nepůvodní jev u mladších generací a ve městech.

Protože v normalizovaných folklorních přepisech apostrofy nevedeme (viz 4.4.3), stačí v těchto textech jednoduché odstranění:

```
Search: ['`´]
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

- **Dvojí mezera:** (Nahrazení dvou mezer za jednu mezeru. Tento postup je třeba opakovat, dokud bude něco nahrazovat.)

```
Search: "  "
Replace: " "
```

(Uvozovky jsou použity k manifestaci běžných mezer, nejsou součástí regulárního výrazu.)

- **Mezera před interpunkcí:**

```
Search: " "([\.:;\?!…\)])
Replace: \1
```

(Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

- **Mezera po závorce:**

```
Search: (\()" "
Replace: \1
```

(Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

- **Měkké zalomení řádku:**

```
Search: \r?\n
Replace: \r\n
```

- **Mezera na začátku odstavce a na konci odstavce:** (Pokud byly odstavce odsazovány pomocí více mezer, krok odstranění dvojích mezer z nich udělal mezeru jedinou, proto už vícenásobné mezery není třeba regulárním výrazem řešit.)

```
Search: (\r\n | \r\n)
Replace: \r\n
```

- **Chybné zalomení pokračujícího textu:** (Při OCR je někdy nový řádek s pokračujícím textem rozeznán jako nový odstavec; je tak rozdělena věta/souvětí.)

  ```
  Search: ([^\.:\?!…])\r\n
  Replace: \1" "
  ```

  (Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

- **Dvojí mezera**: (Nutné opakování jednoho z předchozích kroků. Nahrazení dvou mezer za jednu mezeru, tentokrát stačí jedno opakování.)

  ```
  Search: "  "
  Replace: " "
  ```

  (Uvozovky jsou použity k manifestaci běžných mezer, nejsou součástí regulárního výrazu.)

- **Dvojí odstavec:** (Mezery mezi odstavci nejsou ve výsledném textu žádoucí, napomůže to též dalšímu kroku spojení slov rozdělených napříč stránkami. Krok je třeba opakovat, dokud něco vyhledá.)

  ```
  Search: \r\n\r\n
  Replace: \r\n
  ```

- **Rozdělení slov napříč stránkami:**

  ```
  Search: ([--])\r\n
  Replace: \1
  ```

- **Rozdělení slov původně napříč řádky:** (Často jsou při OCR spojovníky užité pro dělení slov v textu zachovány, nebo jsou rozpoznány jako pomlčky, nacházíme tak v textu případy jako „dob-ře“, „dob- ře“, „dob–ře“ a „dob– ře“. Spojovník se jinak užívá v nářečních textech jen u částice „-li“ a jejích variant „-ly“, „-lę“, „-le“, „-lə“ a „-i“ (případně jiných znakových forem těchto částic, srov. tabulka 4.15), další užití spojovníku jsou v nářečních textech natolik řídká, že je možné je zanedbat. Proto postupujeme následujícími kroky.)

  Nejprve zafixujeme částici „-li“, případně její varianty, které se v textech vyskytují, tj. nahradíme u nich spojovník za unikátní kombinaci znaků. Počítáme i s případy, kdy byla daná skupina rozpoznána nekorektně:

  ```
  Search: ([--])( ?)(li|ly|lę|le|lə|i)\b
  Replace: §§§\3
  ```

  Poté spojíme slova rozdělená spojovníkem nebo pomlčkou:

  ```
  Search: \b([--])( ?)\b
  Replace:
  ```

  (Výraz v pozici replace zůstává prázdný.)

  Nakonec vrátíme spojovník výrazům s „-li“:

  ```
  Search: §§§(li|ly|lę|le|lə|i)\b
  Replace: -\1
  ```

- **Odstranění majuskulí uprostřed a na konci slova psaného malými písmeny:** (Chybou OCR někdy dochází k situaci, kdy je minuskule rozpoznána jako majuskule /„tenČí“, „sVážej“/. U znaků „V“, „C“, „Č“, „S“, „Š“, „Z“, „Ž“ se to děje nejčastěji a současně právě u těchto znaků je rozpoznání kvality znaku také většinou správné, neboť jsou svým minuskulím nejpodobnější. Oproti tomu ostatní takové znaky bývají většinou rozpoznané chybně, nebo se ocitly uprostřed nepřerušeného řetězce vlivem nerozpoznání mezery mezi slovy. Jejich změna na minuskule však má většinou smysl, protože umožní další hromadné úpravy v textu, které vycházejí z předpokladu, že velká písmena jsou jen na začátku slova.)

  Nejprve si ve všech takovýchto slovech nepatřičná velká písmena označíme:

```
Search: \b([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ]?[a-ěščřžýáíéóúůďťň]+)
([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])([a-ěščřžýáíéóúůďťň]*)\b
Replace: \1{\2}\3
   Options: case sensitive

Search: \b([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])
([a-ěščřžýáíéóúůďťň]+)\b
Replace: \1{\2}\3
   Options: case sensitive

Search: \b([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ]?[a-ěščřžýáíéóúůďťň]+)
([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])([a-ěščřžýáíéóúůďťň]*)([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])
([a-ěščřžýáíéóúůďťň]*)\b
Replace: \1{\2}\3{\4}\5
   Options: case sensitive

Search: \b([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])
([a-ěščřžýáíéóúůďťň]+)([A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇ])([a-ěščřžýáíéóúůďťň]*)\b
Replace: \1{\2}\3{\4}\5
   Options: case sensitive
```

  V jednotlivých regulárních výrazech je třeba dosadit sady znaků podle situace v daném textu.

  Poté postupně procházíme celou abecedu a majuskule ve složených závorkách nahrazujeme za minuskule. Tímto způsobem můžeme opravit i některé soustavné chyby, které v rozpoznání znaků proběhly:

```
Search: \{A\}
Replace: a
   Options: case sensitive

Search: \{Á\}
Replace: á
   Options: case sensitive

Search: \{B\}
Replace: b
   Options: case sensitive
```

Uvedený postup kroků typografického sjednocení textů má svou logickou strukturu a je vhodné ho zachovat v pořadí, které je uvedeno.

Je možné, že proces OCR v některých textech vytvoří i jiné soustavné chyby, které půjde odstranit regulárními výrazy, zde jsou řešeny pouze chyby a nejednotnosti nejtypičtější a nejčastější.

4.4 Normalizace folklorního textu

4.4.0 Úvod

Poté, co očistíme data na čistě nářeční text a sjednotíme je po typografické stránce, je třeba sjednotit také způsob zápisu nářečí, normalizovat ho. Je při tom nutné oddělit texty ve folklorním přepisu a v přepisu dialektologickém, které jsou založeny na odlišném principu. Folklorní přepis, o který nám jde zde, je založen zejména na fonologickém principu, ale obecně vlastně na všech principech českého pravopisu. Dialektologický přepis stojí především na principu fonetickém.

4.4.1 Problém sjednocení folklorních textů

Oproti dialektologickému přepisu pracuje folklorní přepis s užším spektrem znaků. V řadě případů si může vystačit se znaky standardní české abecedy. Zejména to platí u většiny nářečí české nářeční skupiny, jejichž fonologie nebývá příliš odlišná od fonologie spisovného jazyka. Ale jsou i nářečí, jejichž fonologie se od spisovného jazyka liší tak, že by potřebovala své vlastní znakové vyjádření. V takovýchto případech autoři volí několik možných řešení:

- zvolí pro dané fonémy znaky v souladu s tradicí českých odborných zápisů nářečí;
- použijí vlastní speciální znaky;
- zůstanou u standardních znaků spisovného jazyka, takže jeden znak použijí pro označení dvou různých fonémů.

Tyto tři přístupy ke grafické reprezentaci fonému se mohou uplatňovat i v jednom textu (u různých fonémů). Největší komplikace nám však způsobí vždy poslední jmenovaný přístup, kdy jeden znak reprezentuje více fonémů, a není tedy jednoznačný. Příkladem může být používání jednoho znaku „l“ pro tvrdé „ł“ i netvrdé „l“. Kdyby takovýchto případů nebylo, normalizace folklorních textů by byla velmi jednoduchá, protože by se pouze sjednotily používané znaky. Pokud ale někteří autoři pro dvojici fonémů používají dva znaky a jiní jeden, musíme rozhodnout, kterému přístupu dáme přednost. Technicky snadným řešením je převést všechny texty na společného jmenovatele; tedy pokud některý autor nebo někteří autoři dvojici fonémů nerozlišují a používají pro ni pouze jeden znak, přikročíme k tomu, že u všech ostatních textů převedeme danou dvojici znaků na znak jediný. Tím však současně ochuzujeme zápis o důležité dialektologické informace, a pokud chceme v budoucnu převádět folklorní transkripci na přepis dialektologický, kde tyto znaky rozlišeny být musí, poškozujeme si závažně data. Jestliže bychom však chtěli jít opačným směrem a do textů, ve kterých nejsou některé dvojice fonémů znakově rozlišeny, tato rozlišení dodatečně doplnit, zjistíme, že to algoritmickými postupy není možné. Není to možné bez relativně úplné databáze slovních tvarů daného nářečí, jejíž absenci a obtížnou vytvořitelnost v této fázi zpracování dat jsme zdůvodnili už v části 4.2.3.1. Zde bychom ji navíc potřebovali i se vzájemnými převody mezi různými typy zápisu. Zdánlivě neřešitelný problém můžeme překlenout prostřednictvím provizorního řešení, které nazýváme termínem „obojetnost“.

4.4.2 Obojetnosti

Pokud text znakově nerozlišuje dvojici fonémů, které jsou v jiných folklorních textech rozlišovány, zavedeme provizorně do textu možnost, že jde o oba fonémy (resp. oba příslušné znaky). Formálně jsme pro ně

zavedli takovýto zápis: „k[l/ł]uk“, „za sk[l/ł]em“, „není [u/u̯]hod“, „tos [u/u̯]hod“, kde první znak v hranaté závorce je znak, který byl v textu původně. Jak je vidět, nejde o případy, u nichž by člověk nedokázal rozhodnout, který znak je adekvátní zvolit, ale o případy, kdy pro toto rozhodnutí nemáme jednoduchý algoritmus a nemáme kapacity v rozsáhlých textech uplatnit lidskou (ruční) opravu.

V dané fázi tedy rozhodnutí dočasně přeskočíme a ponecháme obě možnosti s tím, že desambiguaci obojetností zajistí až strojové učení. Část materiálu totiž sice bude v daných případech obsahovat obojetnosti („za sk[l/ł]em“), ale druhá část bude jednoznačná („za skłem“). V pozdějších fázích také budeme schopni převádět prostřednictvím strojového učení dialektologický přepis na přepis folklorní a v dialektologických přepisech dané obojetnosti nebudou, nebudou tudíž ani ve folklorních transkripcích z nich vzniklých. Strojové učení by tak mělo mít dostatečný trénovací materiál k tomu, aby umělo tyto obojetnosti desambiguovat.

4.4.3 Normalizovaný folklorní text

U normalizovaného folklorního přepisu budeme tedy do velké míry respektovat jeho fonologický princip a pro specifické nářeční fonémy užívat zvláštní znaky, odlišené od znaků jiných. Na druhou stranu je třeba respektovat i jiné principy, kterými se běžný folklorní přepis v souladu s českým pravopisem řídí. Je to především princip historický a princip tradiční. Na podkladě principu historického se například rozlišují dvojice znaků, které sice historicky označovaly dva fonémy, v současnosti však označují foném jediný, který vznikl jejich splynutím. Platí to pro znaky „i-y“ a „í-ý“ i pro znaky „ú-ů“.[37] Na historických principech stojí i kombinační zápis měkkých konsonantů „di“, „ti“, „ni“, „dě“, „tě“, „ně“ a zápis jotace po labiálách prostřednictvím „bě“, „pě“, „vě“, „fě“, na části území „mě“, to se však na většině území místo s jotací vyslovuje s „ň“. Pro úplnost doplňme, že i znaky „e-ě“ označují v současnosti, oproti stavu historickému, jediný foném, pouze v případě „ě“ jde o znak kombinační. Princip tradiční pak určuje pravopis přejatých slov a jejich odvozenin, který si zachovává všechny nebo některé prvky svého původního jinojazyčného pravopisu („hobby“ = *hobi*, „harmonikář“ = *harmonikář*, „helium“ = *hélijum*). Na podobných principech pak stojí také tradice nerozlišovat ve folklorním přepisu i ty fonémy, jejichž měkkostní opozice má fonologickou povahu, což platí především pro měkké labiály na Valašsku („b“ = *b-b′*, „p“ = *p-p′*, „m“ = *m-m′*, „v“ = *v-v′*, „f“ = *f-f′*) a sykavky a polosykavky ve slezských nářečích („s“ = *s-s′/s-ś*, „z“ = *z-z′/z-ź*, „c“ = *c-c′/c-ć*, „š“ = *š-š′*, „ž“ = *ž-ž′*, „č“ = *č-č′*).[38] V tomto směru nemá smysl jít proti tradici a proti způsobu zápisu prakticky všech existujících textů ve folklorním přepisu.

Soubor nářečních znaků, který doporučujeme používat v normalizovaném folklorním přepisu nad rámec standardní české spisovné abecedy, je následující:

Tabulka 4.1 Soubor nářečních znaků normalizovaného folklorního přepisu

nářeční znak	název
w	obouretné *v*
u̯	neslabičné *u* / obalované *l*
ę	široké *e*
ǫ	široké *o*
ł	tvrdé *l*

nářeční znak	název
ł́	dlouhé tvrdé *l*
ĺ	dlouhé netvrdé *l*
ŕ	dlouhé *r*
dz	dz, asibilované *ď*
dž	dž, asibilované *ď*

37 V případě „ú-ů“ šlo historicky o dlouhý vokál a diftong, které byly těmito znaky původně označovány.

38 U tvrdých *dz*, *dž* nejde o fonémy, ale o alofony tvrdých *c*, č, fonologickou platnost mají pouze jejich měkké varianty, vzniklé asibilací *ď*.

4.4.4 Normalizace folklorního textu podle nářečních podskupin

Trénování strojového učení by mělo probíhat na podkladu materiálu uspořádaného podle nářečních podskupin. Nářeční podskupiny jsou totiž dostatečně velké na to, aby z nich bylo možné získat patřičně rozsáhlá trénovací data, a současně dostatečně malé na to, aby byly ještě relativně nářečně jednotné. Proto pravidla normalizované transkripce definujeme pro jednotlivé nářeční podskupiny.

Je třeba přitom vzít v úvahu nejen to, že nářeční podskupiny jsou jazykově nejednotné, ale že v nich také probíhá permanentní jazykový vývoj, který ovlivňuje jak jejich hranice, tak zejména výskyt jednotlivých znaků v přepisech těchto nářečí. Protože chceme pokrýt větší časový rozsah trénovacích dat v souladu s materiálem, který máme k dispozici, počítáme i s vývojovými posuny během tohoto období a vyznačujeme u konkrétních znaků jejich obecné vývojové tendence. Vždy takto označujeme pouze znaky, které jsou relativně řídké.

označení vývojové tendence	význam
(w)	znak, který je řídký a zánikový, objevuje se jen v nejstarších textech
{g}	znak, který je řídký a progresivní, zpravidla proniká do zápisu nářečí nebo zvyšuje svou četnost s průnikem spisovnosti, slov cizího původu nebo obecné češtiny

Je důležité zdrůraznit, že v přehledech normalizovaných znaků nejde o popis hláskosloví v dané nářeční podskupině, ale o sadu znaků užívanou k normalizované transkripci. Ta samozřejmě s hláskoslovím souvisí, ale není nikterak jeho popisem.

V každé tabulce rozlišujeme tři kategorie znaků. Standardní znaky/grafy, digrafy a nářeční znaky. Standardní znaky jsou běžné znaky spisovné české abecedy mimo digrafy. Digrafy jsou ustálená spojení dvou znaků, které popisují jeden foném. Nerozlišovali jsme tu už, zda dané fonémy jsou, či nejsou ve spisovném jazyce, v některých dialektech jsme rozlišovali i taková tautosylabická spojení hlásek, která se v jiných vyvíjela jako jeden foném. Vždy však jde o digrafy složené ze standardních znaků. S výjimkou „ch“ je třeba brát v úvahu, že všechny uvedené digrafy mohou v textech označovat jeden i dva fonémy, a mít tak dvě interpretace: „v**ej**ška“ vs. „s**ej**em“, „**au**to“ vs. „n**au**čit“, „l**eu**koplast“ vs. „n**eu**dav se“, „cho**dz**ili“ vs. „po**dz**emsky“ ap. Ve třetí části tabulky jsou pak uvedeny jednopísmenné nářeční znaky, které jsou mimo sadu standardních znaků.

Materiálově ve všech následujících tabulkách a poznámkách vycházíme z *Databáze souvislých nářečních textů* (DSNT, 2017–2024), *Archivu lidového jazyka* (ALJ, 1952–2024), *Dotazníku pro ČJA* (1964–1976), geoportálu *DiaMa* (Štrubl, Nétek a Stupňánek, 2022) a *Českého jazykového atlasu* (Balhar a kol., 1991–2011).

1 Česká nářeční skupina

1-1 Severovýchodočeská nářeční podskupina

Tabulka 4.2 Soubor znaků normalizovaného folklorního přepisu pro severovýchodočeskou nářeční podskupinu

1-1 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	{g}			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

1-1 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky		
znaky samohlásek	*krátké*							
	dlouhé	ej	ou	{au}	{eu}			
znaky souhlásek	*znělé*							
	neznělé	ch						
	jedinečné					u̯	(w)	(ł)

Poznámky:

- Transkripce nezohledňuje podkrkonošskou vokalizaci původně slabikotvorného *r* a slova jako *parňí*, *daržeť* zapisuje s „e“ jako „perní“, „deržeť“, ə nemá fonologickou platnost. Až na vzácné výjimky se ve folklorních transkripcích neobjevuje.
- Znak pro neslabičné „u̯“ bývá běžně zapisován jako „u“, a je proto potřeba ho v těchto případech nahrazovat obojetností [u/u̯].
- Hlásky obouretné *w* a tvrdé *ł* jsou zánikové a znaky pro ně se objevují jen v nejstarších textech.
- V souladu se všemi ostatními nářečními podskupinami se zde objevuje i řídký znak „w“ z přejatých slov. V této nářeční podskupině však konkuruje znaku pro staré obouretné *w*.
- V souladu se všemi ostatními nářečními podskupinami se zde objevuje řídký znak „x“ z přejatých slov pro neznělé i znělé spojení hlásek *ks*, *gz*.
- V souladu se všemi ostatními nářečními podskupinami jsou označeny jako řídké a progresivní i digrafy „au“ a „eu“, které se objevují výhradně ve slovech cizího původu.
- V souladu s ostatními nářečími české nářeční skupiny se znak „g“ objevuje pouze v přejatých slovech, a proto je označen jako jev progresivní, který je zmnožován přejímkami z cizích jazyků (prostřednictvím jazyka spisovného). Tato slova bývají spíše při periferii slovní zásoby. Navíc mají české dialekty obecnou tendenci k neznělosti a zvlášť v kontrastu s moravskými dialekty vynikne jejich adaptace výrazů jako: „kuláš“, „cikán“, „katě“, „inkoust“, „brikáda“, „kuma“ (srov. Balhar a kol., 2005, s. 316–323), ale i „hankár“, „chulikán“, „helikon“ aj. „G“ je proto označeno jako progresivní pouze v dialektech české nářeční skupiny.
- Vokály „é“ a „ó“ nejsou označeny jako znaky řídké, a to vzhledem k neemfatickému dloužení typu „póle“, „dóle“, „Józef“, „léžet“, „vjéřit“, „zéli“. Emfatické dloužení, typ „neé“, citoslovce a cizí slova nejsou brány v úvahu.

1-2 Středočeská nářeční podskupina

Tabulka 4.3 Soubor znaků normalizovaného folklorního přepisu pro středočeskou nářeční podskupinu

1-2 FOLKLORNÍ PŘEPIS		**standardní znaky**												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	{g}			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

1-2 FOLKLORNÍ PŘEPIS		**digrafy**				**nářeční znaky**
znaky samohlásek	*krátké*					
	dlouhé	ej	ou	{au}	{eu}	
znaky souhlásek	*znělé*					
	neznělé	ch				
	jedinečné					(u̯)

Poznámky:

- Znaky „é“ a „ó“ jsou o něco vzácnější než v podskupině 1-1, přesto se ještě objevují poměrně často („nahóře“, „Bóží“, „péři“, „Véna“, sufix „-ové“). Nejsou tedy označeny jako řídké.
- Znak „u̯“ se může v rámci středočeské nářeční skupiny objevovat, ale spíše už jen v lexikalizované podobě v malém počtu slov („zrou̯na“, „prau̯divej“, „vejstau̯nost“).

1-3 Jihozápadočeská nářeční podskupina

Tabulka 4.4 Soubor znaků normalizovaného folklorního přepisu pro jihozápadočeskou nářeční podskupinu

1-3 FOLKLORNÍ PŘEPIS		**standardní znaky**												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	{é}		í	{ó}	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	{g}			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

1-3 FOLKLORNÍ PŘEPIS		**digrafy**						**nářeční znaky**	
znaky samohlásek	*krátké*								
	dlouhé	ej	(yj)	ou	(uu)	{au}	{eu}		
znaky souhlásek	*znělé*								
	neznělé	ch							
	jedinečné							(u̯)	(ł)

Poznámky:

- Znaky „é“, „ó“ jsou na rozdíl od ostatních podskupin české nářeční skupiny označené jako řídké a progresivní, protože se prakticky neobjevují v žádných domácích slovech mimo typ „neé“, emfatické dloužení a citoslovce.
- Na Chodsku, případně i v jeho okolí se může objevovat digraf „yj“ namísto „ej“, velmi vzácně též digraf „uu“ (který v daných pozicích může splývat v „ú“). Pokud autor v dané lokalitě zapisuje „ej“, respektujeme to a digraf nenahrazujeme, neboť první složka diftongu může zaznívat jako úzké *e*, směrem k jihozápadním hranicím se toto úžení zintenzivňuje, může však přesto zůstávat v *e*-ové podobě „ej“, zvláště směrem k současnosti.
- „U̯“ a „ł“ jsou archaické jevy pozůstatků tvrdého *ł*, které se mohou objevovat jen v nejstarších zdrojích z jižního Doudlebska, a to zcela reliktně.
- Ve stejné oblasti jsou zánikové a ve folklorním přepisu nijak nereflektované měkké retnice (*b′, p′, m′, v′, f′*).
- Hláska *ə* (*prvňə*, *zedňak*, ale také *bəl*, *Pəlzeň*) se ve folklorním přepisu reflektuje zpravidla jako znak „i“, nebo případně „y“ dle historického principu.
- Prakticky se v této podskupině téměř neobjevují případy, kdy je nutno zavádět obojetnosti.

1-4 Českomoravská nářeční podskupina

Tabulka 4.5 Soubor znaků normalizovaného folklorního přepisu pro českomoravskou nářeční podskupinu

1-4 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	{g}			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

1-4 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky
znaky samohlásek	*krátké*					
	dlouhé	ej	ou	{au}	{eu}	
znaky souhlásek	*znělé*					
	neznělé	ch				
	jedinečné					(u̯)

Poznámky:

- Znaky „é“, „ó“ se v těchto dialektech opět objevují v domácích slovech (*céra*, *péři*, *léži*, *chódi*, *hóji*, *dóm*).
- Hláska *ə* se nepravidelně objevuje zejména ve slově *bəl*, případně i v jiných slovech se skupením *il* > *əl* (*vobəlí*). Lze ji považovat za alofon *i*. Ve folklorních textech bývá zapisována jako „y“ nebo „i“ dle kontextu.
- Hláska *y* ve skupení *dý*, *tý*, *ňy* a ojediněle i mimo něj není ve folklorním přepisu reflektována a distribuce znaků „i“ a „y“ se drží historických principů. Jde o zánikové jevy, které nemají fonologickou povahu.
- V oblasti okolo Třeště jsou zánikové a ve folklorním přepisu nijak nereflektované měkké retnice (*b′, p′, m′, v′, f′*).
- Znak „u̯“ se objevuje vzácně a pouze ve starších textech, i zde jde pouze o lexikalizované případy („kreu̯“, „zrou̯na“, „prau̯da“; blíže k nim a dalším lexikalizovaným případům viz Utěšený, 1960, s. 146–151).

2 Středomoravská nářeční skupina

2-1 Centrální středomoravská nářeční podskupina

Tabulka 4.6 Soubor znaků normalizovaného folklorního přepisu pro centrální středomoravskou nářeční podskupinu

2-1 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		{í}	ó	{ú}	{ů}	{ý}					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

2-1 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*					ę̯	ǫ̯
	dlouhé	{ej}	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*						
	neznělé	ch					
	jedinečné						

Poznámky:

- Na rozdíl od českých dialektů v užším smyslu a v souladu se všemi následujícími podskupinami je znak „g“ vyňat z řídkých znaků. Nejenže mají moravské a slezské dialekty mnohem větší tendenci k výskytu „g“ v přejatých slovech („cigán“, „gatě“, „ingóst/ingúst/ingust“, „gořalka“), ale současně tato slova bývají mnohem blíž centru slovní zásoby („legát“, „gazda“, „cugr“), „g“ se vyskytuje hojně nejenom v substantivech, ale v mnohem větší míře i v dalších slovních druzích, např. ve slovesech („flágnót/flágnút/flagnut‘“, „džgat/džgac“, „mignót/mignút/mignuč“, „hongat“ aj.) a v expresivech domácího původu („glgat“, „gagotat“, „mňágat“, „mégat“ aj.).
- Funguje zde relativně důsledné krácení, takže znaky „í“, „ú“, „ů“ a „ý“ představují progresivní jev ústupu tohoto krácení.
- Diftongy „ej“ a „ou“ se zde sice historicky monoftongizovaly na „é“ a „ó“, ale od 20. století zde funguje progresivní tendence k opětné diftongizaci, která zprvu postupovala od západu, později se objevuje rozptýleně na celém území.
- Původní *u* a *y* se historicky vyvinuly v široké vokály, nynější grafém „u“ je výsledkem krácení z dlouhého *ú*, proto jsou znaky „ú“, „ů“ řídké. „Y“ je výsledkem úžení *„é“* v *„í“* po tvrdých nebo obojetných („to je hlópy“).
- Široké vokály bývají někdy ve folklorních textech označovány „ê“, „ô“ nebo řidčeji „è“, „ò“. Tyto znaky se dají jednoduše nahradit.
- U textů, které nerozlišují široké vokály, je třeba zavést obojetnosti „[e/ę]“, „[o/ǫ]“.

2-2 Jižní středomoravská nářeční podskupina

Tabulka 4.7 Soubor znaků normalizovaného folklorního přepisu pro jižní středomoravskou nářeční podskupinu

2-2 FOLKLORNÍ PŘEPIS		**standardní znaky**												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

2-2 FOLKLORNÍ PŘEPIS		**digrafy**			
znaky samohlásek	*krátké*				
	dlouhé	{ej}	{ou}	{au}	{eu}
znaky souhlásek	*znělé*				
	neznělé	ch			
	jedinečné				

Poznámky:

- Diftongy „ej“ a „ou“ jsou výrazným progresivním jevem v této podskupině.
- Redukovaný vokál *ə* ve znojemském typu („ťəcho“, „ňəc“) a rozptýleně i jinde se ve folklorní transkripci manifestuje jako znak „e“ nebo absencí znaku. Jde o zanikající jev.
- Vokál zaokrouhlené dlouhé *ǫ́* se reflektuje jako „á“.

2-3 Západní středomoravský okrajový úsek

Tabulka 4.8 Soubor znaků normalizovaného folklorního přepisu pro západní středomoravský okrajový úsek

2-3 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

2-3 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*						
	dlouhé	{ej}	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*						
	neznělé	ch					
	jedinečné					(u̯)	(ł)

Poznámky:

- Diftongy „ej“ a „ou“ pronikají do této podskupiny nejvýrazněji ze všech středomoravských dialektů.
- Zábřežský redukovaný vokál *ə* za původní *y* se ve folklorní transkripci zapisuje jako „y“.
- Na Zábřežsku se omezeně objevuje neslabičné „u̯“ za „v“ v lexikalizovaných případech.
- Zánikové je na Zábřežsku rovněž tvrdé „ł“.

2-4 Východní středomoravský okrajový úsek

Tabulka 4.9 Soubor znaků normalizovaného folklorního přepisu pro východní středomoravský okrajový úsek

2-4 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

2-4 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky
znaky samohlásek	*krátké*					
	dlouhé	{ej}	{ou}	{au}	{eu}	
znaky souhlásek	*znělé*					
	neznělé	ch				
	jedinečné					ł

Poznámky:

- V oblasti postupně dochází k ústupu tvrdého *ł*. V případě, že není dvojí *l* zapisováno, je na zvážení, zda zavádět obojetnost „[l/ł]" pro danou dobu a lokalitu. V průběhu 20. století už vyslovovali dvojí *l* jen někteří mluvčí. Určitým vodítkem může být mapa ústupu dvojího *l* (Stupňánek a Vondráková, 2022d), je však třeba brát v úvahu, že za vyznačenou hranicí zachovaného dvojího *l* nebylo užívání zcela důsledné.
- V oblasti Holešovska jsou zánikové a ve folklorním přepisu nijak nereflektované měkké retnice (*b'*, *p'*, *m'*, *v'*, *f'*).

3 Východomoravská nářeční skupina

3-1 Jižní východomoravská nářeční podskupina

Tabulka 4.10 Soubor znaků normalizovaného folklorního přepisu pro jižní východomoravskou nářeční podskupinu

3-1 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

3-1 FOLKLORNÍ PŘEPIS		digrafy					nářeční znaky				
znaky samohlásek	*krátké*										
	dlouhé	aj	ej	ou	{au}	{eu}					
znaky souhlásek	*znělé*										
	neznělé	ch									
	jedinečné						u̯	ł	ł́	Ĺ	ŕ

Poznámky:

- V souladu s celou východomoravskou nářeční skupinou se v rámci folklorního přepisu rozlišuje dlouhé „ł́ “, „Ĺ“, „ŕ“. V případě, že je folklorní přepis nerozlišuje, není vhodné je doplňovat prostřednictvím obojetností.
- V oblasti dochází k postupnému ústupu tvrdého *ł* a obalovaného *u̯*. Pokud není v textu dvojí *l* zapisováno, je otázkou dialektologického vyhodnocení, zda zavádět obojetnost „[l/ł]“, resp. „[l/u̯]“ pro danou dobu a lokalitu. Jistým vodítkem může být mapa ústupu dvojího *l* (Stupňánek a Vondráková, 2022d), je však třeba brát v úvahu, že za vyznačenou hranicí zachovaného dvojího *l* nebylo užívání zcela důsledné.
- V případě obalovaného *u̯* bývá ve folklorních textech rozeznávána dvojice „l-u“. V takovém případě je nutné zavést pro znak „u“ obojetnost „[u/u̯]“. Daná dvojice je ale rozlišována i pomocí znaků „l-ľ“, kde „ľ“ značilo obalované *u̯* (Kynčl, 1928; týž, 1943–1945). V takovém případě lze znaky nahrazovat dokonale, stejně jako v případě, že je rozlišována dvojice „ľ-l“ nebo „ľ-ł“, případně trojice „ľ-l-ł“ (srov. např. Slavičínský, 1927).
- V oblasti se objevuje jak tautosylabické *aj* („daj“, „najlepší“), tak tautosylabické *ej* („dej“, „nejlepší“). Druhý jmenovaný případ je záležitostí typu dolského. V celé oblasti se pak objevuje *ej* i v dalších případech („tej našej“, „vjecej“, „najmilejší“). Digraf „ej“ byl v minulosti posilován i mizejícím nadměrným dolským *ej* („čejst“, „dobrej“, „řejkat“).
- Nadměrné dolské *ou* („nouž“, „rouža“, „na houře“) je rovněž jev výrazně mizející, je však na druhou stranu posilováno progresivním „ou“ spisovným a obecně českým („táhnou“, „se mnou“) i obecně moravským („majou“, „dělajou“).
- Hláska *ə* se nepravidelně objevuje zejména ve slově *bəl*, případně i v jiných slovech se skupením *il* > *əl*. Lze ji považovat za alofon *i*. Ve folklorních textech bývá zapisována jako „y“ nebo „i“ dle kontextu.
- V oblasti Břeclavska jsou zánikové a ve folklorním přepisu nijak nereflektované měkké retnice (*b′, p′, m′, v′, f′*).

3-2 Severní východomoravská nářeční podskupina

Tabulka 4.11 Soubor znaků normalizovaného folklorního přepisu pro severní východomoravskou nářeční podskupinu

3-2 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

3-2 FOLKLORNÍ PŘEPIS		digrafy					nářeční znaky				
znaky samohlásek	*krátké*										
	dlouhé	aj	ej	{ou}	{au}	{eu}					
znaky souhlásek	*znělé*										
	neznělé	ch									
	jedinečné						(u̯)	ł	ł́	Í	ŕ

Poznámky:

- Pokud text nerozlišuje dvojí *l*, zpravidla je třeba zavést obojetnost „[l/ł]“. Pokud je rozlišována dvojice „ľ-ł“, převádíme měkké „ľ“ na „l“.
- Pouze v nejstarších textech bychom se mohli setkat s dvojicí „l-u̯“.
- Valašské měkké retnice (*b′, p′, m′, v′, f′*) mají sice fonologickou platnost, ale ve folklorním přepisu je od tvrdých nerozlišujeme.
- Rozdíl mezi „i“ a „y“ je na Valašsku vyslovován, ale přepis se řídí spíše historickým a tradičním principem, protože rozdíl „i-y“ nemá povahu fonologickou, ale spíše kombinační (měkké *i* následuje po měkkých souhláskách, tvrdé *y* po tvrdých, nemají tudíž schopnost rozlišit význam). Proto zapisujeme např. „Amerika“, přestože je vyslovováno *Ameryka*.
- Zatímco digrafy „aj“ a „ej“ jsou v nářečí relativně běžné („aj“, „najprv“, „naprotivaj“, „gajdy“, „tej druhej“, „zasej“, „včilej“, „veselejší“), digraf „ou“ se dostává do nářečí pouze vlivem spisovnosti.
- Kombinační povahu má i spojení sykavek a polosykavek s „i“: vyslovuje se *cy, zy, sy, čy, žy, šy* (přesněji většinou spíše *cï, zï, sï, čï, žï, šï*), zapisujeme však v souladu s historickým principem „ci“, „zi“, „si“, „či“, „ži“, „ši“.
- Hláska *ə*, která se nepravidelně objevuje ve slově *bəł*, je alofonem fonému *y* a ve folklorním přepisu se rovněž zapisuje jako „y“.

3-3 Kopaničářská nářeční podskupina

Tabulka 4.12 Soubor znaků normalizovaného folklorního přepisu pro kopaničářskou nářeční podskupinu

3-3 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u		y					
	dlouhé	á	é		í	ó	ú	ů	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	{ř}	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	{ř}	{x}	
	jedinečné	r	l	m	n	ň	j							

3-3 FOLKLORNÍ PŘEPIS		digrafy								nářeční znaky				
znaky samohlásek	*krátké*													
	dlouhé	aj	ej	ia	iá	ie	ié	{au}	{eu}					
znaky souhlásek	*znělé*	dz												
	neznělé	ch												
	jedinečné									u̯	ł	ł́	Ĺ	ŕ

Poznámky:

- V oblasti se objevují specifické diftongy „ia“, „iá“, „ie“, „ié“, které bývají zapisovány buď takto, nebo formou „ja“, „já“, „je“, „jé“. Mnohdy obě formy existují vedle sebe v jednom textu („mlynárja“ vs. „v poriadku“, srov. Frolec a Holý, 1967, s. 126–127). Ve folklorním přepisu dáváme přednost digrafům typu „ia“, na něž variantu s „j“ snadno převedeme, neboť digrafy typu „ja“ se vždy objevují za konsonantem (s výjimkou prefixů zakončených na konsonant).
- V oblasti dochází k pozvolnému ústupu tvrdého *ł* a obalovaného *u̯*. Pokud není v textu dvojí *l* zapisováno, je na dialektologickém zvážení, zda zavádět obojetnost „[l/ł]“, resp. „[l/u̯]“ pro danou dobu a lokalitu. Jistým vodítkem může být mapa ústupu dvojího *l* (Stupňánek a Vondráková, 2022d).
- Na Hrozenkovsku se objevuje foném *dz*, který vznikl asibilací *ď* (obdobně i *c* z původního *ť*). Zapisujeme ho „dz“.
- Podskupina je charakteristická absencí „ř“ (namísto něj je zde „r“), přesto „ř“ částečně v těchto pozicích do jazyka proniká jako progresivní jev.

4 Slezská nářeční skupina

4-1 Slezskomoravská nářeční podskupina

Tabulka 4.13 Soubor znaků normalizovaného folklorního přepisu pro slezskomoravskou nářeční podskupinu

4-1 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u	y						
	dlouhé													
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	ť		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

4-1 FOLKLORNÍ PŘEPIS		digrafy					nářeční znaky	
znaky samohlásek	*krátké*							
	dlouhé	(aj)	ej	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*	(dz)	(dž)					
	neznělé	ch						
	jedinečné						u̯	ł

Poznámky:

- Tvrdé *y* se ve Slezsku fonologizovalo, proto je ve folklorním přepisu zapisováno přesně tam, kde je vyslovováno. I skupiny *cy, zy, sy, čy, žy, šy, dży, řy*, případně *cï, zï, sï, čï, žï, šï, dżï, řï* jsou transkribovány v normalizovaném folklorním textu prostřednictvím „y" („řykała", „łupežnicy").
- Pokud text nerozlišuje dvojí *l*, zpravidla je třeba zavést obojetnost „[l/ł]", v severní části území se zvyšuje pravděpodobnost dvojice „l-u̯", zvláště u generace, která chodila do školy v meziválečném období. Ostatní generace, ať už starší, nebo mladší, mají zpravidla „ł". Pokud je rozlišována dvojice „ľ-ł", případně jiná dvojice (např. „ľ-l", „ľ-u̯"), případně trojice (např. „ľ-l-ł"), převádíme na adekvátní variantu „l-ł" nebo „l-u̯".
- V celém Slezsku se objevuje ustupující tautosylabické „aj" a za něj progresivní tautosylabické „ej" („daj" vs. „dej"), „ej" se ve slezských dialektech jinak vyskytuje běžně („tej jednej", „zasej", „šejšč", „mundřejši").
- Vzácněji a jen ve slezskomoravských dialektech se objevuje progresivní „ou" („sou", „vedoucy", „zešroubovane").
- Jako výsledek asibilací se vyskytují fonémy *dz'*, *dž'* zapisované v normalizované transkripci digrafy „dz", „dž" bez vyznačení jejich měkkosti, podobně to platí pro další výsledky asibilací, které nejsou odlišovány od tvrdých variant („c", „č"). Asibilace jsou ve slezskomoravské nářeční skupině ustupujícím jevem, který se na části území vůbec nevyskytuje.
- V grafice nejsou rozlišovány ani měkké a tvrdé sykavky („s", „š").

4-2 Slezskopolská nářeční podskupina

Tabulka 4.14 Soubor znaků normalizovaného folklorního přepisu pro slezskopolskou nářeční podskupinu

4-2 FOLKLORNÍ PŘEPIS		standardní znaky												
znaky samohlásek	*krátké*	a	e	ě	i	o	u	y						
	dlouhé													
znaky souhlásek	*znělé*	b	d	{ď}	{h}	z	ž	g			v	ř	{x}	{w}
	neznělé	p	t	{ť}		s	š	k	c	č	f	ř	{x}	
	jedinečné	r	l	m	n	ň	j							

4-2 FOLKLORNÍ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*						
	dlouhé	(aj)	ej	{au}	{eu}		
znaky souhlásek	*znělé*	dz	dž				
	neznělé	ch					
	jedinečné					u̯	ł

Poznámky:

- Zápis tvrdého „y“ je shodný se slezskomoravskou skupinou, přibývá však k němu zde i foném *ů̊* vzniklý z původního *e*, který se tradičně zapisuje rovněž jako „y“, i když je vyslovován jinak.
- Pokud text nerozlišuje dvojí *l*, zpravidla je třeba zavést obojetnost „[l/ł]“ nebo „[l/u̯]“, a to podle oblasti výskytu. Pokud je v rámci dvojího *l* znakově rozlišováno měkké „l‘“, převádíme ho na standardní znak „l“.
- Znak „u̯“ se používá i pro protetickou hlásku, která ve fonologickém systému splynula s obalovaným *l*.
- Výsledky asibilací původních „d‘“, „t‘“ jsou u slezskopolských nářečí většinou ještě poměrně dobře zachovány, proto jsou znaky „d‘“, „t‘“ označeny jako řídké a progresivní. U digrafů „dz“ a „dž“ není značena jejich měkkost. Měkkost není rozlišována ani u sykavek („s“, „š“).
- V některých ojedinělých textech lidového původu jsou zapisovány (tvrdé) hlásky *č*, *š* polským pravopisem „cz“, „sz“. Tyto spřežky lze velmi jednoduše a neproblematicky nahrazovat za „č“, „š“.
- Jako řídký a progresivní je označen i grafém „h“, neboť slezskopolské dialekty namísto něj mají tradičně „g“.
- Slezskopolské měkké retnice a měkké veláry (*b‘, p‘, m‘, v‘, f‘, k‘, g‘*) nejsou ve folklorním přepisu rozlišovány.

4.5 Normalizace dialektologického přepisu

4.5.0 Úvod

Dialektologický přepis je založen na fonetickém principu, jehož podstatou je snaha co nejlépe grafickými prvky postihnout zvukovou stránku jazyka. Je vesměs využíván v odborných dialektologických pracích, a přestože tu byla snaha dialektologický přepis sjednotit (Hála, Vážný a kol., 1944; rozšířená verze 1951), fakticky tu existuje poměrně dost velká variabilita v odborných dialektologických transkripcích. To je dáno na jedné straně různými odbornými zájmy a cíli jednotlivých autorů, ale na druhou stranu také snahami o zjednodušení, neboť přesnější a detailnější přepisy vyžadují více času a není obvykle praktické jimi zpracovávat rozsáhlejší materiál.

4.5.1 Problém sjednocení dialektologických textů

V rámci materiálu, který můžeme v českém prostředí sesbírat (srov. Stupňánek a Vondráková, 2022a, s. 37–45), nacházíme dialektologické přepisy s různým stupněm přesnosti. Přesnější a detailněji zpracované texty, které používají pro týž dialekt širší paletu znaků, lze obvykle snadno převést na přepis jednodušší. Opačným směrem to však většinou možné není. Hrubší popis zvukové stránky většinou nelze pouze na základě dat obsažených v textu převést na jemnější zvukové a znakové rozlišení. Přirozenou cestou normalizace je tedy zjednodušení zápisu, při něm nenarážíme na žádné převodní obtíže. Znamená to sice přicházet o mnohá přesná data, obsažená v rozličných dialektologických zápisech, ale ani to ve skutečnosti není na škodu. Vytváříme totiž materiál pro strojové učení, které je fakticky pokročilou statistikou, a statistika vyžaduje opakování, aby mohla fungovat. A čím přesnější je dialektologický přepis, tím méně pravděpodobné je opakování téhož slova či jevu, protože detailnost zvukového popisu ho umožňuje zapsat ve více zvukových variantách. Folklorní přepis, jehož hlavní princip je fonologický, od konkrétních realizací abstrahuje, takže zapíše jediné slovo „odsud“ tam, kde ho může jednoduchý dialektologický přepis zapsat podle různých fonetických kontextů ve čtyřech variantách: *otsut*, *otsud, ocut, ocud*. Pokud budeme rozlišovat ještě znak pro znělou výslovnost před pauzou (*ḍ*), přibydou další dvě varianty; budeme-li rozlišovat různé výslovnostní odstíny vokálů, s každou novou variantou se nám počet zdvojnásobí. Jedno slovo tedy má v jednom dialektu tím víc realizací, čím přesněji se jeho konkrétní zvuková realizace zapíše. Zjednodušení zápisu je tak pozitivním krokem, který fakticky zvyšuje efektivitu strojového učení tím, že zvyšuje množství opakování v textu.

4.5.2 Normalizované náhrady za znaky dialektologického přepisu

Normalizace dialektologického textu tak není v zásadě ničím komplikovaným. Pokud samo OCR proběhlo v pořádku a pokud jsme při něm dodrželi zásadu zachovávat každý znak jedna ku jedné, je potom následující normalizace už jen přímočarým procesem hromadného hledání a nahrazování znaků. V naprosté většině případů českých dialektologických textů si vystačíme s následující tabulkou náhrad speciálních znaků za znaky normalizované.

Jsou i publikace, v nichž nalezneme ještě další varianty nářečních znaků, vesměs podle individuální úpravy autora. Např. Mazlová (1949) na Zábřežsku rozeznává jen krátkých *e*-ových hlásek sedm a k jejich rozlišení používá mimo jiné i řez písma („**e:**“, „**e**“, „**e.**“, „*e*“, „*ę.*“, „*ę*“, „*ę:*“). Takovéto specifické individuální notace nezahrnujeme, a pokud se u takovýchto zdrojů zdaří OCR, je třeba si pro ně vytvořit individuální seznam náhrad.

Někdy se také může stát, že text obsahuje ještě speciální nehláskové fonetické informace nad rámec těch, které jsou uvedeny v tabulce (tj. mimo přízvuky a ráz). Většinou jde o vyznačení pauz nebo intonace, které také bývá individuální a vyskytuje se ojediněle (srov. Bachmann, 2001). Je potřeba i tyto znaky z textu odstranit.

Tabulka 4.15 obsahuje také používané speciální modifikační znaky podobné hornímu indexu, ale neobsahuje přímo znaky horního indexu, které se občas používají pro oslabenou realizaci nějaké hlásky, prakticky jakékoli. Horní index, jakožto formátování písma, bývá odstraněn většinou už před vlastní normalizací, což ale není na škodu. Normalizovaný dialektologický přepis totiž oslabené znaky nerozlišuje a zapisuje je jako klasické znaky. Fakticky tedy není třeba jejich převod, a proto nejsou uváděny v tabulce.

Tabulka 4.15 Soubor speciálních znaků dialektologického přepisu a jejich normalizovaných náhrad

znak	normalizace	popis hlásky
ạ	a	krátké široké *a*
ä	a	krátké široké *a*
ạ	a	krátké široké *a*
ạ́	á	dlouhé široké *a*
ä́	á	dlouhé široké *a*
ạ́	á	dlouhé široké *a*
ạ	a	krátké labializované *a*
å	a	krátké labializované *a*
ǻ	ạ́	dlouhé labializované *a*
ḅ	b	prodloužená realizace *b* (znělá výslovnost před pauzou)
b'	b′	měkké *b*
b´	b′	měkké *b*
ḃ	b′	měkké *b*
ḃ	b′	měkké *b*
c'	c′	palatalizované *c*, asibilované *ť*
c´	c′	palatalizované *c*, asibilované *ť*
ċ	c′	palatalizované *c*, asibilované *ť*
ć (záp. od Ostravy)	c′	palatalizované *c*, asibilované *ť*
č'	č	palatalizované *č*, asibilované *ť*
č′	č	palatalizované *č*, asibilované *ť*
č´	č	palatalizované *č*, asibilované *ť*
ć	č	palatalizované *č*, asibilované *ť*
č̊	č	palatalizované *č*, asibilované *ť*
ḍ	d	prodloužená realizace *d* (znělá výslovnost před pauzou)
ḍ	ď	prodloužená realizace *ď* (znělá výslovnost před pauzou)
ê	ę	krátké široké *e*
è	ę	krátké široké *e*
ę́	é	dlouhé široké *e*
ẹ	e	krátké zavřené *e*
ẹ́	é	dlouhé zavřené *e*
ẹ	e	krátké, velmi zavřené *e*
ę	e	krátké nazální *e*
ᵉ	e	ultrakrátké *e*
f'	f′	měkké *f*
f´	f′	měkké *f*

znak	normalizace	popis hlásky
f́	f′	měkké *f*
ḟ	f′	měkké *f*
g'	g′	měkké *g*
g´	g′	měkké *g*
ǵ	g′	měkké *g*
ġ	g′	měkké *g*
ḥ	h	prodloužená realizace *h* (znělá výslovnost před pauzou)
ʰ	h	přídech u souhlásky / slabé protetické *h*
x	ch	*ch*
ɣ	h	znělé *ch*
ǐ	i	polodlouhé *i*
ï	y	krátké široké *i*
ǐ́	ý	dlouhé široké *i*
i̯	i	krátké otevřené *i*
í̯	í	dlouhé otevřené *i*
ĩ	i	krátké nazální *i*
ⁱ	i	ultrakrátké *i*
i̯	j	neslabičné *i*
k'	k′	měkké *k*
k´	k′	měkké *k*
ḱ	k′	měkké *k*
ḳ	k′	měkké *k*
l′	l	měkké *l*
l'	l	měkké *l*
l´	l	měkké *l*
ľ	l	měkké *l*
l̥	l	slabikotvorné *l*
l̥′	l	slabikotvorné měkké *l*
l̥'	l	slabikotvorné měkké *l*
l̥´	l	slabikotvorné měkké *l*
l̥ˇ	l	slabikotvorné měkké *l*
ł̥	ł	slabikotvorné tvrdé *l*
ĺ̥	ĺ	dlouhé slabikotvorné *l*
ĺ̥′	ĺ	dlouhé slabikotvorné měkké *l*
ĺ̥'	ĺ	dlouhé slabikotvorné měkké *l*
ĺ̥´	ĺ	dlouhé slabikotvorné měkké *l*
ĺ̥ˇ	ĺ	dlouhé slabikotvorné měkké *l*
ł̥́	ł́	dlouhé slabikotvorné tvrdé *l*
m̩	m	retozubné *m*
m̤	m	prodloužená realizace *m*
m̥	m	slabikotvorné *m*
m'	m′	měkké *m*

znak	normalizace	popis hlásky
m´	m′	měkké *m*
ḿ	m′	měkké *m*
ṁ	m′	měkké *m*
ŋ	n	velární *n*
n̩	n	prodloužená realizace *n*
n̥	n	slabikotvorné *n*
n′	ň	palatalizované *n*
n'	ň	palatalizované *n*
n´	ň	palatalizované *n*
ń	ň	palatalizované *n*
ň̩	ň	prodloužená realizace *ň*
ň̥	ň	slabikotvorné *ň*
ŏ	o	polodlouhé *o*
ô	ǫ	krátké široké *o*
ò	ǫ	krátké široké *o*
ǫ́	ó	dlouhé široké *o*
ọ	o	krátké zavřené *o*
ọ́	ó	dlouhé zavřené *o*
o̤	o	krátké, velmi zavřené *o*
ǫ̨	o	krátké nazální *o*
p'	p′	měkké *p*
p´	p′	měkké *p*
ṕ	p′	měkké *p*
ṗ	p′	měkké *p*
r̥	r	slabikotvorné *r*
ŕ̥	ŕ	dlouhé slabikotvorné *r*
r̭	ř	neznělé *ř*
ř̥	ř	neznělé *ř*
s̄	ss	prodloužená realizace *s* (přechod ke geminaci)
s̄	ss	prodloužená realizace *s* (přechod ke geminaci)
s'	s′	palatalizované *s*
s´	s′	palatalizované *s*
ṡ	s′	palatalizované *s*
ś (záp. od Ostravy)	s′	palatalizované *s*
s̄̄	šš	prodloužená realizace *š* (přechod ke geminaci)
s̄̄	šš	prodloužená realizace *š* (přechod ke geminaci)
š′	š	palatalizované *š*
š'	š	palatalizované *š*
š´	š	palatalizované *š*
ś	š	palatalizované *š*
ṡ	š	palatalizované *š*
ŭ	u	polodlouhé *u*

znak	normalizace	popis hlásky
ų̰	u	krátké nazální *u*
ů	u	krátké, velmi otevřené *u*
ų̊	u	krátké, nazální, velmi otevřené *u*
v̩	v	prodloužená realizace *v* (znělá výslovnost před pauzou)
v̥	u̯	labializované, zaokrouhlené *v*
v'	v′	měkké *v*
v´	v′	měkké *v*
v́	v′	měkké *v*
v̇	v′	měkké *v*
y͡	y	průvodní vokál *y* vokalizovaného *r*
y̨͡	y	krátké nazální *y*
y̯	y	krátké, velmi otevřené *y*
y̨̯	y	krátké, nazální, velmi otevřené *y*
z̩	z	prodloužená realizace *z* (znělá výslovnost před pauzou)
z'	z′	palatalizované *z*
z´	z′	palatalizované *z*
ż	z′	palatalizované *z*
ź (záp. od Ostravy)	z′	palatalizované *z*
ž̩	ž	prodloužená realizace *ž* (znělá výslovnost před pauzou)
ž′	ž	palatalizované *ž*
ž'	ž	palatalizované *ž*
ž´	ž	palatalizované *ž*
ž̯	ž	palatalizované *ž*
ž̯	ž	palatalizované *ž*
ʒ	dz	*dz*, znělé *c*
ʒ′	dz′	palatalizované *dz*, asibilované *ď*
ʒ'	dz′	palatalizované *dz*, asibilované *ď*
ʒ´	dz′	palatalizované *dz*, asibilované *ď*
ʑ	dź	polské palatální *dž*, asibilované *ď*
ǯ	dž	*dž*, znělé *č*
ǯ′	dž	palatalizované *dž*
ǯ'	dž	palatalizované *dž*
ǯ´	dž	palatalizované *dž*
ʒ̯	dž	palatalizované *dž*
ь	i	ultrakrátké *i*
ъ	ə	ultrakrátké *ə*
ᵊ	ə	ultrakrátké *ə*
ʔ		ráz
ˈ		hlavní přízvuk
ˌ		vedlejší přízvuk

Doplňme tabulku ještě seznamem speciálních dialektologických digrafů a jejich normalizací:

Tabulka 4.16 Soubor speciálních dialektologických digrafů a jejich normalizovaných náhrad

digraf	normalizace	popis hlásky
a͡i	aj	diftong *aj*
a͡u	au	diftong *au*
a͡u	au	diftong *au*
a͡u	au	diftong *au*
d͡z	dz	*dz*, znělé *c*
d͡z	dz	*dz*, znělé *c*
d͡ž	dž	*dž*, znělé *č*
d͡ž	dž	*dž*, znělé *č*
e͡i	ej	diftong *ej*
ę͡i	ej	diftong *ej* s otevřenou *e*-ovou složkou
ę͡j	ej	diftong *ej* s otevřenou *e*-ovou složkou
ẹ͡i	ej	diftong *ej* se zavřenou *e*-ovou složkou
ẹ͡j	ej	diftong *ej* se zavřenou *e*-ovou složkou
ḛ͡i	ej	diftong *ej* s velmi zavřenou *e*-ovou složkou
ḛ͡j	ej	diftong *ej* s velmi zavřenou *e*-ovou složkou
e͡u	eu	diftong *eu*
e͡u	eu	diftong *eu*
e͡u	eu	diftong *eu*
i͡i	ij	diftong *ij*
ḭ͡i	ij	diftong *ij* s otevřenou *i*-ovou složkou
ḭ͡j	ij	diftong *ij* s otevřenou *i*-ovou složkou
i͡a	ja	diftong *ja*
i͡á	já	diftong *já*
i͡e	je	diftong *je*
i͡é	jé	diftong *jé*
o͡u	ou	diftong *ou*
o͡u	ou	diftong *ou*
o͡u	ou	diftong *ou*
ǫ͡u	ou	diftong *ou* s otevřenou *o*-ovou složkou
ǫ͡u	ou	diftong *ou* s otevřenou *o*-ovou složkou
ọ͡u	ou	diftong *ou* se zavřenou *o*-ovou složkou
ọ͡u	ou	diftong *ou* se zavřenou *o*-ovou složkou
ṏ͡u	ou	diftong *ou* s velmi zavřenou *o*-ovou složkou
ṏ͡u	ou	diftong *ou* s velmi zavřenou *o*-ovou složkou
u͡u	uu	diftong *uu*

4.5.3 Přehled nářečních znaků normalizované dialektologické transkripce

Spektrum speciálních nářečních znaků normalizované transkripce je naproti tomu několikanásobně menší. Souhrnně jde o následující znaky (tabulka 4.17).

Tabulka 4.17 Soubor speciálních nářečních znaků normalizované dialektologické transkripce

normalizovaný speciální znak	popis hlásky
ą̊	dlouhé labializované *a*
b′	měkké *b*
c′	palatalizované *c*, asibilované *ť*
ć	polské palatální *č*, asibilované *ť*
ę̆	krátké široké *e*
f′	měkké *f*
g′	měkké *g*
k′	měkké *k*
Í	dlouhé *l*
ł	krátké tvrdé *l*
ł́	dlouhé tvrdé *l*
m′	měkké *m*
ǫ̆	krátké široké *o*
p′	měkké *p*
ŕ	dlouhé *r*
s′	palatalizované *s*
ś	polské palatální *š*
u̯	neslabičné *u* / obalované *l*
v′	měkké *v*
w	bilabiální *v*
z′	palatalizované *z*
ź	polské palatální *ž*
ə	krátký redukovaný vokál

Digrafy v dialektologické transkripci (bez odlišení toho, zda jsou nářeční, či nikoli) jsou následující:

Tabulka 4.18 Soubor digrafů v dialektologické transkripci

normalizovaný diagraf	popis hlásky
aj	diftong *aj* / tautosylabické *aj*
au	diftong *au*
dz	*dz*, znělé *c*
dz′	palatalizované *dz*, asibilované *ď*
dź	polské palatální *dž*, asibilované *ď*
dž	*dž*, znělé *č*, asibilované *ď*
ej	diftong *ej* / tautosylabické *ej*
eu	diftong *eu*
ch	*ch*
ij	diftong *ij*
ja	diftong *ja*
já	diftong *já*
je	diftong *je*
jé	diftong *jé*
ou	diftong *ou*
uu	diftong *uu*

A pro úplnost vypočtěme i užívané standardní znaky (tabulka 4.19).

Tabulka 4.19 Soubor standardních znaků dialektologické transkripci

standardní znaky normalizované transkripce		
a	i	ř
á	í	s
b	j	š
c	k	t
č	l	ť
d	m	u
ď	n	ú
e	ň	v
é	o	y
f	ó	ý
g	p	z
h	r	ž

4.5.4 Přehled znaků normalizované dialektologické transkripce podle nářečí

Vzhledem k tomu, že trénování strojového učení by mělo probíhat na základě nářečních podskupin, opět podáváme přehledy znaků normalizované transkripce pro každou jednotlivou nářeční podskupinu. Znaky v ní obsažené nemusí být charakteristické pro všechny dialekty dané nářeční podskupiny. Každá tabulka je opět rozdělena na standardní znaky, digrafy a nářeční znaky. Jde tedy znovu o znakový systém, nikoli fonetický popis dané skupiny dialektů, přestože se k fonetice váže. Dialektologická transkripce je založena na fonetickém principu, její normalizovaná podoba však obsahuje řadu zjednodušení někdy založených na fonologickém principu (např. nerozlišování znělého a neznělého *ř*, velárního *n*, retozubného *m*), často však na prostém zjednodušení grafiky (*dz*, *dž* jakožto asimilované znělé varianty *c*, *č* jsou zapisovány digrafy, protože se shodují s většinovou realizací *dz*, *dž* ve slovech jako *podzim*, *džbán*; palatalizované *č*, *dž*, vzniklé asibilací *ť*, *ď*, není rozlišováno od tvrdých variant *č*, *dž*, protože v některých dialektech postupně splývají, atd.).

V tabulkách opět označujeme speciálně znaky, které jsou (v souvislosti s hláskami, které reprezentují) řídké, a to buď na ústupu, nebo naopak na postupu:

označení vývojové tendence	význam
(w)	znak, který je řídký a zánikový, objevuje se jen v nejstarších textech
{g}	znak, který je řídký a progresivní, zpravidla proniká do zápisu nářečí nebo zvyšuje svou četnost s průnikem spisovnosti, slov cizího původu nebo obecné češtiny

U digrafů opět platí, že až na *ch* mohou mít všechny dvojí interpretaci, tj. mohou reprezentovat jednu hlásku i dvě. Tautosylabická spojení typu *aj*, *ej* (*najlepší*, *dej*), která se diachronně měnila společně, pojímáme též jako diftong, tudíž i digraf.

1 Česká nářeční skupina

1-1 Severovýchodočeská nářeční podskupina

Tabulka 4.20 Soubor znaků normalizovaného dialektologického přepisu pro severovýchodočeskou nářeční podskupinu

1-1 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

1-1 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**				**nářeční znaky**		
znaky samohlásek	*krátké*					(ə)		
	dlouhé	ej	ou	{au}	{eu}			
znaky souhlásek	*znělé*	dz	dž					
	neznělé	ch						
	jedinečné					u̯	(w)	(ł)

1-2 Středočeská nářeční podskupina

Tabulka 4.21 Soubor znaků normalizovaného dialektologického přepisu pro středočeskou nářeční podskupinu

1-2 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

1-2 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**				**nářeční znaky**
znaky samohlásek	*krátké*					
	dlouhé	ej	ou	{au}	{eu}	
znaky souhlásek	*znělé*	dz	dž			
	neznělé	ch				
	jedinečné					(u̯)

1-3 Jihozápadočeská nářeční podskupina

Tabulka 4.22 Soubor znaků normalizovaného dialektologického přepisu pro jihozápadočeskou nářeční podskupinu

1-3 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	{é}	í	{ó}	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

1-3 DIALEKTOLOGICKÝ PŘEPIS		digrafy						nářeční znaky		
znaky samohlásek	*krátké*							(ə)		
	dlouhé	ej	yj	ou	(uu)	{au}	{eu}			
znaky souhlásek	*znělé*	dz	dž					(b′)	(v′)	
	neznělé	ch						(p′)	(f′)	
	jedinečné							(u̯)	(ł)	(m′)

1-4 Českomoravská nářeční podskupina

Tabulka 4.23 Soubor znaků normalizovaného dialektologického přepisu pro českomoravskou nářeční podskupinu

1-4 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u	(y)					
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

1-4 DIALEKTOLOGICKÝ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*					(ə)	
	dlouhé	ej	ou	{au}	{eu}		
znaky souhlásek	*znělé*	dz	dž			(b′)	(v′)
	neznělé	ch				(p′)	(f′)
	jedinečné					(u̯)	(m′)

2 Středomoravská nářeční skupina

2-1 Centrální středomoravská nářeční podskupina

Tabulka 4.24 Soubor znaků normalizovaného dialektologického přepisu pro centrální středomoravskou nářeční podskupinu

2-1 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	é	{í}	ó	{ú}						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

2-1 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**				**nářeční znaky**	
znaky samohlásek	*krátké*					ę̥	ǫ̥
	dlouhé	{ej}	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*	dz	dž				
	neznělé	ch					
	jedinečné						

2-2 Jižní středomoravská nářeční podskupina

Tabulka 4.25 Soubor znaků normalizovaného dialektologického přepisu pro jižní středomoravskou nářeční podskupinu

2-2 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

2-2 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**				**nářeční znaky**
znaky samohlásek	*krátké*					(ə)
	dlouhé	{ej}	{ou}	{au}	{eu}	ą̊
znaky souhlásek	*znělé*	dz	dž			
	neznělé	ch				
	jedinečné					

2-3 Západní středomoravský okrajový úsek

Tabulka 4.26 Soubor znaků normalizovaného dialektologického přepisu pro západní středomoravský okrajový úsek

2-3 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u	(y)					
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

2-3 DIALEKTOLOGICKÝ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*					(ə)	
	dlouhé	{ej}	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*	dz	dž				
	neznělé	ch					
	jedinečné					(u̯)	(ł)

2-4 Východní středomoravský okrajový úsek

Tabulka 4.27 Soubor znaků normalizovaného dialektologického přepisu pro východní středomoravský okrajový úsek

2-4 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u	(y)					
	dlouhé	á	é	í	ó	ú	(ý)					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

2-4 DIALEKTOLOGICKÝ PŘEPIS		digrafy				nářeční znaky	
znaky samohlásek	*krátké*						
	dlouhé	{ej}	{ou}	{au}	{eu}		
znaky souhlásek	*znělé*	dz	dž			(b')	(v')
	neznělé	ch				(p')	(f')
	jedinečné					ł	(m')

3 Východomoravská nářeční skupina

3-1 Jižní východomoravská nářeční podskupina

Tabulka 4.28 Soubor znaků normalizovaného dialektologického přepisu pro jižní východomoravskou nářeční podskupinu

3-1 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u	(y)					
	dlouhé	á	é	í	ó	ú	(ý)					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

3-1 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**					**nářeční znaky**					
znaky samohlásek	*krátké*						ə					
	dlouhé	aj	ej	ou	{au}	{eu}						
znaky souhlásek	*znělé*	dz	dž				(b′)	(v′)				
	neznělé	ch					(p′)	(f′)				
	jedinečné						u̯	ł	ł́	Í	ŕ	(m′)

3-2 Severní východomoravská nářeční podskupina

Tabulka 4.29 Soubor znaků normalizovaného dialektologického přepisu pro severní východomoravskou nářeční podskupinu

3-2 DIALEKTOLOGICKÝ PŘEPIS		**standardní znaky**										
znaky samohlásek	*krátké*	a	e	i	o	u	y					
	dlouhé	á	é	í	ó	ú	ý					
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

3-2 DIALEKTOLOGICKÝ PŘEPIS		**digrafy**					**nářeční znaky**					
znaky samohlásek	*krátké*						ə					
	dlouhé	aj	ej	{ou}	{au}	{eu}						
znaky souhlásek	*znělé*	dz	dž				b′	v′				
	neznělé	ch					p′	f′				
	jedinečné						(u̯)	ł	ł́	Í	ŕ	m′

3-3 Kopaničářská nářeční podskupina

Tabulka 4.30 Soubor znaků normalizovaného dialektologického přepisu pro kopaničářskou nářeční podskupinu

3-3 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u						
	dlouhé	á	é	í	ó	ú						
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	{ř}
	neznělé	p	t	ť		s	š	k	c	č	f	{ř}
	jedinečné	r	l	m	n	ň	j					

3-3 DIALEKTOLOGICKÝ PŘEPIS		digrafy							nářeční znaky				
znaky samohlásek	*krátké*												
	dlouhé	aj	ja	já	je	jé	{au}	{eu}					
znaky souhlásek	*znělé*	dz	dž										
	neznělé	ch											
	jedinečné								u̯	ł	ł́	Ĺ	ŕ

4 Slezská nářeční skupina

4-1 Slezskomoravská nářeční podskupina

Tabulka 4.31 Soubor znaků normalizovaného dialektologického přepisu pro slezskomoravskou nářeční podskupinu

4-1 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u	y					
	dlouhé											
znaky souhlásek	*znělé*	b	d	ď	h	z	ž	g			v	ř
	neznělé	p	t	ť		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

4-1 DIALEKTOLOGICKÝ PŘEPIS		digrafy					nářeční znaky			
znaky samohlásek	*krátké*						ə			
	dlouhé	(aj)	ej	{ou}	{au}	{eu}				
znaky souhlásek	*znělé*	dz′	dź	dž			z′	ź		
	neznělé	ch					s′	ś	c′	ć
	jedinečné						u̯	ł		

4-2 Slezskopolská nářeční podskupina

Tabulka 4.32 Soubor znaků normalizovaného dialektologického přepisu pro slezskopolskou nářeční podskupinu

4-2 DIALEKTOLOGICKÝ PŘEPIS		standardní znaky										
znaky samohlásek	*krátké*	a	e	i	o	u	y					
	dlouhé											
znaky souhlásek	*znělé*	b	d	{ď}	{h}	z	ž	g			v	ř
	neznělé	p	t	{ť}		s	š	k	c	č	f	ř
	jedinečné	r	l	m	n	ň	j					

4-2 DIALEKTOLOGICKÝ PŘEPIS		digrafy				nářeční znaky				
znaky samohlásek	*krátké*					ə				
	dlouhé	(aj)	ej	{au}	{eu}					
znaky souhlásek	*znělé*	dz	dź	dž		b′	v′	g′	ź	
	neznělé	ch				p′	f′	k′	ś	ć
	jedinečné					u̯	ł	m′		

4.6 Převod folklorního přepisu na dialektologický přepis

4.6.0 Úvod

Převod z normalizovaného dialektologického do normalizovaného folklorního přepisu je zdaleka nejnáročnější fází zpracování textových dat. Přitom k tomuto převodu nemůžeme využít strojové učení, neboť sám tento převod je nutným předpokladem, abychom mohli strojové učení pro podobný úkol natrénovat. Cílem převodu jsou totiž trénovací data, kde na jedné straně budeme mít text ve folklorním přepisu a k němu dotvoříme týž text v přepisu dialektologickém. Na takovéto dvojici textů pak může strojové učení trénovat převod oběma směry.

Jestliže ale tento iniciální převod chceme dělat na základě algoritmů a pravidel, zjistíme, že není prakticky možné převádět texty z dialektologického přepisu do přepisu folklorního. Dialektologický přepis je totiž mnohem konkrétnější, stojí na fonetickém principu, takže zachycuje příslušné alofony fonémů, nikoli samy fonémy. Ty jsou naopak podstatné pro přepis folklorní. Na základě alofonů nelze dosti často foném s jistotou určit, pokud jsme v situaci, kterou jsme již popisovali: nemáme kompletní „slovník“, tedy relativně úplnou databázi tvarů určitého nářečí (viz 4.2.3.1). Další principy folklorního přepisu, např. historický a tradiční, nám převod ještě dál komplikují (např. rozhodováním mezi „i/y“, „ú/ů“, „di/dy“, „ti/ty“, „ni/ny“). Máme-li tedy dialektologický zápis slova *gdiš*, máme fakticky osm způsobů, jak by slovo mohlo vypadat ve folklorním přepisu, pokud to nevíme („gdiš“, „gdiž“, „gdyž“, „gdyš“, „kdyš“, „když“, „kdiž“, „kdiš“). Všech těchto osm zápisů by mohlo být vysloveno jako *gdiš*.

Převod opačným směrem, z folklorního přepisu do dialektologického, je o poznání snazší. Konkrétní zvuková realizace folklorního zápisu (výběr alofonů) se většinou řídí pravidly, která se dají odvodit z příslušného

nářečí, zákonitostí jeho výslovnosti a z hláskového okolí, je mnohem lépe algoritmizovatelná. Dialektologický přepis folklorního zápisu „když" je tedy zcela jednoznačný, víme-li, o jaké nářečí jde, máme-li slovo v kontextu věty a předpokládáme-li, že jde o realizaci standardní. A nepotřebujeme k tomu slovo „když" znát. Je sice pravda, že většinu našeho převodu nakonec tvoří řada „slovníkových" výjimek, ale ty jsou fakticky jen dočištěním několika stovek pravidel, která platí obecně a tvoří naprostou většinu formujících znaků dialektologického textu.

Pravidla vyjadřujeme prostřednictvím regulárních výrazů typu vyhledání a nahrazení (angl. search & replace), jejichž zřetězení vede k přebudování textu ve folklorním zápisu na zápis dialektologický. Regulární výrazy zde však není možné uvést v plném rozsahu jednoduše proto, že jejich množství je příliš velké. Jedna sada regulárních výrazů jich zahrnuje několik tisíc a vztahuje se jen na část dialektů v rámci jedné podskupiny. Kdybychom měli pokrýt všechny dialekty kompletními sadami, šel by počet regulárních výrazů do stovek tisíc. Uvádíme tak pouze vzory, které bývají někdy ukázkově konkretizovány pro určitou skupinu nářečí. Vzory nejsou explicitně prováděny pro všechny znaky všech nářečí a jejich možné kombinace. U jednotlivých vzorů je však zmíněno, kam do regulárního výrazu je třeba dosadit jakou sadu znaků. A sady znaků normalizovaného folklorního přepisu jsou pak vždy uváděny v tabulkách znaků pro jednotlivé nářeční podskupiny (viz 4.4.4). Tabulky tedy určují dosazení do pozice search regulárního výrazu. Konkrétní rozsahy jednotlivých jevů, tedy to, jaký mají v daném nářečí výsledek a jak v návaznosti na to nastavit pozici replace regulárního výrazu, je pak možné přehledně najít na geoportálu *DiaMa* (Štrubl, Nétek a Stupňánek, 2022), v databázi *InteGra* (Nétek, Štrubl a Stupňánek, 2022; geoportál i související databáze jsou velmi podrobné, ale zahrnují jen některé jevy) nebo v *Českém jazykovém atlase* (Balhar a kol., 1991–2011), zejména v jeho 5. díle (atlas není zcela přesný, ale zahrnuje většinu jevů).

V celé této podkapitole jsme materiálově čerpali z právě uvedených publikací, dále z ALJ (1952–2024), *Dotazníku pro ČJA* (1964–1976), PSJČ (1935–1957) a SSJČ (1960–1971), přihlíželi jsme též k *České výslovnostní normě* (Hůrková, 1995) a IJP (2008–2024). Nejzásadněji jsme ovšem vycházeli ze zpracovávaného textového materiálu, který je součástí *Databáze souvislých nářečních textů* (DSNT, 2017–2024).

4.6.1 Číslice

I když je v nářečních textech zvykem používat pro číslovky spíše slovní než číselnou formu, přesto se ve folklorních textech číslice v malém množství objevují. Jejich převod na slovní formu prostřednictvím regulárních výrazů však přináší řadu možných komplikací, a proto lze doporučit číslice na slovní formu nepřevádět. Je pro to několik důvodů:

- už při normalizaci některých folklorních textů může být nejrozumnější cestou odstranit z textu regulárními výrazy všechny číslice (čísla stránek, číslice ze záhlaví a zápatí, jakékoli číslice napomáhající členění textu atd.) a spolu s tím obětovat i několik číslic přímo v nářečním textu; je to možný postup např. při dočišťování textů, jejichž OCR není příliš dokonalé a u nichž se sofistikovanějšími způsoby nepodařilo všechny nadbytečné číslice odstranit;
- regulární výrazy samy o sobě nejsou vhodným nástrojem pro určení pádu číslovky vyjádřené ve větě číslicí, bylo by zapotřebí použít parser, který pro nářečí češtiny zatím neexistuje. Analýza čistě prostřednictvím regulárních výrazů by byla značně nespolehlivá, přitom na provedení náročná a generovala by značné množství obojetností;
- doplňkovým důvodem je pak dubletnost u samotných číslovek; týká se:
  - hláskosloví: *sedm* vs. *sedum*, *dvje* vs. *dje*, *dvanáct* vs. *dvanác* vs. *dvanást*, *devíťi* vs. *deviťi*, *devjet* vs. *deujet*, *pjet* vs. *pet* atd.;
  - morfologie: *dvje/dje sta* vs. *dvje/dje sťe*, *dvaced druhej* vs. *dvacátej druhej*, *dvaceťi* vs. *dvaceťima* atd.;
  - lexika: *jednadvacet/jedenadvacet* vs. *dvacet jedna*.

Zápis prostřednictvím číslice nemůže indikovat jazykovou realizaci dané číslovky, i když by bylo možné s určitou mírou nejistoty předpokládat, že zvláštní a méně obvyklé nářeční varianty by byly v textu spíše zachyceny slovem než číslicí.

Převod číslic do fonetické podoby (dialektologického zápisu) by v kombinaci všech uvedených důvodů byl enormně náročný, přičemž by ve výsledku trénovací materiál zatěžoval nadměrným množstvím obojetností.

4.6.2 Zkratky

4.6.2.1 Typy zkratek

Ačkoli zkratky nejsou častým prvkem folklorních textů, přesto se v nich vyskytují a je třeba se s nimi nějakým způsobem vypořádat. Pokud je necháme mimo naši pozornost, uplatní se na nich např. znělostní asimilace nebo neutralizace („V KSČ byl“ > *V GZDŽ bil*, „atd.“ > *att.*). Problematiku zkratek však nelze prostřednictvím regulárních výrazů vyřešit zcela uspokojivě a bezchybně vzhledem k variantnosti výslovnosti zkratek i dalším souvisejícím problémům, o nichž pojednáme. Zvuková realizace zkratek je obecně velmi odlišná od jejich zápisu, takže nám v grafice zcela chybí indicie k rozpoznání formy jejich výslovnosti v daném projevu. Při konverzi z folklorní do dialektologické transkripce tedy z praktických důvodů nezbývá než volit přibližné řešení, kdy se řešení přesnému snažíme co nejvíce přiblížit.

Mohlo by se také nabízet zkratky na celý proces konverze fixovat a do výsledného dialektologického přepisu je přenést v původní (nefonetické) formě, ale to by snižovalo kvalitu převodu a vylučovalo materiál týkající se zkratek z trénovacích dat. Obhajobou pro takový přístup by mohla být kolize iniciálových zkratek s texty psanými verzálkami (nadpisy, texty návěstí a nápisů, vzkazů, hlasitých výkřiků, příp. jinak motivované texty psané majuskulemi). Jak již však bylo konstatováno v podkapitole 4.3.1, je vhodné již v rámci čištění textu vyloučit všechny nadpisy a názvy kapitol a také odstranění jiných souvislejších textů psaných majuskulemi se jeví jako nanejvýš vhodné, protože:

- návěstí a nápisy nebývají v nářečí, obecně se v takovémto textu mohou objevit neobvyklé, pro nářečí neautentické jevy;
- texty psané verzálkami zpravidla nebývají dlouhé, a jejich ztráta tak není podstatná;
- pokud chceme zachovávat velká a malá písmena, znamenalo by to převádět texty psané verzálkami odděleně od všeho ostatního textu, což představuje velký nárůst počtu regulárních výrazů a další značné komplikace s velmi nízkými výtěžky.

Naproti tomu převod zkratek do fonetické podoby zajistí, že tato část textu bude zapsána podle stejných pravidel jako jeho zbytek, může tedy podstupovat teritoriálně podmíněné asimilace a jiné úpravy regulárními výrazy.

- **Iniciálové zkratky**

  Iniciálové zkratky (zkratky velkými písmeny tvořené z počátečních písmen víceslovných názvů) tvoří většinu zkratek, se kterými se ve folklorních textech musíme potýkat. Jejich výslovnost není jednoznačná. U iniciálových zkratek je sice dominantní vokalizovaná výslovnost, ale vedle toho existuje i výslovnost redukovaná a výslovnost jako zkratkové slovo.

  1. Vokalizovaná výslovnost má určitá pravidla, zkratka se vyslovuje prostřednictvím vokalizovaných názvů českých písmen:
     - *á*, *é*, *í*, *ó*, *ú* (případné ypsilon se mimo zkratku „XY“ vyslovuje jako *í*, nebo, kde je to možné, snad i *ý*[39]) – dlouhá výslovnost krátkých i dlouhých iniciálových vokálů je vesměs pravidlem (mimo Slezsko);

39 Jde však nepochybně ve zkratkách o zcela řídký jev, na který se nám nepodařilo narazit.

- *bé, cé, čé, dé, ďé, gé, jé, pé, kvé, té, ťé, vé* (i za *w*, „BMW" = *béemvé*) – ve Slezsku zpravidla krátké *e*;
- *ef, el, em, en, eň, er, eř, es, eš*;
- *há, chá, ká* – ve Slezsku zpravidla krátké *a*;
- *iks, zet, žet*.

Problémem u těchto zkratek je nezjistitelnost přítomnosti rázů, která ovlivní způsob zápisu („USA" = *ú es á*, „MNV" = *emenvé* vs. *em en vé*) stejně jako nejistota ohledně vnitřních asimilací a splynulin („KSČ" = *káesčé* vs. *káešče*, „JZD" = *jézédé* vs. *jézeddé*, „SNB" = *esembé* vs. *esenbé*), případně i jiné nepravidelnosti a psychologické změny („MDŽ" = *médéžé*). Jinak zde platí běžná pravidla spodoby.

2. Redukovaná výslovnost je proti tomu vzácnější a u většiny zkratek není užívána vůbec, ale vyskytuje se a může být jak menšinovou, tak i převládající výslovnostní variantou („Kčs" = *kačəsə*, „SSSR" = *səsəsərə*, „MDŽ" = *mədažə*). Vlivem průniku zkratek ze spisovných útvarů jazyka a vlivem spisovné normy (výrazně preferující vokalizovanou výslovnost) nebývá zřejmě redukovaná výslovnost nikdy jedinou uzuální výslovnostní možností, byť ojediněle se k ní blíží (např. „VŘSR" = *vařəsərə*). U redukované výslovnosti platí tato pravidla: všechny iniciálové konsonanty bývají následovány redukovaným vokálem *ə*, iniciálové vokály jsou vyslovovány převážně s rázem a mnohdy si ponechávají původní krátkost, zvláště na první pozici nebo před jiným vokálem („OSVČ" = *osəvəčə*, „NOÚZ" = *nə o ú zə* vs. „ČSA" = *čə sə á*, „ČSAD" = *čə sə á də*). Redukovaná výslovnost zkratky se obvykle nevyskytuje, pokud lze iniciálovou zkratku vyslovit jako zkratkové slovo. Většina zkratek je pak vžitá pouze ve vokalizované podobě. Distribuci vokalizované a redukované výslovnosti ovlivňují faktory uzuální i individuální, faktory teritoriální nikdy nebyly pro nářečí češtiny soustavně zkoumány, písmenný sklad zkratky ani původ mluvčího však zjevně není dostatečným vodítkem pro jednoznačnou volbu mezi vokalizovanou a redukovanou výslovností.
3. Výslovnost jako zkratkové slovo – jde relativně o řídký jev, který je možný pouze tam, kde lze zkratku vyslovit jako slovo („ARO" = *áro*, „UEFA"= *Uefa*). Řada takových zkratek je však stále vyslovována vokalizovaně („ROH" = *er ó há* nebo *eróhá*), příčiny opět nejsou jiné než uzuální, z hlediska regulárních výrazů tedy nelze dospět k obecné apriorní formulaci a je v celé oblasti iniciálových zkratek zapotřebí řešit maximum případů empiricky, tedy na základě dokladů a zkušenosti s jednotlivými zkratkami.

- **Jiné zkratky**

Ostatní typy zkratek, které bývají v mluvě nebo při hlasitém čtení nahrazovány celými slovy, se ve folklorních textech vyskytují zcela minimálně a bývají zpravidla rozepisovány. Pokud už se objeví, je spíše vyšší pravděpodobnost, že je míněna výslovnost zkratky jakožto zkratky (tedy většinou výslovnost redukovaná nebo výslovnost jako zkratkové slovo, ve specifických případech i výslovnost vokalizovaná). Jde o následující typy zkratek:

1. Zkratky čistě grafické (typu „č.", „atd.", „mj.") mohou být v ojedinělých případech zapisovány a vyslovovány jako zkratka („atd. atd." = *a tə də a tə də*, „př. n. l." = *pə řə nə lə*). Spadají sem i zkratky, u kterých je v běžné mluvě ustálená i vokalizovaná výslovnost („s. r. o." = *es er ó*, „a. s." = *á es*) a také zkratky kontrahované (cca = *cé cé á / cə cə á*).
2. Zkratky titulů (*PhDr.*, *Ing.*, *plk.*) mohou být výjimečně takto zapsány a vysloveny, většinou buď redukovaně, nebo jako zkratkové slovo („nějakej Bc. Ryšánek" = *ňejakej bə cə Rišánek*, „eště není JUDr." = *ešťe neňí judr*).
3. Mluvené anglické internetové zkratky (typu „wtf", „lol") – ačkoli v nám dostupném materiálu nejsou přítomny, mohou se objevit v záznamech současné mladé generace ve folklorním zápisu. Zkratky bývají (mimo občasnou plnou anglickou výslovnost) vyslovovány jako zkratkové

slovo, a pokud to není možné, pak redukovaně („omg" = *o mə gə*, „wtf" = *və tə fə*, „rofl" = *rofl*, „asap" = *asap*). V ojedinělých případech zaznívají zkratky i vokalizovaně („thx" = *té há iks*).

Jak plyne z uvedeného, většinu známých zkratek je třeba převádět jednotlivě, na základě konkrétní znalosti nebo předpokladu konkrétní výslovnosti, obecnější nebo univerzální pravidla většinou uplatnit nelze. Hojně se v těchto případech budou objevovat obojetnosti.

Nejprve je potřeba oddělit iniciálové zkratky od textů psaných verzálkami. To je současně poslední krok čištění folklorního přepisu i první krok převodu na normalizovaný dialektologický přepis. Pokud totiž souvislejší text psaný verzálkami budeme mazat (a je to rozumnější postup než jakékoli další složité nakládání s ním), měl by chybět jak ve výsledném dialektologickém přepisu, tak ve výchozím přepisu folklorním, aby byly dobře srovnatelné. Měli bychom si tedy text velkými písmeny označit a tam, kde bude delší než jedno slovo (s maximálně pěti znaky), ho vymazat, protože s největší pravděpodobností nepůjde o zkratku. Ani to ovšem není triviální a bezvýjimečné. Existuje několik případů, v nichž by mohl být text zkratek chybně vyhodnocen jako souvislý text psaný verzálkami nebo naopak:

- zápis zkratek ve folklorních textech není ustálený, zapisují se nejen zkratkami („LTO"), ale i vokalizovaně („EL TÉ Ó", „el té ó", i „eL Té Ó"). První typ vokalizovaného zápisu by však regulární výraz vyhodnotil jako souvislý text psaný verzálkami;
- může nastat případ výčtu zkratek (např. „ČSSD, ODS a jiné strany"), který by byl opět detekován chybně;
- existují i členěné zkratky, které pro regulární výrazy nejsou rozeznatelné od více slov („RB OSN", „FI MU");
- v textu se mohou objevit zápisy jako „HNED!", „TABÁK", „CLO", které zkratkami nejsou, ale mohou tak snadno být vyhodnoceny.

Na základě zkušenosti s dostupným materiálem a četnosti různých konkurujících si případů lze doporučit postup, který následuje.

4.6.2.2 Zpracování vokalizovaných zkratek majuskulemi

Dříve než budeme mazat text majuskulemi, nejprve je potřeba sjednotit všechny už existující vokalizované zápisy zkratek (které jsou relativně časté) a převést je všechny na minuskule krom prvních písmen ve větě. Zohledňujeme různé varianty zápisu/výslovnosti.

Celou skupinu zafixujeme do provizorních složených závorek „{}", přičemž by neměla přesáhnout pět vokalizovaných liter. Pro vokalizované jednoslovné zkratky to uděláme takto:

```
Search: \b(Á|É|Í|Ó|Ú|A|E|I|O|U|Ů|Y|Ý|BÉ|CÉ|ČÉ|DÉ|ĎÉ|GÉ|JÉ|PÉ|KVÉ
|TÉ|ŤÉ|VÉ|BÉ|CE|ČE|DE|ĎE|GE|JE|PE|KVE|TE|ŤE|VE|EF|EL|EM|EN|EŇ|ER
|EŘ|ES|EŠ|HÁ|CHÁ|KÁ|HA|CHA|KA|IKS|ZET|ŽET|ZÉ|ŽÉ|ZE|ŽE){2,5}\b
Replace: {\1}
   Options: case sensitive
```

Pro vokalizované zkratky víceslovné to následně uděláme tímto způsobem:

```
Search: (\b(Á|É|Í|Ó|Ú|A|E|I|O|U|Ů|Y|Ý|BÉ|CÉ|ČÉ|DÉ|ĎÉ|GÉ|JÉ|PÉ|KVÉ
|TÉ|ŤÉ|VÉ|BÉ|CE|ČE|DE|ĎE|GE|JE|PE|KVE|TE|ŤE|VE|EF|EL|EM|EN|EŇ|ER
|EŘ|ES|EŠ|HÁ|CHÁ|KÁ|HA|CHA|KA|IKS|ZET|ŽET|ZÉ|ŽÉ|ZE|ŽE)\b ){1,4}
(Á|É|Í|Ó|Ú|A|E|I|O|U|Ů|Y|Ý|BÉ|CÉ|ČÉ|DÉ|ĎÉ|GÉ|JÉ|PÉ|KVÉ|TÉ|ŤÉ|VÉ|BÉ
|CE|ČE|DE|ĎE|GE|JE|PE|KVE|TE|ŤE|VE|EF|EL|EM|EN|EŇ|ER|EŘ|ES|EŠ|HÁ|CHÁ
|KÁ|HA|CHA|KA|IKS|ZET|ŽET|ZÉ|ŽÉ|ZE|ŽE)\b
Replace: {\1\3}
   Options: case sensitive
```

Ve složených závorkách „{}“ už pak můžeme nahrazovat pomocí následujícího výrazu jednotlivé verzálkové znaky:

```
Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive
```

Tento regulární výraz nahrazuje „S“ za „s“. Mohli bychom sice postupovat i tím způsobem, že bychom pomocí daného výrazu nahrazovali celé vokalizované iniciály („ES“ za „es“), ale může zde potom docházet k nežádoucím překryvům („TE“ za „te“: „ZETES“ > „ZEteS“; „E“ za „e“, „I“ za „i“: „EF IKS“ > „eF iKS“), které by nám zablokovaly další převod do minuskulí. Takže buď musíme pořadí náhrad vokalizovaných liter náležitě upravit, aby k nežádoucím překryvům nedocházelo, nebo přistoupíme k jednoduššímu postupu, který však budeme muset vícekrát opakovat.

Jednodušší řešení nahrazování jednotlivých znaků je mnohem více otevřené pozdějším úpravám, protože nemusíme sledovat vzájemné vztahy mezi regulárními výrazy, když objevíme nový jev. Pokud nalezneme nový typ vokalizace iniciály, který jsme v sadě regulárních výrazů nezohlednili, stačí přidat tuto vokalizaci bez dalšího do fixujícího regulárního výrazu a není třeba dalších změn. Při náhradách celých vokalizovaných liter („ES“ za „es“) bychom museli zařadit pro nový typ vokalizace nový regulární výraz nahrazující majuskule za minuskule a promyslet znovu celé řazení, zda nedochází k nechtěným kolizím.

Jak už bylo zmíněno, uvedený regulární výraz je třeba pro každý znak několikrát opakovat. Bylo by to nutné i pro každou vokalizovanou iniciálu, neboť v jedné iniciálové zkratce se může objevit více shodných iniciál („KSSS“) a regulární výraz vždy nahrazuje ve fixované skupině jen jeden prvek. Při nahrazování prostých znaků se počet opakování ještě více zmnoží. Počítáme-li, že zkratka vyslovovaná jako jedno slovo nepřesahuje pět iniciál („NSDAP“), pak je nejlepší každý regulární výraz opakovat pětkrát, i když pětinásobný výskyt shodného znaku (mimo „É“, „E“ jakožto frekventovaný znak většiny vokalizací) už není příliš pravděpodobný.

Náležitá podoba náhrady pro znak „S“ by tedy byla následující:

```
Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive

Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive
```

```
Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive

Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive

Search: (\{[^{}]+)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive
```

A takto pro každé písmeno české abecedy. Není zde třeba obohacovat inventář znaků podle nářečí, neboť iniciálové zkratky vesměs pocházejí ze spisovného jazyka, takže využívají pouze klasické české znaky bez obohacení.

Pro počáteční písmeno vokalizované iniciálové zkratky pak použijeme následující výraz:

```
Search: (\{)S([^{}]*\})
Replace: \1s\2
   Options: case sensitive
```

Nyní, když máme převedeno na minuskule, můžeme závorky odstranit:

```
Search: \{|\}
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

V případě, že v textech detekujeme vokalizované zkratky, které jsou psány s počátečními velkými písmeny (např. „JéZéDé“ nebo „Jé Zé Dé“), můžeme použít týž postup, pouze v úvodních regulárních výrazech nahradíme vokalizované litery typu „JÉ“ za „Jé“. V takovémto případě je ideálním postupem tuto sekvenci regulárních výrazů od zavedení po odstranění složených závorek zařadit jako první, tj. ještě před ty, které řeší vokalizované zkratky psané čistě majuskulemi.

4.6.2.3 Odstranění nezkratkových textů psaných verzálkami

Teď už můžeme přistoupit k odstranění souvislejšího textu psaného verzálkami. Uděláme to následujícími kroky.

Označíme pomocí složených závorek všechna slova psaná verzálkami včetně možných předcházejících a následujících interpunkčních znamének a případného číslování arabskými nebo římskými číslicemi:

```
Search:
((\(|[0-9]+[.\)] |[IVXC]+[.\)] )?\b[A-Ž]+\b(\.|…|,|:|!|\?|;| [---] |\)))
Replace:{\1}
   Options: case sensitive
```

Výraz „[A-Ž]“ je většinou vhodnější nahradit přesnějším výrazem „[A-ZÁČĎÉÍŇŘŠŤÚŽ]“, tedy „A-Z“ + výčtem českých znaků připadajících v úvahu, neboť výraz „[A-Ž]“ obvykle obsahuje některé nežádoucí znaky, jiné žádoucí zase neobsahuje.

A poté vymažeme všechny souvislé řady výrazů ve složených závorkách:

```
Search: (\{[^{}]\} )(\{[^{}]\} )*(\{[^{}]\})
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

Dále vymažeme všechny výrazy ve složených závorkách umístěné na samostatných řádcích:

```
Search: \r\n\{[^{}]\}\r\n
Replace: \r\n
```

Pak do složených závorek umístíme pouze čistý text bez interpunkce a současně prodloužíme jeho minimální délku o jeden znak (v souvislém textu psaném verzálkami bylo třeba zachytit i slova o jednom znaku, což znamenalo, že se označila i slova o jednom znaku na začátku vět, nyní už prodloužíme minimální délku souvislého řetězce verzálek na dva znaky):

```
Search: (\{|\})
Replace:

Search: (\b[A-ZÁČĎÉÍŇŘŠŤÚŽ][A-ZÁČĎÉÍŇŘŠŤÚŽ]+\b)
Replace:{\1}
   Options: case sensitive
```

A nakonec vymažeme všechny osamocené řetězce majuskulí delší než 5 znaků:

```
Search: \{\b[A-ZÁČĎÉÍŇŘŠŤÚŽ]{6,30}\b\}
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

Tímto jsme odstranili souvislé texty psané verzálkami, ale také případné výčty zkratek a zkratky členěné na více řetězců liter. V nářečních textech ovšem jde o extrémně vzácné případy. Dále jsme naopak neodstranili osamocená krátká slova psaná verzálkami. I jich ale bude velmi malé množství.

V tuto chvíli jsme fakticky ukončili čištění folklorního textu, může proběhnout jeho normalizace (aniž nutně budeme odstraňovat složené závorky) a následně můžeme začít s vlastním převodem folklorního textu na dialektologický, kde ponechané složené závorky opět využijeme. Stejně tak ale uvedené čištění textu můžeme provést až po normalizaci folklorního textu a po uložení folklorního textu lze volně pokračovat převodem.

4.6.2.4 Převod zkratek

Samotný převod zkratek pak bude mít tři fáze.

- **Převod jednotlivých zkratek**. V první vypočítáme konkrétní a známé zkratky, u kterých si budeme relativně jisti způsobem jejich čtení. S ohledem na frekventovanost zkratky můžeme a nemusíme zavádět obojetnosti, spíše je vhodnější převést zkratku v jedné verzi, pokud ta převažuje. Budeme tedy postupovat takovýmto způsobem:

  ```
  Search: {KSČ}
  Replace: káesčé
     Options: case sensitive
  ```

```
Search: {MDŽ}
Replace: [mədəžə/médéžé}
   Options: case sensitive
```

Iniciálové zkratky vyslovované jako zkratková slova převedeme na slova s prvním velkým písmenem, aby se nám s nimi v dalším převodu dobře pracovalo:

```
Search: {NATO}
Replace: Nato
   Options: case sensitive

Search: {IČO}
Replace: Ičo
   Options: case sensitive
```

V této „slovníkové" fázi můžeme rovnou převést i zkratky jiné než psané (výhradně) verzálkami:

```
Search: s. r. o.
Replace: es er ó
   Options: case sensitive

Search: s. r. o.
Replace: Es er ó
   Options: case sensitive

Search: ([\.…\?!] |\r\n|„)MUDr.
Replace: Mudr
   Options: case sensitive

Search: MUDr.
Replace: mudr
   Options: case sensitive
```

Vždy je při těchto převodech třeba dbát na jazykové podmínky v daném nářečí.

- **Vokalizace iniciál.** Zkratky, které se v předchozí fázi nepřevedly, můžeme automaticky vokalizovat. Tato procedura může přinést i nežádoucí výsledky tam, kde fakticky nejde o zkratku, nebo i v případě, že je vyslovována jako zkratkové slovo. Může též být zvykem vyslovovat zkratku redukovanou výslovností.

  Nejprve umístíme každý znak zbylých iniciálových zkratek do vlastní složené závorky:

```
Search: \{([A-ZÁČĎÉÍŇŘŠŤÚŽ])([A-ZÁČĎÉÍŇŘŠŤÚŽ])([A-ZÁČĎÉÍŇŘŠŤÚŽ])?
([A-ZÁČĎÉÍŇŘŠŤÚŽ])?([A-ZÁČĎÉÍŇŘŠŤÚŽ])?\}
Replace: {\1}{\2}{\3}{\4}{\5}
   Options: case sensitive
```

```
Search: \{\}
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

Nyní již stačí postupovat abecedou a nahrazovat hromadně iniciály za jejich vokalizace, jejichž přehled jsme podali v části 4.6.2.1. Je možné zároveň nahrazovat i obdobné zkratky s tečkou (zkratky křestních jmen typu „F. X. Šalda“):

```
Search: \{A\}|\bA\.
Replace: á
   Options: case sensitive

Search: \{Á\}|\bÁ\.
Replace: á
   Options: case sensitive

Search: \{B\}|\bB\.
Replace: bé
   Options: case sensitive
```

- **Převod skloňovaných iniciálových zkratek.** Pro případy, kdy je iniciálová zkratka v textu skloňována a zapsána tak kombinací majuskulí a minuskulí („od ČEZu“, „v IKEMu“) včetně pravopisně nekorektních variant („s IČem“, „do SAPy“), vytvoříme regulární výrazy, které převedou zkratku mimo první písmeno do minuskulí.

  Nejprve si všechna velká písmena z daných zkratek s výjimkou prvního dáme opět do složených závorek:

```
Search: \b([A-ZÁČĎÉÍŇŘŠŤÚŽ])([A-ZÁČĎÉÍŇŘŠŤÚŽ])([A-ZÁČĎÉÍŇŘŠŤÚŽ])?
([A-ZÁČĎÉÍŇŘŠŤÚŽ])?([A-ZÁČĎÉÍŇŘŠŤÚŽ])?([a-záčďěéíňóřšťúůýž]+)\b
Replace: \1{\2}{\3}{\4}{\5}\6
   Options: case sensitive

Search: \{\}
Replace:
```

  Poté opět procházíme abecedu a nahrazujeme majuskule ve složených závorkách za minuskule:

```
Search: \{A\}
Replace: a
   Options: case sensitive

Search: \{Á\}
Replace: á
   Options: case sensitive

Search: \{B\}
Replace: b
   Options: case sensitive
```

Výsledkem bude podoba zkratkového slova s velkým počátečním písmenem.

Tímto bychom měli mít zkratky převedeny do dialektologického přepisu.

4.6.3 Slova cizího původu

Slova cizího původu mají jiná pravidla zápisu než slova domácího původu, proto je nutné zpracovat je dřív, než budou regulární výrazy provádět změny na slovech domácího původu (nebo i na slovech cizího původu, která jsou z hlediska zápisu neproblematická). V celé této části jsme částečně přihlíželi k publikaci J. Hůrkové (1995), byť se zabývá spisovnou výslovností.

4.6.3.1 Citáty z jiných jazyků

Prakticky nemožné je podchytit výslovnost citátů z jiných jazyků. Ve folklorních textech se občas objevují, ve starších textech jde nejčastěji o němčinu, ale ve vyprávěních o cestování nebo setkání s cizinci (např. s vojáky cizích vojsk) se může objevit téměř jakýkoli jazyk. Aby bylo možné tyto jazyky převést, bylo by potřeba vytvořit jejich detekci a pravidla přepisu pro každý jazyk zvlášť, což by bylo neúměrně náročné (např. pro jazyky jako angličtina, kde je grafika velmi odlišná od výslovnosti) a navíc zbytečné. O tyto jazyky a jejich správnost nám v trénovacích datech nejde, proto je logické na jejich správný přepis rezignovat.

Ve folklorních textech také často nebývá daný jazyk přepisován podle pravidel pravopisu daného jazyka, ale mnohdy víceméně foneticky, zvlášť v případech, kdy mluvčí či autor danému jazyku špatně rozuměl. Z hlediska samotného převodu je to lepší varianta, ale z hlediska jeho účelu na tom nezáleží, neboť k trénování nářečí češtiny neposlouží. Bylo by nejlepší jakékoli cizojazyčné pasáže ve folklorním i dialektologickém přepise odstranit, ale jejich detekce není jednoduchá, nemáme-li ani slovník výchozího nářečí. A detekovat množství jiných jazyků, navíc v různě deformovaných fonetických přepisech při absenci slovníku a dat výchozího jazyka, je rovněž mimořádně obtížné, a proto je rozumnější na tyto snahy v této fázi přípravy dat zcela rezignovat.

4.6.3.2 Citátové výrazy

Citátové výrazy definujme pro naše účely jako ustálená spojení dvou a více slov, pocházející z jiného jazyka, která se zapisují jeho původním pravopisem a jsou neohebná („à propos“, „déjà vu“). Ve folklorních textech se citátové výrazy objevují pouze výjimečně. V běžném nářečním projevu, uplatňujícím se zejména v soukromé neformální komunikaci, jsou vlastně nepatřičné a z toho důvodu i řídké a těžko předvídatelné. Lze je očekávat spíše jen v městské mluvě a u vzdělanějších mluvčích. Při normalizaci folklorního textu je možné jejich přítomnost zachytit, neboť bývají obvykle vysázeny kurzívou. V opačném případě jsou zpravidla přepisovány foneticky („fo pa“ vs. kurzívou zapsané „*faux pas*“), a jsou tak mimo náš nynější zájem. Citátové výrazy mohou pocházet z různých jazyků, proto je opět nelze převádět prostřednictvím obecných pravidel, ale pouze skrze jednotlivé případy. Vzhledem k jejich kontingentnímu charakteru v nářečních promluvách nelze ani stanovit jejich systematicky pojatý seznam. Je tudíž možné tuto sekci regulárních výrazů buď vůbec vypustit, nebo zahrnout pouze nejčastější, nejfrekventovanější případy a později případně doplňovat o další nalezené citátové výrazy.

Citátové výrazy jsou také oblastí, kde se nejčastěji objevují pravopisné chyby. Děje se to pochopitelně v textech, které neprošly náležitými korekturami a editorskou kontrolou, jichž bývá mezi nářečními texty obecně velké množství. Pokud jsou tedy ve zpracovávaném materiálu texty lidového původu, lokální nářeční tisky, obecní zpravodaje, internetové texty nebo texty (svobodně) ineditní, můžeme očekávanou chybovost v této sekci regulárních výrazů zohlednit a pokusit se ji podchytit.

```
Search: \bf(aux|oux|au|ou)\b \bpa(s)?\b
Replace: f{o/ó} pa
```

Regulární výraz zahrnuje varianty „faux/foux/fau/fou pas/pa“. Vynechává možnosti „fo/fó pas“, pro tyto dva zápisy je třeba vytvořit následující regulární výraz tak, aby zohledňoval zapsanou vokalickou délku.

```
Search: \bf(o/ó)\b \bpas\b
Replace: f\1 pa
```

Varianta „fo/fó pa“ není zohledněna vůbec, protože odpovídá výslovnosti, a není proto třeba ji převádět.

```
Search: \br(a|e)(nd|dn)e(s|z)( |-|-|)v(ous|ouz|ou|u)\b
Replace: randevú
```

Regulární výraz zohledňuje nejen různé varianty zápisu sousloví „rendez-vous“, ale i možnost zapsat výraz namísto spojovníku s pomlčkou, mezerou nebo bez mezery.

```
Search: \bpar\b \bex(c)?e(ll|l)(e|a|á)n(c|s)(e)?\b
Replace: par ek{s/}celán{s/c}
```

Výraz „par excellence“ s různými variantami chybného zápisu.

```
Search: \bs(cie|ai)nc(e)?( |-|-|)fict(i)?on\b
Replace: sajns fikšn
```

U slov anglického původu, jako je „science fiction“, jsou chyby v zápisu řidší, přesto se občas objevují. Je zohledněn i zápis s pomlčkou nebo zápis jednoslovný.

4.6.3.3 Výrazy graficky neadaptované

Podobným případem jako citátové výrazy jsou výrazy cizího původu graficky neadaptované („hobby“, „café“). Oproti citátovým výrazům jde o výrazy jednoslovné. Pokud jsou to substantiva, ojediněle se mohou skloňovat, čímž ale přecházejí k cizím slovům, která jsou zčásti adaptovaná a jimž jsou věnovány zbylé oddíly této podkapitoly. Platí zde jinak vše, co bylo řečeno o citátových výrazech, pouze pravděpodobnost jejich výskytu ve folklorních textech je o něco větší. Je možné tuto část regulárních výrazů opět buď zcela vypustit, nebo se pokusit vybrat ty, které by nejspíš mohly připadat v úvahu.

Je potřeba mít stále na paměti, že regulární výraz popisující výraz graficky neadaptovaný by měl být dostatečně dlouhý nebo grafotakticky specifický, aby bylo silně pravděpodobné, že nekoliduje s žádným nářečním výrazem. V opačném případě je lepší jej vyřadit nebo vyřadit některé varianty jeho chybného zápisu. Vždy je třeba daný výraz předem otestovat na co nejrozsáhlejším a nejrozmanitějším nářečním materiálu.

```
Search: \bmenu\b
Replace: meny

Search: \br(a|e)(ll|l)(y|i)(e)?\b
Replace: rel{í/i}
```

Výraz „rallye“ s variantami zápisu např. „rellye“, „rally“, „ralye“, „rallie“, ale také „raly“, „reli“ je právě případem, kdy některé varianty jsou natolik krátké a grafotakticky nespecifické (tj. sestava znaků by se snadno mohla vyskytnout i v českých nářečích). Konkrétně varianta „reli“ může kolidovat s nářečními výrazy „relija“ (s významem ‚zmatek, pokřik‘), „relik“ (‚cop, cůpek‘), „relikvíje“ aj., proto není v regulárním výrazu zohledněna možnost skloňování, která však bývá poměrně vzácná.

```
Search: \brequiem
Replace: rekvijem
```

Regulární výraz bere v úvahu možnost ojedinělého skloňování tohoto slova. Je možný i zápis „rekviem“, který v této fázi není třeba převádět, skupina „ie“ se převede na *ije* v pozdějších fázích převodu.

4.6.3.4 Německá příjmení[40]

Dlouholetá koexistence němčiny s češtinou měla vliv i na náš jazyk a naše prostředí. Většina slov převzatých z němčiny se přizpůsobila české gramatice i grafice, ale neplatí to zejména o německých příjmeních (nebo příjmeních německého původu). Velká část z nich se také ovšemže přizpůsobila („Huml“, „Rajchrt“, „Fajt“)[41] nebo adaptovala alespoň částečně či zdánlivě[42] a různou měrou („Ciesler“, „Šlais“, „Štich“). Tato jména i jména zcela nepřizpůsobená grafice češtiny („Hübschmann“ = *Hipšman*, „Thiel“ = *Til*, „Streit“ = *Štrajt*) je třeba přepisovat jinými pravidly než česká slova.

Ačkoli by sám převod němčiny nebyl tak složitý, komplikovaná je detekce němčiny, zvláště u částečně adaptovaných jmen. Proto není – až na výjimky – vhodné aplikovat obecná pravidla převodu němčiny ani na slova začínající velkým písmenem, a to kvůli možným interferencím se jmény českého, případně i jiného původu. Nezbývá tedy než vytvořit regulární výraz pro každé jméno nebo skupinu jmen, s pravidlem, že obecnější regulární výrazy musí následovat až po konkrétnějších.

Např. nejprve:

```
Search: Hüb(sch|š)ma(nn|n)
Replace: Hypšman
   Options: case sensitive
```

a až poté:

```
Search: Hüb(sch|š)
Replace: Hypš
   Options: case sensitive
```

Kdybychom pořadí přehodili, museli bychom podruhé už hledat (neexistující) „Hypšma(n|nn)“, což obecně komplikuje bezprostřední chápání daného regulárního výrazu i pozdější editace sady regulárních výrazů v případě, že bychom např. nahrazování „Hüb(sch|š)“ z nějakých důvodů museli vyřadit. Ypsilon místo měkkého *i* (*Hypš* vs. *Hipš*) vkládáme do nahrazujícího výrazu proto, že v dialektech s tvrdým *y* v hláskoslovném systému se na tomto místě tvrdé *y* vyslovuje. V dialektech, kde zapisujeme pouze splynulé *i*, se pak toto „y“ společně s originálními „y“ převede na *i* v pozdějších fázích převodu. Rozlišovat v ranějších fázích převodu měkčící a neměkčící *i* je velmi užitečné.

---

40 V celé této části o německých příjmeních jsme materiálově vycházeli ze zdrojů: Beneš (1998), Matúšová (2015) a Malačka (2011–2020).

41 Jména pro jednoduchost uvádíme pouze v nepřechýlené podobě.

42 Svou roli tu hraje i pravopis, resp. způsob zápisu českých Němců a české němčiny, který neodpovídal pravopisu německému. Tato problematika však přesahuje náš zájem, podstatné je, že výsledná grafika původně německých příjmení je mnohdy jakýmsi kompromisem mezi zápisem německým a českým.

V německých příjmeních je třeba se zaměřit na následující jevy:

- **„ei“ = *aj***

```
Search: Un(z|c)(ei|aj)t(i|y)(g|k)
Replace: Uncajty\4
   Options: case sensitive
```

Regulární výraz nahrazuje existující příjmení „Unzeitig“, „Uncajtik“, „Uncajtyk“, „Uncajtig“ a pokrývá i další možné kombinace zápisu. Nahrazuje výraz i se všemi myslitelnými příponami a koncovkami. Respektuje přitom koncovou souhlásku základu („Unzeitiga“ vs. „Uncajtika“ = *Uncajtyga* vs. *Uncajtyka*).

```
Search: Wei(s|ss)ma(nn|n)
Replace: Vajsman
   Options: case sensitive
```

Omezeně se může skupina „ei“ vyslovovat též jako „ej“, zvláště u jmen, u nichž se ztratilo povědomí o německém původu.

- **„ey“ = *ej***

```
Search: Bey((v|b)l)
Replace: Bej\1
   Options: case sensitive

Search: Beyr
Replace: Bejr
   Options: case sensitive

Search: Meyer
Replace: Mejer
   Options: case sensitive
```

- **„ai“ = *aj***

```
Search: Sailer
Replace: Sajler
   Options: case sensitive

Search: Krain
Replace: Krajn
   Options: case sensitive
```

- **„ie“ = *í/ý***

```
Search: Klier
Replace: Klír
   Options: case sensitive
```

Zde dáváme měkké *í*, protože *l* cizího původu bývá v nářečích reflektováno jen jako netvrdé *l*, nemůže tedy po něm následovat tvrdé *y*.

```
Search: Nied(e)?rl
Replace: Nýdrl
   Options: case sensitive
```

Jde o jméno „Niederle/Niedrle". Zde nahrazujeme tvrdým *ý*, protože nedochází ke změkčení *N* (ostatně existuje i příjmení „Nýdrle"). Protože jméno „Niederle/Niedrle" v dalších pádech nemusí mít „-e" (gen. *Nýdrla*), není koncové „-e" obsaženo v regulárním výrazu. Teprve po jménech jako „Niederberger", „Niederhofer", „Niederm(aj|ai|ay|ei)er" aj. může následovat jméno „Nieder". Jak ještě uvidíme, kvůli ostatním pádům i přechýleným formám a dalším odvozeninám je třeba počítat v regulárním výrazu s výpustkou předkoncového „-e-" („Niedrová", „Niederbergrovi" vs. většinou „Niederbergerová" atd.).

```
Search: (Sch|Š)m(ie|í)d
Replace: Šmíd
   Options: case sensitive
```

Jde v prvé řadě o jméno „Schmied". Regulární výraz nahrazuje i existující formy „Šmied" a „Schmíd", vyhledává, ale nemění i formu „Šmíd". Pro formy s krátkým vokálem a formy dlouhé se zakončením na „dt" = *t* je třeba samostatný regulární výraz. Regulární výraz by měl následovat až po vyřešení jmen, kde je nějaká forma tohoto jména složkou (při rozlišování velkých a malých písmen stačí předem podchytit jména typu „Schmiedmayer", nezasáhne do jmen typu „Hammerschmied").

- **kvantita**

Délka samohlásek není často v německých příjmeních nijak značena, i když bývá v češtině vyslovována.

```
Search: Ober(s|š)tein
Replace: Óbrštajn
   Options: case sensitive
```

Délka však není závazná, někdy se u týchž jmen nevyslovuje, výsledkem by tedy měla být obojetnost. Je potom na zvážení, zda je materiál natolik rozsáhlý a jméno a jeho krátká výslovnost natolik frekventované, aby se pro daný případ vyplatilo obojetnost zavádět.

```
Search: Wagner
Replace: V[a/á]gner
   Options: case sensitive

Search: Kroner
Replace: Króner
   Options: case sensitive
```

- **„h“ po samohlásce, před souhláskou = dlouhá samohláska**

```
Search: Mahler
Replace: Máler
   Options: case sensitive

Search: Ohma(nn|n)
Replace: Óman
   Options: case sensitive
```

U jmen, která jsou kratší, jako „Ohm“ = *Óm* nebo „Uhl“ = *Úl* je nutné dávat velký pozor na interferenci s českými výrazy. Platí to obecně o jakémkoli kratším jménu. V tomto případě hrozí interference s českými výrazy jako „ohmatat“, „ohmejzdat“, „uhladit“, „uhlí“ aj. na začátku věty i s českými příjmeními jako „Uhlík“, „Uhlíř“, „Uhlář“ nebo toponymy jako „Uhliště“, „Uhlířské Janovice“ atd. Protože tyto výrazy jsou v úhrnu pravděpodobnější než výskyt daného jména, lze správný převod příjmení zajistit pouze přidáním možných přípon a koncovek, specifikovaných pro danou nářeční oblast. Následující dva regulární výrazy jsou lokalizovány do oblasti dialektů severovýchodočeských (1-1).

Vzhledem k tomu, že je nutné nahrazovat právě první velké písmeno, není možné převést současně i případné výskyty fyzikální jednotky „ohm“ (jinak by to šlo přidáním jejích koncovek, vytvořením referenční skupiny z prvního písmene a odstraněním podmínky „case sensitive“).

```
Search: Uhl((((|a|ovi|e|em|ov[eé]|u|ů|um|ům|om|ech|[áa]ch|y)|
((k|ou̯k|ovk|o[vw])(a|y|je|u|o|ou|[áa]ch|[áa]m|ama]))|
(ce|ou̯ce|ovce)|(ek|ou̯ek|o[vw]ek)|ic(|e|i|í|[ií]ch|[ií]m|ema)|o[vw]
(i|á|ý|ou|ejch|ech|ých|ejm|em|ým|ejma|ýma))[ \.,:;\?!])
Replace: Úl\1
   Options: case sensitive

Search: Ohm((((|a|ovi|e|em|i|ov[eé]|u|ů|um|ům|om|ech|[áa]ch|y)|
((k|ou̯k|ovk|o[vw])(a|y|je|u|o|ou|[áa]ch|[áa]m|ama]))|
(ce|ou̯ce|ovce)|(ek|ou̯ek|o[vw]ek)|ic(|e|i|í|[ií]ch|[ií]m|ema)|o[vw]
(i|á|ý|ou|ejch|ech|ých|ejm|em|ým|ejma|ýma))[ \.,:;\?!])
Replace: Óm\1
   Options: case sensitive
```

Regulární výraz podchycuje nejen tvary příjmení „Uhl“ a „Ohm“ nebo „Uhlová“, „Ohmovi“, ale i různé oblastní (severovýchodočeské) formy a tvary příjmení jako „Uhlice“ (= *Úlice*), „Ohmka“ (= *Ómka*), „Uhlou̯ka“, „Ohmowa“ aj. i s jejich lokálními pádovými koncovkami. Kvůli interferenci s apelativem „uhli“ je zanedbán řídký tvar nominativu plurálu maskulina „Uhli“, který se v případě výskytu nepřevede, protože by náhradami apelativa materiál spíše poškozoval. Podobný, avšak komplikovanější případ je jméno „Pohl“ (= *Pól*), kde jde ve stejném případě o interferenci se slovesem „pohnout“ v jižních Čechách a na Moravě, a to v různých tvarech minulého příčestí: „pohl“, „pohla“, „pohli“ a „pohly“. Krom toho v jižní podskupině středomoravských dialektů může kolidovat tvar vokativu singuláru příjmení „Pohle“ a minulé příčestí v plurálu všech rodů slovesa „pohnót“ („pohle“). V takovýchto případech je prakticky vždy proprium cizího původu vzácnější, a musí ustoupit. Každý regulární výraz je nutné testovat na co nejrozsáhlejším materiálu, pokud nevyhledává prvky zcela nečeské.

- **zdvojené vokály = dlouhý vokál**

```
Search: Haas
Replace: Hás
   Options: case sensitive

Search: Beer
Replace: Bér
   Options: case sensitive
```

- **zdvojené konsonanty = jednoduchý konsonant**

```
Search: Klemm
Replace: Klem
   Options: case sensitive

Search: Miller
Replace: Miler
   Options: case sensitive
```

Pro zdvojené konsonanty, zejména pro skupinu „nn“, je třeba počítat s tím, že v některých oblastech se může vyslovovat jako gemináta. Územní rozsah geminované výslovnosti „nn“ určuje přibližně *Český jazykový atlas* (Balhar a kol., 2005, s. 426–428). V oblasti východomoravských a slezských nářečí dochází místy ke geminaci i u slov jako *masso*, *kašša*, nemáme ale dostatek dat, zda a jak to ovlivňuje i výslovnost německých příjmení se zdvojenými konsonanty. Jde o mizející jev, takže je možné jej zanedbat.

- **„v“ = f**

```
Search: (V|F)(ei|ai|aj)t
Replace: Fajt
   Options: case sensitive
```

Zahrnuje existující jména „Veit“, „Vait“, „Vajt“, „Feit“, „Fait“, „Fajt“.

```
Search: Hauptvog(e)?l
Replace: Hauptfógl
   Options: case sensitive
```

- **přehlásky**

Zohledňujeme pět přehlásek:

- „ä“ > *e/é*
  Krátké i dlouhé.

```
Search: Grätz
Replace: Gréc
   Options: case sensitive
```

```
Search: K(äs|ess)tner
Replace: Kestner
   Options: case sensitive
```

- „ö“ > *e/é*
  Krátké i dlouhé.

```
Search: Höger
Replace: Hégr
   Options: case sensitive

Search: Körner
Replace: K[é/e]rner
   Options: case sensitive
```

- „ü“ > *i/í/y/ý*
  Krátké i dlouhé.

```
Search: Kurfürst
Replace: Kurfirst
   Options: case sensitive

Search: Frühauf
Replace: Frýhauf
   Options: case sensitive
```

- „eu“ > *oj*

```
Search: Neuma(nn|n)
Replace: Nojman
   Options: case sensitive

Search: Freund
Replace: Frojnd
   Options: case sensitive
```

- „äu“ > *oj*

```
Search: Häusler
Replace: Hojzler
   Options: case sensitive

Search: Ohnhäuser
Replace: Ónhojz[e/]r
   Options: case sensitive
```

- **„th“ = *t***

```
Search: Thomas
Replace: Tomas
   Options: case sensitive
```

```
Search: Barth
Replace: Bárt
   Options: case sensitive
```

- **„sch“ = š**

```
Search: (Sch|Š)(w|v)ar(z|tz|c)
Replace: Švarc
   Options: case sensitive

Search: Hirsch
Replace: Hirš
   Options: case sensitive
```

- **„tsch“ = č**

```
Search: Fritsch
Replace: Fryč
   Options: case sensitive

Search: Deutsch
Replace: Dojč
   Options: case sensitive
```

- **„ck“ = *k***

```
Search: Schi(ck|k)
Replace: Šik
   Options: case sensitive

Search: E(ck|k)erth?
Replace: Ekrt
   Options: case sensitive
```

- **výpustky *e* („er“, „el“, „en“)**

Ve víceslabičných jménech často dochází v souladu s německou výslovností k tomu, že se skupiny „er“, „el“, „en“ před koncem slova nebo původním morfologickým švem vyslovují jako slabikotvorné hlásky *r*, *l*, *n*, nebo je *e* nahrazeno redukovaným vokálem *ə* (zvl. v případě skupiny „en“). Česká výslovnost německých jmen však v některých případech *e* zachovává. Nečiní tak důsledně, svou velkou roli jistě hraje i lokální a rodinná tradice výslovnosti příjmení, ale přesto se tak děje v dobře definovatelných hláskoslovných okolích.

- „er“

  Pokud zakončení *-er* předchází sonanta *r*, *l*, *n*, *j* nebo samohláska, pak hláska *e* prakticky vždy bývá vyslovována.

```
Search: Lederer
Replace: Léderer
   Options: case sensitive
```

```
Search: Koller
Replace: Koler
   Options: case sensitive

Search: Bogner
Replace: B[o/ó]gner
   Options: case sensitive

Search: G(aj|ai||ay|ei)er
Replace: Gajer
   Options: case sensitive

Search: Neubauer
Replace: Nojbauer
   Options: case sensitive
```

Pravidelně se vyslovuje *e* ještě v příjmení „Fischer/Fišer“:

```
Search: Fischer
Replace: Fišer
   Options: case sensitive
```

V ostatních případech se *e* spíše nevyslovuje, ale je třeba počítat s oběma možnostmi:

```
Search: Laichter
Replace: Lajcht[e/]r
   Options: case sensitive

Search: Balzer
Replace: Balc[e/]r
   Options: case sensitive

Search: Gr(u|ú|ů)ber
Replace: Grúb[e/]r
   Options: case sensitive

Search: Wimmer
Replace: Vim[e/]r
   Options: case sensitive
```

- „el“

V koncové skupině „-el“ se většinou *e* vypouští, ojediněle se však objevuje i výslovnost s *e*:

```
Search: Mein(|e)l
Replace: Majnl
   Options: case sensitive
```

```
Search: Men(z|tz|c)(|e)l
Replace: Mencl
   Options: case sensitive

Search: Blümel
Replace: Bl[i/í]m[e/]l
   Options: case sensitive
```

- „en"

  Na konci slova se *e* v naprosté většině vyslovuje („Erben"), před morfologickým švem se někdy vyslovuje, někdy ne. Zde je tedy potřeba počítat s obojetností („Rosenberg").

```
Search: Hagen
Replace: H[a/á]gen
   Options: case sensitive

Search: Wagenknecht
Replace: Vág[e/]nknecht
   Options: case sensitive
```

- **„di", „ti", „ni" > *dy*, *ty*, *ny***

```
Search: Di(t|tt)r(i|y)ch
Replace: Dytrych
   Options: case sensitive

Search: (S|Š)tifter
Replace: Štyft[e/]r
   Options: case sensitive

Search: Honig
Replace: H[o/ó]nyg
   Options: case sensitive
```

- **„*s*" = [z/s]**

  Jednoduché „s" po samohlásce (včetně dvojhlásek s klouzavou složkou *j*, tj. vyslovovaných *aj*, *ej*, *oj*) se zpravidla vyslovuje jako *z*, ale výslovnost může kolísat mezi *z* a *s*. Máme zde samozřejmě na mysli „s", které není ovlivněno dalšími asimilačními nebo jinými pravidly („Halsbach") a které nebývá zpravidla v jiné verzi příjmení zdvojené. Proto například nenahrazujeme příjmení „Hes" za „He[s/z]", ale nahrazujeme:

```
Search: Kraus
Replace: Krau[s/z]
   Options: case sensitive

Search: Eisenberger
Replace: Aj[s/z][e/]nberg[e/]r
   Options: case sensitive
```

```
Search: Eisenbergr
Replace: Aj[s/z][e/]nbergr
   Options: case sensitive
```

Jméno „Hes“ máme i ve variantě „Hess“, tvary a odvozeniny budou tedy znít *Hese*, *Hesová*, *Hesovi* atd. Naproti tomu např. jméno „Kraus“ sice na první pohled působí tak, že vstupuje do asimilací s koncovým *-s*, ale není to tak, v okolí neovlivněném asimilacemi se nám většinou objevuje základ zakončený na *-z*, tedy *Krauze*, *Krauzová*, *Krauzovi*, i když se řidčeji objevují i formy *Krause*, *Krausová*, *Krausovi*.

- **„tz“, „z“ = *c***

```
Search: Dietz
Replace: Dýc
   Options: case sensitive

Search: Kreut?zmann?
Replace: Krojcman
   Options: case sensitive
```

4.6.3.5 Cizí slova částečně graficky adaptovaná

Řada slov cizího původu se v nářečích češtiny adaptovala foneticky, morfologicky i slovotvorně, ale přesto v jejich grafice zůstávají prvky jejich původního pravopisu. Nejčastěji jde o tvrdou výslovnost skupin „di“, „ti“, „ni“ („statika“), často také neoznačování kvantity vokálů („definitiva“). Existují sice určité vzorce, podle kterých lze takové skupiny najít a hromadně nahradit, ale ve většině případů je třeba postupovat v převodu víceméně po jednotlivých slovech, nebo spíš po jejich základech, jejichž pomocí lze převést celou skupinu slov příbuzných a všechny jejich morfologické tvary.

Platí zde opět pravidlo, že musíme postupovat od konkrétnějších případů k obecnějším, které by je zahrnovaly. I když představíme jednotlivé jevy, které budeme převádět, odděleně, je vždy u každého konkrétního základu slova zapotřebí převést všechny přítomné jevy naráz. Hledat podruhé již převodem pozměněný základ slova je zbytečně komplikující. Vytvořili bychom nežádoucí vzájemné vazby mezi jednotlivými kroky a jakýkoli pozdější zásah do regulárních výrazů by vyžadoval kontroly všech výrazů navazujících. Tento postup také usnadňuje čitelnost a přehlednost regulárních výrazů, neboť základy slov jsou ve svém zápisu totožné se spisovným nebo folklorním zápisem.

- **„di“, „ti“, „ni“ = *dy*, *ty*, *ny***

  Jak již bylo dříve uvedeno, znak „y“ na tato místa dosazujeme, ať už dialekt má, nebo nemá tvrdé *y*, neboť v pozdějších fázích převodu budou případně „y“ nahrazena. Nyní nám pomohou izolovat česká „di“, „ti“, „ni“, která budou následně převedena na *dí*, *tí*, *ňi*.

  Regulární výrazy stavíme následujícím způsobem:

```
Search: (harmon)i
Replace: \1y
```

  Tento regulární výraz např. zahrnuje slova jako „harmonicky“, „harmonie“, „harmonika“, „harmonikář“, „harmonirovat‘“ a jiné včetně jejich tvarů. Ale vyhovuje třeba i slovům jako „neharmonickej“, „disharmonie“, „filharmonie“ ap. U slov, která nevymezíme zepředu, je třeba vyzkoušet, zda nemohou vyhovovat i jinému než žádoucímu řetězci v nářečních textech, zvláště pokud jsou kratší.

Pakliže víme, že existuje sloveso „harmonírovat“, musíme pro něj vytvořit zvláštní regulární výraz, který bude mít dlouhé „í“. Opět ho nebudeme omezovat zepředu, protože sloveso může mít prefixy a výraz je dostatečně jednoznačný:

```
Search: (harmon)í
Replace: \1ý
```

Pokud chceme vytvořit regulární výraz pro slovo „diplomat“, měli bychom ho nejprve vytvořit pro slovo „diplomatický“ (případně i „diplomatika“, „diplomatismus“), kde máme dva jevy, které musíme převést. Kdybychom pracovali pouze s úsekem „diplomati“, byl by homonymní s nominativem plurálu slova „diplomat“, kde se poslední „t“ vyslovuje jako *ť*. Proto musíme postupovat takto, přičemž druhý regulární výraz obsáhne nejen slovo „diplomat“, ale i „diplom“, „diplomovaný“ apod.:

```
Search: (diplomat)i([cksz])
Replace: \1y\2

Search: (d)i(plom)
Replace: \1y\2
```

V případě, že musíme ve slově nahradit více jevů, postupujeme např. jako u slova „begonie“, které bývá zapisováno i „begónie“. Vzhledem k dvojímu zápisu a dubletní délce *o/ó* při zápisu krátkém nemůžeme vyhledávat pomocí výrazu „(beg[oó]n)i“, ale musíme jej rozdělit na dva:

```
Search: (beg)oni
Replace: \1[o/ó]ny

Search: (begón)i
Replace: \1y
```

Zepředu musíme omezovat některá slova nebo části jejich řetězců, neboť jsou homonymní s řetězci ve výrazech domácích:

```
Search: \b(ant)i
Replace: \1y

Search: \b(n)i(kl)
Replace: \1y\2
```

Zde například předpona „anti-“ se zakončením nominativu plurálu substantiv typu „haranti“, „parchanti“, ve druhém případě „podnikl“, „vynikl“ apod. Regulární výraz typu „\b(ant)i“ by však měl být v pořadí až mezi posledními a nejobecnějšími.

K dalším takovým obecným regulárním výrazům patří např.:

```
Search: ([dtn])isti(k|ck)
Replace: \1ysty\2

Search: (ist)i(k|ck)
Replace: \1y\2
```

```
Search: ([dtn])i(vist)
Replace: \1y\2

Search: ([dtn])i([sz]m)
Replace: \1y\2

Search: ([dtn])i([sz]a[cč])
Replace: \1y\2

Search: \b(d)i(š)
Replace: \1y\2
```

U regulárního výrazu:

```
Search: ([dtn])i(a|á|e|i|í|o|ó|u|ú|ů)
Replace: \1y\2
```

je třeba nejen přizpůsobit sadu vokálů konkrétnímu nářečí, ale také se některým kombinacím, konkrétně „ia“, „iá“, „ie“, vyhnout u kopaničářské nářeční podskupiny, kde se objevuje v domácích slovech. Regulární výraz tam tedy bude vypadat takto:

```
Search: ([dtn])i(i|í|o|ó|u|ú|ů)
Replace: \1y\2
```

Hiátové *j* v této fázi ještě doplňovat nebudeme.

Podobně to platí pro regulární výraz, který popisuje typ „bomboniéra“, „moskytiéra“, který se vyslovuje dubletně, měkce i tvrdě (*bomboňijéra/bombonijéra*). V kopaničářských dialektech se vyskytuje „ié“ i ve slovech domácího původu, proto je třeba jej v těchto dialektech vyloučit. Pro zbylé nářeční podskupiny použijeme tento výraz:

```
Search: ([dtn])i(é)
Replace: [\1y/\1i]\2
```

„D“, „t“, „n“ se neměkčí, také když předcházejí příponám „-ista“, „-istický“, „-istika“ apod. Regulární výraz ale musíme rozdělit kvůli ojedinělým případům složenin „pětistovka“, „šestistovka“, „devítistovka“. Slova typu „devítistěn“ můžeme zanedbat.

```
Search: ([dn])i(st)
Replace: \1y\2

Search: ([^ěsíi]t)i(st)
Replace: \1y\2
```

- **kvantita ve slovech cizího původu**

  Typicky jde u slov cizího původu o neznačenou délku vokálu, která navíc bývá zpravidla dubletní. Existují sice typická zakončení slov, u nichž k tomu dochází, např. „-on“, „-ivní“, „-ura“, „-ilie“, „-emie“, „-erium“, „-erie“, „-log“, „-nom“, „-ion“ (podrobněji viz Hůrková, 1995, s. 51–54), ale nejsou prakticky nikdy dostatečně specifická, aby nekolidovala se slovy domácího původu. Opět tedy postupujeme metodou „slovníkového“ výčtu tam, kde jsme na slovo nenarazili už v rámci předchozího oddílu.

```
Search: (citr)o(n)
Replace: \1[o/ó]\2

Search: (benz)i(n)
Replace: \1[i/í]\2

Search: (vag)o(n)
Replace: \1[o/ó]\2

Search: (termof)o(r)
Replace: \1[o/ó]\2
```

- **další případy u slov přejatých**

  Existuje řada dalších jevů ve slovech přejatých, které je nutné řešit pomocí regulárních výrazů vyhledávajících konkrétní slova.

  - tautosylabické „ai“: Tato skupina je v přejatých slovech poměrně častá („detail“, „email“, „kokain“, „mozaika“), ale nelze ji dobře automaticky rozeznat od skupiny heterosylabické („Kain“, „archaický“, „naivní“, „zainteresovat“) nebo případně jinak vyslovované („fair“, „airbag“). Proto je nutné vycházet opět z jednotlivých slov:

```
Search: (med)ai(l)
Replace: \1aj\2

Search: \b(l)ai(k|c)
Replace: \1aj\2
```

  - další tautosylabická spojení „ei“, „oi“, „ui“: Řidčeji se objevují další tautosylabická spojení s *i* jako klouzavou složkou.

```
Search: (kof)ei(n)
Replace: \1ej\2

Search: (Al)ois
Replace: \1ojz
   Options: case sensitive

Search: (r)ui(n)
Replace: \1uj\2
```

- změna „s“ > *z*: V cizích slovech je opět poměrně častá, do značné míry se to projevilo v současné spisovné normě, text však může vzhledem k datu svého vzniku dodržovat starší pravopis. Často, pokud je vyslovováno *z* místo *s*, se to projeví v zápisu, protože je to obvykle zapisovatelem vnímáno jako nestandardní, nářeční prvek („do servizu“). Existuje však několik případů, kdy „s“ můžeme nahrazovat, včetně jednoho obecného, a to přípony „-ismus“. Daná skupina hlásek „ism“ se však může potenciálně objevovat i v českých slovech, proto je potřeba přidat další její určení, včetně koncovek, které musí být přizpůsobeny danému nářečí.

```
Search: (dy)s(le(x|kt))
Replace: \1z\2

Search: (pul)s
Replace: \1z

Search:
\b([a-zěščřžýáíéďťňúůó]{3,20}i)s(m(u|em|y|ů|u|ům|um|ech|ách|
ama|ami))\b
Replace: \1z\2
```

- skupina „th“: Obvykle se vyslovuje jako *t*, pokud je ve slovech německého nebo řeckého původu (ve starším pravopisu), může se však vyskytnout i ve slovech domácích nebo ve slovech např. původu anglického, kde se vyslovuje jinak. Proto není možné nahrazovat čistě jen tuto skupinu hlásek a je třeba opět přistoupit k převodu na jednotlivých slovech:

```
Search: (met)h(anol)
Replace: \1\2

Search: (Furt)h
Replace: \1
   Options: case sensitive
```

- ostatní případy: Existuje krom toho celá řada případů, které vzhledem k jejich nízké frekvenci nemá smysl typologizovat, je však možné je řešit právě po jednotlivých slovních základech:

```
Search: (z)oo
Replace: \1ó

Search: (l)eas(in[gk])
Replace: \1íz\2

Search: airbag
Replace: érbeg

Search: \b(po)i(nt)
Replace: \1e\2
```

4.6.3.6 Izolované znaky

Po provedení změn na konkrétnějších případech je možné aplikovat některé obecnější zásady, jak se vyslovují znaky cizích jazyků.

- „w“ = *v*

  Výslovnost znaku „w“ jako *v* má svou výjimku ve starších podkrkonošských dialektech; pokud tomu odpovídá zpracovávaný materiál, je třeba tento regulární výraz bez náhrady vypustit.

  ```
  Search: w
  Replace: v
  ```

- „ö“ = *e*

  ```
  Search: ö
  Replace: e
  ```

- „ü“ = *i*

  ```
  Search: ü
  Replace: i
  ```

- „ä“ = *e*

  ```
  Search: ä
  Replace: e
  ```

- intervokalické „i“

  ```
  Search: [aeiouyáěéíioóuúů]i[aeiouyáěéíioóuúů]
  Replace: \1j\2
  ```

  (Druhou skupinu vokálů v regulárním výrazu je třeba upravit dle nářečí.)

4.6.4 „Di“, „ti“, „ni“, „dě“, „tě“, „ně“ > *ďi, ťi, ňi, ďe, ťe, ňe*

Nyní můžeme přistoupit k převodu „di“, „ti“, „ni“ na *ďi*, *ťi*, *ňi* a „dě“, „tě“, „ně“ na *ďe*, *ťe*, *ňe*. Jde o rozsáhlou a zásadní část převodu, která proměňuje celý text. Po stránce regulárních výrazů však bude poměrně stručná, veškeré problémy, které by se v ní mohly skrývat, jsme vyřešili v předchozích částech.

```
Search: d(i|í|ě)
Replace: ď\1
   Options: case sensitive

Search: D(i|í|ě)
Replace: Ď\1
   Options: case sensitive

Search: t(i|í|ě)
Replace: ť\1
   Options: case sensitive
```

```
Search: T(i|í|ě)
Replace: Ť\1
   Options: case sensitive

Search: n(i|í|ě)
Replace: ň\1
   Options: case sensitive

Search: N(i|í|ě)
Replace: Ň\1
   Options: case sensitive

Search: ([ďťň])ě
Replace: \1e
```

4.6.5 Znaky „x“ a „ch“

Z jazykového hlediska patří řešení znaku „x“ (*iks*) ještě do kategorie cizích slov, neboť se ve slovech domácího původu nevyskytuje. Ale současně souvisí i s náhradou digrafu „ch“ za grafém „x“, kterým si pomůžeme v dalším průběhu převodu. Pomáháme si jím především proto, abychom nemuseli digraf „ch“ uchopovat složitějšími postupy. Tento krok je dosti důležitý na to, abychom ho vyčlenili jako samostatný oddíl a nezanořili do oddílu předchozího nebo následujícího. Právě na tomto místě musí být z toho důvodu, že jednotlivé slovní výjimky bylo nejvhodnější nahrazovat ještě ve fázi, kdy jsme do textu nezasahovali jinými rozsáhlejšími náhradami, a slova tak stále vypadala jako ve „slovníku“, tj. nebyla na půli cesty v přeměně mezi folklorním a dialektologickým přepisem. Nyní, když už zahajujeme obecnější a důsažnější změny, musíme sáhnout k tomuto kroku, který nám manipulaci s *ch* výrazně ulehčí.

Protože chceme *ch* nahradit v textech provizorně za *x*, musíme nejprve „x“ nahradit za jeho zvukovou podobu. Není to úplně jednoduché, protože „x“ se většinou reflektuje jako *ks*, ale v některých slovech jako „existovat“ = *egzistovat*, „exekutor“ = *egzekutor* se většinou vyslovuje jako *gz*. Ale ani to ne vždycky, v nářečích české skupiny (1) se často setkáváme s tím, že se i v těchto kontextech objevuje dubletně výslovnost *ks*. V těchto nářečích tedy musíme zavést obojetnost. Názorně to ukážeme na následujících regulárních výrazech.

Museli jsme v nich zohlednit nejen formace začínající na „ex“ + samohláska, ale i jejich prefixace, které je nutné upravit dle konkrétního nářečí. Existují i případy, kdy je „ex-“ vnímáno i v našem prostředí jako předpona a vokál po ní následující může mít proto ráz, ale jde o tak vzácná slova („exadmirál“), že tuto možnost nezohledňujeme. Případy, kdy dochází k asimilaci před znělou („exhumace“ = *egzhumace*, „Rixdorf“ = *Rigzdorf*), pokud se objeví, budou ošetřeny až při převodu běžných asimilací znělosti. Nynější převod všech „x“ mimo jmenované výjimky na *ks* tedy nevadí.

```
Search: X
Replace: Ks
   Options: case sensitive

Search: \b((do|na|ná|vo|po|pro|pře|při|pří|s|u|ú|vy|vý|za|zá|ne)?e)
x(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[ks/gz]\3
```

```
Search: \b(na|vo|po|pře)dex(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[d/t]e[ks/gz]\2

Search: \b(v?o)bex(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[b/p]e[ks/gz]\2

Search: \b(ro|v|be)?zex(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[z/s]e[ks/gz]\2

Search: \bvex(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[v/f]e[ks/gz]\2

Search: x
Replace: ks
   Options: case sensitive

Search: \b((na|vo|po|pře))dex(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: \1[d/t]e[ks/gz]\3
```

V regulárních výrazech je současně zohledněna asimilace znělosti na morfologickém švu před vokálem *e*. Zde je provedena pro případ dublety.

Následně už jen převedeme digraf „ch“ na jednoduchý znak „x“, který budeme používat až do konce našeho převodu, kde jej nahradíme zpět za „ch“.

```
Search: C[hH]
Replace: X
   Options: case sensitive

Search: ch
Replace: x
   Options: case sensitive
```

4.6.6 Příprava před asimilacemi u jedinečných souhlásek, *v* a *h* a před vokály

Jedinečné souhlásky jsou souhlásky *j, r, l, m, n, ň* a k nim v různých nářečích přistupují ještě *ł* a *u̯* (regionální obměny *l*). Dále k nim v širším smyslu můžeme přičleňovat ještě *v* a *h*, které měly svou jedinečnou pozici v systému spíše historicky a jejichž asimilace se řídí vlastními pravidly. K asimilacím před *v* počítáme *i* asimilace před *u̯, w* (regionální obměny *v*). Jedinečnými souhláskami jsou i dlouhé *ŕ, ĺ* a *ḱ*, které ale jako slabikotvorné hlásky nikdy nebývají v pozici, kdy by jejich působením docházelo ke znělostní asimilaci.

Asimilace znělosti můžeme pro naše účely rozdělit do čtyř druhů:

- vnitroslovní: dochází k nim mezi sousedícími hláskami uvnitř jednoho slova;
- na morfologickém švu: uvnitř slova na morfologickém švu;
- předložkové: na rozhraní mezi slovem a předložkou;
- mezislovní: na rozhraní slov.

Nás v tuto chvíli zajímají pouze asimilace předložkové, které musíme ošetřit, protože zahrnují výjimky. Do vnitroslovních asimilací klasické jedinečné souhlásky ani vokály nevstupují vůbec, *v* a *h* naproti tomu mohou být vnitroslovně asimilovány regresivně i progresivně. Na morfologickém švu dochází v případě jedinečných souhlásek typicky jen k asimilaci před sufixem *-me*, *-my* (v 1. os pl. indikativu, imperativu i kondicionálu aktiva). Vokály (či spíše ráz před nimi) pak naopak mohou mít vliv na prefixy („podučitel" = *potučitel,* „podorat" = *potorat*). Těmito typy asimilací se však budeme zabývat až v pozdějších fázích převodu. Pokud jde o asimilace mezislovní, v této pozici dochází k asimilacím všech typů, přípravu provádíme právě kvůli nim, protože je bez ošetření konkrétních případů nelze prostřednictvím regulárních výrazů rozpoznat od asimilací předložkových.

V rámci předložkových asimilací platí, že u většiny předložek k asimilaci před jedinečnou nedochází. Jde vesměs o předložky zakončené na znělý konsonant: „nad", „(v)od", „pod", „zpod", „před", „zpřed", „bez", „v", „h", „z". „H" je přitom varianta předložky „v" v dialektech Podkrkonoší. Zvláštní výjimkou je předložka „přes", která většinou má podobu *přez*, ale bývá dubletní, zvláště v dialektech české nářeční skupiny (srov. Bělič, 1972, s. 63).

```
Search: \b(nad|[v]?od|pod|[zs]pod|před|[zs]před|bez|v|h|z) (j|r|l|m|n|ň)
Replace: \1§§\2

Search: \b(naď|[v]?oď|poď|[zs]poď|přeď|[zs]přeď) (ň)
Replace: \1§§\2

Search: \b(pře)s (j|r|l|m|n|ň)
Replace: \1[s/z]§§\2
```

Skupinu „§§" zavádíme proto, abychom spojení dočasně odlišili od mezislovní asimilace. U výčtu předložek je třeba zohlednit pravidelné regionální obměny, v tomto případě např. varianty „pud" (obměna „pod"), „vud" (obměna „od") v jižní středomoravské podskupině (2-2) nebo „pred" (obměna „před") u kopaničářských dialektů (3-3). Zohledněny jsou ve folklorních textech občas zapsané varianty typu „přeď ňú", kde je zachycena asimilace měkkosti. Mohou se objevit na celém území. Může zde docházet k homonymiím („pod'", „před'" jako imperativy, které se mohou asimilovat jinak), ale pravděpodobnost jejich výskytu před „ň-" je nesrovnatelně nižší než u předložek. Skupinu „(j|r|l|m|n|ň)" je třeba upravovat podle převáděného dialektu, zvláště proto, že znak „u̯" může být dle nářečí jednou obměnou *v*, jindy *l*. Obojetnost u předložky „přes" nezavádíme u dialektů s vysokou mírou znělosti.

Není vhodné pro skupinu „(j|r|l|m|n|ň)" používat hranaté závorky „[jrlmnň]", neboť v případě grafémů, které jsou složeny z kombinačních znaků, jako je např. znak „u̯", je tento grafém při provádění regulárního výrazu rozeznán jako znaky dva „u" a „̯" a vyhledává je zvlášť. Jakmile tedy může být mezi znaky i jen potenciálně speciální nářeční znak, je lépe používat formu „(||||)". Při jakémkoli doplňování nebo rozšiřování regulárního výrazu pak nemusíme měnit jeho skladbu.

Dalšími typy předložkové asimilace jsou asimilace neslabičných předložek *k* a *s* před jedinečnou, před *v*, *h* a před samohláskou. V rámci těchto skupin jsou výjimkou osobní zájmena, před nimiž se předložky až na ojedinělé a nepravidelné výjimky neasimilují (srov. Bělič, 1972, s. 64). Tyto skupiny fixujeme prostřednictvím jiné skupiny znaků (QQ), neboť je budeme potřebovat odlišit.

```
Search: \b(k) ((ňemu|ňí|ňi|ňím|ňim|[nv]ám|[nv]am)\b
Replace: \1QQ\2
```

```
Search: \b(s) (ňim|ňím|ňí|ňi|ňíma|ňima|ňema|[nvV][aá]m[ai])\b
Replace: \1QQ\2
```

Je třeba zohlednit tvary zájmen podle nářečí. V regulárních výrazech výše jsou upraveny pro severovýchodočeskou nářeční podskupinu (1-1). Pro severní východomoravskou nářeční podskupinu (3-2) by regulární výrazy zajišťující tutéž funkci vypadaly takto:

```
Search: \b(k) (ňemu|ňem|ňí|ňi|ňý|ňéj|ňej|ňim|ňím|ňm|[nv]ám|[nv]am)\b
Replace: \1QQ\2

Search: \b(s) ((ň[íi]m|ňm|ň[úíuů]|ňou|ň[uúů]m|ň[iye]ma|ň[ií]m[iy]|
ň[ií]m'i|[nv][aá]m[ayi]|[nv][aá]m'i)\b
Replace: \1QQ\2
```

Vzhledem k tomu, že některé laické texty zapisují východomoravské původní *ú* jako „ů“, zahrnujeme i varianty „ňů“, „ňům“.

Dále potřebujeme fixovat další případy předložkových asimilací u neslabičných předložek, které mají specifické výsledky dle nářečí. Jsou to:

- Spojení předložek *s* a *k* s náslovnými jedinečnými, s náslovným *h* a *v* nebo náslovnými vokály.

  ```
  Search: \b([sk]) (j|r|l|m|n|ň|a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|h|v)
  Replace: \1§§\2
  ```

  Opět platí, že výčet jedinečných, vokálů i variant *v* musí být upraven dle dialektu.

- Spojení předložky *v* s vokály.

  ```
  Search: \b([vhx]) (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
  Replace: \1§§\2
  ```

  V tomto případě je regulární výraz lokalizován pro severovýchodočeská nářečí (1-1), kde může mít předložka *v* místy podobu *h*, případně asimilovaného *ch*.

- Spojení předložky *z* s vokály.

  ```
  Search: \b(z) (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
  Replace: \1§§\2
  ```

Ještě musíme ošetřit jeden případ, a to tvary slovesa „být“ začínající na „js-“, kde se „j“ nevyslovuje, a tvoří tak „falešnou“ jedinečnou, před kterou by při převodu mohlo docházet k chybným asimilacím. Proto ještě převedeme:

```
Search: \bJs
Replace: S
   Options: case sensitive

Search: \bjs
Replace: s
   Options: case sensitive
```

Nyní už můžeme přistoupit k provedení mezislovních a předložkových asimilací před jedinečnými, *v*, *h* a vokály, aniž bychom narušili výsledek těchto výjimek.

4.6.7 Mezislovní asimilace u jedinečných souhlásek, *v* a *h* a před vokály

4.6.7.1 Mezislovní asimilace před jedinečnými

Asimilace před jedinečnými lze rozdělit do dvou typů, z nichž každý má v nářečích jiné výsledky.

- **typ „už mám“** (znělá před jedinečnou)

  Jde o případ, kdy je slovo zakončeno na kterýkoli znělý foném a následuje slovo začínající na kteroukoli jedinečnou. Spektrum příslušných znaků je dáno konkrétním nářečím. Je třeba nezapomínat na digrafy, zvláště tam, kde výsledkem může být neznělá („medvědz něvi“ = *medvjeć ňevi*), jinak bychom jako výsledek po provedení všech asimilací měli spojení dvou neznělých (např. *ts*), které zpravidla mohou mít dvojí výslovnost: jako asibilanta i jako dvě hlásky. V případě digrafu typu „dz“ označujícího jeden foném je však výsledkem vždy pouze asibilanta.

  ```
  Search: b (j|r|l|m|n|ň)
  Replace: p@@\1

  Search: d (j|r|l|m|n|ň)
  Replace: t@@\1
  ```

  Takovéto regulární výrazy je potřeba vytvořit pro každý znak, který odpovídá kategorii znělých souhlásek v daném nářečí. Není třeba zohledňovat velké znaky, protože náhrada se děje na konci slova. Dvojice znaků „@@“ blokuje skupinu proti dalším změnám.

- **typ „kus masa“** (neznělá před jedinečnou)

  Jde o zcela obdobný případ, pouze je slovo zakončeno na neznělý foném. Opět je třeba provést náhradu pro každý neznělý znak zvlášť.

  ```
  Search: c (j|r|l|m|n|ň)
  Replace: c@@\1

  Search: č (j|r|l|m|n|ň)
  Replace: č@@\1
  ```

4.6.7.2 Mezislovní asimilace před *v*

Opět jde o dva typy.

- **typ „už vím“** (znělá před *v*)

  Je třeba dle dialektu počítat i s regionálními obměnami „v“ („w“ a „u̯“).

  ```
  Search: b (v)
  Replace: p@@\1

  Search: d (v)
  Replace: t@@\1
  ```

  Opět je třeba provést pro všechny znaky znělých souhlásek daného dialektu.

- **typ „moc veselá“** (neznělá před *v*)

```
Search: c (v)
Replace: dz@@\1

Search: č (v)
Replace: dž@@\1
```

4.6.7.3 Mezislovní asimilace u *h*

Tento typ asimilace je fakticky dosti nepravidelný, zvláště v oblasti severovýchodočeských nářečí. Jeho výsledkem může být regresivní i progresivní asimilace (*pjed hoďin, pjet choďin*), ale také k asimilaci vůbec nemusí dojít nebo dochází k oslabení až vypuštění *h*, případně k jeho náhradě za ráz (*jejich husi*, *tak o mňela taki*). Naposled zmíněné možnosti se však objevují spíše nepravidelně, proto pro převod lze počítat spíše jen s prvními dvěma.

Mezislovní asimilace u *h* může být třech typů, všechny pro kombinaci neznělá + *h*:

- **typ „pět hodin“** (exploziva před *h*)

```
Search: t h([^§ ])
Replace: [t@@x/d@@h]\1
   Options: case sensitive

Search: t H
Replace: [t@@X/d@@H]
   Options: case sensitive
```

  Je třeba vytvořit regulární výrazy pro znaky „t“, „ť“, „p“, „k“.

  Regulární výrazy zavádějí obojetnost, platí tedy pro oblasti, kde se vyskytuje dubleta. V takovémto případě je zapotřebí rozlišovat při náhradách velikost písma, protože může dojít jak k regresivní, tak k progresivní asimilaci, tudíž je třeba nahrazovat i náslovné „h“, které může být malé i velké. První regulární výraz počítá s možností, že v textu je možné také narazit na předložku „h“ (ve funkci předložky „v“), která nepodléhá progresivní asimilaci, regulární výraz brání aplikaci na tento případ i prostřednictvím znakové fixace „§§“, kterou jsme zavedli v kroku 4.6.4. U druhého regulárního výrazu je prakticky vyloučeno, že by o tento případ šlo.

- **typ „moc hezká“** (sykavka, polosykavka před *h*)

```
Search: c h([^§ ])
Replace: [c@@x/dz@@h]\1
   Options: case sensitive

Search: c H
Replace: [c@@X/dz@@H]
   Options: case sensitive
```

  Aplikuje se na všechny sibilanty a asibilanty v daném nářečí.

- **typ „jejich husy“** (*ch* před *h*)

```
Search: x h([^§ ])
Replace: [x@@x/h@@h]\1
   Options: case sensitive

Search: x H
Replace: [x@@X/h@@H]
   Options: case sensitive
```

Jde pouze o uvedené dva případy regulárních výrazů, výsledek v pozici replace je třeba upravit dle nářečí.

4.6.7.4 Mezislovní (a předložkové) asimilace před vokálem

Jde o dva typy asimilací, které se liší podle toho, zda je před vokálem znělá, nebo neznělá souhláska. Opět každá z těchto asimilací má v dialektech jiné územní rozšíření. Zahrnuje nejen běžné mezislovní asimilace, ale také některé předložkové. Konkrétně jde o slabičné předložky zakončené na konsonant před vokálem. Tyto jsme nezafixovali v části 4.6.6, proto u nich nyní náhrada bez problémů proběhne.

- **typ „pes a kočka“** (neznělá před vokálem)

```
Search: c (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: dz@@\1

Search: č (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: dž@@\1
```

Je třeba opět provést pro všechny neznělé souhlásky dialektu a v závorce regulárního výrazu vypočítat všechny normalizované vokály daného nářečí.

- **typ „vůz a koně“** (znělá před vokálem)

```
Search: b (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: b@@\1

Search: dž (a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý)
Replace: dž@@\1
```

Vše platí analogicky dle předchozího případu, jen pro znělé souhlásky.

4.6.8 Předložkové asimilace

U předložkových asimilací podáme kompletní sadu vzorů pro všechny případy, tyto vzory je však potřeba přizpůsobit nářečním podskupinám nebo konkrétním dialektům stejným způsobem, jako to bylo u regulárních výrazů výše. Vzory podávají příklady různých výsledků asimilace.

4.6.8.1 Předložkové asimilace před jedinečnými

- **typ „k matce“** (předložka *k* před jedinečnou)

```
Search: k§§(j|r|l|m|n|ň|J|R|L|M|N|Ň)
Replace: g \1
   Options: case sensitive
```

```
Search: K§§(j|r|l|m|n|ň|J|R|L|M|N|Ň)
Replace: G \1
   Options: case sensitive
```

- **typ „s matkou“** (předložka *s* před jedinečnou)

```
Search: s§§(j|r|l|m|n|ň|J|R|L|M|N|Ň)
Replace: s \1
   Options: case sensitive

Search: S§§(j|r|l|m|n|ň|J|R|L|M|N|Ň)
Replace: S \1
   Options: case sensitive
```

4.6.8.2 Předložkové asimilace před *v*

- **typ „k vodě“** (předložka *k* před *v*)

```
Search: k§§([vV])
Replace: k \1
   Options: case sensitive

Search: k§§([vV])
Replace: K \1
   Options: case sensitive
```

- **typ „s vodou“** (předložka *s* před *v*)

```
Search: s§§([vV])
Replace: z \1
   Options: case sensitive

Search: S§§([vV])
Replace: Z \1
   Options: case sensitive
```

4.6.8.3 Předložkové asimilace u *h*

- **typ „k holiči“** (předložka *k* před *h*)

```
Search: k§§h
Replace: [k x/g h]
   Options: case sensitive

Search: K§§h
Replace: [K x/G h]
   Options: case sensitive

Search: k§§H
Replace: [k X/g H]
   Options: case sensitive
```

```
Search: K§§H
Replace: [K X/G H]
   Options: case sensitive
```

- **typ „s holí“** (předložka *s* před *h*)

```
Search: s§§h
Replace: [s x/z h]
   Options: case sensitive

Search: S§§h
Replace: [S x/Z h]
   Options: case sensitive

Search: s§§H
Replace: [s X/z H]
   Options: case sensitive

Search: S§§H
Replace: [S X/Z H]
   Options: case sensitive
```

4.6.8.4 Předložkové asimilace před vokálem

- **typ „k autu“** (předložka *k* před vokálem)

```
Search: k§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: k \1
   Options: case sensitive

Search: K§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: K \1
   Options: case sensitive
```

- **typ „s autem“** (předložka *s* před vokálem)

```
Search: s§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: z \1
   Options: case sensitive

Search: S§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: Z \1
   Options: case sensitive
```

- **typ „v autě“** (předložka *v* před vokálem)

```
Search: v§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: [v/f] \1
   Options: case sensitive
```

```
Search: V§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: [V/F] \1
   Options: case sensitive
```

- **typ „z auta“** (předložka *z* před vokálem)

```
Search: z§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: [z/s] \1
   Options: case sensitive

Search: Z§§(a|e|i|o|u|y|á|é|ě|í|ó|ú|ů|ý|A|E|I|O|U|Y|Á|É|Ě|Í|Ó|Ú|Ů|Ý)
Replace: [Z/S] \1
   Options: case sensitive
```

4.6.8.5 Odstranění zbytků po fixaci předložkových asimilací

Po provedení všech těchto kroků můžeme odstranit fixace předložkových asimilací, neboť už další mezislovní a předložkové asimilace před jedinečnými, před *v*, *h* a před vokály nebudeme provádět:

```
Search: (§§|QQ|@@)
Replace: " "
   Options: case sensitive
```

(Uvozovky jsou použity k manifestaci běžné mezery, nejsou součástí regulárního výrazu.)

4.6.9 Příprava na asimilace znělosti před znělými a neznělými

Asimilace znělosti před znělými a neznělými probíhají celkem velmi pravidelně ve všech typech asimilací: ve vnitroslovních, na morfologickém švu, v předložkových i mezislovních. Výjimkou jsou pouze asimilace před *v* a *h*, které jsou obě počítány mezi znělé souhlásky. U těchto hlásek už jsme vyřešili jejich asimilace mezislovní a předložkové; pokud jde o asimilace vnitroslovní a na morfologickém švu, chovají se opět jinak než hlásky znělé. Na morfologickém švu se hláska *v* chová po znělých stejně jako jedinečná souhláska, po neznělých ve shodě s asimilacemi vnitroslovními; hláska *h* se může dle nářečí po neznělé chovat buď jako znělá a způsobit regresivní asimilaci (*zhoďit*, *taghle*), nebo může sama podléhat progresivní asimilaci (*schoďit*, *takchle*). Vnitroslovní asimilace u *v* mohou dle nářečí dopadnout buď tak, že dojde k progresivní asimilaci *v* (*tfaroch*, *chfálit*), nebo k žádné asimilaci znělosti nedochází (*tvaroh*, *chválit*); *h* do vnitroslovních asimilací vstupuje jen vzácně ve slovech cizího původu („kuthan“, „lejthar“), obvykle dopadají v souladu s asimilacemi *h* na morfologickém švu. Jinak platí jednoduché pravidlo, že u znělých a neznělých souhlásek dochází k regresivním asimilacím znělosti vnitroslovně i přes jakoukoli hranici mimo pauzu.

V rámci příprav na asimilace znělosti před znělými a neznělými musíme dořešit všechny jevy, které ovlivní znělost těch hlásek, které mohou způsobovat asimilace znělosti. A dále i zapracovat na formálních prvcích textu, které by zabránily správné detekci asimilace.

4.6.9.1 Znělé souhlásky před pauzou

Na většině území dochází u znělých souhlásek před pauzou ke změně na neznělé. Neplatí to však pro území celé. Na značně rozsáhlém areálu severovýchodočeských nářečí s přesahem do nářečí českomoravských a středomoravských se v těchto případech zachovávají znělé konsonanty. Jev je možné zachytit i v současnosti, ale pouze nepravidelně a ani před šedesáti lety, kdy se konal výzkum pro *Český jazykový atlas*, se nejednalo o pravidelný jev (srov. Balhar a kol., 2005, s. 404). Pokud se však v některých lokalitách výrazněji zachovává, je třeba s ním počítat.

Tento jev je potřeba nyní zpracovat kvůli případům dvou nebo tří znělých souhlásek na konci slova před pauzou. Od znělosti poslední souhlásky se bude na základě asimilací odvozovat znělost celé souhláskové skupiny, proto je třeba tuto změnu zavést ještě před těmito asimilacemi („drozd" = *drost*, „hvíždž" = *hvíšč*). Pro každou znělou souhlásku v daném nářečí tedy vytvoříme následující regulární výraz:

```
Search: b(( )?[\.…,!\?;:\---])
Replace: [b/p]\1

Search: d(( )?[\.…,!\?;:\---])
Replace: [d/t]\1
```

Za indikátor pauzy bereme běžnou interpunkci: tečku, trojtečku, čárku, vykřičník apod., ale také pomlčku. Regulární výrazy nemusíme tvořit pro měkké retnice, které se na konci slova nevyskytují.

4.6.9.2 Neznělá před *h* v rámci slova

Ať už jde o asimilaci vnitroslovní nebo přes morfologický šev, regresivní asimilace před *h* může ovlivnit výsledky dalších asimilací. Proto je nutné tyto případy převést ještě před asimilacemi před znělými a neznělými. Při té příležitosti je vhodné převést všechny asimilace v rámci jednoho slova související s *h*. Nejčastějším typem je asimilace *s* a *h* na morfologickém švu. Mimoto dochází ke styku neznělé a *h* uvnitř slova jen velmi vzácně, ve slovech cizího původu. U těchto případů je třeba dávat pozor na spojení *th*, které se může objevit jednak ve jménech německého původu („Werther", „Thüringer", „Barth"), jednak ve starším pravopisu slov řeckého původu („thema", „methanol", „atheista"). Můžeme-li obojí vyloučit (to znamená, že byl proveden převod německých jmen podle oddílu 4.6.3.4 a slov cizího původu s „th" podle oddílu 4.6.3.5 nebo je přepis převáděných textů novějšího charakteru), můžeme *t* do následujících regulárních výrazů bez problémů zahrnout. Pro případ progresivní asimilace můžeme použít tento regulární výraz:

```
Search: (p|t|ť|x|s|š|c|č|f|k)h
Replace: \1x
```

Je třeba vložit do regulárního výrazu sadu všech neznělých souhlásek daného nářečí.

Pro případ regresivní asimilace musíme použít sadu regulárních výrazů pro každý konsonant a minimálně pro spojení *s* + *h* i variantu pro velké a malé písmeno:

```
Search: sh
Replace: zh
   Options: case sensitive

Search: Sh
Replace: Zh
   Options: case sensitive
```

4.6.9.3 Umožnění asimilace přes závorku obojetností

Současně je potřeba mít na paměti i to, že dosavadními převody mezislovních asimilací znělosti jsme pravděpodobně zanesli do textů řadu obojetností. Přitom jsme u těchto asimilací vždy změnili jen první hlásku, kterou by asimilace zasáhla. Dál by převod pokračoval dle zásad asimilací před znělými a neznělými, ale pokud je teď v textu namísto párové souhlásky obojetnost, bude mít další asimilace dosti pravděpodobně

minimálně dvě verze. Proto je potřeba přesunout do závorek obojetností celé skupiny párových souhlásek, které budou asimilacím podléhat. Uděláme to pomocí následující sady regulárních výrazů:

```
Search: ((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)( )?)\[((b|d|ď|h|z|ž
|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p
|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*\])
Replace: [\1\4/\1\6/\1\8

Search: (b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)\[((b|d|ď|h|z|ž|g|v|c|č
|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)
[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*\])
Replace: [\1\2/\1\4/\1\6

Search: (b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)\[((b|d|ď|h|z|ž|g|v|c|č
|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)
[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*\])
Replace: [\1\2/\1\4/\1\6

Search: ((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)( )?)\[((b|d|ď|h|z|ž
|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť
|p|x|k|f)[^/\[\]]*\])((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f))\
[((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*)/
((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)[^/\[\]]*\])
Replace: [\1\4/\1\6

Search: (b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)\[((b|d|ď|h|z|ž|g|v|c|č
|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)
[^/\[\]]*\])
Replace: [\1\2/\1\4

Search: ((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f))\[((b|d|ď|h|z|ž|g|v|c|č
|s|š|t|ť|p|x|k|f)[^/\[\]]*)/((b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f)
[^/\[\]]*\])
Replace: [\1\2/\1\4
```

Regulární výrazy jsou sestaveny pro případy, kdy jsou v rámci formálního zápisu obojetnosti uvedeny tři nebo dvě varianty („jeji[x h/x x/h h]usy“, „jeji[x x/h h]usy“). Jestliže může nastat případ, že v závorce obojetnosti jsou více než tři varianty, je potřeba přidat regulární výrazy pro tyto počty variant. Je do nich vždy potřeba dosadit kompletní sadu znělých a neznělých souhlásek pro dané nářečí.

4.6.10 Asimilace znělosti před znělými a neznělými souhláskami

Nyní už mohou proběhnout vlastní asimilace před párovými souhláskami. Bude k tomu zapotřebí několik sad regulárních výrazů, neboť párové souhlásky vytvářejí skupiny, jejichž znělost ovlivňuje znělost poslední párové souhlásky ve skupině. Ke změně oproti folklornímu přepisu dochází vždy na místě, kde se stýkají znělá a neznělá souhláska v jakémkoli pořadí. Tehdy druhá z nich vždy ovlivní první. Proto je nejjednodušším postupem vyhledávat pomocí regulárních výrazů dvojice znělá-neznělá a neznělá-znělá a změnit znělost u první z páru. Takovýto postup však nezmění naráz celé skupiny párových souhlásek, a proto je potřeba více iterací. Je otázka, kolik je jich potřeba. Vzhledem k tomu, že asimilace může probíhat i přes

hranici slov, je třeba sečíst nejvyšší možný počet párových hlásek na začátku (jakéhokoli) slova a na konci (jakéhokoli) slova. Nejdelší souvislá skupina párových souhlásek mívá v českých nářečích čtyři členy. Ty se mohou vyskytovat minimálně na začátku slova, avšak není nám znám případ, kdy by bylo potřeba zde řešit asimilace („bzdzina“, „gzdzic se“), nebo je poslední hláskou *v*, které asimilace nezpůsobuje, ale pouze podstupuje („vzkvasit se“, „zčtvernásobit“). Na konci slova známe pouze případy skupin tří párových souhlásek („hynkšt“, „zibst“), ty však mohou být teoreticky doplněny koncovým „-s“ namísto „jsi“ („ten smradlavý chebzds mohł vysekat“) a nezáleží u nich na zrůznění znělých a neznělých, protože mohou být ovlivněny následujícím slovem. Fakticky se nám u žádného textu zatím nestalo, aby čtyři iterace regulárních výrazů nestačily, ale vzhledem k tomu, že nemáme kompletní sadu všech tvarů v nářečí, je lepší preventivně udělat osm iterací nebo v iteracích pokračovat, dokud něco nahrazují. První iterace přitom bude poněkud odlišná od iterací následujících a celý úsek regulárních výrazů věnovaných asimilacím před znělými a neznělými musíme opět zakončit řešením zápisu obojetností.

4.6.10.1 První sada regulárních výrazů

První sadu je potřeba udělat poněkud odlišnou od sad ostatních. Je to proto, že regulární výrazy postupují při vyhledávání odpředu dozadu, tedy zachycují první dvojici s odlišnou zněloností zleva, další vyhledání se následně odehraje nejdříve na třetí a čtvrté pozici, poté na páté a šesté atd. Takto může dojít k tomu, že klíčová první dvojice zprava je o jednu pozici minuta (zablokuje ji dvojice na druhé a třetí pozici zprava), a první iterace tak v dané skupině párových souhlásek proběhne naprázdno. V každé skupině párových souhlásek se jejich znělost odvozuje (z hlediska převodu) od první dvojice s rozdílnou zněloností zprava. Abychom tedy měli jistotu, že první sada regulárních výrazů u této pozice začíná, bude tato sada o něco složitější.

Je také nutné nezapomenout na to, že ve všech dialektech máme znělé a neznělé *ř*, které sice uvnitř slova přijímá neznělost od jakékoli sousední neznělé souhlásky (tj. asimiluje progresivně i regresivně), ale stojí-li *ř* na začátku slova, samo způsobuje mezislovní asimilace. Pokud po *ř* na začátku slova následuje neznělá („řknút“, „řpytit se“), asimiluje *ř* mezislovně i předchozí konsonant na neznělý, pokud následuje po náslovném *ř* cokoli jiného, bude *ř* znělé a předchozí konsonant bude také znělý. Takováto skupina náslovného *ř* + konsonant/vokál je vždy v dané nepřerušené skupině párových konsonantů na první pozici, proto musí být v první sadě regulárních výrazů, a v dalších už ne.

Pro znělé konsonanty před neznělými budou regulární výrazy vypadat následovně:

```
Search: b(( )?(c|č|s|š|t|ť|p|x|k|f|C|Č|S|Š|T|Ť|P|X|K|F)(?!b|d|ď|h|z|ž|g)|
[řŘ](c|č|s|š|t|ť|p|x|k|f))
Replace: p\1
   Options: case sensitive

Search: B(( )?(c|č|s|š|t|ť|p|x|k|f|C|Č|S|Š|T|Ť|P|X|K|F)(?!b|d|ď|h|z|ž|g)|
[řŘ](c|č|s|š|t|ť|p|x|k|f))
Replace: P\1
   Options: case sensitive
```

Vždy je třeba vytvořit zvláštní výraz pro malý a velký znak konsonantu a použít normalizovanou sadu znělých i neznělých konsonantů z daného nářečí nebo nářeční podskupiny.

Pro neznělé konsonanty platí totéž a jejich regulární výrazy budou vypadat takto:

```
Search: c(( )?(b|d|ď|h|z|ž|g|B|D|Ď|H|Z|Ž|G)(?!c|č|s|š|t|ť|p|x|k|f)| [řŘ]
(?!c|č|s|š|t|ť|p|x|k|f))
Replace: dz\1
   Options: case sensitive

Search: C(( )?(b|d|ď|h|z|ž|g|B|D|Ď|H|Z|Ž|G)(?!c|č|s|š|t|ť|p|x|k|f)| [řŘ]
(?!c|č|s|š|t|ť|p|x|k|f))
Replace: Dz\1
   Options: case sensitive
```

4.6.10.2 Druhá až osmá sada regulárních výrazů

Následně už stačí jednoduché nahrazování dvojic konsonantů s rozdílnou zněloslí, které v sedmi iteracích bezpečně provede nutný převod:

```
Search: b(( )?(c|č|s|š|t|ť|p|x|k|f|C|Č|S|Š|T|Ť|P|X|K|F))
Replace: p\1
   Options: case sensitive

Search: B(( )?(c|č|s|š|t|ť|p|x|k|f|C|Č|S|Š|T|Ť|P|X|K|F))
Replace: P\1
   Options: case sensitive

Search: c(( )?(b|d|ď|h|z|ž|g|B|D|Ď|H|Z|Ž|G))
Replace: dz\1
   Options: case sensitive

Search: C(( )?(b|d|ď|h|z|ž|g|B|D|Ď|H|Z|Ž|G))
Replace: Dz\1
   Options: case sensitive
```

4.6.10.3 Oprava nesoustavností v obojetnostech

Obojetnosti mají zaznamenávat různé varianty, k nimž může docházet, přičemž mají zachycovat přesně variantní skupinu znaků. Po provedení všech asimilačních změn uvedených výše však dojde k tomu, že některé obojetnosti dopadnou nesoustavně, někdy do závorky pojímají i text, který se napříč variantami nemění, někdy jsou naopak rozděleny do dvou sousedních zápisů obojetností, které naznačují větší množství kombinací, než kolik jich může nastat. Tyto nesoustavnosti je třeba opravit.

- **typ „povla[k v/k f] auťe"**[43]

  V tomto typu má korektně být v závorce obojetnosti pouze „povlak [v/f] auťe", nikoli „povla[k v/k f] auťe". Opravu pro dvě a tři varianty v závorce (a jakýkoli počet znaků v nich) provedeme takto:

  ```
  Search: \[([^/[\]]+)([^/[\]]+)/\1([^/[\]]+)\]
  Replace: \1[\2/\3]
  ```

[43] Jde o podobu, jak by se daný text zobrazoval už po provedení asimilací a dalších změn. Nejde tedy ani o folklorní, ani o dialektologický přepis, ale přechod mezi nimi.

```
Search: \[([^/[\]]+)([^/[\]]+)/\1([^/[\]]+)/\1([^/[\]]+)\]
Replace: \1[\2/\3/\4]
```

- **typy „br[zd/st] [v/f] auťe“, „pí[st/zd] [v/f] auťe“**

  V tomto typu má korektně být jedna závorka „br[zd v/st f] auťe“, „pí[st f/zd v] auťe“, nikoli dvě závorky „br[zd/st] [v/f] auťe“, „pí[st/zd] [v/f] auťe“. Opravu pro jakýkoli počet asimilovaných znaků v první závorce provedou následující regulární výrazy:

```
Search: \[([bddďhzžgv]+)/([cčsštťpxkf]+)\] \[v/f\]
Replace: [\1 v/\2 v/\2 f]

Search: \[([cčsštťpxkf]+)/([bddďhzžgv]+)\] \[v/f\]
Replace: [\2 v/\1 v/\1 f]
```

4.6.11 Nářeční změny

V následující fázi doplníme do textů nářeční jevy, které nejsou obsažené v normalizovaném folklorním zápisu. Všechny tyto jevy jsou nářečně podmíněny, a proto se neaplikují na celém území, ale pouze na základě konkrétního nářečí nebo nářečního typu nebo podskupiny. Obecnou zásadou tu je neměnit nic, co bylo zapsáno, tedy nedoplňovat nářeční jevy, které nejsou v souladu s folklorním zápisem, ale pouze takové, které jsou implicitní, tedy daným zápisem nevyjádřitelné. Nebudeme tedy měnit např. kvantitu vokálů, i když očekáváme kvantitu jinou, nebudeme měnit kvalitu hlásek, nebudeme doplňovat jevy, o kterých víme, že by v daném nářečí měly být, ale v zápise ani v autorově metatextu nejsou nijak naznačeny. Další důležitou zásadou je měnit pouze relativně pravidelné jevy. Pokud o výskytu jevu spolehlivě víme, ale současně víme, že byl na povážlivém ústupu (kolem 25 % případů a méně), do textů ho vůbec nezavádíme, protože tím data více poškozujeme, než jim pomáháme. V případě, že se jev vyskytuje kolem 50 %, můžeme zavést obojetnost. Tím si ovšem data také mírně poškozujeme, protože komplikujeme jejich následné zpracování. Proto je i zde na zvážení, zda u případů s četností znatelně pod 50 % vůbec jev evidovat.

4.6.11.1 Měkké retnice, jotace a *ň* po retnicích

Dosud jsme neřešili převod skupin typu „bě“, „pě“, „vě“, „fě“, „mě“ a v textu nám zůstávaly. Tyto skupiny tedy převedeme nyní na základě konkrétních nářečí. V některých nářečích k nim totiž může přistoupit ještě skupina „u̯ě“ nebo „wě“, v některých nářečích je výsledkem vždy retnice + *je*, v jiných se u *m* vyslovuje *mňe*, v dalších se tato skupina vyslovuje jako *b’e, p’e, v’e, f’e, m’e*, v jiných jsou různé kombinace dublet, které můžeme zavést jako obojetnosti.

Nejvhodnějším postupem je nahradit nejdřív „ě“ u všech retnic mimo *m* podle způsobu, jakým se v nářečí měkkost u retnic reflektuje.

```
Search: ([bpvf])ě
Replace: \1[’/j]e
```

Hlásku *m* potom vyřešíme zvlášť. V textu se obvykle budou vyskytovat zápisy „mě“ i „mňe“ (ten druhý je již výsledkem převodu u „dě“, „tě“, „ně“). Oba opět nahradíme podle příslušného stavu v nářečí.

```
Search: (m)(ě|ňe)
Replace: \1ňe
```

V nářečích, v nichž se objevují měkké retnice, je pak většinou potřeba pokračovat, neboť se relativně pravidelně vyskytují nejen před původním *ě*, ale i před měkkým *i*.

```
Search: ([bpvfm])i
Replace: \1'i
```

Takovouto změnu pochopitelně nezavádíme tam, kde měkké retnice silně ustupují. Případy vydělené jotace před *i* (*bjič*, *pjivo*) není třeba nijak ošetřovat, protože tento zápis se, v případě přítomnosti tohoto jevu, objevuje už ve folklorní transkripci.

4.6.11.2 Geminace souhlásek

Geminace je jev, který se v určité míře odráží ve folklorním přepisu, ale ve většině nářečí se geminované hlásky prakticky nevyslovují. V některých nářečích nalezneme geminaci naopak ve větším rozsahu, než bychom ji našli ve spisovném jazyce, jehož způsob zápisu folklorní přepis do značné míry přebírá, a tudíž některé geminace v těchto případech neeviduje (*žłttý*, *necco*). Jde ovšem o jev vesměs mizející, proto geminaci souhlásek nikdy do textu sami nezavádíme, pouze ji ponecháváme nebo nahrazujeme za prostou souhlásku.

Nejjednodušším případem je geminace *nn*, resp. *nň/ňň*, který se objevuje v řadě slov (*Anna*, *panna*), zejména pak v adjektivech (*kamennej*, *deňňí*). Zde můžeme dle nářečí snadno nahrazovat, případně ponechávat geminaci, pouze musíme rozhodnout, zda budeme převádět slova typu „denňe" (v této fázi převodu jen takto), „holenňí" (bylo-li ve folklorním textu v souladu s nářečím zapsáno „holenní") na „ňň", nebo je takto necháme. Považujeme tento převod za doporučeníhodný, neboť při geminaci se obvykle měkkost přenáší více či méně i na první člen dvojice.

```
Search: nn
Replace: [nn/n]

Search: nň
Replace: [ňň/ň]
```

Veškeré další případy jsou už poněkud složitější. Často se např. setkáme s geminací u sykavek. „Sušší", „vyšší", „jednodušší", ale také „užší", které se nám nyní po asimilacích bude manifestovat jako „ušší", „tužší" – „tušší", „těžší" – „ťešší" atd. Rovněž např. předpona „roz-" + „s-" nebo „-z" + přípona „-ský" bude mít nyní podobu „ross-", „-sský". Případy jako „snazší" – „snasší" nebo předpona „roz-" + „š-" – „rosš" se často také ve výsledku mohou vyslovovat jako šš nebo š (*rošščípať*, *roščípat*). I ve znělé kombinaci předpony „roz-" + „z/ž" dochází k podobnému efektu. Častěji se ve folklorních textech objevuje ještě kombinace „kk" ve slovech typu „měkkej", „naměkko" nebo „jakkerak" ap. Méně už další geminace.

Pokud se v nářečí geminace jako běžný jev nevyskytuje, je nejlepším postupem převést všechny tyto případy, včetně kombinace sykavek, na jednoduchou souhlásku, tedy:

```
Search: (b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f|j|r|l|m|n|ň)\1
Replace: \1

Search: sš
Replace: š

Search: zž
Replace: ž
```

Po předponě „roz-“ + sykavka často bývá dubletní výslovnost, je tedy možné vyřešit situaci i takto:

```
Search: (ro)(z|s)([zž]|[sš])
Replace: \1[\2§\3/\3]

Search: (b|d|ď|h|z|ž|g|v|c|č|s|š|t|ť|p|x|k|f|j|r|l|m|n|ň)\1
Replace: \1

Search: §
Replace:
```

(Výraz v pozici replace zůstává prázdný.)

Pokud se v daném nářečí geminátу vyskytují, je potřeba pracovat v jednotlivých případech s dubletami dle uvážení nářeční situace.

4.6.11.3 Dvojí „i“, „y“

Přestože český pravopis, potažmo folklorní přepis „i“ a „y“ rozlišují, fonologicky rozlišovány jsou v současné době jen v nářečích na východě území (3-2, 4-1, 4-2). V minulosti bychom našli širší výskyt (srov. Stupňánek a Vondráková, 2022c), takže vždy je třeba zvážit dialektologické okolnosti, ale systematické ponechání rozdílu mezi „i“ a „y“ bude vždy otázkou menšiny dialektů. U většiny musíme tento rozdíl v této fázi odstranit. Z předchozích kroků pro to máme připravenou půdu, takže nám stačí jednoduché regulární výrazy:

```
Search: y
Replace: i
   Options: case sensitive

Search: Y
Replace: I
   Options: case sensitive

Search: ý
Replace: í
```

V případě dialektů severní východomoravské podskupiny (3-2), kde je tvrdost *i/y* kombinační vlastností jednoho fonému *i/y*, nestačí ponechat rozložení „i“ a „y“, jak je uvádí folklorní přepis, ale je třeba u cizích slov nahradit po tvrdé souhlásce „i“ za „y“. Jde o případy jako *Ameryka*, *kylo*, *energyje* ap. U hlásek *d*, *t*, *n* bychom měli mít cizí slova ošetřená, stačí nám tedy *h*, *ch*, *k*, *g*, *r*:

```
Search: ([hxkgr])i
Replace: \1y

Search: ([hxkgr])í
Replace: \1ý
```

4.6.11.4 Skupiny *čy*, *žy*, *šy*, *ry*, *cy*, *zy*, *sy*

Stejným případem jako posledně zmíněný je i výsledek skupin „či“, „ži“, „ši“, skupiny „ři“ a skupin „ci“, „zi“, „si“ s krátkým i dlouhým *i*-ovým vokálem. Tyto skupiny se v mnohých nářečích vyslovují s tvrdým *y*, nejde

však o *y*, které by se vždy fonologizovalo jako samostatná hláska, ale často o alofon fonému *i*, který se objevuje v kombinaci právě s těmito hláskami. Děje se tak na části území severní východomoravské nářeční podskupiny (3-2), na východním Telečsku a na Jemnicku, které jsou součástí českomoravské nářeční podskupiny (1-4). Výsledky pro kombinace „č/ž/š“, „ř“ a „c/z/s“ s „i“ a „í“ jsou pokaždé územně jiné, ale zvláště v minulosti jsou dosti pravidelné, proto je možné provádět soustavný převod. Především v oblasti českomoravských dialektů se může skupina „č/ž/š“ v závislosti na konkrétní lokalitě ještě rozpadat na jednotlivé typy podle těchto hlásek, ale praktické je spíš zavést zde obojetnosti.

V oblasti Slezska k těmto jevům také dochází, ale hláska *y* se tam fonologizovala, a proto ji ve folklorních přepisech najdeme distribuovanou povětšinou již adekvátně.

Regulární výrazy lze použít podle tohoto vzoru:

```
Search: ([čžš])i
Replace: \1y

Search: ([čžš])í
Replace: \1ý

Search: (ř)i
Replace: \1y

Search: (ř)í
Replace: \1ý

Search: ([czs])i
Replace: \1y

Search: ([czs])í
Replace: \1ý
```

4.6.11.5 Progresivní asimilace *v* ve skupinách *kf*, *tf*, *sf*, *šf*, *chf*

Ve většině nářečí k této progresivní asimilaci *v* nedochází, tudíž není třeba oproti folklornímu přepisu cokoli měnit. Souhláska *v* ponechává znělost či neznělost párové souhlásky předchozí. V mnohých dialektech však docházelo u *v* k progresivní asimilaci po neznělé. Tento jev už v minulém století dosti patrně mizel a v současnosti zůstává relativně důsledný jen v některých oblastech slezské nářeční skupiny.

Pro každou z uvedených skupin je územní rozšíření, případně rozšíření dublet jiné, je proto nutné přizpůsobit skladbu regulárních výrazů příslušnému nářečí.

```
Search: (k)v
Replace: \1[v/f]

Search: (t)v
Replace: \1[v/f]

Search: (s)v
Replace: \1[v/f]

Search: (š)v
Replace: \1[v/f]
```

```
Search: (x)v
Replace: \1[v/f]
```

4.6.11.6 Typ „se sestrou"

Různě dopadá v dialektech též kombinace předložky „se" a následující sykavky. Výsledkem může být podoba předložky „ze". Zavedení této změny mírně komplikuje homonymie této předložky se zvratným zájmenem „se". K tomu nedochází v oblastech, kde toto zájmeno má podobu „sa", přesto musíme tuto skutečnost na většině území (a nejlépe na území celém) ošetřit. Můžeme toho dosáhnout kombinací „se" s náslovnou sykavkou a zakončením instrumentálové koncovky v daném nářečí nebo skupině nářečí (je třeba použít instrumentálové koncovky singulárové i plurálové od všech skloňovatelných slovních druhů). Takový způsob vyhledávání se téměř nemýlí.

Pro celou českou nářeční skupinu vypadá regulární výraz takto:

```
Search: \bs(e [zžsšZŽSŠ][^ ,\.…:\?!;][^ ,\.…:\?!;]+?
(m|ma|mi|ou|í|uu|ú)\b)
Replace: [s/z]\1

Search: S(e [zžsšZŽSŠ][^ ,\.…:\?!;][^ ,\.…:\?!;]+?(m|ma|mi|ou|í|uu|ú)\b)
Replace: [S/Z]\1
```

Pro slezskou nářeční skupinu pak takto:

```
Search: \bs(e [zžsšZŽSŠ][^ ,\.…:\?!;][^ ,\.…:\?!;]+?(m|ma|mi|m'i|my|u)\b)
Replace: z\1

Search: S(e [zžsšZŽSŠ][^ ,\.…:\?!;][^ ,\.…:\?!;]+?(m|ma|mi|m'i|my|u)\b)
Replace: Z\1
```

Je třeba přizpůsobit přehled instrumentálových zakončení vždy dle daného dialektu nebo skupiny dialektů.

4.6.11.7 Měkkosti u sykavek a polosykavek

Měkkosti u sykavek (*s', z', ś, ź*) a polosykavek (*c', dz', ć, dź*) nalézáme v dialektech slezské nářeční skupiny (k jejich územnímu rozsahu a vytrácení viz Stupňánek a Vondráková, 2022b). Ve folklorním přepise nebývají s měkkostí zapisovány, v dialektologické transkripci ano. Objevují se důsledně před „i", velmi často před „e" a na konci slova, ale mohou se objevovat i před „a", „e", „o", dokonce ve slezskopolské nářeční podskupině i před „y" vzniklým z *e*. Ocitáme se zde tedy ve velmi obtížné situaci, neboť nemáme vodítko, jak rozpoznat, kde je měkká (polo)sykavka, nemáme-li postupovat po jednotlivých slovech a tvarech. Je tudíž nevyhnutelné zavést pro tyto případy obojetnosti.

Pro dialekty s palatalizovaným *s', z', c', dz'* by regulární výrazy vypadaly takto:

```
Search: (s|z|c)(i)
Replace: \1'\2

Search: (s|z|c)(a|e|o|u)
Replace: \1[/']\2
```

Pro dialekty s polským palatálním ś, ź, ć, dź pak takto pro každou hlásku z trojice *s*, *z*, *c*, při zohlednění malých a velkých písmen:

```
Search: s(i)
Replace: ś\1
   Options: case sensitive

Search: S(i)
Replace: Ś\1
   Options: case sensitive

Search: s(a|e|o|u|y)
Replace: ś\1
   Options: case sensitive

Search: S(a|e|o|u|y)
Replace: Ś\1
   Options: case sensitive
```

U dialektů slezskopolské nářeční podskupiny (4-2), kde zvratné „se" je vždy *śe*, je možné převod poněkud zpřesnit předřazením následujícího regulárního výrazu týkajícího se tohoto frekventovaného slova. Předpokládáme, že předložka „se" byla už v rámci typu „se sestrou" (4.6.11.6) nahrazena za „ze".

```
Search: \bse\b
Replace: śe
   Options: case sensitive
```

Śe je příklonka, takže se v nářečních textech nevyskytuje na začátku věty.

Nepříjemná skutečnost nutnosti zavádění tak frekventovaných obojetností je přece jen do značné míry kompenzována faktem, že obecně v nářečích slezské nářeční skupiny je velmi málo textů zapsaných folklorním přepisem, a naopak relativně větší množství textů zapsaných přepisem dialektologickým. Strojové učení tedy může tuto nedokonalost poněkud vyvážit.

4.6.11.8 Měkkosti u velár *k*, *g*

V nářečích slezskopolských (4-2) pak pravidelně doplňujeme do dialektologického přepisu měkkosti u velár *k*, *g*:

```
Search: (k|g)i
Replace: \1'i
```

4.6.12 Závěrečné změny

Na závěr musíme ošetřit už jen několik drobností, které je nejvýhodnější zavést až po všech hotových změnách.

4.6.12.1 Výslovnost skupiny *t* + neznělá sykavka, polosykavka

Ve většině případů existuje v takovýchto případech dubletní výslovnost s převahou výslovnosti jedné hlásky nad dvěma hláskami („větší" = *vječí* vs. *vjetší*). Je třeba zvážit, zda v těchto případech zavádět obojetnosti,

nebo menšinovou výslovnost zanedbat. Tyto změny zavádíme už po asimilacích, proto i kombinace *d* + neznělá (polo)sykavka už má podobu *t* + neznělá (polo)sykavka („potšitej", „tcera", nikoli „podšitej", „dcera").

- **typ „větší"**

```
Search: tš
Replace: [tš/č]
```

- **typ „předseda"**

```
Search: ts
Replace: [ts/c]
```

- **slovo „dcera"**

```
Search: \btc
Replace: c
   Options: case sensitive

Search: \bTc
Replace: C
   Options: case sensitive
```

- **typ „otcův"**

```
Search: tc
Replace: [tc/c]
```

- **typ „matčin"**

```
Search: tč
Replace: [tč/č]
```

4.6.12.2 Hiátové *j*

Hiátové *j* doplňujeme všude, kde sousedí krátké nebo dlouhé *i* s dalším vokálem. Při tvorbě regulárního výrazu je nejbezpečnější využít celou sadu vokálů daného nářečí, i když většinou zvláštní nářeční znaky do těchto souvislostí nevstupují. Přesto mohou ojediněle nastat případy typu „familiǫ", „s komedią ", i když tyto tvary obvykle už bývají zapisovány s hiátovým *j*. Také musíme zohlednit, že nářeční znak „ụ" je složen ze dvou částí, jeho druhou, kombinační část tedy musíme v regulárním výrazu vyloučit, aby se nám hiátové *j* chybně neobjevovalo ve slovech jako *dïụní* nebo *žiụ*.

```
Search: ([iíyý])((a|e|i|o|u|y|á|é|í|ó|ú|ů|ý)[^̣])
Replace: \1j\2
```

4.6.12.3 Kroužkované „ů"

Nyní nahradíme také kroužkované „ů" za *ú*:

```
Search: ů
Replace: ú
```

4.6.12.4 Opětné zavedení digrafu „ch"

Po všech změnách můžeme nyní vrátit změnu „ch" za „x":

```
Search: x
Replace: ch
   Options: case sensitive

Search: X
Replace: Ch
   Options: case sensitive
```

Tím je převod z normalizovaného folklorního přepisu na normalizovaný dialektologický přepis hotov.

4.7 Shrnutí

Cílem této kapitoly bylo demonstrovat, jakým způsobem můžeme zpracovávat surová textová nářeční data tak, abychom je zformovali do jednotné, konzistentní a vědecky zpracovatelné podoby a abychom z nich vytvořili trénovací data pro účely strojového učení. Jednotlivé postupy zde popsané ve svých dílčích cílech a výsledcích jsou ověřenými postupy pro jakékoli odborné zpracování textových nářečních dat. Ať už jde o výběr nářečních textů, jejich digitalizaci, čištění a formální sjednocení, normalizaci folklorního a dialektologického přepisu nebo převod jednoho typu přepisu na druhý, vždy jde o problémy, které řešíme nejen při přípravě trénovacích dat pro strojové učení, ale i při zpracování většího množství textových dat v nářečí pro jakékoli jiné myslitelné účely. Tato kapitola tedy ukazuje specifika hromadných úprav nářečního textu tak, aby sloužil dalšímu počítačovému zpracování nebo vědeckému využití. Při tom prakticky vždy musíme respektovat vyvinuté postupy, které jsme představili:

- Vybrat ty nejautentičtější, ale současně materiálově nejvydatnější nářeční texty tak, aby jejich zpracování bylo časově únosné (část 4.1).
- Tyto texty digitalizovat při optimálním poměru nákladů na digitalizaci a její přesnosti (část 4.2).
- Vyčistit texty od všech nenářečních součástí a sjednotit je formálně tak, aby interpunkce a úprava byly napříč texty homogenní a předvídatelné (část 4.3).
- Sjednotit pravidla folklorního přepisu a uvést v soulad lidové i etnografické zápisy jednotlivých nářečí (část 4.4).
- Normalizovat také dialektologický přepis, ujednotit a zjednodušit partikulární označování hlásek v rozmanitých dialektologických pracích a vytvořit z různých přepisů jednotnou, konzistentní transkripci (část 4.5).
- Najít způsob, jak převést jeden typ zápisu na druhý (v našem případě folklorní na dialektologický), a tím sjednotit prakticky všechny kvalitně zapsané texty, které jsme schopni digitalizovat (část 4.6).

Ve chvíli, kdy dovedeme automaticky nebo poloautomaticky projít celý tento postup a zpracovávat tímto způsobem nářeční texty ve velkém množství, máme díky tomu přístup k široké paletě počítačových nástrojů založených na statistice. Takovým nástrojem je i strojové učení, k němuž jsou tyto postupy primárně cíleny. Jejich aplikací vytvoříme trénovací data využitelná pro vzájemnou konverzi dialektologického a folklorního přepisu (kapitola 5) i jako podpůrný datový zdroj při rozpoznávání mluvené řeči (kapitola 6).

Konverze textových dat pomocí strojového učení

5

KONVERZE TEXTOVÝCH DAT POMOCÍ STROJOVÉHO UČENÍ

5.0 Úvod

V předchozí kapitole byly názorně ukázány a detailně popsány přístupy pro vzájemný převod mezi folklorní a dialektologickou formou. Jak je patrné, jedná se o náročnou a kvalifikovanou práci s vysokými nároky ve znalosti dialektologie a lingvistiky vůbec. S pomocí navržených pravidel byla s využitím regulárních výrazů vytvořena cenná paralelní data (folklorní a dialektologický zápis stejné věty). Tato data můžeme pojmout jako nositele expertních dialektologických znalostí, které mohou být použity pro trénování systému strojového překladu. Ten tyto znalosti nejen absorbuje, ale může přinést do této oblasti nové poznatky a zjednodušení. V této kapitole budou vysvětleny použité přístupy umožňující trénovat a aplikovat systém pro automatický strojový překlad jednotlivých druhů zápisu používaných v dialektologii.

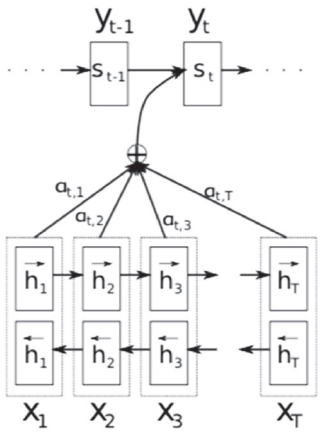
Strojový překlad je jednou z nejvýznamnějších a nejrozšířenějších aplikací umělé inteligence a strojového učení. Přestože tradiční metody strojového překladu, jako jsou statistické metody a rekurentní neuronové sítě (angl. recurrent neural networks, RNN), v minulosti dosáhly významných pokroků, mají své limity. Ty spočívají v zachycení dlouhodobého kontextu (závislosti slov v rámci dlouhých vět a souvětí) v překládaném textu. Problém dlouhodobého kontextu vedl k vývoji pokročilejších metod, jako jsou attention mechanismus a jeho využití v transformer modelech. Ty pak zásadně zlepšily přesnost strojového překladu.

Tato kapitola se zabývá možnostmi strojového učení pro **automatický převod mezi folklorní a dialektologickou formou zápisu** (podrobněji k oběma přepisům viz 3.4.1, 4.1.2.1, 4.4 a 4.5). Nejprve je vysvětlena teorie současných technologií pro strojový překlad, mezi něž se řadí attention mechanismus (5.1), vysvětleno je jeho rozšíření (5.1.1) a zejména přechod do transformer modelů (5.2). Problém s nedostatkem trénovacích dat, typický pro dialektologická data obecně, a jeho řešení v podobě předtrénování jsou objasněny v sekci 5.3. Dále je vysvětlen crosslinguální jazykový model (5.4), který je využit v systémech strojového překladu cílových dialektologických dat (5.5). V závěru kapitoly se věnujeme rozdělení dat a nastavení systému pro předtrénování (5.5.1), též finálnímu doladění a testování systému na nářečních datech (5.5.2).

5.1 Attention mechanismus

Attention mechanismus byl poprvé představen v roce 2014 (Bahdanau a kol., 2014) a rychle se stal klíčovým prvkem mnoha modelů zpracování přirozeného jazyka (angl. natural language processing, NLP). Před zavedením attention mechanismu byly populární sequence-to-sequence modely (Seq2Seq), které využívaly rekurentní neuronové sítě pro kódování vstupní sekvence do fixní vektorové reprezentace a následně dekódování této reprezentace do výstupní sekvence. Nevýhodou těchto modelů byla ztráta informací v případě příliš dlouhé vstupní sekvence (dlouhého kontextu), protože vektor fixní délky nemohl zachytit všechny relevantní informace. Toto bylo efektivně řešeno pomocí attention mechanismu. Základní myšlenka tohoto mechanismu spočívá v tom, že se **model při generování každého slova ve větě zaměřuje na různé části vstupní věty s různou intenzitou**, čímž se efektivně zachycují závislosti mezi vzdálenými slovy. Je tedy měřena míra „pozornosti" nebo vlivu každého vstupního slova ve větě a každého nově generovaného slova. Tento přístup selektivního „věnování pozornosti" různým částem vstupu při generování výstupu umožňuje modelu lépe se vyrovnávat s dlouhými větami a složitými strukturami.

Struktura attention mechanismu je znázorněna na obrázku 5.1.



Obrázek 5.1 Attention mechanismus v systému pro strojový překlad (převzato z Bahdanau, 2014)

Vstupem systému je sekvence slov $x_1, \ldots, x_T$, která je dále zpracovávána v dopředném a i zpětném směru neuronovou sítí, tzv. enkodérem, generujícím sekvenci příznaků $h_1, \ldots, h_T$. Tyto příznaky nesou informaci o kontextu každého slova díky zpracování vstupní sekvence v obou směrech (Schuster a Paliwal, 1997). Výstupní slovo y_t je generováno projekcí vnitřního stavu attention dekodéru s_t a kontextového vektoru c_t pomocí neuronové sítě. Kontextový vektor je vážený průměr vstupních příznaků podle attention skóre. V tomto případě jsou tato skóre počítána mezi vstupní sekvencí $h_1, \ldots, h_T$ a posledním vygenerovaným slovem s_{t-1}.

Výpočet kontextového vektoru a attention váhy

1. **Vstupy a výstupy**: Předpokládejme, že máme vstupní sekvenci slov $\mathbf{X} : x_1, x_2, \ldots, x_t, \ldots, x_T$ a výstupní sekvenci slov $y_1, y_2, \ldots, y_n, \ldots, y_N$.
2. **Enkodér**: Vstupní sekvence slov je zpracována neuronovou sítí, tzv. enkodérem, generující sekvenci skrytých stavů (tzv. embeddingů) h_t.
$$\mathbf{H} = \text{encoder}(\mathbf{X})$$
3. **Attention skóre**: Pro každé slovo ve výstupní sekvenci se počítají tato skóre vzhledem ke každému slovu ve vstupní sekvenci. Attention skóre určuje, jak moc by měl model věnovat pozornost jednotlivým slovům na vstupu při generování aktuálního slova ve výstupu za předpokladu, že známe předchozí vygenerované slovo s_{n-1}.

   Attention skóre lze vypočítat různými způsoby: například pomocí skalárního součinu (angl. dot-product). V tomto případě je pak attention skóre mezi skrytým stavem vstupního slova h_t a skrytým stavem posledního výstupního slova s_{n-1} dáno jako:
$$\text{score}(h_t, s_{n-1}) = h_t \cdot s_{n-1}$$
4. **Normalizace (softmax)**: Attention skóre pro celou větu jsou dále normalizována pomocí softmax funkce, která transformuje skóre do pravděpodobnostního rozdělení (jejichž součet je roven jedné):
$$\alpha_{tn} = \frac{\exp(\text{score}(h_t, s_{n-1}))}{\sum_{t=1}^{T} \exp(\text{score}(h_t, s_{n-1}))},$$
   kde α_{tn} je normalizovaná attention váha pro výstupní slovo s_n při generování výstupního slova y_n.

5. **Kontextový vektor**: Kontextový vektor c_n se vypočítá jako vážený součet skrytých stavů vstupních slov h_t, kde váhy jsou dány hodnotami α_{tn}:

$$c_n = \sum_{t=1}^{T} \alpha_{tn} h_t$$

6. **Generování výstupu**: Kontextový vektor c_n se pak použije spolu s aktuálním skrytým stavem posledního vygenerovaného slova s_{n-1} k vygenerování nového výstupního slova y_n.

5.1.1 Rozšíření a vlastnosti attention mechanismu

Attention mechanismus lze efektivně paralelizovat a zobecnit pomocí maticového násobení.

Definujeme si tři vstupní matice pro celý attention proces: dotazy (angl. query) Q, klíče (angl. keys) K a hodnoty (angl. values) $\mathbf{V}$. Pak $\mathbf{V}$ jsou hodnoty generující kontextové vektory zprůměrováním normalizovaných attention skóre (matice $\mathbf{H}$ v předchozí sekci), Q je matice dotazů, které vstupují do výpočtu attention skóre (matice všech skrytých stavů s_{n-1} z předchozí kapitoly) a matice klíčů K popisuje hodnoty, ke kterým je attention měřena (matice $\mathbf{H}$ v předchozí sekci).

Jednotlivé váhy α_{tn} mohou být uloženy v řádcích matice α (velikost $T \times N$), jejímž „roznásobením" do matice $\mathbf{V}$ mohou být vygenerovány všechny kontextové vektory. Každý řádek této matice se počítá pomocí funkce nad skalárními součiny hodnot v řádcích matice K a matice Q. Matice vah určující podobnosti mezi dotazy a $\mathrm{softmax}$ klíči je pak efektivně počítána:

$$\alpha = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{\boldsymbol{d}_k}}\right)$$

a celý attention proces může být kompaktně zapsán pomocí maticové notace jako

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Hlavní vlastnosti a výhody:

- **Zachycování dlouhodobých závislostí**: Schopnost zaměřit se na relevantní části vstupu bez ohledu na jejich pozici.
- **Paralelismus**: Umožňuje paralelní zpracování, což zrychluje trénování a predikci.
- **Flexibilita**: Lze jej snadno přizpůsobit různým úlohám a datovým strukturám.

Nevýhodou tohoto přístupu je nutnost vidět celý vstup, což komplikuje využití v online a streaming aplikacích.

Attention může pracovat nad různými vstupy, jednou z možností jsou již zmíněná slova, obecně se ale může jednat o výrazně kratší jednotky, například grafémy nebo jejich shluky („podslova"), což výrazně zmenšuje množství vstupních/výstupních symbolů. V dalším textu budeme z výše uvedených důvodů vstup attention procesu nazývat jako token.

Vícehlavý attention (angl. multi-head attention)

Další důležité rozšíření tohoto mechanismu je „vícehlavý" attention, který se skládá z několika paralelních attention mechanismů (hlav). Každá hlava má své vlastní matice dotazů, klíčů a hodnot. Jejich výstupy

jsou následně sloučeny, projektovány a rovněž použity pro generování dalšího slova. Výpočet multi-head attention je dán jako:

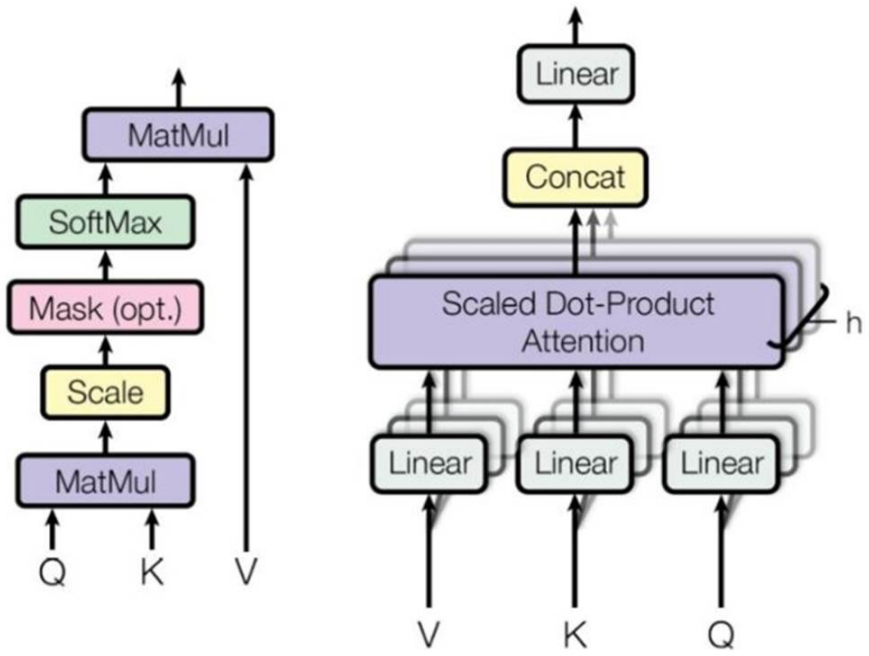
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)W^O,$$

kde jednotlivé hlavy jsou dány jako:

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V),$$

přičemž W_i^Q, W_i^K, W_i^V, W^O jsou trénovatelné váhy.

Obrázek 5.2 popisuje rozdíl mezi jednoduchým a vícehlavým attention přístupem.



Obrázek 5.2 Základní a vícehlavý attention přístup (převzato z Vaswani a kol., 2017)

Poziční kódování

Pozični kódování je klíčovou součástí attention modelů, protože tyto modely nemají jak rozlišovat pořadí vstupních tokenů, neboť sumy ve výše uvedených rovnicích nejsou schopny reflektovat pořadí sčítaných prvků. Tyto modely zpracovávají všechny vstupní tokeny paralelně, a proto potřebují explicitní způsob, jak zakódovat pozici každého tokenu ve vstupní sekvenci. Tento úkol řeší poziční kódování.

Poziční kódování přidává informace o pozici tokenů ve vstupní sekvenci tak, aby model mohl využívat nejen obsah jednotlivých tokenů, ale také jejich relativní nebo absolutní pozici ve vstupní sekvenci. Většina současných attention modelů používá tzv. sinusové poziční kódování, i když existují i jiné varianty.

Sinusové poziční kódování využívá funkce sinus a kosinus pro zakódování pozice tokenů. Pro každý token v sekvenci se generuje poziční vektor se stejnou dimenzionalitou, tento vektor se přičítá k původnímu vektoru před dalším zpracováním.

Poziční vektory jsou definovány následovně:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right),$$

kde:

- PE je poziční enkodér;
- pos je pozice tokenu v sekvenci;
- i je index vektorové složky;
- d je dimenze vstupního vektoru.

Ukázka použití pozičního kódování v transformer modelech (podkapitola 5.2) je vidět na obrázku 5.3.

5.2 Self-attention mechanismus a transfomer modely

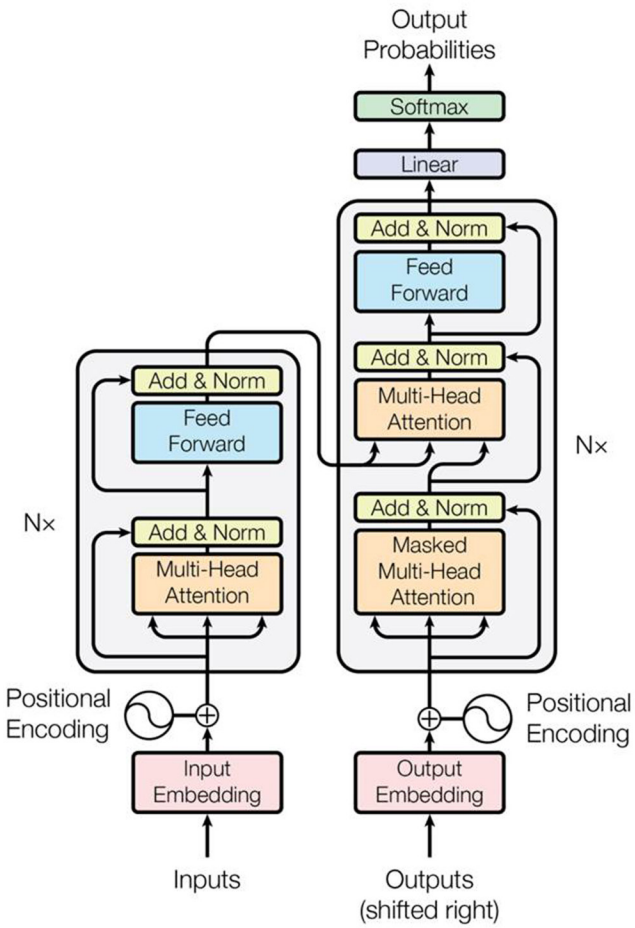
Self-attention je speciální typ attention mechanismu, kde každý prvek sekvence (slovo) věnuje pozornost všem ostatním prvkům ve stejné sekvenci. To umožňuje modelu efektivně zachytit závislosti mezi slovy bez ohledu na jejich vzdálenost ve vstupní sekvenci. V podstatě to znamená, že všechny tři matice (dotazy Q, klíče K a hodnoty V) jsou stejné. Tento typ attention mechanismu je používaný v takzvaných transformer modelech, kde umožňuje modelu posuzovat vztahy mezi všemi slovy ve větě najednou. To značně zlepšuje schopnost modelu chápat kontext a dlouhodobé závislosti. Díky této schopnosti dosahují transformery vynikajících výsledků v různých úlohách zpracování přirozeného jazyka, včetně strojového překladu.

Tato architektura byla poprvé představena v článku *Attention is All You Need* (Vaswani a kol., 2017). Transformery představují revoluční krok v oblasti zpracování přirozeného jazyka (NLP), protože se obejdou bez rekurentních (RNN) a konvolučních neuronových sítí (CNN) a spoléhají se výhradně na attention mechanismy. Transformery se skládají ze dvou hlavních částí: enkodéru (angl. encoder) a dekodéru (angl. decoder). Enkodér zpracovává vstupní větu a převádí ji na sérii reprezentací, zatímco dekodér využívá tyto reprezentace ke generování výstupní věty.

Klíčové komponenty transformeru

1. **Encoder-decoder architektura**:
   - **Enkodér**: Skládá se z vrstev neuronové sítě (typicky 6 vrstev). Každá vrstva obsahuje 2 hlavní podvrstvy:
     - **Self-attention mechanismus**: Umožňuje každému slovu v sekvenci vyhodnotit podobnost ke všem ostatním slovům v sekvenci a přiřadit jim váhy.
     - **Dopředná (angl. feed-forward) neuronová síť**: Aplikuje nelineární transformace na každé pozorování nezávisle.
   - **Dekodér**: Skládá se také z vrstev neuronové sítě (typicky 6 vrstev). Každá vrstva má 3 hlavní podvrstvy:
     - **Masked self-attention mechanismus**: Podobné jako self-attention v enkodéru, ale maskuje budoucí pozice, aby nedošlo k „podvádění“ během trénování a systém byl koherentní s reálnou aplikací, kde budoucnost není známa.
     - **Encoder-decoder attention**: Umožňuje dekodéru zaměřit se na relevantní části vstupu z enkodéru.
     - **Dopředná neuronová síť**: Stejná jako v enkodéru.

Celý proces je zobrazen na obrázku 5.3.



Obrázek 5.3 Transformer model (převzato z Vaswani a kol., 2017)

Transformery jsou schopny lépe zpracovávat paralelně data, což vede k rychlejšímu trénování a efektivnějšímu využití výpočetních zdrojů. Kromě strojového překladu jsou úspěšně aplikovány i v dalších oblastech, jako je např. přepis řeči, sumarizace textu a generování textu.

5.3 Předtrénování transformer modelů

Trénování transformer modelů vyžaduje obrovské množství dat, což může být v některých případech problematické. Hlavní problémy zahrnují:

1. **Nedostatečné pokrytí:** Malé množství dat nemusí dostatečně pokrývat všechny jazykové jevy, což vede k tomu, že model nedokáže dobře generalizovat.
2. **Overfitting:** Při trénování na malém množství dat se model může začít příliš přizpůsobovat na trénovací data, což způsobuje, že jeho výkon na nových datech je špatný.
3. **Nedostatek kontextových informací:** Transformery se spoléhají na bohaté kontextové informace, které mohou chybět, pokud je dataset nebo vstupní věta příliš krátká.
4. **Nedostatečná diverzita:** Malý dataset může mít omezenou diverzitu, což může způsobovat, že model nebude schopen dobře pracovat s různými typy dat a generalizovat.

Tyto problémy řeší předtrénování na obrovském množství dat a následné doladění (angl. fine-tuning) na malém množství dat z cílové domény. Potom lze i s malým množstvím dat dosáhnout vysoké přesnosti díky vnitřním reprezentacím naučeným během předtrénování.

Nejčastější způsob předtrénování spočívá v „zamaskování" (zakrytí) náhodně zvolené části trénovacích dat. Model je potom trénován tak, aby tuto část „uhodl" na základě kontextu, čímž je „nucen" se naučit vnitřní strukturu dat, aniž by byla známa cílová aplikace.

Systém strojového překladu lze vystavět pomocí jazykového modelu, který predikuje následující token v sekvenci na základě předchozích. S tím, že výstupní sekvence je v jiném jazyce než vstupní. Nicméně pro učení vnitřních struktur (předtrénování) obou jazyků překladu není nutné mít pouze paralelní data (přeložené věty). Systém lze „předtrénovat" i na datech z jednotlivých jazyků, kterých je mnohem více. Na následujících řádcích si vysvětlíme maskování u standardních jazykových modelů včetně rozšíření pro strojový překlad, kde tato paralelní data máme k dispozici.

Maskované jazykové modely (MLM)

Maskovaný jazykový model (MLM) je jednou z klíčových úloh použitých při předtrénování. Na rozdíl od tradičních jazykových modelů, kde je predikován následující token v sekvenci na základě tokenů předchozích, MLM náhodně maskuje některé tokeny ve vstupní sekvenci a model se učí předpovídat tyto maskované tokeny na základě jejich kontextu. Tento přístup umožňuje modelu získat bidirekcionální kontextové reprezentace.

Proces MLM

1. **Maskování tokenů:** Náhodně se vybere 15 % tokenů ve vstupní sekvenci. Z těchto vybraných tokenů:
   - 80 % je nahrazeno speciálním tokenem [MASK].
   - 10 % je nahrazeno náhodným tokenem.
   - 10 % zůstává nezměněno.
2. **Předpovídání tokenů:** Model se trénuje, aby předpověděl původní hodnotu maskovaných tokenů na základě kontextu, který poskytují ostatní tokeny ve vstupní sekvenci.

5.4 Translingvální modely (XLM)

Translingvální (angl. cross-lingual) modely mají schopnost pracovat s více jazyky současně a jsou zvláště užitečné pro úlohy, jako je strojový překlad. Tyto modely jsou navrženy tak, aby dokázaly chápat a generovat text ve více jazycích, přičemž mohou přenášet znalosti získané z jednoho jazyka do druhého. XLM model použitý v této metodice je založen na transformer architektuře, jejíž klíčové komponenty jsou:

1. **Enkodér:**
   - Transformer enkodér, který se skládá z několika vrstev (typicky 12 nebo 24). Každá vrstva obsahuje vícehlavý attention mechanismus a dopředné neuronové sítě.
   - Enkodér zpracovává vstupní texty a převádí je na skryté reprezentace.
2. **Poziční kódování:**
   - Poziční kódování je přidáno k embeddingům vstupních slov, aby se zakódovala informace o jejich pozici v sekvenci.
3. **Jazykové identifikační tokeny:**
   - XLM používá speciální tokeny k identifikaci jazyka vstupního textu, což pomáhá modelu přiřazovat správné jazykové kontexty k jednotlivým slovům.

Předtrénování

XLM využívá dva základní přístupy pro předtrénování efektivní multilingvální reprezentace:

1. **Maskovaný jazykový model (angl. masked language model, MLM)**:
   - Tento způsob je již zmíněný výše pro standardní předtrénování transformer modelů.
   - Například pro větu *The cat sat on the mat.* může být slovo *cat* zamaskováno jako „[MASK]" → „The [MASK] sat on the mat." a model se učí predikovat, že „[MASK]" je „cat".
2. **Překladový jazykový model (angl. translation language model, TLM)**:
   - TLM je rozšíření MLM, které využívá paralelní věty v různých jazycích. Model je trénován na predikci zamaskovaných slov, přičemž využívá kontext z obou jazyků.
   - Například pokud máme větu *The cat sat on the mat.* v angličtině a její překlad *Kočka seděla na podložce.* v češtině, model může zamaskovat slovo *cat* v anglické větě a slovo *Kočka* v české větě a učit se je predikovat na základě kontextu obou vět.

Proces trénování

1. **Data**:
   - XLM je trénován na obrovských množstvích textových dat z různých jazyků, včetně paralelních korpusů (překlady textů mezi jazyky) a monolingválních korpusů (texty v jednotlivých jazycích).
2. **Optimalizace**:
   - Model je trénován pomocí jedné z technik matematické optimalizace.
3. **Multilingvální vektorový prostor**:
   - Během trénování se model učí reprezentovat slova z různých jazyků ve stejném vektorovém prostoru, což umožňuje efektivní přenos znalostí mezi jazyky.

5.5 Trénování cross-lingual jazykového modelu na dialektologických datech

Česká dialektologická data v našem případě obsahovala 141 119 vět v dialektologickém formátu získaných skenováním odborné literatury. Spisovná čeština, která je z dostupných dat nejblíže k folklorním datům, byla získána z textu z TED talks.[44] Celkem bylo získáno 170 611 vět.

Paralelní data získaná pomocí pečlivě vytvořených regulárních výrazů (viz kapitola 4) byla rozdělena na trénovací, testovací a vývojové datové sady. Příliš dlouhé odstavce byly rozděleny na kratší úseky o maximální délce cca 100 slov.

Tabulka 5.1 Množství dat (počet vět) v rámci jednotlivých datových sad

Trénování	Vývoj	Test
42 122	2 000	2 000

5.5.1 Předtrénování

Veškerá jednojazyčná data jsou použita k trénování tokenizéru, který vstupní slova efektivně rozloží do „podslov", čímž se efektivně zmenší velikost slovníku, v našem případě na 8 000 tokenů ve slovníku.

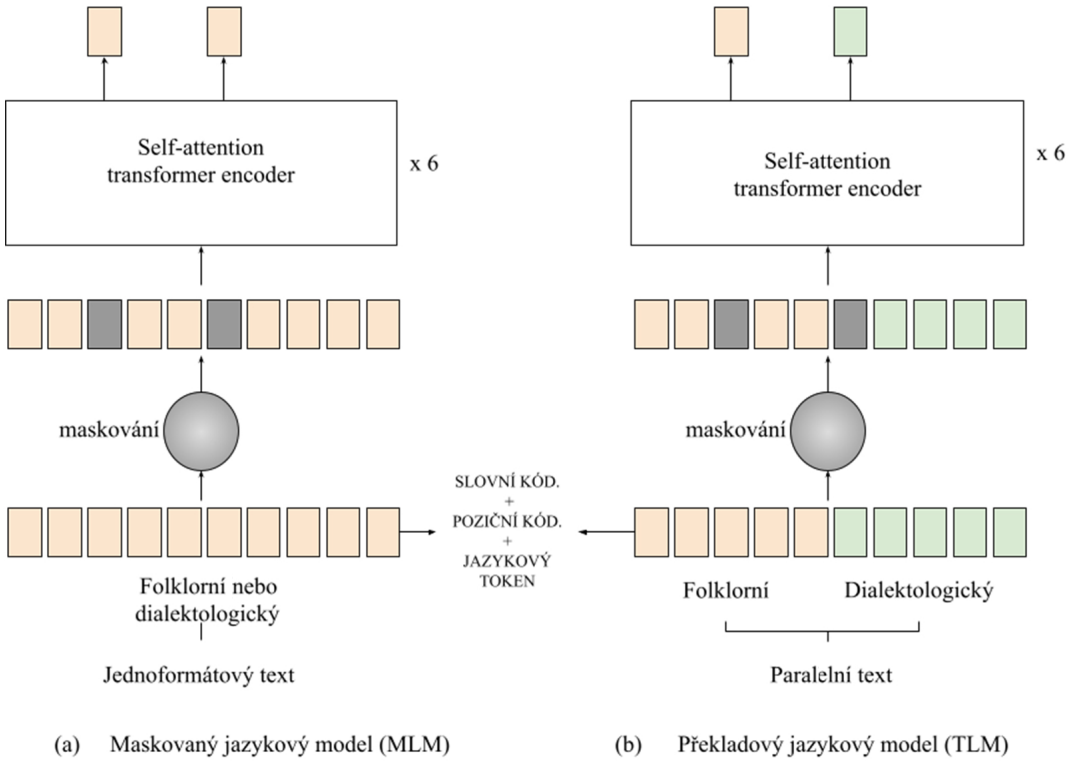
44 TED2020, 2020.

Předtrénování využívá enkodér architekturu transformeru s 6 vrstvami, 512 rozměry embedovacích vektorů, 4 hlavami a dropoutem 0,3 (23M parametrů). Model je trénován po dobu 1 000 epoch s velikostí trénovacích dávek 256, přičemž se používá Adam optimizer s koeficientem učení 2e-4. K vstupním slovním vektorům se přidávají poziční vektory a také vektory popisující způsob zápisu (dialektologický nebo folklorní).

Předtrénování využívá jak maskované modelování jazykového modelu (MLM), tak i překladové modelování (angl. translation language model, TLM) inspirované modelem XLM. Po každé trénovací dávce se střídá MLM a TLM přístup.

- **MLM:** Model dostane buď dialektologický, nebo folklorní text. V obou případech je 15 % vstupu maskováno a model je trénován k předpovědi těchto maskovaných tokenů.
- **TLM**: Model obdrží pár textů v obou formátech. Na zřetězeném textu je maskováno 15 % tokenů a model je trénován pro jejich předpovědi.

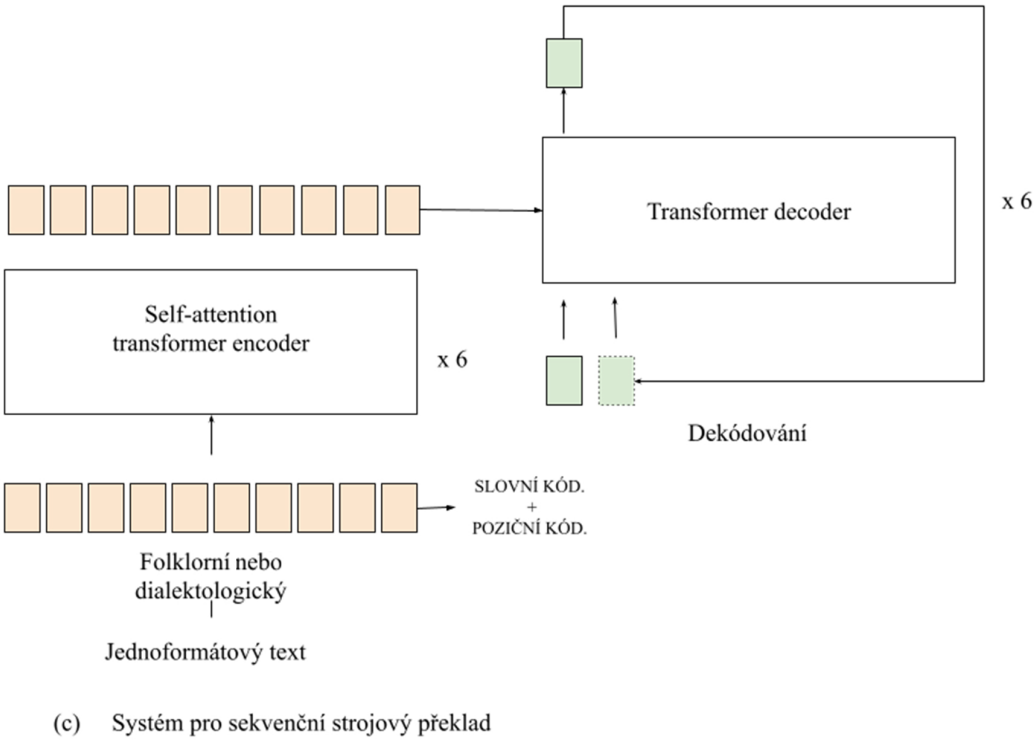
Obrázek 5.4 ukazuje obě schémata MLM a TLM.



Obrázek 5.4 Maskovaný a překladový jazykový model

5.5.2 Doladění modelu

Předtrénovaný enkodér je použit k inicializaci architektury transformer enkodér-dekodér modelu, kde enkodér i dekodér mají stejnou kopii parametrů. Parametry pro cross-attention jsou inicializovány náhodně. Následně jsou trénovány dva modely v obou směrech překladu (dialektologický → folklorní, folklorní → dialektologický), viz blokové schéma modelu na obrázku 5.5.



Obrázek 5.5 Finální systém pro strojový překlad

Doladění probíhalo se stejnou optimalizační konfigurací jako při předtrénování, pouze s nižší hodnotou koeficientu učení (1e-5) a s menší velikostí trénovací dávky (128).

Výsledný model byl vyhodnocen pomocí slovní chybovosti WER (angl. word error rate) na testovací sadě o 2 000 větách.

Tabulka 5.2 Výsledky systému automatického převodu

Směr převodu	WER
Dialektologický → folklorní	1.47
Folklorní → dialektologický	1.13

Příklady výstupu systému:

Reference v **dialektologickém formátu:**

> *Anka, gdiš to slišala, chiťila paňimámu okolo krku a tak se k ňí přiťiskla, že se paňimáma zasmála: „Jen mňe neušk[e/ə]rť, holka, diť mňe mořiš, jako bih bila tvá máma." – „Diť tadi taki ste, gdiš se mnou tak pjekňe mluvíte a ve fšem mňe raďíte."*

Reference ve **folklorním formátu:**

> *Anka, když to slyšala, chytila panimámu okolo krku a tak se k ní přitiskla, že se panimáma zasmála: „Jen mě neuškerť, holka, dyť mě moříš, jako bych byla tvá máma." – „Dyť tady taky ste, když se mnou tak pěkně mluvíte a ve všem mně radíte."*

Výstup překladu: **Dialektologický → folklorní:**

> *Anka, když to slyšala, chytila panimámu okolo krku a tak se k ní přitiskla, že se panimáma zasmála: „Jen mě neuškerť, holka, dyť mně moříš, jako bych byla tvá máma." – „Dyť tady taky ste, když se mnou tak pěkně mluvíte a ve všem mně radíte."*

Výstup překladu: **Folklorní → dialektologický:**

> *Anka, gdiš to slišala, chiťila paňimámu okolo krku a tak se k ňí přiťiskla, že se paňimáma zasmála: „Jen mňe neušk[e/ə]rť, holka, diť mňe moříš, jako bih bila tvá máma." – „Diť tadi taki ste, gdiš se mnou tak pjekňe mluvíte a ve fšem mňe raďíte."*

5.6 Shrnutí

V této kapitole byly vysvětleny základní principy moderních technik jako attention a jeho varianta self-attention, která se využívá v transformer modelech. Tyto modely nalezly široké uplatnění v oblastech zpracování přirozeného jazyka stejně jako v systémech pro přepis mluvené řeči (viz kapitola 6).

Je zde názorně ukázáno, **jak tyto modely využít v oblasti dialektologie** pro učení systému pro automatický převod mezi folklorní a dialektologickou formou zápisu a zpět. Nejprve je navržena architektura podobně jako u systémů pro strojový překlad vycházejících z neurálních jazykových modelů. Pak je na reálném případě předvedeno (včetně nastavení parametrů učení), jak:

- tyto modely nejprve efektivně předtrénovat, aby byl minimalizován problém nedostatku dialektologických dat;
- dotrénovat systém na cílových datech;
- vyhodnotit výsledky.

Tento automatický přístup strojového učení ukazuje schopnost absorbovat dialektologické znalosti nutné pro přípravu dat (viz kapitola 4) s excelentními výsledky.

Strojový přepis mluvené řeči do textu

6

STROJOVÝ PŘEPIS MLUVENÉ ŘEČI DO TEXTU

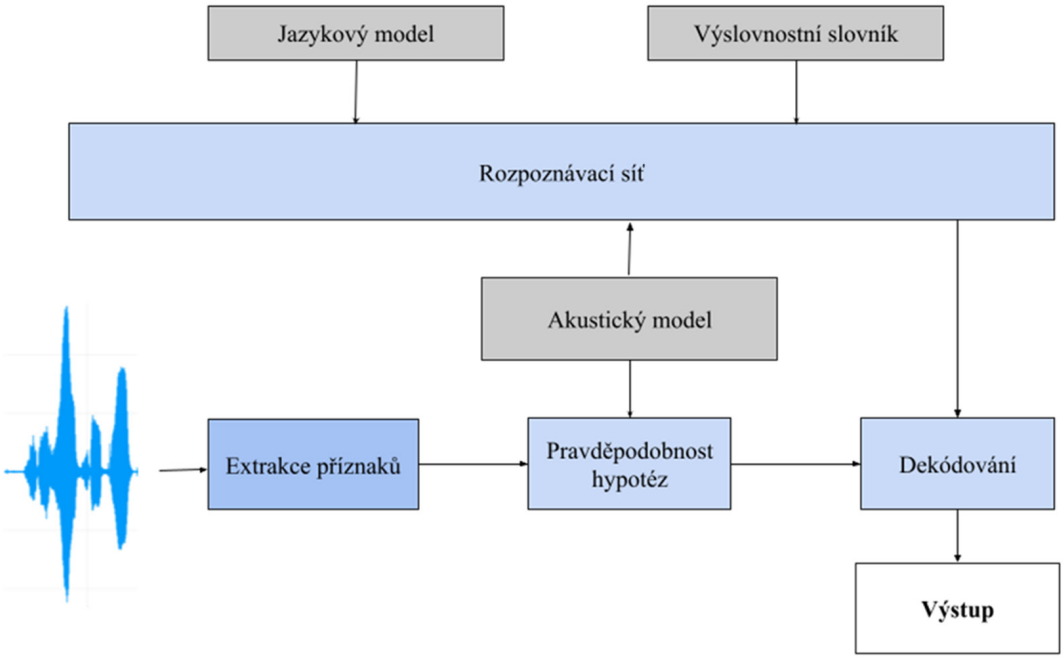
6.0 Úvod

Tato kapitola stručně uvádí základní teorii automatického strojového přepisu klasickým přístupem (6.1), a to včetně rozšíření o neuronové sítě (6.1.1) a jejich vlastnosti (6.1.2). Tento přístup se označuje jako hybridní, neboť kombinuje přístup tradiční a moderní (podrobně viz Hinton, Deng a Yu, 2012; Dahl a kol., 2012; Yu a Deng, 2016). Dále se kapitola stručně zabývá moderními tzv. End-to-End metodami (6.2), jako jsou Connectionist Temporal Classification (CTC; Graves a Jaitly, 2014; Hannun, Case a Casper, 2014) a Attention (Chan a kol., 2016; Radford a kol., 2022). Poté je ukázán a řešen problém s přípravou dialektologických transkripcí pro strojový přepis (6.3).

6.1 Tradiční přístup

Automatický přepis spontánní mluvené řeči z nahrávek (angl. automatic speech recognition, též speech to text translation, ASR/STT) je určen pro převod audia do textové podoby, včetně časové informace o poloze slov. Tím je zpřístupněn obsah jinak než jen pouhým poslechem, což je časově zdlouhavé a omezující, zvláště v případě, pokud je předmětem zájmu obsah, a ne audiální forma. Výsledky přepisu jsou tak určeny pro použití v indexačních systémech v oblasti plně automatizovaného zpracování textu, pro rychlé vyhledávání či automatický překlad nebo pro rychlé přečtení obsahu řeči přímo člověkem.

Základní architektura tradičního tzv. hybridního systému je zachycena na obrázku 6.1.



Obrázek 6.1 Blokové schéma tradičního systému přepisu řeči na text (ASR/STT)

Hybridní systém se skládá z několika základních funkčních bloků:

- **Akustický model** – reprezentuje pravděpodobnost, že určitá sekvence zvuků odpovídá specifickým fonémům. Systém používá skryté Markovovy modely (angl. hidden Markov models, HMM), které modelují časově proměnlivou povahu řeči. K modelování akustických jednotek bylo hojně využíváno statistické modelování pomoci směsí Gaussových rozdělení pravděpodobnosti (angl. Gaussian mixture models, GMM).
- **Jazykový model** – odhaduje pravděpodobnost výskytu sekvence slov. N-gramové modely byly standardem, přičemž určovaly pravděpodobnost výskytu určitého slova na základě předchozích slov.
- **Výslovnostní slovník** – obsahuje mapování slov na sekvence fonémů, které modely používají při rozpoznávání.
- **Dekodér** – využívá statickou rozpoznávací síť modelující všechny posloupnosti slov, které může rozpoznávač vygenerovat.

6.1.1 Využití neuronových sítí v hybridním přístupu

S příchodem hlubokého učení začaly neuronové sítě postupně nahrazovat tradiční komponenty v ASR systémech:

1. **Neuronové akustické modely (DNN-HMM):** Místo tradičních GMM se pro modelování akustických vlastností začaly používat hluboké neuronové sítě (DNN, viz Hinton, Deng a Yu, 2012). Tyto sítě lépe zachycují složité vztahy mezi akustickými prvky a fonémy díky své schopnosti modelovat nelineární vztahy v datech. HMM však zůstávají důležitou součástí modelování časové dynamiky řeči.
2. **Rekurentní a konvoluční sítě (RNN, CNN):** Pro modelování časových závislostí a frekvenčních rysů v akustických datech se začaly používat rekurentní neuronové sítě (RNN) a konvoluční neuronové sítě (CNN). Rozšířená varianta rekurentních sítí Long Short-Term Memory (LSTM) a Gated Recurrent Units (GRU) jsou zvláště účinné v zachycování dlouhodobých závislostí v řeči.
3. **Jazykové modely s RNN:** Místo n-gramových modelů se začaly používat jazykové modely založené na RNN, které lépe zachycují kontextové informace a umožňují modelu lépe rozumět složitým jazykovým strukturám.

6.1.2 Výhody hybridního přístupu

Hybridní přístup kombinuje výhody tradičních metod s výkonem hlubokého učení, což přináší několik klíčových výhod:

1. **Lepší výkon na malých datech:** Neuronové sítě vyžadují velké množství dat pro trénování, což může být limitujícím faktorem. Kombinace s tradičními metodami však umožňuje využít i menší datové sady efektivně.
2. **Flexibilita a adaptabilita:** Neuronové sítě mohou být snadno přizpůsobeny různým jazykům a dialektům, což zlepšuje schopnost ASR systému rozpoznávat řeč v různých kontextech.
3. **Robustnost:** Hybridní modely jsou obecně robustnější vůči šumu a jiným variacím v řeči, protože neuronové sítě dokážou lépe generalizovat a odolávat těmto rušivým vlivům.
4. **Zlepšení jazykového porozumění:** Jazykové modely založené na RNN umožňují lepší zachycení dlouhodobých kontextových závislostí, což zlepšuje celkové porozumění jazyku v přepisu.

6.2 End-to-End přístup

End-to-End modely eliminují potřebu explicitního dělení na akustický, jazykový a výslovnostní model. Místo toho se učí přímo mapovat akustické vstupy na textové výstupy. Tento přístup nejenže zjednodušuje celý proces, ale také vede k vyšší úspěšnosti a robustnosti, protože umožňuje trénování všech částí systému společně, čímž se zlepšuje schopnost modelu generalizovat a adaptovat se na různá prostředí.

Klíčové End-to-End architektury

Současně dominující architektury v oblasti End-to-End systémů:

1. **Connectionist Temporal Classification (CTC):** CTC je metoda používaná k trénování neuronových sítí pro sekvenční data, jako je řeč (Graves a Jaitly, 2014). Umožňuje modelu naučit se sekvenci vstupů a jejich mapování na výstupy, i když délka sekvencí se liší.
2. **Sequence-to-Sequence modely s attention mechanismem:** Tyto modely, původně vyvinuté pro strojový překlad, byly adaptovány pro ASR (Chan a kol., 2016). Kombinují encoder, který převádí vstupní akustický signál na skryté reprezentace (tzv. embeddingy), a decoder, který generuje text. Attention mechanismus umožňuje modelu zaměřit se na relevantní části vstupu při generování každého znaku nebo slova.
3. **Transformer modely:** Modely založené na transformer architektuře, které se spoléhají na self-attention mechanismus, se ukázaly jako velmi účinné v různých sekvenčních úlohách, včetně ASR (Radford a kol., 2022). Tyto modely jsou schopné efektivně zpracovávat dlouhé sekvence a zachycovat složité závislosti mezi různými částmi vstupu.

Více detailů o attention přístupu a transformer modelech lze nalézt v kapitole 5, věnované strojovému překladu.

Nedostatek trénovacích dat a výzvy s tím spojené

Trénování End-to-End ASR modelů sice zjednodušuje tradiční ASR architektury a zvyšuje přesnost přepisu, avšak přináší i nové výzvy, zejména v oblasti dostupnosti a kvality trénovacích dat. Model vyžaduje rozsáhlé trénovací datové sady zahrnující různé jazyky, dialekty, akcenty a různé typy akustických podmínek (např. různé úrovně šumu, přerušení a další variace). Kvalitní trénovací data jsou klíčová pro dosažení vysoké přesnosti a schopnosti modelu generalizovat na nové, dosud neviděné vstupy.

Hlavní problémy spojené s nedostatkem dat:

1. **Nízká dostupnost dat pro menšinové jazyky a dialekty:** Většina dostupných dat je zaměřena na hlavní světové jazyky, což znamená, že menšinové jazyky jsou často nedostatečně zastoupeny. Totéž platí i o teritoriálních dialektech, tedy o cílových datech této metodiky. Zde je navíc obtížné získat cílová audiální data (viz kapitola 3). Tyto problémy vedou k nižší přesnosti ASR systémů pro tyto jazyky či dialekty.
2. **Variabilita v akcentu a výslovnosti:** I v rámci jednoho jazyka, popřípadě i nářečí, může existovat významná variabilita v akcentu a výslovnosti, což může ovlivnit přesnost modelu, pokud není trénován na dostatečně různorodých datech.
3. **Náklady na anotaci dat:** Ruční anotace velkých množství zvukových dat je časově náročná a nákladná, což omezuje dostupnost kvalitních trénovacích dat (obtížnost přepisů je detailně zdokumentována v kapitole 3.4.2).

Předtrénování modelů

Aby bylo možné překonat výše uvedené problémy, byly vyvinuty různé techniky předtrénování, které umožňují efektivnější využití dostupných dat a zlepšují schopnost modelů generalizovat. Předtrénování spočívá v trénování modelu na velkém množství dat, často z různých domén nebo jazyků, a následném „doladění" (fine-tuningu) modelu na konkrétním úkolu. Základní princip je podobný tomu, který byl ukázán v předchozí kapitole (viz 5.3). Spočívá v tom, že se část dat „zamaskuje" a model je nucen tuto část doplnit. Mezi nejznámější přístupy patří Wav2vec 2.0 (Baevski a kol., 2020). Tato technika umožňuje předtrénování na velkém množství řečových akustických dat, která jsou snadno dostupná. Výsledný model je poté, podobně jako v předchozí kapitole, dotrénován na cílových datech.

6.3 Příprava dialektologických dat pro trénování přepisu řeči

Jak bylo zmíněno výše, pro trénování systému automatického přepisu je nutné mít dostatečné množství anotovaných akustických dat z cílové domény. Problém u nářečních dat spočívá v jiném způsobu zápisu, než jaký je vhodný pro akustické modelování. Zde je vhodné pracovat s kratšími časovými úseky (maximálně řádově desítky sekund), což trénovacímu algoritmu usnadňuje nalezení mapování mezi akustickou částí a grafemickým zápisem. Dialektologická data se však v minulosti přepisovala v časově dlouhých blocích, řádově i několik minut.

K rozdělení dat na kratší časové celky je nutné definovat mapování mezi grafemickou a akustickou částí. Pak lze data snadno dělit v místech s úseky ticha delšími než 300 ms, což je maximální délka uzavření hlasového ústrojí v plozivách (fonémech obsahujících ticho).

K nalezení **akustického mapování** je použit následující postup:

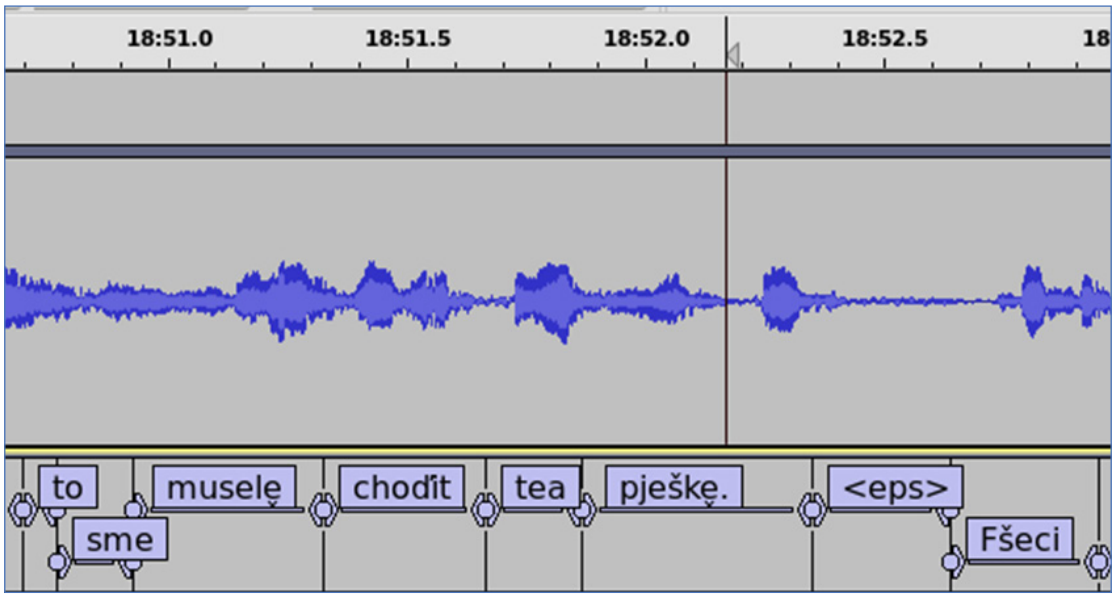
1. Trénování klasických GMM-HMM modelů pro běžnou češtinu. Toho lze snadno dosáhnout s využitím veřejně dostupných dat[45] pomocí nástroje Kaldi.[46]
2. Použití mapování dialektologického zápisu do fonetického zápisu X-SAMPA (Wells, 1995), standardně využívaného pro názvy akustických jednotek (viz tabulka 3.2 v sekci 3.4.2).
3. Nalezení nejpravděpodobnější shody (mapování) mezi dialektologickým zápisem (jednotlivými znaky) a korespondující sekvencí akustických modelů pomocí Viterbiho zarovnání. Mezery mezi slovy jsou mapovány na model ticha.
4. Rozdělení původních dlouhých úseků textu podle detekovaných a namapovaných segmentů ticha přesahujících 300 ms na kratší úseky. Pokud jsou vedle sebe dva úseky kratší než 10 s, jsou tyto úseky opětovně sloučeny do větších celků ne delších než 10 s.

Takto segmentovaná data mohou být přímo využita pro trénování libovolného systému automatického přepisu řeči, tedy hybridní i End-to-End architektury.

Výsledek zarovnání dialektologických dat je zobrazen na obrázku 6.2, kde model „<eps>" je symbol pro mezeru.

45 Viz Datasets, nedat.; ParCzech 3.0, 2013–2017, 2017–2021.

46 Viz KALDI, nedat.



Obrázek 6.2 Automatické zarovnání dialektologických dat

6.4 Shrnutí

V této kapitole byly vysvětleny **dva základní přístupy využívané v oblasti automatického přepisu mluvené řeči**:

1. hybridní, využívající skryté Markovovy modely a neuronové sítě;
2. End-to-End, které jsou kompletně neurální.

V popisu každé metody byly shrnuty její výhody a nevýhody a obě mohou být přirozeně využity pro tvorbu systému pro automatický přepis dialektologických dat.

Pro oba přístupy jsou nicméně klíčová **kvalitně přepsaná nářeční data** (získání těchto dat bylo prezentováno v kapitole 3) a pro obě výše představené metody je nutná i segmentace nahrávek na kratší časové úseky. Proto je tento problém v podkapitole 6.3 detailně řešen:

- nejprve je ukázáno, jak pomocí automatického zarovnání získat časové informace jednotlivých slov.
- pomocí znalosti úseků ticha jsou nahrávky rozsegmentovány.

Takto připravená data mohou být přímo využita pro trénování libovolného systému.

Pro systémy automatického přepisu řeči jsou klíčová i textová data zapsaná dialektologickým, popř. folklorním přepisem (viz kapitola 4). Na rozdíl od pracně získávaných přepisů audiálních dat jsou tato data k dispozici ve větším rozsahu. U hybridních systémů je využití těchto dat přímočaré – je možné pomocí nich natrénovat statistický nebo neurální jazykový model, který modeluje pravděpodobnosti sekvencí jednotlivých slov (v tomto případě zapsaných dialektologickou transkripcí). U End-to-End systémů, ve kterých není explicitní dělení na jednotlivé modely, je využití náročnější, ale i zde lze postupovat např. reskórováním variant vyprodukovaných End-to-End systémem nebo pomocí tzv. mělké fúze (angl. shallow fusion; viz Zhao a kol., 2019), která interpoluje skóre získaná z End-to-End modelu s hodnotami z externího jazykového modelu. V případě aplikace v dialektologii lze u obou variant uvažovat o různých úrovních jazykových modelů, od dialektologického zápisu jako celku až po využití pro danou úzkou dialektologickou oblast.

Zpracování prostorových dat

7

ZPRACOVÁNÍ PROSTOROVÝCH DAT

7.0 Úvod

Jednou z nejvýznamnějších vlastností teritoriálních dialektů je jejich **územní rozsah**, tedy skutečnost, že se určitým dialektem hovoří na konkrétním území. Dialektologický výzkum proto vždy jako jeden z hlavních parametrů uvádí lokalitu, ve které probíhá. I u dialektologických audionahrávek a potažmo i nářečních textů je proto vždy uvedeno místo, kde byla daná nahrávka pořízena nebo ke kterému místu se vztahuje. Jedná se tedy o prostorová data, která umožňují nejen prostorovou vizualizaci (geovizualizaci), ale také další prostorové analýzy.

V první podkapitole (7.1) jsou vysvětleny obecné zásady identifikace geografického umístění prostorových jevů (objektů a procesů) označované pojmem geolokalizace, se zaměřením na dialektologické audionahrávky. Podkapitola popisuje, jakým způsobem probíhá přiřazení prostorové složky u starých nahrávek, u nichž mohou nastat mnohá úskalí vlivem územních změn, a jak by měla probíhat geolokalizace u nahrávek nových. Druhá podkapitola (7.2) je zaměřena na výběr vhodného datového formátu prostorových dialektologických dat, včetně jeho struktury, a doporučení pro vizualizaci nářečních nahrávek v mapách. V podkapitole je nejprve představena problematika datových formátů prostorových (7.2.1) a dialektologických (7.2.2) dat. Po vysvětlení základních metod vizualizace prostorových dat (7.2.3) se zaměřením na vizualizaci dialektologických dat jsou v závěru podkapitoly popsány hlavní zásady interaktivní a multimediální vizualizace nářečních nahrávek (7.2.4).

V jednotlivých podkapitolách jsou popsány praktické příklady čerpající z existující spolupráce odborníků na geoinformatiku a kartografii z Katedry geoinformatiky Přírodovědecké fakulty Univerzity Palackého a odborníků z dialektologického oddělení Ústavu pro jazyk český Akademie věd ČR. V rámci této spolupráce byla zpracovávána mj. i audiální data, byly navrženy různé metody vizualizace a jako výstupy byly vytvořeny jednotlivé tematické mapy, nářeční atlasy i multimediální online mapy. K dalšímu studiu této problematiky lze proto doporučit nářeční atlasy českého jazyka (Ireinová, Voženílek a kol., 2020, 2021, 2022, 2023) a v souvislosti s publikováním webových map i publikaci *Webová kartografie* (Nétek, 2020).

7.1 Geolokalizace audiálních dat

Jednou z velmi důležitých vlastností (atributů), které jsou u audiálních dat zaznamenávány a uchovávány, je jejich **geografické umístění**, tedy například přiřazení názvu obce, ke které je daná nahrávka platná. Nemusí se nutně jednat o místo pořízení nahrávky, protože pokud je nahrávka s mluvčím pořizována v jiném místě, než je jeho dlouhodobé bydliště, ke kterému náleží daná mluva, může jít nikoliv o místo pořízení, ale právě o místo původu/bydliště mluvčího. Tento aspekt se samozřejmě může různit podle účelu a obsahu nahrávek, nicméně v dialektologickém sběru dat je primární právě místní příslušnost nářeční mluvy, tj. zpravidla je zaznamenáno místo původu, nikoliv dlouhodobé bydliště (vztáhnutelné i k dospělému věku). Jedná se tedy o atribut vztahující se k autochtonnosti mluvčího (k této problematice v souvislosti s autorstvím textů viz 4.1.2.3).

Pojem **geolokalizace** vychází ze slovního spojení geografická lokalizace. Jedná se o postup, kterým je zjišťována geografická poloha určitého objektu, zařízení nebo v případě dialektologických dat místo, kde

se mluvčí naučil nářečí, které používá. Geolokace může být realizována na různých úrovních podrobnosti, od záznamu přesných zeměpisných souřadnic po přiřazení existujících míst, jako je obec, oblast, stát apod. Nejčastěji je tento pojem používán v informatice, kde na základě technických parametrů přenosu identifikuje lokaci uživatele využívajícího internetové služby pomocí IP adresy (číslo, které jednoznačně identifikuje síťové rozhraní v počítačové síti), případně v oblasti mobilních telefonů a služeb, kdy na základě geolokace probíhá například automatické nastavení časového pásma, jazyka nebo jsou uživateli nabízeny cíleně reklamy z okolí místa, kde se nachází. V dialektologii nejsou používány automaticky zaznamenávané parametry nahrávky, protože ty často žádnou možnost geolokace neumožňují, ale je využíváno metadat[47], která zaznamenává dialektolog o dané nahrávce. Geolokalizace v těchto metadatech je tedy vytvářena manuálně, přičemž podrobnost záleží na metodice sběru dat, kterou explorátor využívá.

7.1.1 Geolokalizace existujících nahrávek na území České republiky

Administrativní členění České republiky se k 1. 1. 2024 dělí území na 14 samosprávných krajů, 76 okresů a hlavní město Prahu, 205 správních obvodů obcí s rozšířenou působností, 392 správních obvodů obcí s pověřeným obecním úřadem, 6 254 obcí a 4 vojenské újezdy, 15 105 částí obcí (z toho je například 142 městských částí a 22 správních obvodů v hlavním městě Praze), 13 076 katastrálních území a 23 582 základních sídelních jednotek (základní skladebná část sídelního útvaru, používaná zejména pro statistické účely). Kromě toho je Česká republika dělena i v rámci Nomenklatury územních statistických jednotek (NUTS), což jsou územní celky vytvořené pro účely statistického hodnocení členských zemí Evropské unie. V rámci této nomenklatury se Česká republika dělí na 8 regionů soudržnosti a dále na jednotky odpovídající krajům, okresům a obcím.

Jak již bylo uvedeno, existující nahrávky dialektologického výzkumu realizovaného pracovníky Ústavu pro jazyk český AV ČR zahrnují **období od 50. let 20. století** (Šimečková, 2024a). Územní členění České republiky za tu dobu dostálo obrovských změn. Po druhé světové válce byl z hlediska územního členění českého území obnoven předválečný stav k roku 1938. I proto docházelo v 50. a 60. letech k **velkým změnám v administrativním členění státu**, především k postupnému slučování obcí. Zákonem č. 367/1990 Sb., o obecním zřízení, byla obnovena právní subjektivita obcí a následně docházelo k další vlně masivních změn v územním vymezení obcí a vyšších administrativních jednotek. Jak uvádí Rychtaříková a kol. (2021), v roce 1990 nově vzniklo 1 684 obcí a v letech 1991–1992 vzniklo dalších 337 obcí. Ještě v roce 1993, při vzniku samostatné České republiky, vzniklo dalších 104 obcí a následně v dalších dvou letech 43 (Vajdová a kol., 2006). Od roku 1995 vznikaly obce již v jednotkách případů, a to většinou spojením nebo rozdělením obcí již existujících, a počet obcí se ustálil na počtu kolem 6 250, jak je tomu do současnosti.

Dalším aspektem, který nejen z historického, ale i ze současného pohledu komplikuje geolokaci na podrobnost obcí, resp. částí obcí, je skutečnost, že názvy obcí nejsou unikátní. Jak uvádí *Historický lexikon obcí České republiky* (Český statistický úřad, 2015a), nejčastějším názvem pro obec nebo část obce je v současné době *Nová Ves* (62 výskytů), následovaná *Lhotou* (31) a *Lhotkou* (28), *Chlumem* a *Petrovicemi* (oba názvy po 25 výskytech). Tyto výsledky navíc zahrnují pouze Lhoty a Lhotky v čisté podobě bez jakýchkoliv přívlastků, jako je např. *Nová Lhota*, *Horní Lhota* apod., což dále markantně zvyšuje jejich výskyt až na 234 Lhot. Docházelo také ke změnám názvů obcí, a to zejména v 60. letech 20. století, kdy byly vyzvány obce, které měly duplicitní názvy v rámci okresu, aby si určily nový název s použitím přídomku. Ke změnám názvů obcí a částí obcí však docházelo průběžně i v posledních desetiletích, jako příklad uvádí Český statistický úřad (2015b) změnu názvu části obce z *Karviná* na *Karviná-Doly*, poté *Karviná 2-Doly* a konečně *Doly*. Pomoc-

47 Metadata jsou data, která poskytují informace o jiných datech, v daném případě o pořízené nahrávce. Mezi tato data patří například identifikace mluvčího (jméno, příjmení, věk, povolání apod.) a geolokalizace nahrávky (zpravidla na podrobnost obcí nebo částí obcí). Například v připravované *Databázi nářečních promluv pro odbornou veřejnost* jsou rozlišovány tři základní druhy metadat: osobní (jméno, věk atd.), lokační (místo původu, pobytu mluvčího, místo sběru) a obsahová (viz 3.2.1.2).

níkem pro případy, kdy se měnily názvy a místní části obcí, jsou podpůrné dokumenty, jako jsou existující lexikony obcí, například *Historický lexikon obcí České republiky 1869–2011* (Český statistický úřad, 2015a).

Jak je uvedeno v podkapitole 7.2, existuje řada datových formátů prostorových dat a řada přístupů k tomu, jak dialektologická prostorová data zpracovat. Pro potřeby kartografické vizualizace je však většinou zapotřebí, aby prostorová data byla zpracována do jednotné topologie. V případě realizovaného zpracování již existujících dialektologických nahrávek popisovaného v této metodice byl zvolen přístup **geolokace na podrobnost současně platných částí obcí** v České republice. Z dostupných dat není reálně možné rekonstruovat přesný záznam zeměpisných souřadnic vztažného místa audionahrávky, proto je zvoleno přiřazení k vhodné existující areálové datové vrstvě při maximálním možném zachování podrobnosti. Z hlediska zeměpisných souřadnic lze k takové nahrávce přiřadit konkrétní zeměpisnou šířku a délku centroidu takto zvoleného areálu, tj. centroidu části obce.

Obec je územní jednotkou vymezenou výčtem katastrálních území, která ji tvoří, a má svůj název sloužící k její nezaměnitelné identifikaci. Část obce má jednoznačný název a je evidenční (ne územní) jednotkou obce dle § 27 zákona č. 128/2000 Sb., o obcích (obecní zřízení), ve znění pozdějších předpisů.

Existující **dialektologické nahrávky** mají v metadatech uveden název obce a identifikátor výzkumné lokality, který je podle číselníku *Českého jazykového atlasu* propojen se jménem obce a její příslušností k okresu podle stavu z roku 1974 (Balhar a kol., 1992, s. 45–47). Pokud byla nahrávka pořízena mimo existující výzkumnou síť lokalit (například novější audionahrávky), je v metadatech uveden název obce nebo části obce a opět její příslušnost k okresu. Uvedení identifikátoru výzkumné lokality významně zjednodušuje správnou geolokaci při převodu do formátu prostorových dat. U nahrávek bez uvedení tohoto identifikátoru, například z důvodu výzkumu mimo definovanou síť, může obecně docházet ke komplikacím, kdy jedním z důvodů je mnohočetný výskyt stejných názvů obcí v rámci okresu (např. v okrese Benešov bylo podle Českého statistického úřadu postupně identifikováno 29 výskytů Lhot a Lhotek). V takovém případě došlo ke zpřesnění prostorového atributu zpravidla odborníky z dialektologického oddělení ÚJČ AV ČR, kteří vycházejí z dokumentace realizovaných výzkumů.

Snahou vždy je, aby byla zachována **maximální možná přesnost**, a to jak z hlediska geolokace nahrávky, tak z hlediska zachování informace o „míře spolehlivosti" dané geolokace. Jak je uvedeno výše, na základě přiřazení audionahrávky k areálu části obce je možné jí přiřadit v případě potřeby i konkrétní zeměpisné souřadnice centroidu tohoto areálu. Míra možné nepřesnosti těchto souřadnic je tak dána rozlohou areálu, protože bydliště respondenta zpravidla nebude přímo v místě centroidu, ale bude zcela určitě v rámci daného areálu části obce. U starších nahrávek navíc často není uveden konkrétní název části obce, ale je uveden název obce celé. V takovém případě je z důvodu **jednotného zpracování prostorových atributů** potřeba rozhodnout, ke které části obce bude nahrávka přiřazena. V některých případech doupřesní informace odborníci pracující s archivní dokumentací provedeného výzkumu, někdy však toto zpřesnění možné není. V takových případech musí dojít k rozhodnutí, jakým jednotným přístupem bude ke geolokaci docházet, v našem případě k přiřazení příslušnosti k existující části obce. Poměrně logickým přístupem zde je, že se bude v takovém případě jednat o přiřazení k „hlavní" části obce, tj. například k centrální části města nebo k části obce, která má stejné jméno jako obec.

PŘÍKLAD Z PRAXE

Nahrávka z 50. let 20. století má uvedeno místo pořízení Praha. Současný číselník částí obcí však žádnou část obce „Praha" v hlavním městě Praze neobsahuje. V době pořízení nahrávek většina měst v České republice nebyla rozdělena na části obce platné v současnosti a ani požadavek na uvedení lokalizace nebyl stanoven na podrobnější úroveň, než je sídlo. Tyto případy proto řeší nastavený metodický postup geolokace starých nahrávek, kdy se záznamy s lokalizací uvedenou pouze na podrobnost sídla přiřazují k té části obce, která má stejné (resp. nejbližší možné) označení. Například pokud je ve staré nahrávce uvedeno sídlo Bystročice, existující části obce jsou Bystročice a Žerůvky, pak je nahrávka přiřazena části obce Bystročice. U měst se zpravidla jedná o výběr centrální části města, proto je audionahrávka s uvedeným místem pořízení „Praha" lokalizována do městské části Praha I. Z důvodu zachování informace o zpřesnění (a potenciálně zavedení chybné lokalizace v rámci sídla) by měla být informace o původní lokalizaci zachována v metadatech nahrávky.

V některých případech došlo v minulosti i ke změně názvu obce. Veřejnosti nejznámější jsou změny názvů související s historickými událostmi, jako jsou poválečné přesuny obyvatelstva a změny německých názvů na české po druhé světové válce. Mezi tato sídla patří například *Zlín* (v letech 1949–1989 *Gottwaldov*), *Jeseník* (do r. 1947 *Frývaldov*) a další. V rámci unifikace názvů sídel, kdy bylo cílem sídla nazývat neduplicitně v rámci území, docházelo také ke změnám ve smyslu upřesnění názvů. Příkladem je *Světlá nad Sázavou* (do r. 1925 *Světlá*), *Bartošovice v Orlických horách* (do r. 1950 *Bartošovice*), *Horka u Staré Paky* (do r. 1950 *Horka*) a další. Velké vyčleňování částí obcí nastalo po r. 1990, a to například resortním předpisem *Přehled změn v územní organizaci, v názvech obcí a jejich částí* (Ministerstvo vnitra, 1991). Kromě vyčlenění částí obcí tyto resortní předpisy zahrnují i konkrétní změny názvů obcí, např. *Mokrovraty* (do r. 1990 *Mokrá Vrata*). V těchto případech je potřeba dohledat změny názvů v lexikonech obcí, resortních příkazech nebo v jiných relevantních zdrojích.

Jak bylo popsáno výše, geolokace může probíhat v **různé míře podrobnosti**. Stanovení vyžadované podrobnosti geolokace přitom probíhá vždy na základě **jasně vymezeného účelu**, ke kterému budou tato data sloužit. Obecně platným doporučením je, aby metadata audionahrávek obsahovala kompletní původní geolokalizační údaje zaznamenané v době pořízení nahrávky a současně je jim jako další atribut přiřazena nová geolokace dle metodického pokynu pro převod do formátu prostorových dat platného pro konktétní projekt, organizaci apod. Uchování původní informace je velmi důležité, neboť s rozvojem informačních technologií a především s nástupem umělé inteligence může v budoucnu existovat nástroj, který geolokaci provede na základě původních údajů přesněji, než je stanoveno současnou metodikou.

7.1.2 Geolokace nových nahrávek na území České republiky

U nově vytvářených audionahrávek je vhodné předem stanovit, jakým způsobem bude geolokace probíhat. K jedné nahrávce je možné přiřadit i **více lokalizačních atributů**, například současné i minulé bydliště respondenta, místo pořízení nahrávky apod. Je potřeba vymezit, na jaké úrovni podrobnosti bude geolokace probíhat.

Současně používané technologie, kdy většina explorátorů disponuje smart zařízeními s umožněním **zjištění přesné polohy prostřednictvím GPS** (Globální polohový systém, angl. Global Positioning System, GPS), se nabízí reálná možnost zaznamenání přesných zeměpisných souřadnic, ke kterým má být nahrávka lokalizována. Toto je nejpodrobnější možný záznam, který umožňuje nejpraktičtější možné využití do budoucna. Ne vždy je však toto použití možné, ať už z důvodu nedostupného zařízení, nebo z důvodu pořízení nahrávky jinde, než kde je její reálná územní příslušnost (mluvčí s explorátorem spolu hovoří v jiném místě, než kde mluvčí bydlí a kde dlouhodobě žije). V takovém případě je vhodnější uvádět **pojmenování části obce**, které odpovídá nejpodrobněji zpracovávané úrovni prostorových dat.

Aby byla zajištěna jednoznačnost údajů, je vhodné využít existující číselníky. Nejvhodnějším zdrojem takového číselníku je Registr územní identifikace, adres a nemovitostí (RÚIAN). RÚIAN je jedním ze základních registrů veřejné správy, je veřejným seznamem a obsahuje kromě adresních informací také údaje o územních prvcích, územně evidenčních jednotkách a jejich vzájemných vazbách (Český úřad zeměměřický a katastrální, 2024). Z hlediska napojení na existující prostorová data Geoportál Českého úřadu zeměměřického a katastrálního obsahuje vrstvu „Adresní body ČSÚ". Tzv. **adresní body** jsou body v databázi prostorových dat, které reprezentují adresní místa. Adresou se rozumí kombinace údajů název okresu, název obce, název městského obvodu nebo městské části, název části obce nebo v případě hlavního města Prahy katastrální území, číslo popisné nebo evidenční, název ulice, číslo orientační a případně údajů potřebných pro účely poštovních služeb, která jednoznačně určuje adresní místo (Geoportál, 2024).

7.2 Příprava pro interaktivní geovizualizaci

7.2.1 Datové formáty prostorových dat

Datovým formátem se rozumí způsob reprezentace určité informace v elektronické podobě a její následné interpretace. V minulosti byly typickými datovými formáty proprietární, binární formáty konkrétních aplikací jednoho výrobce softwaru. Tyto formáty neměly a doposud často nemají existující nebo přístupnou dokumentaci a práce s daty v jiných aplikacích byla a je omezená. Moderní datové formáty jsou jasně definovány a nejsou svázané s konkrétními aplikacemi. Důraz se klade na dostupnost dokumentace formátu a možnost jejího bezplatného využití. Soubory v těchto formátech lze zpracovávat různými aplikacemi. Národní úřad pro informační a kybernetickou bezpečnost uvádí výčet povolených datových formátů pro příjem dokumentů v elektronické podobě (Národní úřad pro kybernetickou a informační bezpečnost, 2024).

Datová sada je tvořena údaji, které spolu souvisí a budou poskytovány jako jeden celek, tj. v jednom souboru ke stažení. Pokud je obsah datové sady příliš velký, je možné jej rozdělit do více datových sad. Každá pak bude mít svoje distribuce, které se liší pouze ve formátu. U datových sad, kde je důležité přesně informovat o provedených změnách, se doporučuje zveřejnit datovou sadu s iniciálním obsahem a poté se seznamem provedených změn (tj. jaké záznamy byly smazány a jaké byly vytvořeny či aktualizovány a jakým způsobem). Dělení na menší datové sady není vhodné provádět tak, že se jednotlivé datové sady poskytují jako jednotlivé záznamy.

Struktura datového souboru se týká způsobu, jakým jsou data v souboru organizována a uložena. Správně navržená struktura datového souboru je klíčová pro efektivní ukládání, vyhledávání, zpracování a analýzu dat. Při řešení struktury datového souboru se zohledňují: typ souboru (textový, binární, formátovaný), struktura záznamů (tabulková, hierarchická, síťová), metadata (data o datech), datový typ (číselný, textový, booleovský, časový). Každý typ datového souboru má své výhody a nevýhody v závislosti na kontextu použití, požadavcích na výkon a složitosti datových vztahů.

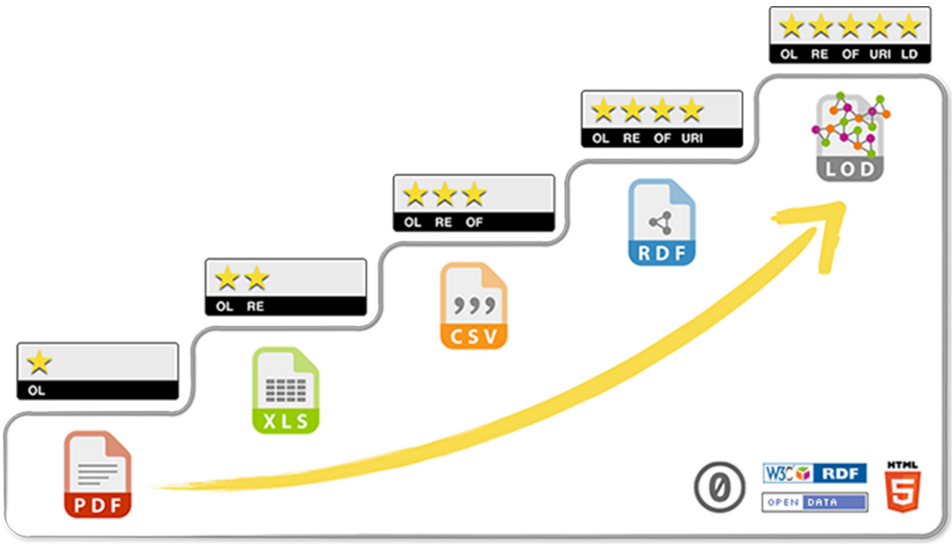
V současné době je i při dialektologickém výzkumu žádoucí, aby se pracovalo v režimu otevřené vědy. **Otevřená věda** (angl. open science) usiluje o snadný, rychlý a bezplatný přístup publikací, dat i jiných výstupů vědeckého bádání v digitální podobě všem zájemcům. **Otevřená data** (angl. open data) jsou (podle § 3 odst. 5 zákona č. 106/1999 Sb., o svobodném přístupu k informacím) informace zveřejňované způsobem umožňujícím dálkový přístup v otevřeném a strojově čitelném formátu, jejichž způsob ani účel následného využití není povinným subjektem, který je zveřejňuje, omezen a které jsou evidovány v národním katalogu otevřených dat.

Dialektologický výzkum často vyžaduje rozsáhlé sběry dat, jako jsou nahrávky, transkripce a analýzy. Otevřená věda zajišťuje, že veškerá data, metodologie a výsledky jsou volně přístupné, což umožňuje ostatním výzkumníkům ověřit a reprodukovat studie. To zvyšuje důvěryhodnost a validitu výzkumu i efektivnější

využití dostupných dat a nástrojů. Díky tomu dialektologové mohou profitovat ze spolupráce s jinými obory. Interdisciplinární přístup usnadňuje výměnu poznatků a metod mezi různými vědeckými disciplínami. Práce s dialektologickými daty v režimu otevřené vědy zvyšuje kvalitu, efektivitu a dopad dialektologického výzkumu, což přispívá k lepšímu pochopení jazykové rozmanitosti a jejímu uchování pro budoucí generace.

Aby mohla být data široce používána, definuje se otevřenost datových sad. Míra otevřenosti dat se vyjadřuje pomocí pěti stupňů otevřenosti (OPENDATA.CZ, 2024):

- Stupeň otevřenosti 1 – datové sady jsou dostupné on-line s jasným vymezením podmínek užití; nejsou kladeny žádné požadavky na datové formáty; zahrnují webové služby OGC WMS a OGC WMTS, které nezpřístupňují vlastní data, ale pouze obrázky generované z těchto dat; tento stupeň není považován za dostatečný stupeň otevřenosti.
- Stupeň otevřenosti 2 – datové sady jsou poskytovány ve strojově čitelném formátu, nejčastěji za účelem zaznamenání určité množiny údajů; jsou používány formáty umožňující co nejsnazší přístup k jednotlivým zaznamenaným údajům pomocí běžných a volně dostupných programovacích prostředků bez nutnosti jakéhokoliv předzpracování, např. jako tabulka nebo textový dokument; data lze poskytovat v komprimovaném tvaru.
- Stupeň otevřenosti 3 – oproti stupni 2 navíc vyžaduje, aby specifikace formátu byla otevřená, tzn. aby byla vyhledatelná a zdarma dostupná v síti WWW a aby existovaly volně dostupné programovací nástroje pro jejich zpracování; příkladem jsou formáty CSV a XML, ne však PDF ani DOC, XLS apod.; pro prostorová data je vhodné zvolit některý z otevřených formátů OGC, jako je GML, GeoJSON, GeoPackage, WKT, ESRI Shapefile.
- Stupeň otevřenosti 4 – pro datové sady je povinnost identifikovat entity, kterých se týkají údaje obsažené v datové sadě, ve tvaru IRI (Internationalized Resource Identifier).
- Stupeň otevřenosti 5 (nejvyšší stupeň otevřenosti) – datové sady splňují standardy propojených dat postavené nad standardy sítě WWW a umožňují vyjadřovat souvislosti mezi různými datovými sadami v podobě strojově zpracovatelných odkazů.



Obrázek 7.1 Stupně otevřenosti datových sad (převzato z OPENDATA.CZ, 2024)

Nastavení struktury datového formátu je zásadní pro efektivní práci s dialektologickými daty a vyžaduje pečlivé plánování a dokumentaci. Dobře navržená struktura usnadňuje správu, analýzu a sdílení dat. Nasta-

vení struktury datového formátu zahrnuje několik kroků, které zajistí, že data jsou organizována a ukládána tak, aby byla snadno použitelná, efektivní a robustní vůči chybám:

- Definování účelu a požadavků – určuje se, jaké informace budou data obsahovat a jaký je jejich účel, jaké jsou technické požadavky na formát, zejména kompatibilita s konkrétním softwarem nebo potřebné rychlosti přístupu.
- Výběr vhodného formátu – vybírá se vhodný datový formát; textové formáty (CSV, JSON, XML) jsou snadno čitelné a kompatibilní s mnoha nástroji, binární formáty (Parquet, Avro) jsou efektivnější pro velké objemy dat a složité struktury, databázové formáty (SQL, NoSQL) umožňují složité dotazy a manipulaci s daty.
- Určení struktury záznamů – definují se atributy a typy dat, u hierarchických dat uzly a poduzly a u objektových dat objekty a jejich vlastnosti.
- Definování datových typů – určují se základní typy (celá čísla, desetinná čísla, text, datum a čas, booleovské hodnoty).
- Stanovení metadat – definují se názvy a popisy polí, jednotky a formáty dat a pravidla validace, např. povolené hodnoty, rozsahy hodnot, pravidla pro unikátnost a povinné položky.
- Validace a integrita dat – řeší se správnost a konzistence dat, požadavky na správnost dat při vkládání nebo načítání, definují se vztahy mezi datovými entitami.
- Optimalizace a indexace – definují se indexy pro rychlejší přístup k datům a kompresní techniky ke zmenšení velikosti souboru a zlepšení výkonu.
- Dokumentace struktury – sestavuje se dokumentace struktury dat se schématy, popisem formátu, datových typů, pravidel validace a dalšími informacemi pro snazší správu, analýzu a sdílení dat.

Prostorová data (geografická data) jsou data, která mají přímou nebo nepřímou souvislost s konkrétními polohami na zemském povrchu. Prostorová data obsahují informace o geografických jevech (objektech a procesech) a jejich umístění v prostoru. Prostorová data mají dvě složky – geometrickou a deskriptivní (atributovou). Geometrická složka nese informace o tvaru a umístění v území (pomocí zvoleného souřadnicového systému), deskriptivní složka obsahuje údaje o vlastnostech (atributech) reprezentovaného jevu, např. název, adresu, výšku budovy, počet pater apod. Rozlišují se dva základní typy prostorových dat – vektorová a rastrová. Vektorová data reprezentují prostorové jevy pomocí bodů, linií a polygonů, případně sítí, povrchů a objemů. Jsou založena na souřadnicovém systému, kde každá geometrická entita má své přesné umístění na zemském povrchu. Rastrová data reprezentují data v mřížce buněk, kde každá buňka má určitou číselnou hodnotu. Typickým příkladem jsou satelitní snímky nebo letecké fotografie. Rastrová data také reprezentují pole hodnot, např. hodnoty nadmořské výšky, teploty vzduchu nebo srážek, kde každá buňka má určitou hodnotu z měření.

Datové formáty prostorových dat jsou speciálně navrženy pro ukládání, správu a analýzu dat, která mají geografickou (prostorovou) složku. Tyto formáty jsou klíčové v geoinformatice, zejména geografických informačních systémech (GIS), dálkovém průzkumu Země a kartografii. Mezi nejběžnější formáty prostorových dat patří:

- GeoPackage – moderní, otevřený, na platformě nezávislý formát, jenž vznikl jako náhrada za Shapefile, je standardem OGC (viz níže). Postavený na souborové databázi a knihovně SQLite, podporující rastrová i vektorová data. Má univerzální použití, je vhodný pro podkladové mapy i tematické vrstvy.
- GeoJSON – formát založený na JSON, který je snadno čitelný jak pro lidi, tak pro stroje; používá se pro reprezentaci jednoduchých geografických prvků (body, linie a polygony) a jejich atributů; je vhodný pro webové aplikace a snadnou integraci s JavaScriptovými knihovnami Leaflet a Mapbox.

- KML (Keyhole Markup Language) – formát založený na XML pro znázorňování geografických dat především v Google Earth a Google Maps; pracuje s body, liniemi, polygony a pokročilejšími funkcemi pro vkládání obrázků a popisů.
- GML (Geography Markup Language) – formát založený na XML a standardizovaný Open Geospatial Consortium (OGC); je velmi flexibilní a rozšiřitelný pro podporu širokého okruhu geografických informací; je vhodný pro výměnu prostorových dat mezi různými systémy a aplikacemi.
- GeoTIFF – rozšíření běžného TIFF formátu pro ukládání rastrových obrázků spolu s prostorovou informací (georeferencí); je používán v dálkovém průzkumu Země a v GIS aplikacích pro ukládání satelitních snímků, leteckých fotografií a dalších rastrových dat.
- ESRI File Geodatabase (.gdb) – proprietární formát vyvinutý společností ESRI pro ukládání velkého množství vektorových a rastrových dat; podporuje složité datové struktury, zejména topologické a síťové.
- PostGIS – rozšíření pro open-source databázový systém PostgreSQL, které přidává podporu pro prostorová data; umožňuje ukládání a analýzu vektorových i rastrových dat přímo v databázi; je velmi výkonným řešením pro prostorové dotazy a analýzy.
- WKT (Well-Known Text) a WKB (Well-Known Binary) – standardizované formáty pro reprezentaci geometrických objektů a pro výměnu prostorových dat mezi různými GIS aplikacemi a databázemi; WKT je textový formát, WKB je binární formát.
- NetCDF (Network Common Data Form) – formát určený pro ukládání a sdílení vícerozměrných časoprostorových datových sad, např. klimatické modely; je hojně používán v meteorologii, oceánografii a klimatologii.
- HDF (Hierarchical Data Format) – formát určený pro ukládání velkých vědeckých datových sad z dálkového průzkumu Země, zejména satelitních snímků.
- Shapefile (.shp) – historicky nejrozšířenější formát pro ukládání vektorových prostorových dat; skládá se z několika souborů, kde hlavní soubor má příponu .shp a obsahuje geometrii, další soubory s příponami .shx a .dbf obsahují indexy a atributová data. Shapefile je příkladem neoficiálního standardu, kdy jej podporují prakticky všechny geografické informační systémy, ale protože byl vyvinut již v 90. letech minulého století, není z řady důvodů vhodný pro použití ve webové kartografii, navíc dnes již není upřednostňován ani pro desktopové nástroje.

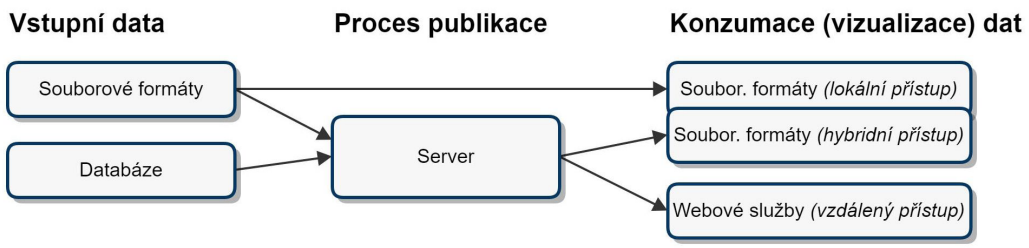
Každý z těchto formátů má své výhody a je vhodný pro různé typy aplikací a analýz prostorových dat. Výběr správného formátu závisí na specifických požadavcích projektu, kompatibilitě s používanými nástroji a požadavcích na výkon.

Webové služby obecně využívají vzdáleného přístupu k datům. Podstatou je architektura klient-server, konkrétně využívají principu SOA (Service Open Architecture), kdy spolu komunikují dva stroje na základě standardizovaných protokolů, založených nejčastěji na jazyku XML. Klient volá vybranou službu zadáním specifické URL s požadovanými parametry, zpět od služby dostává odpověď formou požadovaného výstupu. Klientem/konzumentem webové služby může být webová mapová aplikace, software či jiná webová služba. Díky dodržování standardizovaných rozhraní a protokolů jsou webové služby ukázkou interoperability a nezávislosti na zvolené technologii či platformě.

Ve webové kartografii lze využít univerzálních služeb a protokolů (WSDL, UDDI, REST, SOAP) nebo řady specifických standardů vyvinutých a rozšiřovaných pod hlavičkou sdružení Open Geospatial Consortium (OGC). Konkrétně se jedná o prostorově založené webové služby, jejichž prostřednictvím lze přistupovat ke geografickým datům, prostorovým operacím i aplikacím bez nutnosti lokálního přístupu.

Pro přenos a konzumaci prostorových dat prostřednictvím WWW a následnou vizualizaci se uplatňují tři odlišné **přístupy k datům**:

- lokální – historicky nejstarší přístup je založený na lokálním ukládání dat; při serverovém řešení jsou data umístěna na totožném (vlastním) serveru jako aplikace, v případě intranetové nebo lokální aplikace jsou uložena v přímém dosahu (lokální síť, na disku); nejčastěji lokální přístup využívá souborové formáty dat (např. shapefile); z pohledu aktualizace dat a sdílení mezi více uživateli není lokální uložení souborů efektivní, protože distribuce aktualizované verze vyžaduje doručení ke všem uživatelům, což vyžaduje šíření skrze externí médium (fyzická média CD/DVD/USB či online datová úložiště), z čehož pramení logistické i bezpečnostní nedostatky (možnost zneužití, GDPR); proto je potřeba zohlednit časovou prodlevu od okamžiku publikace skrze zapracování aktualizace na straně klienta až po vizualizaci;
- vzdálený – modernější řešení využívá vzdáleného přístupu, vycházejícího z principu SOA, kdy jsou data uložena na vzdáleném (cizím) serveru, který je dostupný skrze podporovaný protokol v síti internetu; příkladem vzdáleného přístupu je využití dat prostřednictvím webových mapových služeb nebo prostorové databáze dostupné online; zásadní charakteristikou je centrální správa dat, umožňující časově, finančně, logisticky, organizačně i technicky efektivnější správu dat;
- hybridní – je kombinací tradičního lokálního souborového formátu uložení dat a vzdáleného přístupu; u technických, legislativních či bezpečnostních omezení webových služeb se jako typická alternativa nabízí využít lokálního uložení souborů, což však popírá výhody centralizovaného sdílení dat; avšak online přístup skrze protokol HTTP(S) zadáním specifické URL umožňuje i vzdálený přístup; tento centrální přístup vycházející z principu webových služeb umožňuje jejich efektivnější správu a aktualizaci; je však potřeba zohlednit, že vzhledem k jejich datové povaze nelze hovořit o webových službách v pravém slova smyslu – jedná se stále o souborové formáty.



Obrázek 7.2 Schéma prostorových dat v kontextu trojice odlišných přístupů (Nétek, 2020)

7.2.2 Prostorová dialektologická data

Prostorová dialektologická data jsou specializovaným typem dat používaným v dialektologii při studiu geografické distribuce nářečních jevů. Prostorová dialektologická data kombinují lingvistické informace s informacemi o poloze. I prostorová dialektologická data se skládají z geometrické a atributové složky. Geometrická složka definuje tvar a polohu dialektologických jevů v prostoru, k čemuž využívá body (např. pro výzkumné lokality), linie (např. pro hranice regionů) a polygony (např. pro nářeční oblasti). Tyto prvky jsou v mapě lokalizovány pomocí souřadnicových systémů, z nichž je pro online mapy využíván výhradně systém WGS84, který využívá zápisu souřadnice zeměpisné délky a šířky (latitude, longitude), což je považováno za standardizované a univerzální řešení. Atributová složka zahrnuje deskriptivní informace o jednotlivých geometrických prvcích ve formě atributů, jakými jsou identifikátor, název, další popisné informace či jiná tematicky relevantní data. Pro bod reprezentující konkrétní polohu, respektive konkrétní obec, může atributová složka obsahovat název obce či okresu, rok sběru dat, počet informátorů, typ pro-

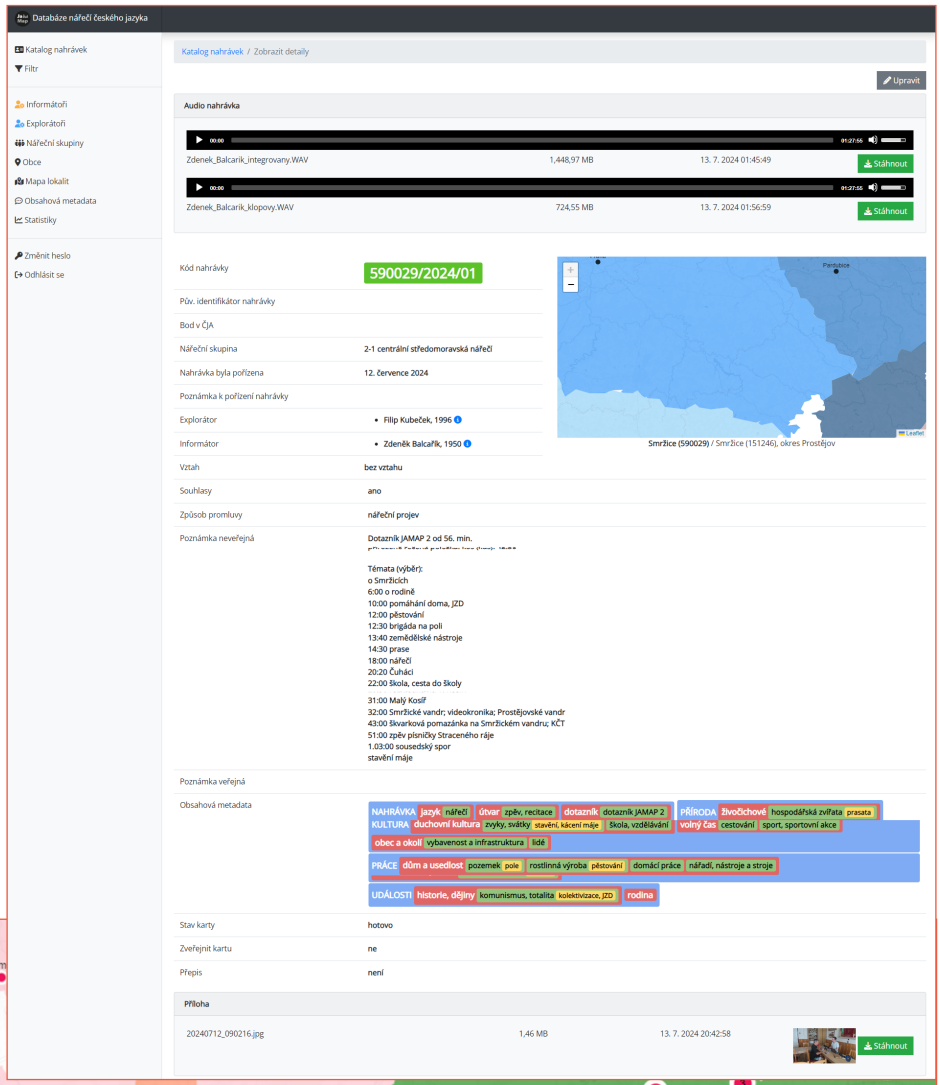
mluvy či příslušnost k nářeční oblasti. Rozšířením atributových dat je navázání na multimediální data. Ta jsou vzhledem ke zcela odlišnému charakteru (formát, velikost, přehrání v případě audiálních dat) uložena samostatně, nicméně atributová složka dat odkazuje na jejich umístění či identifikátor, který zajišťuje jejich implementaci do mapy.

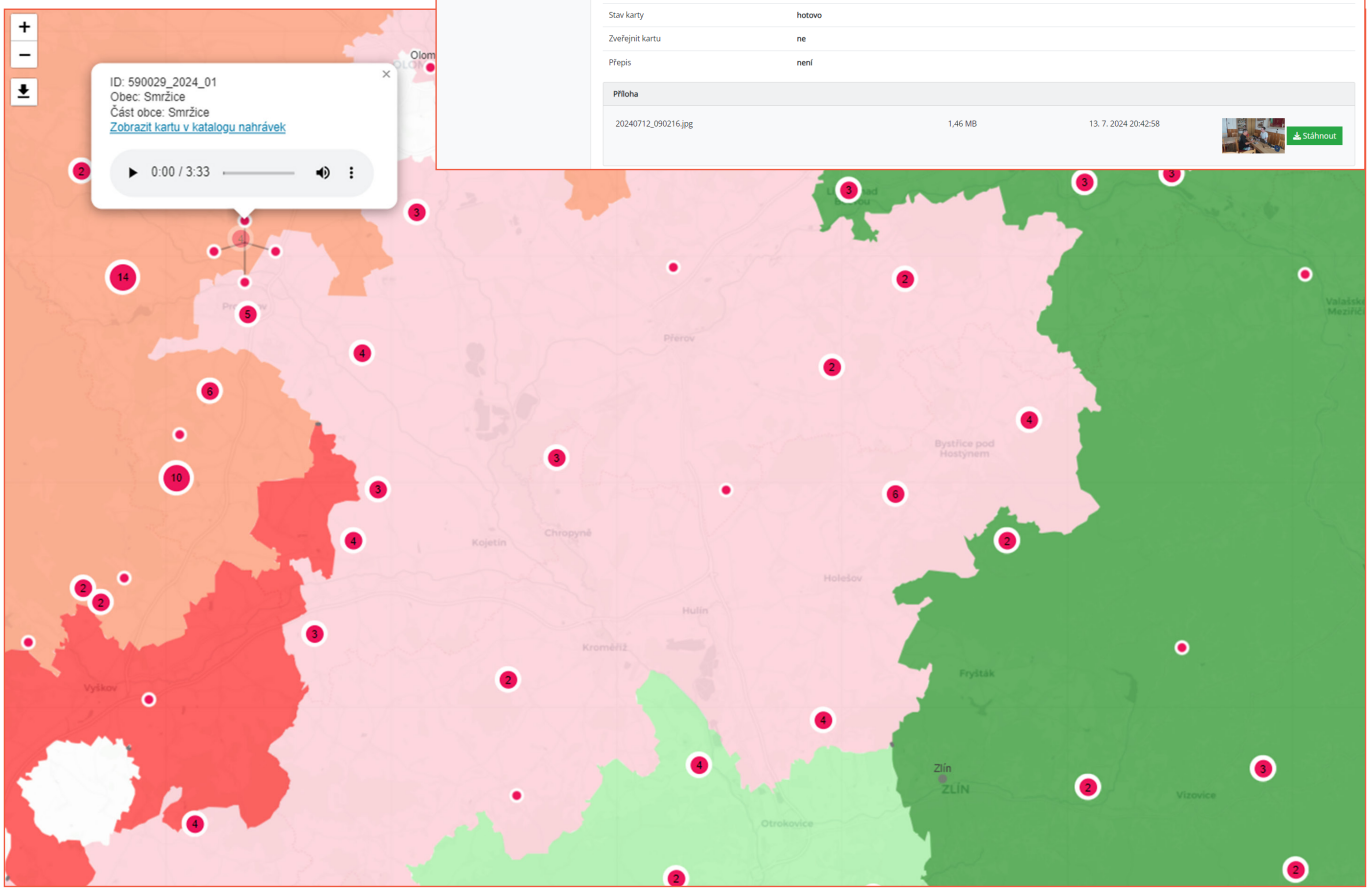
K práci s prostorovými dialektologickými daty se používají různé **datové formáty** a způsoby ukládání. V tabulkových formátech mohou být dialektologická data uložena v tabulkách (např. CSV, Excel), kde řádky reprezentují jednotlivé záznamy a sloupce různé lingvistické a geografické informace. V geografických formátech (Shapefile, GeoJSON, KML apod.) se ukládají geografická data do prostorové složky a lingvistická data do atributových tabulek. V databázových formátech jsou lingvistické a geografické informace organizovány v relačních tabulkách, které jsou propojeny pomocí primárních a cizích klíčů.

Pro prostorová dialektologická data se používají všechny datové formáty uvedené v podkapitole 7.2.1. Pro ukládání zvukových nahrávek se používají audio formáty, např. WAV, MP3. Specifický přístup vyžadují audiální dialektologická data, která kombinují zvukové nahrávky nářečních ukázek s geografickými informacemi o jejich lokalizaci. Použití různých formátů a nástrojů pro ukládání a analýzu těchto dat podporuje komplexní výzkum a aplikace v oblasti dialektologie a geolingvistiky. Geometrickou složku tvoří nejčastěji souřadnice míst, kde byly audiální nahrávky pořízeny, což pomáhá v mapování a analýze nářečí v určitém regionu. Atributovou složku tvoří nejčastěji zvukové nahrávky rozhovorů, výslovností slov, frází nebo celých textů. Do atributů se ukládají i demografické informace o mluvčích (věk, pohlaví, vzdělání, sociální status) pro analýzu vlivu těchto faktorů na nářeční jevy.

Kombinace audiálních a prostorových dat umožňuje studium vzorů a trendů v dialektech a pochopit, jak se zvukové charakteristiky jazyka vyvíjejí a mění v různých geografických a sociálních kontextech.

Obrázek 7.3
Prostorová dialektologická data (vpravo) strukturovaná pro webovou dialektologickou mapu (dole) (čerpáno z Databáze nářečních promluv pro odbornou veřejnost*)*





7.2.3 Metody vizualizace prostorových dat

Podstatou vizualizace prostorových dat je jejich transformace do grafických reprezentací, které umožňují přenos informací obsažených v datech pro jejich pochopení, analýzu a interpretaci. Vizualizace zahrnuje použití map, grafů, diagramů a dalších vizuálních nástrojů k prezentaci geografických vztahů, vzorců a trendů.

Vizualizace prostorových dat je doménou **kartografie**, vědy, technologií a umění tvorby a užití map a mapám příbuzných děl. Vizualizace umožňuje transformaci složitých dat do přehledných a srozumitelných kartografických děl, což usnadňuje identifikaci prostorových vzorů, podporuje prostorové analýzy a zvyšuje angažovanost uživatelů (Kraak a Ormeling, 2010). Díky pokročilým vizualizačním technologiím a nástrojům je možné vytvářet statické i interaktivní mapy a aplikace, které poskytují hlubší vhled do geografických dat a jejich souvislostí. Kvalitní produkty vizualizace zjednodušují složité geografické informace a činí je srozumitelnějšími pro široký okruh uživatelů.

Mezi základní produkty vizualizace prostorových dat patří:

- Mapy – základní nástroj vizualizace prostorových dat znázorňující geografické jevy a jejich atributy; mapy pomáhají identifikovat prostorové vzory a trendy, které mohou být obtížně odhalitelné v datových souborech, a umožňují lepší pochopení prostorových vztahů, zejména závislosti mezi geografickými jevy a sociálními faktory; vedle konvenčních papírových map existují pokročilé digitální kartografické produkty; interaktivní mapy a aplikace umožňují uživatelům prozkoumávat data podle svých potřeb a zájmů prostřednictvím interakce mezi uživatelem a databázemi digitálních dat; dynamické mapy zahrnují vyjádření časových změn; multimediální mapy zprostředkovávají informace uživatelům pomocí zvuku, fotografií, videa a virtuálních scén; webové mapy a mapové služby, např. Google Maps, Leaflet aj., umožňují vytvářet a publikovat interaktivní mapy na internetu prostřednictvím webových aplikací a zpřístupnit jejich obsah širokému publiku.
- Grafy – geometrická znázornění závislosti mezi dvěma nebo více proměnnými, ve kterých je hodnota znázorňovaného jevu závislou proměnnou na jiné nezávislé proměnné; grafy umožňují porovnání geografických informací s jinými metrikami.
- Diagramy – geometrické obrazce se snadno měřitelným parametrem, jehož velikost umožňuje pomocí stupnice určit hodnotu vlastnosti znázorňovaného jevu; na rozdíl od grafu není diagram vázán na souřadnicové osy a neznázorňuje závislost mezi dvěma nebo více proměnnými.

Digitální kartografické produkty jsou úzce spojeny s analytickými nástroji, které umožňují provádět prostorové analýzy, např. výpočty vzdáleností, analýzy hustoty, vyhledávání blízkých objektů aj., a tím pomáhají odpovídat na různé otázky týkající se geografických dat a pomáhají rozhodování na základě prostorových informací.

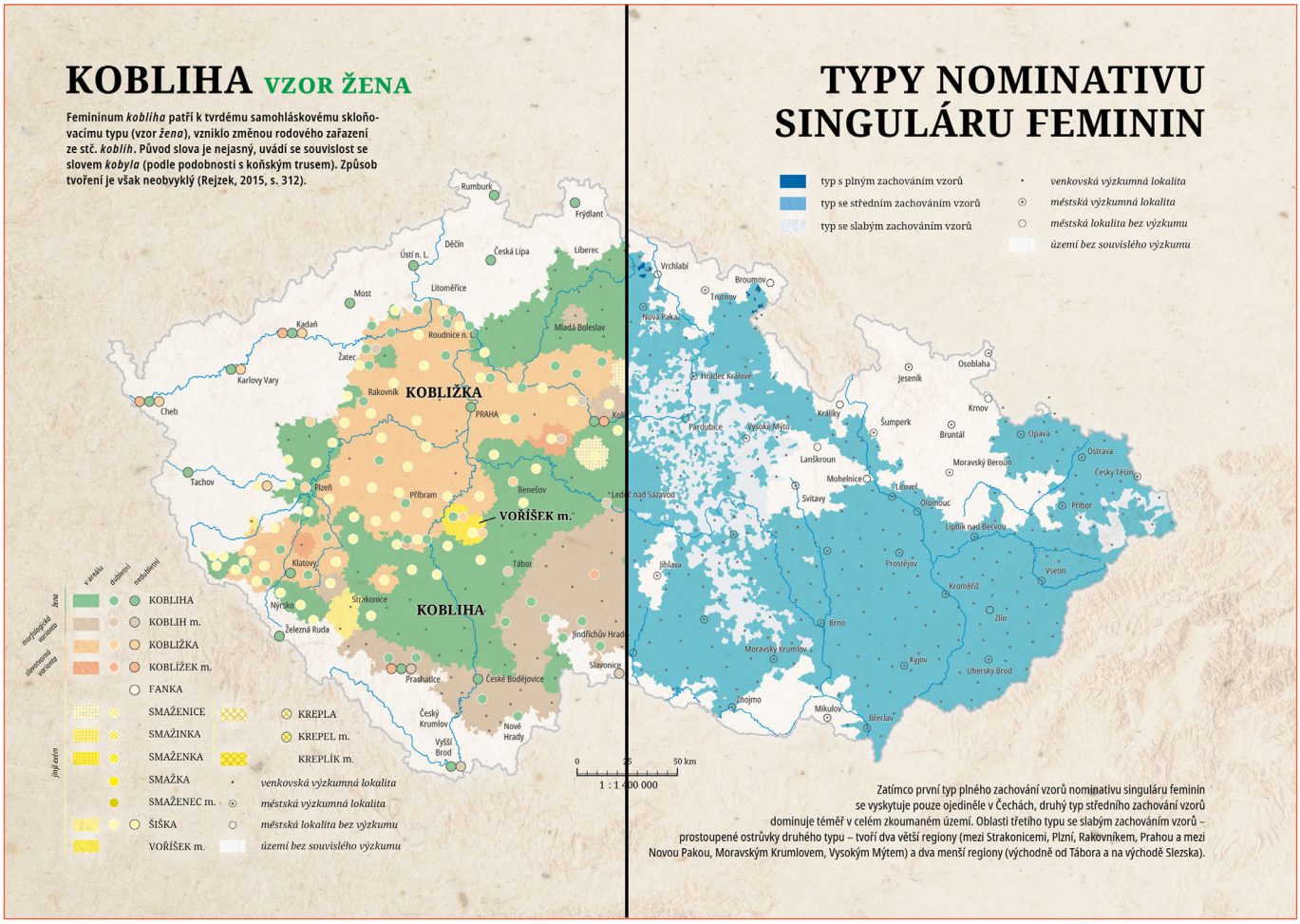
Tematická kartografie je součástí kartografie, která se zabývá tvorbou tematických map. Tematické mapy jsou specifické svým obsahem. V něm převládají prvky jednoho nebo více příbuzných témat nad prvky jinými, které jsou z hlediska zaměření tematické mapy druhořadé. Tento výběr nejvíce ovlivňuje zastoupení fyzickogeografických a socioekonomických prvků v obsahu tematické mapy. Je-li obsah tematické mapy zaměřen na fyzickogeografické prvky, jsou obvykle potlačeny prvky socioekonomické a naopak.

Na rozdíl od obsahu topografických map, jehož prvky se seskupují do polohopisu a výškopisu, se obsah tematických map skládá z tematického obsahu a topografického podkladu. **Tematický obsah** je souhrn prvků obsahu mapy tvořící mapovanou tematiku nebo s ní úzce související. Tematický obsah tvoří hlavní část obsahu tematických map. Tvoří jej jeden nebo více prvků, jimiž mohou být libovolné fyzickogeografické nebo socioekonomické jevy (objekty a procesy). V legendě tematické mapy je tematický obsah umís-

těn na začátku, je náležitě strukturován a je seřazen sestupně od nejdůležitějšího tématu. **Topografický podklad** slouží k prostorové lokalizaci jednotlivých prvků tematického obsahu a k určení jejich vzájemných topologických vztahů. Topografický podklad obsahuje pouze prvky topologicky důležité, zejména vodstvo, komunikace, sídla, politicko-administrativní hranice a prvky s vazbou na tematiku mapy, např. kóty, hranice katastrů atd. Topografický podklad se na různých tematických mapách liší. Jeho základem je říční síť, která tvoří jeho kostru a s výjimkou některých statistických map (kartogramy, kartodiagramy) se vyskytuje téměř na všech tematických mapách. Výběr dalších prvků topografického podkladu závisí na tématu mapy. V legendě tematické mapy je topografický podklad umístěn na její závěr nebo nemusí být vůbec uveden (pokud je jednoduchý a provedený srozumitelnými znaky).

Podle koncepce tematického obsahu vymezují V. Voženílek a J. Kaňok (2011) tři základní **typy tematických map**, a to analytické, komplexní a syntetické:

- **Analytické tematické mapy** obsahují prvky jednoho nebo nanejvýše několika málo témat tak, jak byly zjištěny mapováním nebo analytickými výpočty. Tematický obsah je minimálně zgeneralizován a nevyjadřuje vzájemné vazby jevů přímo, nýbrž se soustřeďuje na „prosté" rozmístění (inventarizaci) vyjádření objektů hlavního tématu mapy. Při vizualizaci tematického obsahu analytických map se nejčastěji používá jedna jednoduchá znázorňovací metoda. Analytickou mapou je například mapa výzkumných lokalit, administrativního členění státu, rozmístění kulturních památek, hustoty zalidnění apod.
- **Komplexní tematické mapy** vyjadřují jevy příbuzného tématu tvořící logický celek, např. městskou dopravu, geologii, obyvatelstvo, kriminalitu. Je pro ně typická precizní strukturovanost tematického obsahu umožňující srovnání vzájemné důležitosti jak jednotlivých objektů téhož druhu, tak různých obsahových prvků mezi sebou. Používají většinou kombinace více jednoduchých znázorňovacích metod. V podstatě představují společné vyjádření obsahů několika analytických map jedné tematiky. Jedná se o velmi užitečný typ tematických map, neboť přinášejí při úspoře místa více informací než několik analytických map dohromady. Díky značné grafické zaplněnosti jsou náročné na sestavení.
- **Syntetické tematické mapy** mají zobecněný tematický obsah (nejčastěji z komplexních map), protože znázorňují v souhrnu více různých jevů s cílem ukázat jejich zásadní souvislosti nebo vztahy. Syntézu těchto jevů znázorňují jako novou kvalitu. Jevy, které by komplexní mapu neúnosně přeplnily, jsou nahrazeny nově definovanými typy, jejichž výskyt se vymezuje obvykle areálovými znaky. Syntetické mapy podávají složitější informace než mapy analytické nebo komplexní, z nichž vznikají syntézou (nejčastěji regionalizací nebo typologizací). Jejich tvorba vyžaduje důkladnou znalost tématu a schopnost provádět jeho syntézu. Čtení syntetických map vyžaduje kvalifikovaného uživatele, protože informace v mapě se vyvozují cestou myšlenkových pochodů, zejména abstrakcí, generalizací a především syntézou vstupních elementárních údajů. Syntetickými mapami jsou například mapy nářečních oblastí, klimatických oblastí, dopravních zón, geomorfologických jednotek nebo hydrologických rajonů.



Obrázek 7.4 Příklady analytické (vlevo) a syntetické (vpravo) dialektologické mapy (Ireinová, Voženílek a kol., 2023; upraveno do koláže)

7.2.4 Multimediální interaktivní vizualizace nářečních nahrávek

Vývoj webových mapových aplikací stojí na dvou základních pilířích: datovém a softwarovém aspektu. Datový aspekt zahrnuje především volbu datových zdrojů, jejich formátů a přístupů k datům. Softwarový aspekt zahrnuje volbu mapové knihovny a funkcionalitu (nástroje, operace a procesy, kterou aplikace poskytuje).

Podkladová mapa vyjadřuje topografický podklad, tedy geografický kontext pro zobrazovanou oblast v mapě. Pokud jsou podkladové mapy dostupné ve standardizovaném formátu webových služeb, lze je použít v různých mapových knihovnách, tedy nejen v knihovně, pro kterou jsou primárně určeny (v praxi téměř výhradně dlaždicové mapové služby všech etablovaných poskytovatelů, např. Esri, Google Maps, Mapy.cz, Mapbox, Here, Bing, OpenStreetMap, ČÚZK atd.). Obsahem tematických vrstev, příp. překryvných vrstev, může být jakýkoliv tematický obsah. Jedná se o hlavní informační obsah celé mapové aplikace dostupné v celém spektru formátu (viz výše).

Základním principem mapových dlaždic je vygenerování originálního datasetu v jednotlivých úrovních měřítek a následně pro každé měřítko rozřezání původně celistvého obrazu do sady dlaždic. Pro využití dlaždic se jako výhodné kartografické zobrazení jeví Web Mercator, neboť po odříznutí polárních oblastí se

mapa celého světa dá zobrazit jako čtverec, což odpovídá dlaždici nulté úrovně. Každá další úroveň rozdělí předchozí dlaždici na čtyři další. Konvencí je velikost jedné dlaždice 256 × 256 pixelů, prakticky majoritním řešením (formátem) rastrových dlaždic je webová služba Web Map Tiles Service (WMTS).

Softwarový aspekt zohledňuje charakteristiky konkrétních mapových knihoven. Pojem software je použit v širším slova smyslu, nejedná se o programy instalované do operačního systému prohlížeče, ale o aplikace běžící v prostředí internetu.

Zvuková mapa

Zvuková mapa je z technologického hlediska příkladem multimediální mapy, která je zpravidla interaktivní, tj. uživatel může svým výběrem měnit například měřítko mapy, zapínat a vypínat tematické vrstvy nebo spouštět přehrávání lokalizovaných dialektologických nahrávek. Multimediální mapa obecně pracuje s více než dvěma „médii", mezi které se řadí animace, zvuk, video, text, grafika a virtuální scéna. Multimediální mapy mohou být vytvořeny mnoha způsoby, avšak v současné době jednoznačně dominuje tvorba webových multimediálních map. Publikování takových map, tedy jejich zpřístupnění široké skupině uživatelů, je mnohem jednodušší nežli tvorba složitých desktopových aplikací (Nétek, 2021). To, že moderní technologie a prostředí internetu přinášejí možnost obohacení dialektologických map o audio- nebo videonahrávky, popisuje ve své práci S. Rabanus (2020).

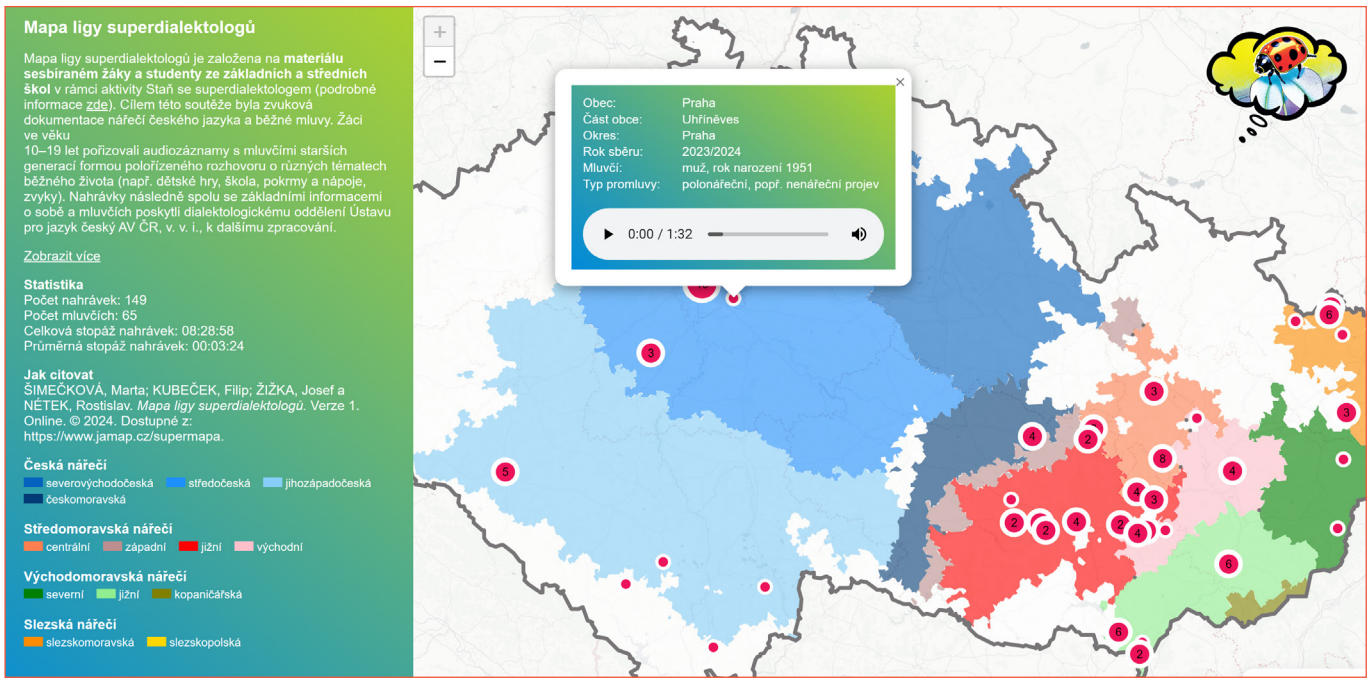
PŘÍKLAD Z PRAXE

Příkladem jedné z prvních internetových zvukových map je mapa *Sounds Familiar? Accents and Dialects of the UK* (Moving Image Gateway, 2024), vytvořená Britskou knihovnou v roce 2007 (mapa již byla umístěna do archivu webových stránek a nefunguje správně). Aktuálně dostupná je z hlediska anglických nářečí webová mapa *Sound Map* (My Learning, 2024), vzniklá v rámci projektu Dialect and Heritage. Současně existují další projekty na mapování anglických nářečí, například *International Dialects of English Archive* (IDEA, 2024), kde je dostupná mapa celého světa s lokalizací nahrávek podle místa pořízení nahrávky, kdy je možné po rozkliknutí zjistit věk a pohlaví mluvčího, včetně místa narození. Následně se po kliknutí na název nahrávky přehraje audionahrávka. Zvukové mapy vznikají po celém světě a zpravidla se věnují jednomu konkrétnímu území nebo jazyku.

Nářeční zvukové mapy jsou vzácným a neocenitelným nástrojem pro lingvisty, historiky i širokou veřejnost zajímající se o vývoj a rozmanitost jazyků. Umožňují detailně sledovat geografické rozšíření a variabilitu jednotlivých dialektů, odhalovat jemné nuance ve výslovnosti, slovní zásobě a gramatice, které jsou jinak snadno přehlédnutelné. Tyto mapy poskytují jedinečný vhled do jazykové krajiny, ukazují, jak se regionální identity odrážejí v mluveném jazyce, a pomáhají zachovávat dialekty, které by jinak mohly upadnout v zapomnění. Zvláště v případě slovanských jazyků, které se vyznačují bohatou nářeční rozmanitostí, jsou takové mapy nezbytným nástrojem pro pochopení a zachování jazykového dědictví. Pro česká nářeční data existuje pouze jediná online platforma Mapka (Goláňová a kol., 2023), vytvořená pro korpusy mluvené češtiny, které jsou součástí Českého národního korpusu. Ačkoliv poskytuje určité informace o českých nářečích, její využití je značně omezené. Obsahuje jen malé množství nahrávek, což neumožňuje komplexní analýzu a srovnávání dialektů. Navíc je tato platforma překonaná a neodpovídá moderním trendům po vizuální ani funkční stránce, což negativně ovlivňuje celkový dojem a efektivitu práce s daty.

Vzhledem k těmto nedostatkům bylo vytvořeno vlastní řešení, které překonává zmíněné nedostatky, ale také nabídne modernější, uživatelsky přívětivější a technologicky pokročilejší platformu pro studium a porovnávání českých nářečí. **Mapa ligy superdialektologů** (Šimečková a kol., 2024) je interaktivní nástroj pro prezentaci dialektologického výzkumu realizovaného v projektu JAMAP, jehož cílem bylo zdokumentovat

regionální rozdíly v češtině za pomoci žáků základních a středních škol prostřednictvím audiozáznamů v jim přívětivém a dostupném prostředí interaktivní multimediální mapy (o sběrové akci viz 3.2.2). Tato webová zvuková mapa využívá kombinaci technologií HTML5 (Hypertext Markup Language), CSS3 (Cascading Style Sheet) a JavaScript. HTML5 se využívá pro strukturování webové stránky a integraci audiozáznamů, CSS pro stylování mapy a uživatelského rozhraní a pro zajištění responzivního designu a JavaScript obstarává interaktivní funkce (vyskakovací okna), vč. interaktivní manipulace s obsahem stránky a multimédii (audiozáznamy). Prostorová vizualizace je postavena na knihovně Leaflet, díky které lze přidávat prostorové tematické vrstvy do mapy, vč. markerování poloh nahrávek. Mapová knihovna Leaflet dále zajišťuje standardní funkcionalitu pro přiblížení umožňující lepší uživatelský zážitek při práci s mapou, je navržená s ohledem na jednoduchost a výkon, nativně podporující mobilní zařízení. Leaflet nativně podporuje formáty WMS, GeoJSON, vektory a mapové dlaždice, podporu dalších formátů zajišťují pluginy (KML, CSV, WKT, TopoJSON, GPX). Spolu s OpenLayers a Google Maps API se jedná o nejpopulárnější mapové knihovny jazyka JavaScript, které využívá většina trhu.



Obrázek 7.5 Uživatelské rozhraní Mapy ligy superdialektologů *(Šimečková a kol., 2024)*

7.2.5 Požadavky na tvorbu multimediální interaktivní nářeční mapy

Tvorba interaktivní multimediální nářeční mapy stojí na několika klíčových požadavcích – datových, technologických, softwarových a uživatelských.

Datové požadavky

První skupina požadavků zajišťuje pro výsledný produkt **kvalitní datovou základnu**. Ta zahrnuje:

- předpřipravená audiální dialektologická data (nahrávky), metadata o lokalitě, explorátorovi, informátorovi, o způsobu pořízení dat aj.; nahrávky musí být kvalitní a poskytovat dostatečné množství příkladů pro každý zkoumaný dialekt, aby bylo možné je efektivně vyhledávat a analyzovat;
- předpřipravená geografická data tematického obsahu, který tvoří hlavní obsah mapy, např. body výzkumu, vygenerované areály výskytu variant, dublety aj., vč. doplňujícího tematického obsahu, např. nářeční oblasti, etnografické oblasti apod.;
- předpřipravená geografická data topografického podkladu, např. řeky, hranice, sídla atd.

Technické požadavky

Požadavky na technickou infrastrukturu, tzv. back-end, se zaměřují u interaktivních map na **serverovou infrastrukturu**, která zajišťuje její funkčnost a spolehlivost. Klíčové je rozhodnutí, zda bude použit vlastní server, nebo sdílený webhosting, vždy je však potřeba zajistit minimální parametry pro podporu nejnovějších funkcí a zabezpečení. Doporučuje se použití PHP verze 8.0 a vyšší, přičemž kapacita úložného prostoru by měla být minimálně v řádu stovek megabajtů. K tomu je nezbytný **FTP (File Transfer Protocol) přístup na server**, který umožňuje nahrávání a správu souborů. Doporučit pro tvorbu multimediálních interaktivních map lze nástroje Filezilla nebo WinSCP. Dalším kritickým požadavkem je zajištění (neprostorové) **databáze**. Mezi vhodné databázové systémy patří MySQL, MariaDB nebo PostgreSQL, které jsou schopny efektivně uchovávat a zpracovávat rozsáhlé množství databázových dat potřebných pro provoz nářeční mapy. K technickým požadavkům patří i **uživatelské rozhraní pro vstup do databáze**, například phpMyAdminer, které umožňuje administrátorům snadno spravovat databázová data. Pro zajištění bezpečnosti dat a komunikace mezi uživateli a serverem je nezbytné implementovat na serveru **SSL certifikát** renomované certifikační autority (např. Lets Encrypt), který umožní provoz zabezpečeného protokolu HTTPS. Tento certifikát zajišťuje šifrování dat přenášených mezi serverem a klienty, čímž chrání citlivé informace před potenciálními útoky a zneužitím, zároveň je nezbytný pro využití funkcionality geolokace v mapové aplikaci.

Programové požadavky

Programové (softwarové) požadavky se týkají zejména té části softwarové aplikace, kterou uživatelé přímo interagují, tzv. front-end. Zahrnuje vše, co uživatelé vidí a používají v prohlížeči, včetně designu, struktury, stylování a interaktivních prvků. Vývojáři používají nejčastěji kombinaci technologií HTML5, CSS3 a JavaScript k vytvoření vizuálně atraktivních a funkčních uživatelských rozhraní dynamických a responzivních webových stránek. V neposlední řadě je nutné zvolit adekvátní mapovou knihovnu, jako je Leaflet, OpenLayers nebo GoogleMaps API, pro efektivní zobrazení a interakci s geografickými daty.

Pro front-end interaktivní multimediální nářeční mapy jsou klíčové softwarové požadavky, které zajistí moderní a **uživatelsky přívětivé rozhraní**. Podpora HTML5, CSS3 a JavaScriptu je nezbytná pro vytváření dynamických a responzivních webových stránek, tedy stránek, které se automaticky přizpůsobují různým velikostem a typům obrazovek bez ohledu na typ zařízení. Jejich rozvržení, obrázky, text a další prvky se dynamicky mění a optimalizují pro nejlepší uživatelský zážitek na různých zařízeních (vedle počítačů a notebooků nejen mobilní zařízení, tj. chytré telefony a tablety, ale také smart TV, terminály apod.). Responzivní design zajišťuje, že stránka zůstane snadno čitelná a použitelná bez ohledu na zařízení. HTML5 umožňuje strukturovat obsah webu, CSS3 poskytuje pokročilé možnosti stylování a JavaScript dodává interaktivitu, což dohromady tvoří základ moderního webového vývoje. Uživatelské rozhraní multimediální mapy musí být intuitivní a snadno ovladatelné. Musí umožňovat uživatelům snadné přiblížení a oddálení obsahu mapy, vyhledávání konkrétních prvků podle zvolených kritérií (lokalit, nahrávek aj.) a přehrávat dostupné audiální záznamy. Uživatelské rozhraní musí zajistit integraci multimediálního obsahu a ostatních prvků obsahu mapy, tedy nejen zvukové nahrávky, ale i doplňující textové, případně též obrazové materiály.

Obrázek 7.6 Princip responsivního designu (Webdevel, 2024)

Do programových požadavků náleží i volba adekvátní **mapové knihovny**, která se používá pro vykreslování a interakci s geografickými daty. Mezi vhodné volby patří Leaflet, OpenLayers nebo GoogleMaps API. Každá z těchto knihoven nabízí různé funkce, možnosti a podporované formáty dat, které umožňují efektivní zobrazení a manipulaci s prvky v mapě a zároveň poskytují uživatelsky přívětivé nástroje pro práci s celou mapou.

Podpora **verzování a automatického zálohování** je rovněž klíčová pro zajištění kontinuity a bezpečnosti dat. Verzovací systémy, nejčastěji Git, umožňují sledovat změny v kódu a usnadňují spolupráci mezi vývojáři. Automatické zálohování zajišťuje, že všechna data a kód jsou pravidelně zálohovány, což minimalizuje riziko ztráty dat v případě technických problémů nebo chyb při vývoji. Tyto prvky dohromady zajišťují, že front-end interaktivní mapy je nejen funkční a esteticky příjemný, ale také robustní a bezpečný.

Uživatelské požadavky

Současné postupy ve vývoji koncových aplikací se neobejdou bez uživatelských požadavků, kterými jsou specifikace a očekávání uživatelů ohledně funkcionality, použitelnosti a výkonu aplikace. Funkcionalita zahrnuje to, co uživatelé chtějí, aby aplikace dělala, včetně specifických funkcí a schopností, např. vyhledávání, filtrace, interaktivní mapy a multimediální obsah. Použitelnost se týká snadné a intuitivní práce s aplikací, což zahrnuje design uživatelského rozhraní, navigaci, responzivní design a přístupnost pro různé uživatele, včetně těch se specifickými potřebami. Výkon se zaměřuje na rychlost a efektivitu, s jakou aplikace funguje, aby poskytovala plynulý a bezproblémový uživatelský zážitek.

Pro úspěšnou tvorbu a implementaci interaktivní multimediální nářeční mapy je klíčové zajistit spolehlivou **komunikaci mezi dialektology, kartografy a vývojáři**. Efektivní spolupráce mezi těmito odborníky umožní přesné a relevantní zpracování jak dialektologických, tak geografických dat. Klíčové je mít k dispozici **skupinu uživatelů pro uživatelské testování**, která poskytuje zpětnou vazbu k identifikaci a řešení případných problémů s použitelností a funkčností aplikace pro lepší celkový uživatelský zážitek.

Procesní postup

Procesní postup při vývoji webové koncové aplikace představuje systematický přístup k plánování, vytváření, testování a nasazení webové aplikace. Tento postup zahrnuje několik fází, které jsou nezbytné pro efektivní a úspěšný vývoj. V první fázi, **shromažďování požadavků**, se identifikují potřeby a očekávání uživatelů a zainteresovaných stran. Tento krok zahrnuje analýzu trhu, rozhovory s uživateli a vytváření specifikací. Shromažďování požadavků je klíčovým krokem, který zajistí, že vývoj multimediální interaktivní mapy bude odpovídat potřebám a očekáváním všech uživatelů a že výsledný produkt bude funkční, uživatelsky přívětivý a technicky proveditelný. Ve druhé fázi, **plánování a návrh**, se vypracuje architektura aplikace, návrh uživatelského rozhraní a databázový model. Vývojový tým vytvoří wireframy a prototypy, které pomohou vizualizovat konečný produkt. **Vývoj** je hlavní fází, během níž se aplikace programuje. Front-end a back-end vývojáři pracují na jednotlivých komponentech aplikace podle stanoveného plánu. Kód se pravidelně testuje a integruje, aby se zajistila jeho funkčnost a kompatibilita. Ve fázi **testování a validace** se aplikace podrobuje důkladnému testování, aby se odhalily chyby a problémy s použitelností. Testují se různé scénáře, zahrnující funkční testy, výkonové testy a bezpečnostní testy. Po úspěšném testování přechází postup do fáze **nasazení aplikace** na produkční server a zpřístupnění uživatelům. Tento krok zahrnuje konfiguraci serveru, nastavení databáze a zajištění potřebné infrastruktury. V závěrečné fázi, **údržba a aktualizace**, se aplikace pravidelně monitoruje a aktualizuje, aby se zajistila její bezpečnost, stabilita a přidávání nových funkcí na základě zpětné vazby od uživatelů. Tento kontinuální proces zajišťuje, že aplikace zůstává relevantní a uživatelsky přívětivá po celou dobu svého životního cyklu.

V případě multimediální interaktivní mapy probíhá procesní postup podle výše uvedeného obecného přístupu:

1 **shromažďování požadavků na multimediální interaktivní mapu** zahrnuje několik klíčových kroků a aktivit, které zajišťují, že výsledný produkt bude odpovídat potřebám uživatelů a zúčastněných stran:

   1.1 **identifikace zainteresovaných stran** – určení všech osob a organizací, které budou mapu používat nebo které jsou na ní závislé, zejména dialektologové, kartografové, výzkumníci, studenti a další odborníci;

   1.2 **analýza uživatelských potřeb** – zjištění požadavků a očekávání od uživatelů prostřednictvím rozhovorů, dotazníků, workshopů a pozorování, vč. pochopení, jaké informace a funkce jsou pro uživatele nejdůležitější:

   - definování datových požadavků;
   - definování technických požadavků;
   - definování programových požadavků;
   - definování uživatelských požadavků.

2 **plánování a návrh** je souborem několika klíčových aktivit a kroků, které zajišťují, že výsledný produkt bude efektivně splňovat stanovené požadavky:

   2.1 **časový harmonogram** – definování jednotlivých fází projektu, vč. milníků, termínů a odpovědností členů;

   2.2 **sestavení týmu** – určení a přiřazení rolí a odpovědností v rámci týmu, vč. vývojářů, designérů, testerů a dalších odborníků;

   2.3 **definování architektury systému a návrh uživatelského rozhraní** – volba serverové architektury, databázového systému, integračního rozhraní pro front-end a back-end, návrh vizuálního vzhledu a funkčnosti mapy (interaktivní prvky, navigace, zobrazení multimediálního obsahu).

3 **vývoj aplikace** představuje soubor kroků a aktivit, během nichž se navržené koncepty a plány přeměňují ve funkční mapovou aplikaci:

3.1 **nastavení vývojového prostředí** – vytvoření a konfigurace vývojového prostředí, vč. instalace potřebných softwarových nástrojů, rámců a knihoven (vývojové servery, databázové servery a nástroje pro správu verzí aj.);

3.2 **vstup dat** – získávání, připravování a ukládání dat, která budou mapovou aplikací zpracovávána a využívána, vč. kontroly a čištění dat, validace formátů a struktur dat:

- příprava vstupních dat podle datových požadavků (viz výše) v pořadí: (i) audiální dialektologická data, (ii) geografická data tematického obsahu, (iii) geografická data topografického podkladu;
- transformace výše uvedených předpřipravených dat do formátů vhodných pro interaktivní vizualizaci – GeoJSON pro tematická data, topografický podklad a metadata, mapové dlaždice pro topografický podklad;

3.3 **programování** – zahrnuje řadu klíčových aktivit, které společně vedou k vytvoření funkční a uživatelsky přívětivé aplikace:

- zajištění technických požadavků (viz výše)
  - vytvoření základní struktury mapy v HTML dokumentu s podporou jazyka JavaScript;
  - nasazení vhodné mapové knihovny umožňující standardní geoinformatické nástroje;
  - implementace standardních uživatelských nástrojů (přiblížení/oddálení, posun mapy, vyskakovací okna aj.);
  - implementace vlastní specifické uživatelské funkcionality (vyhledávání ve vlastních tematických datech);
  - propojení prostorové, atributové a multimediální složky do funkčního celku;
- návrh a implementace grafického stylu (ikony, barvy, logo, celkový vizuální styl aj.).

4 **testování a validace** v iterativním vývoji multimediální interaktivní mapy zajišťují, že mapa je funkční, spolehlivá a uživatelsky přívětivá:

4.1 **testování** – může zahrnovat různé druhy testů, např. jednotkové testy (testování jednotlivých komponent aplikace, aby se ověřilo, že každá část funguje správně izolovaně), integrační testy (k ověření, že jednotlivé komponenty aplikace správně spolupracují), funkční testy (zda aplikace plní všechny požadované funkce a specifikace, např. uživatelské rozhraní, navigaci, interaktivitu map a přehrávání zvukových nahrávek), výkonové testy (měření, jak aplikace reaguje pod zátěží, aby se zajistilo, že bude schopna obsloužit očekávaný počet uživatelů a objem dat), bezpečnostní testy (k ověření, že je aplikace odolná vůči různým bezpečnostním hrozbám); nejdůležitější je **uživatelské testování**, které zapojuje skutečné uživatele do testování, aby se ověřilo, že aplikace splňuje jejich potřeby a očekávání – uživatelé testují aplikaci v reálných scénářích a poskytují zpětnou vazbu;

4.2 **validace dat** – je kontrola správnosti a konzistence dat v aplikaci pro zajištění, že geografická data, audiální nahrávky a další obsah jsou správně integrovány a zobrazovány;

4.3 **sestavení dokumentace** k aplikaci, vč. uživatelského manuálu a rad pro případné problémy s aplikací – po celkové **harmonizaci** finálního mapového produktu je nezbytné sestavit kompletní dokumentaci jak z pohledu vývojářů, tak i uživatelů;

4.4 principem **iterativního vývoje** je vývoj v dílčích cyklech – verzích, kdy každá dílčí verze je otestována, zvalidována a zdokumentována (viz výše) tak, aby celý proces vývoje mohl efektivně pokračovat v dalším cyklu.

5 **nasazení** multimediální interaktivní mapy zahrnuje několik činností, které zajišťují, že aplikace bude úspěšně uvedena do produkčního prostředí a bude dostupná uživatelům:

5.1 **příprava produkčního prostředí** – konfigurace produkčních serverů (webových a databázových), zajištění jejich dostatečné kapacity pro očekávaný počet uživatelů a objem dat, implementace bezpečnostních opatření (firewally, SSL certifikáty aj.);

5.2 **migrace dat** – přenos všech potřebných dat z vývojového a testovacího prostředí do produkčního prostředí a zajištění integrity a konzistence dat během migrace a po ní;

5.3 **nahrání aplikace** – nahrání kódu a souborů aplikace na produkční servery – spuštění skriptů pro inicializaci databáze a dalších služeb;

5.4 **monitorování a optimalizace** – nastavení nástrojů pro monitorování výkonu a dostupnosti aplikace, průběžné sledování provozu aplikace, identifikace a řešení případných problémů, optimalizace výkonu na základě reálného používání a zatížení;

5.5 **zálohování** – nastavení pravidelných zálohovacích procesů pro data a kód aplikace;

5.6 **publikování mapy** – zaindexování webového produktu pro dohledatelnost ve webových vyhledávačích (Seznam, Google) a definování marketingové strategie (výběr médií, distribuce, formáty atd.).

6 **údržba a aktualizace** multimediální interaktivní mapy zajišťují, že aplikace zůstane funkční, bezpečná a relevantní pro uživatele:

6.1 **pravidelné monitorování** – neustálé sledování výkonu a dostupnosti aplikace v kontextu návštěvnosti pomocí monitorovacích a analytických nástrojů a identifikace a řešení problémů s výkonem nebo dostupností, např. zpomalení načítání, výpadky serverů, problémy s databází;

6.2 **bezpečnostní aktualizace** – pravidelné aktualizace softwarových komponent a bezpečnostních opatření na ochranu proti novým hrozbám;

6.3 **přidávání nových funkcí** – rozšíření mapy o nové funkce a vylepšení na základě zpětné vazby od uživatelů a na základě jejich nových požadavků;

6.4 **aktualizace dokumentace** – udržování technické dokumentace a uživatelských příruček, aby odrážely změny a nové funkce v aplikaci;

6.5 **uživatelská podpora a školení** – poskytování podpory uživatelům prostřednictvím helpdesku a nabízení školení či vzdělávacích materiálů.

7.3 Shrnutí

Cílem této kapitoly bylo představit možnosti geolokace, tedy přiřazení prostorové složky nářečním nahrávkám, a to jak u nově vytvářených nahrávek, tak také nabídnout možnosti přiřazení geografické polohy nahrávkám již existujícím a starším. Dále bylo cílem představit existující datové formáty prostorových dat, jejich specifika a zásady výběru. Třetím dílčím tématem je představení metod vizualizace prostorových dat se zaměřením na tvorbu map s nářeční problematikou. Zájemce o danou problematiku tak získá představu o tom, jaká prostorová data je potřeba sbírat, jakým způsobem je možné je zpracovat a jaké se nabízejí geovizualizační výstupy.

Srovnání novosti postupů

8

SROVNÁNÍ NOVOSTI POSTUPŮ

Metodika sestává ze dvou základních částí, a to jednak z detailně popsaných strategií pro **sestavení souboru nářečních dat** optimalizovaných pro strojové učení (kapitola 3, 4), jednak z popisu postupů vedoucích ke **zpracování těchto dat** pomocí umělé inteligence (kapitola 5, 6) a metodami geoinformatiky za účelem jejich geolokace a geovizualizace (kapitola 7). Na řešenou problematiku je tak nahlíženo komplexně, od počátku celého procesu po jeho završení, díky čemuž lze metodiku označit za jedinečnou nejen v českém prostředí, ale také v zahraničí.

Středobodem metodiky jsou jazyková data rozrůzněná regionálně. Mohou být přitom charakteru audiálního, textového, popř. kombinovaného (v případě, že je ke zvukovému záznamu zhotoven přepis). Pořízení **audiálních dat** bylo dosud vázáno na zapojení již existujícího zvukového archivu, u nářečních dat jsou ale takové archivy většinově nedostupné; díky metodice se však otevírá cesta k tvorbě vlastních sbírek s kvalitním obsahem, a to prakticky komukoli, včetně zájemců mimo obor dialektologie. V kapitole 3 je dosud nejkomplexněji propracována **metoda osobního přímého rozhovoru v dialektologickém zkoumání**, jehož cílem je získání nářečních zvukových záznamů. Je v ní navržena řada strategií pro navození nářečního kódu u mluvčího (např. metoda zprostředkovaného explorátora, metoda skupinového rozhovoru, metoda akomodace, metoda reminiscence), dále jsou v ní představeny různé techniky kladení otázek (včetně práce s egodokumenty nebo fotointerview) nebo potřebné úkony ze strany explorátora podle fází rozhovoru, včetně těch vlastnímu rozhovoru předcházejících nebo po něm následujících (úprava prostoru, příprava nahrávacího zařízení, problematika odpovědnosti vůči jazykovému společenství aj.). Metoda rozhovoru je běžně využívána i v jiných humanitních a sociálněvědních disciplínách, např. v sociologii, etnologii, antropologii, historii nebo psychologii. Vybrané strategie, jež jsou součástí příslušné kapitoly, tak byly částečně čerpány z již publikovaných prací, přičemž šlo většinově o literaturu mimo okruh dialektologie, potažmo lingvistiky (s výjimkou sociolingvistiky, v jejímž rámci jsou některé metody propracovány o něco více než v jiných jazykovědných oblastech, a korpusové lingvistiky, v níž je metoda užívána při sběru dat pro mluvené korpusy). V českém prostředí mají k popsaným postupům nejblíže metodické příručky J. Chromého (2014, 2021) a vybrané části publikace M. Šimečkové (2024a). U většiny zdrojů však platí, že v nich předložené postupy jednak nejsou podány natolik zevrubně, jako je tomu v této metodice, jednak – a to především – jsou v nich podané návody optimalizovány majoritně na jiná data než jazyková, popř. regionálně rozrůzněná. Zde zpřístupněné postupy jsou přitom založeny na bohaté zkušenosti odborníků z dialektologického oddělení Ústavu pro jazyk český AV ČR, v. v. i., s intenzivními nářečními výzkumy napříč českojazyčným územím; výklad je tak na mnoha místech opřen nejen o příklady dobré praxe, ale také o ukázky praxe chybné či zastaralé jakožto varování před případnými nežádoucími kroky.

Praktickou součástí metodiky je mj. uvedení do základních technik nahrávání, etické a právní problematiky (včetně uchovávání citlivých a osobních údajů, anonymizace, informovaného souhlasu či mikroetiky oboru), vytvoření archivu, přípravy textových dat v podobě transkriptů nebo sběru formou občanské vědy, popř. transfer cizích dat do vlastního archivu. I zde platí, že postupy představené v metodice jsou buďto vylepšené (minimálně ve smyslu aplikování rad na data pořízená v českém prostředí, např. zohlednění české legislativy u tajného pořizování nahrávek), nebo zcela nové (např. strukturace digitálního nářečního archivu včetně systému osobních, geolokačních či obsahových metadat, čerpaná ze zkušeností s budováním unikátní *Databáze nářečních promluv pro odbornou veřejnost* nebo sběr dat prostřednictvím veřejnosti v souvislosti s akcí *Staň se superdialektologem*).

Rovněž u **textových dat**, pojednaných v kapitole 4, je metodika soustředěna na popis jejich sběru, konkrétně na jejich výběr, prioritizaci a získání pomocí **digitalizace tištěných nářečních textů**. U daných textových zdrojů nebyl postup digitalizace dosud publikován, a to ani v zahraničí, neboť jde o převod – oproti digitalizaci textů spisovných – velmi specifický. Popsána je situace, kdy digitalizátor nemá k dispozici použitelný nářeční slovník, který by automaticky korigoval optické rozpoznání znaků. Představený způsob digitalizace vychází výhradně z postupů, které autor dané kapitoly vyvíjel v dialektologickém oddělení ÚJČ AV ČR od roku 2017 a které nejsou známy ani digitalizačním firmám. V nich řeší dané problémy klasickým proofreadingem, který je časově i finančně náročný, navíc bývá nutně zajišťován korektory, kteří daný dialekt a transkripci neznají, tudíž jim řada chyb unikne a další do textů vnáší nově. Metodika umožňuje zvládat tutéž činnost bez korektur, přičemž práci zvládne jediný člověk, nadto za výrazně kratší čas a v podobné kvalitě.

S digitalizací úzce souvisí **normalizace textu**. Při pořizování zápisů lze vycházet z *Pravidel pro vědecký přepis dialektických zápisů českých a slovenských* (Hála, Vážný a kol., 1943; rozšířená verze 1951), která ovšem v praxi vedla jen k částečnému sjednocení dialektologických transkriptů. Metodika nabízí soustavu znaků aktualizovanou pro potřeby strojového učení, a to nejen pro přepisy dialektologické, ale také folklorní (pro ty dosud žádná pravidla vydána nebyla). Představené postupy normalizace přitom nejsou direktivním předpisem, který by měl být dodržen (a nebude možné jej vynutit stejně jako zmiňovaná *Pravidla*), nýbrž funkčním návodem, jak rozmanité folklorní a dialektologické přepisy převést na společný, konzistentní základ. Normalizace digitalizovaných textů u folklorních přepisů překlenuje problém sjednocení metodou obojetností, která je v daném kontextu unikátní; obdobné problémy na jiné úrovni řeší třeba korpusová lingvistika (ambiguity, desambiguace). U dialektologických přepisů je totéž řešeno metodou zjednodušení zápisu a algoritmického převodu znaků, který je novem minimálně v českém prostředí.

Pro **převod folkorního přepisu na přepis dialektologický** existují tzv. G2P převodníky (angl. grapheme to phoneme conversion), které převádějí grafémy na fonémy. Existují pro spisovnou češtinu a pro řadu dalších spisovných jazyků, pro dialekty však velmi omezeně, a pokud ano, je jejich fungování nepříliš úspěšné. Z toho důvodu byl sestaven vůbec první návod na převod, který vykazuje chybovost pouze v řádech promile (oproti obvyklým až desítkám procent). G2P převodníky určené pro dialekty jsou vesměs založeny na strojovém učení a vycházejí z materiálu vytvořeného člověkem (vzniklého ručním přepisem); oproti tomu bylo v našem případě zapotřebí tento materiál pomocí převodu teprve vytvořit, a to ve vysoké kvalitě. Dosaženo toho bylo pomocí pravidel/algoritmů. Jde tak o první G2P převod vytvořený pro dialekty češtiny, který je jedinečný svou přesností. Postupy normalizace a převodu přepisů, jak jsou podány v metodice, tedy řeší vůbec poprvé problémy nejednotnosti zápisu nářečí češtiny. Zároveň ukazují cestu ke sjednocení většiny existujících textů, čímž se otevírají zcela nové možnosti ve zpracování českých nářečních dat. Příslušná část metodiky mj. umožňuje tvorbu rozsáhlých textových korpusů, které pro dialekty češtiny dosud nevznikly a nebyly realizovatelné, a vůbec poprvé zpřístupňuje možnost jakéhokoli rozsáhlého a systematického zpracování nářečních textů.

Kapitoly 5 a 6 přinášejí pohled do **teorie strojového učení** a jeho využití v dialektologii jak v oblasti zpracování přirozeného jazyka, tak pro automatický přepis řeči. Zabývají se automatickým trénováním systému pro převod mezi folklorní a dialektologickou formou zápisu a přípravou dat pro akustické trénování systému pro přepis řeči. Oba přístupy jsou v české dialektologii novinkou, neboť byly dosud aplikovány výhradně na velké jazyky, nikoliv na jazyky malé či na teritoriální dialekty.

Popisovaná **geolokace nahrávek** (kapitola 7) nabízí všeobecně uznávané postupy pro přiřazení prostorové složky prostorovým objektům a jevům, včetně konkrétní aplikace na nářeční data. Navrhuje se využití existujících prostorových databází, jako je datová vrstva částí obcí České republiky nebo adresní body z RÚIAN. Inovativní postupy pak nabízí metodika pro geolokaci starých dialektologických dat, která nejsou v potřeb-

né podrobnosti dostupná pro další využití. Následuje vysvětlení **formátů prostorových dat** s doporučením pro jejich výběr a implementaci, umožňující následné pokročilé zpracování v prostředí geografických informačních systémů. Pro tvorbu **dialektologických map** byly aplikovány nejnovější přístupy webové kartografie, vycházející z publikací V. Voženílka a kol. (2021) a R. Nétka (2020). Nově byly do dialektologických map implementovány nejen primární audionahrávky, ale i výsledky konverze audiálních a textových dialektologických dat pomocí metod strojového učení.

Uplatnění metodiky v praxi

9

UPLATNĚNÍ METODIKY V PRAXI

Komplexnost metodiky, orientované na sběr a zpracování nářečních dat, zaručuje její **mnohostranné využití**. Předložené postupy jsou sice primárně určeny odborné komunitě, avšak upotřebení mohou nalézt i u široké veřejnosti, a to včetně vybraných institucí a firem.

S ohledem na neexistenci souborných metodologických příruček věnovaných sběru audiálních dat lze očekávat poptávku po podrobném **návodu na realizaci terénního (nářečního) výzkumu**, podaném v kapitole 3, a to jak mezi dialektology a lingvisty vůbec, tak mezi badateli dalších vědních oborů, v nichž je uplatňována metoda inverview, včetně vysokoškolských studentů, kteří jinak mnohdy tápou při hledání vhodných strategií aplikovatelných v empirickém výzkumu. Při praktikování popsaných metod mohou tito zájemci získat kvalitní data, co se týče zachyceného jazyka a také techniky nahrávání, a v případě dialektologických sběrů tak dokumentovat mnohem efektivněji teritoriální dialekty coby součást nehmotného kulturního dědictví. Tatáž skupina uživatelů uvítá i obecné návody týkající se dalšího nakládání se získanými daty, přičemž popis **archivace a katalogizace digitálních audiálních dat** může oslovit i další potenciální uživatele vytvářející zvukové archivy ať už soukromé, nebo v rámci konkrétní instituce. Ve druhém případě jsou to archivy zakládané na vysokoškolských (nejen) bohemistických pracovištích, též hudební/zvukové archivy budované pod záštitou knihoven, muzeí, obecně databázových či informačních center, paměťových institucí aj. Lze doplnit, že již během přípravy metodiky proběhl transfer znalostí týkajících se této složky péče o audiální data se čtyřmi vědeckými pracovišti, a to se Slovanským ústavem Akademie věd ČR, v. v. i., dialektologickým oddělením Jazykovedného ústavu Ľudovíta Štúra (Slovensko), Pracownou Dialektologicznou Uniwersyty im. Adama Mickiewicza w Poznaniu (Polsko) a etymologicko-onomastickou sekcí Inštitutu za slovenski jezik Frana Ramovša (Slovinsko).

Obdobně rozsáhlý okruh uživatelů nalezne inovativní postup **digitalizace nářečních textů** (OCR; kapitola 4). Digitalizace dokumentů prováděná v knihovnách a dalších informačních institucích zpravidla probíhá nesprávnými metodami, které nářeční data deformují a poškozují. Kvalitní digitalizace nářečních textů bude vždy náročnější, a tudíž i finančně nákladnější než digitalizace textů spisovných. Zavedení doporučené metodiky by však přineslo podstatné snížení nákladů na OCR. Běžnou praxí, a to i ve firmách specializujících se na digitalizaci textů, je propojení OCR s proofreadingem, přičemž korekturní zásahy jsou časově i finančně velmi náročné, nadto bývá výsledný digitalizát do velké míry chybový z důvodu neznalosti nářečně obměněných slov ze strany zařízení nebo korektora. Naopak při zavedení navržených postupů zvládne týž proces jeden člověk za mnohem kratší dobu, a to v téměř srovnatelné kvalitě (čím delší text, tím větší úspora času, tudíž i peněz, přičemž danou činnost může provádět jakýkoliv pracovník po krátkém zaškolení). Úspora času i financí může být až pětinásobná. Je potřeba pouze hardwarová a softwarová výbava, v obou případech v řádech několika tisíc korun, z níž přinejmenším část je stejně nutná k digitalizaci svépomocí. Při uplatnění zásad na **čištění a formální sjednocení textu** pak lze dosáhnout snížení nákladů v pozdějších fázích zpracování textu (normalizace, převod), a to tak, že toto následné zpracování zvládne jeden člověk namísto týmu lidí.

Metodologický postup **normalizace a převodu folklorní a dialektologické transkripce** řeší problémy nejednotnosti zápisu nářečních textů a jejich automatického nebo poloautomatického převodu. Tento návod prakticky umožňuje **vytvořit rozsáhlé korpusy nářečních textů**, což dosud nebylo v českém pro-

středí možné bez vysokých nákladů (plynoucích z ručních oprav, převodů textů apod.) – proto také dosud nevznikly žádné veřejné (a systematicky pojaté) korpusy českých nářečních textů. Daná část metodiky tak umožní vytvořit soustavné, konzistentní korpusy v dialektologickém přepisu, přičemž urychlí práce minimálně o jeden řád oproti ručnímu převodu, a to při prvním převodu daného nářečí. Při dalších převodech téhož nářečí se úspora navyšuje již o mnoho řádů, neboť odpadá sestavování sady regulárních výrazů a výsledek libovolně objemné práce (i několikaleté) je hotový v řádu vteřin nebo desítek vteřin.

Metodika rovněž přináší postup, jak **zpracovávat a připravovat dialektologická data pro strojové učení**, a to včetně systémů pro přepis řeči (kapitola 5, 6). Součástí je mj. praktická ukázka trénování systému pro automatický převod mezi textovými formáty užívanými v dialektologii, včetně architektury, nastavení parametrů a demonstrace očekávaných výsledků v případě replikování. Metodika byla vyvinuta a prototypována pro teritoriální dialekty českého jazyka, ale jedná se de facto o metody, které jsou použitelné nezávisle na zvoleném jazyce. Po prezentaci metod na odpovídajících odborných konferencích (např. ICPhS, INTERSPEECH) očekáváme zájem ze strany zahraničních odborníků na strojové učení. Metodika je přenositelná rovněž do oblasti psaných textů obecně, v níž bude využito spolupráce s týmem Ing. Michala Hradiše, Ph.D., z FIT VUT zabývajícím se OCR technikami pro historické a ručně psané dokumenty.

Kapitola 7 nachází uplatnění při práci odborníků, kteří vytvářejí, archivují a jinak zpracovávají dialektologická data se zaměřením se na jejich **prostorovou složku**. Uplatnění této části metodiky lze očekávat ve vědeckých a akademických praxích pracujících s textovými i audiálními dialektologickými daty, případně u zájmové odborné i laické veřejnosti, která vytváří mapy s dialektologickou tematikou. Metodika bude dále aktivně využívána při spolupráci odborníků na geoinformatiku a kartografii z Katedry geoinformatiky Přírodovědecké fakulty Univerzity Palackého v Olomouci a odborníků z dialektologického oddělení Ústavu pro jazyk český Akademie věd ČR. Prostřednictvím prezentací a odborných příspěvků bude povědomí o metodice a jejím obsahu rozšiřováno mezi odborné i laické zájemce. Metodika poslouží také jako výchozí materiál pro další diskuzi k problematice vizualizace dat z výzkumů společenských věd, vizualizace multimediálních dat, zpracování jiných geolingvistických výzkumů apod.

Ekonomický impakt nebyl primárním cílem metodiky. Využívání představených postupů však může vést ke značnému zefektivnění zpracování dialektologických dat, čímž dojde k významné úspoře osobních nákladů dotčených institucí. Kromě již uvedené úspory v souvislosti s digitalizací a normalizací textových dat lze zmínit zejména zásadní usnadnění transkripčních prací v případě vyvinutí automatického přepisovače optimalizovaného právě na česká nářeční data. Za předpokladu, že je hodinový úsek převáděn zkušeným transkriptorem coby člověkem do textové podoby minimálně 5 hodin (viz 3.4.2), se při strojovém převodu snižuje čas strávený daným úkolem několikanásobně (je však nutné počítat s lidskou revizí, jejíž délka závisí na úspěšnosti softwaru, resp. na chybovosti výsledného textu, jež přímo souvisí s počtem diferenčních jevů vyskytujících se v nahrávce/přepisu).

10

Pro koho je metodika určena

PRO KOHO JE METODIKA URČENA

Subjekty, jimž je metodika určena, byly částečně přiblíženy již v kapitole 9. Vzhledem k obsahu metodiky lze předpokládat, že jejími hlavními uživateli budou odborníci zejména z řad lingvistů, strojových inženýrů a geoinformatiků, mnohé postupy však mohou být využity také v jiných vědních disciplínách, ve veřejném sektoru nebo i laickou veřejností.

U kapitoly 3 lze modelové uživatele rozdělit do pěti základních skupin – jednak jsou to lingvisté zpracovávající audiální data, nikoliv nutně regionálně rozrůzněná (kromě dialektologů může jít o sociolingvisty, odborníky na konverzační analýzu, korpusovou lingvistiku ap.), hledající návod na metody terénního jazykového výzkumu, též katalogizaci a archivaci dat. Druhou skupinou jsou odborníci spřízněných humanitních a sociálněvědních oborů (etnologové, orální historikové, sociologové aj.), kteří hledají tytéž odpovědi, a to ve spojitosti s obecnými metodami řízeného rozhovoru. Třetí skupinu tvoří odborníci na počítačové zpracování přirozeného jazyka, pro něž je získání jazykových dat (třeba i formou aktivního sběru) podmínkou pro základní i aplikovaný výzkum a vývoj. Postupy věnované uspořádání audiálních dat, jejich uložení a případnému zveřejnění (včetně etických zásad) se dotýkají čtvrté skupiny uživatelů, jimiž jsou instituce spravující zvukové dokumenty (archivy, knihovny, paměťové instituce, univerzity aj.). Lze počítat též se zájmem té části laické veřejnosti, která se aktivně zasazuje o uchování nářečí coby součásti regionální identity, a to cestou audiální dokumentace. Veškeré zájmové skupiny naleznou v dané kapitole návody celého procesu zpracování audiálních dat, a to od jejich sběru po uložení ve strukturovaném archivu, umožňujícím mj. zveřejnění dat.

Obdobně široký je záběr kapitoly 4. Postupy digitalizace nářečních publikací jsou využitelné pro archivy, (digitální) knihovny, muzea, databázová centra a další informační instituce. Normalizace folklorních a dialektologických přepisů (včetně jejich převodu) jsou stěžejní pro dialektology, potažmo lingvisty (včetně tvůrců jazykových korpusů), též etnografy a odborníky na zpracování přirozeného jazyka. Normalizace folklorních přepisů je nezbytná pro vydávání nářečních textů, tato část metodiky tak najde uplatnění jak u autorů těchto textů, tak u pracovníků nakladatelství a vůbec osob pracujících s regionálně rozrůzněným jazykem v textové podobě (nakladatelé, editoři, korektoři, novináři aj.).

Kapitola 5 je určena odborníkům z oblasti zpracování přirozeného jazyka, stejně tak jako odborníkům z oblasti dialektologie sondujícím využití moderních technik strojového učení. Přináší návrh systému, včetně teorie a experimentální části. Je zde exaktně změřena chybovost navrženého řešení využívajícího strojový překlad pro učení automatické konverze mezi formáty zápisu nářečí. Kapitola 6 osloví odborníky v oblasti automatického přepisu řeči i studenty a zájemce o ni. Kromě úvodu do problematiky, obsahujícího srozumitelný přehled metod automatického přepisu řeči (jak založených na komponentech, tak na tzv. End-to-End metodách zahrnujících pouze jeden komplexní model produkující ze vstupu přímo výstup), přináší také návod, jak připravit nevelká a jazykově rozrůzněná dialektologická data tak, aby bylo možné i na těchto relativně malých datových sadách systémy strojového učení trénovat.

Kapitola 7, respektive podkapitola 7.1, je určena odborníkům se zájmem o řešenou problematiku, kterým nabízí možnosti geolokace existujících i nově vznikajících nahrávek a rozšíření popisných dat o geografickou polohu a další prostorové atributy. Podkapitola 7.2 je určena pro odborné i laické tvůrce map, papírových i webových, jejichž obsahem je dialektologický materiál. Mapy mohou být sestavovány za účelem doprovodu k odbornému textu, popularizačním a vzdělávacím aktivitám či jako dokumentace nářečního kulturního dědictví, umožňující například i zvýšení turistického potenciálu a rozvoj cestovního ruchu.

11

Seznam použité literatury

SEZNAM POUŽITÉ LITERATURY

ABBYY = *ABBYY® FineReader PDF 15. User's Guide* (2019). Charlotte: ABBYY Production LLC. Online. Dostupné z: https://pdf.abbyy.com/media/1676/users_guide.pdf.

ALJ = *Archiv lidového jazyka* (1952–2024). Brno: Archiv dialektologického oddělení Ústavu pro jazyk český AV ČR.

ARDILA, Rosana, BRANSON, Megan, DAVIS, Kelly a kol. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, s. 4218–4222. ISBN 979-10-95546-34-4. Též online. Dostupné z: https://commonvoice.mozilla.org/en/datasets.

BAEVSKI, Alexei, ZHOU, Henry, MOHAMED, Abdelrahman a AULI, Michael (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, s. 12449–12460. ISBN 978-1-7138-2954-6.

BAHDANAU, Dzmitry, CHO, Kyunghyun a BENGIO, Yoshua (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. Online. Dostupné z: https://arxiv.org/abs/1409.0473.

BACHMANN, Luděk (2001). *Nářečí na Vysokomýtsku*. Praha: Academia. ISBN 80-200-0859-4.

BACHMANNOVÁ, Jarmila (2008). *Za života se stane ledacos. Vyprávěnky ze Železnobrodska*. Liberec: Bor, 2008. ISBN 978-80-86807-90-4.

BALHAR, Jan a kol. (1992–2011). *Český jazykový atlas 1–5. Dodatky*. Praha: Academia. ISBN 80-200-0013-5. [1. díl, 1992; 2. díl, 1997; 3. díl, 1999; 4. díl, 2002; 5. díl, 2005; 6. díl (Dodatky), 2011.] Též online (HTML verze, 2018). Dostupné z: https://cja.ujc.cas.cz/e-cja/.

BALHAR, Jan (2010). Vydávat? Rozhodně nevydávat všechno! *Naše řeč*, roč. 93, č. 2, s. 106–108. ISSN 0027-8203.

BALHAR, Jan (1974). *Skladba lašských nářečí.* Praha: Academia.

BALHAR, Jan (1957). K charakteristice lašského okrajového nářečí na jihozápadním Opavsku. *Slezský sborník*, roč. 55, s. 105–114.

BARTOŠ, František (2006). *Kytice z lidového básnictva našim dětem*. 5. vydání. Zlín: Muzeum jihovýchodní Moravy ve Zlíně. ISBN 80-903411-7-9.

BARTOŠ, František (1895). *Dialektologie moravská II. Nářečí hanácké a české*. Brno: Matice moravská.

BARUCH, Josef (1948). *Pod Junákovem. Všelijaké povjedačky*. Valašské Meziříčí: Osvěta.

BARUCH, Jožka (1934–1937). *Ludé z poza Junákova I–III*. Praha, Přerov: Jožka Baruch.

BECKER, Howard S. (1964). Problems in the publication of field studies. In: VIDICH, Arthur J., BENSMAN, Joseph a STEIN, Maurice R. (eds.). *Reflections on Community Studies*. New York: Wiley, s. 267–284.

BĚLIČ, Jaromír (1954). *Dolská nářečí na Moravě*. Praha: Nakladatelství ČSAV.

BĚLIČ, Jaromír (1972). *Nástin české dialektologie*. Praha: Státní pedagogické nakladatelství.

BENEŠ, Josef (1998). *Německá příjmení u Čechů 1–2*. Ústí nad Labem: Univerzita J. E. Purkyně v Ústí nad Labem. ISBN 80-7044-212-3.

BENEŠOVÁ, Lucie, KŘEN, Michal a WACLAWIČOVÁ, Martina (2015). Korpus spontánní mluvené češtiny ORAL2013. *Časopis pro moderní filologii*, roč. 97, č. 1, s. 42–50. ISSN 2336-6591.

BORNAT, Joanna (2003). A Second Take: Revisiting Interviews with a Different Purpose. *Oral History*, roč. 31, č. 1, s. 47–53. ISSN 0143-0955.

BOROCKÝ, Josef (2003). *Hanácké slovňik (Hanácký slovník) středohanáckyho nářeči*. Hrubčice: Jiří Dokoupil. Online. Dostupné z: http://www.banda.cz/webs/k/kom/usr_files/file/hanackyslovnik.pdf.

BOULIANNE, Gilles, BURGET, Lukáš, GLEMBEK, Ondřej a kol. (2011). The Kaldi Speech Recognition Toolkit. In: *Proceedings of the IEEE 2011. Workshop on Automatic Speech Recognition and Understanding*, s. 1–4. ISBN 978-1-4673-0366-8. Též online. Dostupné z: https://kaldi-asr.org/.

BRABER, Natalie a DAVIES, Diane (2016). Using and creating oral history in dialect research. *Oral history*, roč. 44, č. 1, s. 98–107. ISSN 0143-0955.

BŘENĚK, Otýn (1921). *Pšišery*. Brno: Kramerius.

BUCHNER-FUHS, Jutta (1977). Die Fotobefragung – eine kulturwissenschaftliche Interviewmethode? *Zeitschrift für Volkskunde*, roč. 93, s. 189–216.

CLEMENTE, Ignasi (2008). Recording Audio and Video. In: WEI, Li a MOYER, Melissa (eds.). *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Malden: Blackwell Publishing, s. 177–191. ISBN 978-1-4051-2607-6.

CODÓ, Eva (2008). Inverviews and Questionnaires. In: WEI, Li a MOYER, Melissa (eds.). *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Malden: Blackwell Publishing, s. 158–176. ISBN 978-1-4051-2607-6.

ČERMÁK, František a kol. (2007). *Frekvenční slovník mluvené češtiny*. Praha: Karolinum. ISBN 978-80-246-1425-0.

ČESKÁ REPUBLIKA (1990). Zákon č. 367/1990, o obcích (obecní zřízení). In: *Sbírka zákonů. 1990, částka 59, číslo 367*. Online. Dostupné z: https://www.zakonyprolidi.cz/cs/1990-367.

ČESKÝ STATISTICKÝ ÚŘAD (2015a). *Historický lexikon obcí České republiky 1869–2011*. Praha: Český statistický úřad. Online. Dostupné z: https://csu.gov.cz/produkty/historicky-lexikon-obci-1869-az-2015.

ČESKÝ STATISTICKÝ ÚŘAD (2015b). *Zajímavosti názvů obcí v České republice*. Praha: Český statistický úřad. Online. Dostupné z: https://csu.gov.cz/zajimavosti_nazvu_obci_v_ceske_republice.

ČESKÝ ÚŘAD ZEMĚMĚŘICKÝ A KATASTRÁLNÍ (2024). *Registr územní identifikace, adres a nemovitostí (RÚIAN)*. Praha: Český úřad zeměměřický a katastrální. Online. Dostupné z: https://cuzk.gov.cz/ruian/RUIAN.aspx.

ČIŽMÁROVÁ, Libuše (2000). *Jazykový atlas jihozápadní Moravy*. Brno: Masarykova univerzita. ISBN 80-210-2488-7.

DAHL, George E., YU, Dong, DENG, Li a ACERO, Alex (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, roč. 20, č. 1, s. 30–42. ISSN 2329-9290.

Databáze nářečních promluv pro odbornou veřejnost (v přípravě). Brno – Olomouc: Ústav pro jazyk český – Vysoké učení technické – Univerzita Palackého v Olomouci. [Dokončení databáze plánováno na r. 2026.]

Dotazník pro ČJA = Dotazník pro výzkum českých nářečí (1964–1976). Brno: Archiv dialektologického oddělení Ústavu pro jazyk český AV ČR.

DSNT = *Databáze souvislých nářečních textů* (2017–2024). Brno: Archiv dialektologického oddělení Ústavu pro jazyk český AV ČR.

DUBĚDA, Tomáš, HAVLÍK, Martin, JÍLKOVÁ, Lucie a ŠTĚPÁNOVÁ, Veronika (2014). Průzkum výslovnostního úzu u výpůjček a cizích vlastních jmen – metodologické otázky. *Jazykovědné aktuality*, roč. 51, č. 3–4, s. 125–141. ISSN 1212-5326.

ELIÁŠ, Jindra a SAINT-EXUPÉRY, Antoine de (2020). *Malé principál*. Brno: Jota. ISBN 978-80-7565-776-3.

ELLIS, Stanley (1974). The Survey of English Dialects and Social History. *Oral History*, roč. 2, č. 2, s. 37–43.

ELMENTALER, Michael, ROSENBERG, Peter a kol. (2015). *Norddeutscher Sprachatlas (NOSA). Band 1: Regiolektale Sprachlagen*. Hildesheim: Olms Verlag. ISBN 978-348-7153-292.

ERNESTUS, Mirjam (2000). *Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface*. Utrecht: LOT. ISBN 90-76864-02-0.

ESCA (2015). *Ten Principles of Citizen Science*. Online. Dostupné z: https://www.ecsa.ngo/documents/#documents.

FIC, Karel (1971). *Drahonínské nářečí*. Disertační práce. Brno: Universita Jana Evangelisty Purkyně.

FOJTÍK, František (2011). *Od vesna do vesna*. Valašské Klobouky: Muzejní společnost ve Valašských Kloboukách ve spolupráci s Městským muzeem. ISBN 978-80-260-0971-9.

FONTANA, Andrea a FREY, James H. (1994). Interviewing: The Art of Science. In: DEZIN, Norman K. a LINCOLN, Yvonna S. (eds.). *The Handbook of Qualitative Research*. Thousand Oaks: Sage Publications, s. 361–376. ISBN 0-8039-4679-1.

FONTANA, Andrea (1977). *The Last Frontier: The Social Meaning of Growing Old*. Beverly Hills, CA: Sage Publications. ISBN 978-0803908321.

FROLEC, Václav a HOLÝ, Dušan (1967). *Lidové povídky ze Slovácka*. Praha: Odeon.

GAJDOŠOVÁ, Katarína a ŠIMKOVÁ, Mária (2014). Slovenský hovorený korpus (2008 – 2012). In: GAJDOŠOVÁ, Katarína a ŽÁKOVÁ, Adriána (eds.). Jazykovedné štúdie XXXI. Rozvoj jazykových technológii a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Bratislava: Veda, s. 65–84. ISBN 978-80-224-1391-6. GAVORA, Peter (2005). *Učitel a žáci v komunikaci*. Brno: Paido. ISBN 80-7315-104-9.

GEOPORTÁL (2024). *Adresní body ČSÚ*. Praha: CENIA. Online. Dostupné z: http://ms.cenia.cz/php/micka/record/basic/56f2afc3-c49c-417b-ab3a-7ae4c0a80137.

GLUCK, Sherna Berger a PATAI, Daphne (eds., 1991). *Women's Words. The Feminist Practice of Oral History*. New York: Routledge. ISBN 978-0415903714.

GOLÁŇOVÁ, Hana, WACLAWIČOVÁ, Martina, PEJCHA, Jakub, ČAPKA, Tomáš a BENEŠOVÁ, Lucie (2023). *Mapka: mapová aplikace pro korpusy mluvené češtiny.* Verze 2.0. Online. Dostupné z: http://korpus.cz/mapka.

GOLÁŇOVÁ, Hana, WACLAWIČOVÁ, Martina a LUKEŠ, David (2021). *DIALEKT: nářeční korpus*. Verze 2 z 23. 12. 2021. Online. Dostupné z: http://www.korpus.cz.

GOLÁŇOVÁ, Hana (2009). PhDr. Jan Balhar, CSc. In: CHROMÝ, Jan a LEHEČKOVÁ, Eva (eds.). *Rozhovory s českými lingvisty II*. Praha: Akropolis, s. 12–35. ISBN 978-80-86903-95-8.

GORDEN, Raymond L. (1980). *Interviewing. Strategy, techniques, and tactics*. Georgetown: Dorsey Press. ISBN 978-0256023701.

GRAVES, Alex a JAITLY, Navdeep (2014). Towards End-to-End Speech Recognition with Recurrent Neural Networks. *Proceedings of the 31st International Conference on Machine Learning*, roč. 32, č. 2, s. 1764–1772. ISSN 2640-3498.

GUILLEMIN, Marilys a GILLAM, Lynn (2004). Ethics, Reflexivity, and „Ethically Important Moment" in Research. *Qualitative Inquiry*, roč. 10, č. 2, s. 261–280. ISSN 1077-8004.

HÁLA, Bohuslav, VÁŽNÝ, Václav a kol. (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských*. Praha: Česká akademie věd a umění v Praze.

HÁLA, Bohuslav, VÁŽNÝ, Václav a kol. (1944). Pravidla pro vědecký přepis dialektických zápisů. *Věstník České akademie věd a umění*, roč. 52, č. 1, s. 63–68.

HAMMERSLEY, Martyn a ATKINSON, Paul (2007). *Ethnography. Principles in practise*. Third edition. London, New York: Routledge. ISBN 0-203-94476-3.

HANNUN, Awni, CASE, Carl, CASPER, Jared a kol. (2014). Deep Speech: Scaling up End-to-End Speech Recognition. Online. Dostupné z: arXiv preprint arXiv:1412.5567.

HAUPTMANN, Gerhart (1898). *Tkalci. Hra z let čtyřicátých*. Přeložil Josef Krušina ze Švamberka. Praha: Vzdělavací bibliotéka.

HAVLÍKOVÁ, Jana (2004). Finanční odměna pro participanty výzkumu a její implikace pro výzkumný vztah. *Biograf*, č. 35, s. 74–84. ISSN 1211-5770.

HEJNAL, Ondřej a LUPTÁK, Ľubomír (2013). Když výzkum, tak skrytý: Serpentinami formalizované etiky. In: PAVLÁSEK, Michal a NOSKOVÁ, Jana (eds.). *Když výzkum, tak kvalitativní. Serpentinami bádání v terénu*. Brno – Praha: Masarykova univerzita – Etnologický ústav AV ČR, s. 133–147. ISBN 978-80-210-6480-5.

HERBEN, Jan (1946). *Brumovice*. Praha: Družstevní práce.

HINTON, Geoffrey, DENG, Li, YU, Dong a kol. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, roč. 29, č. 6, s. 82–97. ISSN 1558-0792.

HOFFMANNOVÁ, Jana, MÜLLEROVÁ, Olga a kol. (2007). *Čeština v dialogu generací*. Praha: Academia, s. 7–23. ISBN 978-80-200-1549-5.

HOFFMANNOVÁ, Jana a ZEMAN, Jiří (2007). Úvod. In: HOFFMANNOVÁ, Jana a MÜLLEROVÁ, Olga (eds.). *Čeština v dialogu generací*. Praha: Academia, s. 7–23. ISBN 978-80-200-1549-5.

HOFFMANNOVÁ, Jana (1992). Metodologie „konverzační analýzy" a transkripční symboly. In: STACHOVÁ, Jiřina (ed.). *Symbol v lidském vnímání, myšlení a vyjadřování*. Praha: Filozofický ústav ČSAV, s. 234–241.

HOFFMANOVÁ, Klára, BACHMANNOVÁ, Jarmila a SAINT-EXUPÉRY, Antoine de (2023). *Malej princ*. Brno: Jota. ISBN 978-80-7689-314-6.

HORÁLKOVÁ, Zdeňka (1962). K jazyku lidových písní. *Naše řeč*, roč. 45, č. 1–2, s. 13–26.

HOŠEK, Ignác (1897). *O poměru jazyka písní národních k místnímu nářečí*. Praha: Česká akademie císaře Františka Josefa pro vědy, slovesnost a umění.

HŮRKOVÁ, Jiřina (1995). *Česká výslovnostní norma*. Praha: Scientia. ISBN 80-85827-93-X.

HURTWORTH, Rosalind (2003). Photo-interviewing for research. *Social research UPDATE*, č. 40, online. Dostupné z: https://sru.soc.surrey.ac.uk/SRU40.html.

CHAMBERS, J. K. a TRUDGILL, Peter (2004). *Dialectology*. Second edition. Cambridge: Cambridge University Press. ISBN 0-521-59378-6.

CHAN, William, JAITLY, Navdeep, LE, Quoc V. a VINYALS, Oriol (2016). Listen, Attend and Spell. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, s. 4960–4964. ISSN 1520-6149.

CHROMÝ, Jan (2021). *Empirické metody v lingvistice. Učebnice pro studenty lingvistických a spřízněných oborů*. Online. Dostupné z: https://dl1.cuni.cz/pluginfile.php/1241622/mod_resource/content/1/Jan%20Chrom%C3%BD%20-%20skripta.pdf. [Průběžná verze připravované učebnice. Poslední aktualizace 12. 10. 2021]

CHROMÝ, Jan (2015a). The use of prothetic /v/ by older speakers in Prague. *Poznan Studies in Contemporary Linguistics*, roč. 51, č. 2, s. 203–225. ISSN 1897-7499.

CHROMÝ, Jan (2015b). Vliv jazykových faktorů na užívání protetického *v-* v pražské mluvě. *Slovo a slovesnost*, roč. 76, č. 1, s. 21–39. ISSN 0037-7031.

CHROMÝ, Jan (2014). *Práce s empirickými daty. Příručka pro studenty Bc. studia ČJL*. Praha: Karolinum. ISBN 978-80-246-2801-1.

CHROMÝ, Jan (2012): Howard Giles a teorie komunikační akomodace. *Studie z aplikované lingvistiky*, roč. 3, č. 1–2, s. 111–120. ISSN 2336-6702.

IDEA (2024). *International Dialects of English Archive*. Online. Dostupné z: https://www.dialectsarchive.com/.

IJP = *Internetová jazyková příručka* (2008–2024). Praha: Ústav pro jazyk český AV ČR. Online. Dostupné z: https://prirucka.ujc.cas.cz/.

IREINOVÁ, Martina, VOŽENÍLEK, Vít, KONÍČEK, Jakub a kol. (2023a). *Atlas nářečí českého jazyka – instrumentál plurálu*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-6244-8.

IREINOVÁ, Martina, VOŽENÍLEK, Vít, KONÍČEK, Jakub a kol. (2023b). *Atlas nářečí českého jazyka – krácení vokálů*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-5846-5.

IREINOVÁ, Martina, VOŽENÍLEK, Vít, KONÍČEK, Jakub a kol. (2023c). *Atlas nářečí českého jazyka – nominativ singuláru feminin*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-6414-5.

IREINOVÁ, Martina, VOŽENÍLEK, Vít, KONÍČEK, Jakub a kol. (2022). *Atlas nářečí českého jazyka – deklinace substantiv*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-6245-5.

JAMES, Nalita a BUSHER, Hugh (2012). Internet Interviewing. In: GUBRIUM, Jaber F., HOLSTEIN, James A., MARVASTI, Amir B. a McKINNEY, Karyn D. (eds.). *The SAGE Handbook of Interview Research. The Complexity of the Craft*. Second edition. Los Angeles: SAGE, s. 177–191. ISBN 978-1-4129-8164-4.

JANČÁKOVÁ, Jana (1987). *Nářečí a běžná mluva na Příbramsku.* Praha: Univerzita Karlova.

JAVOŘICKÁ, Vlasta (2004). *Ticho po pěšině*. 2. vydání. Brno: Jota. ISBN 80-7217-249-2.

JECH, Jaromír (1959). *Lidová vyprávění z Kladska*. Praha: Státní nakladatelství krásné literatury, hudby a umění.

JEŘÁBEK, Richard (ed., 1997). *Počátky národopisu na Moravě. Antologie prací z let 1786–1884*. Strážnice: Ústav lidové kultury. ISBN 80-86156-05-2.

JIRSÁK, František (1932–1934). Lidová rčení. *Naše řeč*, roč. 16, č. 8, s. 253–255; roč. 17, č. 1, s. 30–31; roč. 17, č. 5, s. 159–160; roč. 17, č. 6–7, s. 220–221; roč. 18, č. 1, s. 29–30.

JOHNSON, D. Chris (2023). *User Guide - KU ScholarWorks Collection of German Dialect Recordings from Kansas and Missouri*. Online. Dostupné z: http://hdl.handle.net/1808/30722.

JOHNSTONE, Barbara (2000). *Qualitative Methods in Sociolinguistics*. New York – Oxford: Oxford University Press.

KADERKA, Petr a SVOBODOVÁ, Zdeňka (2006). Manuál pro přepisovatele televizních diskusních pořadů. *Jazykovědné aktuality*, roč. 43, č. 3–4, s. 18–51. ISSN 1212-5326.

KADOCH, František (2008). *Lidová mluva na Šumavě a v Pošumaví*. Černá v Pošumaví: Tiskárna FOP. ISBN 978-80-254-2808-5.

KALDI, nedat. Online. Dostupné z: https://kaldi-asr.org/.

KAŠÍK, Antonín (1908). *Popis a rozbor nářečí středobečevského.* Praha: Nákladem České akademie císaře Františka Josefa pro vědy, slovesnost a umění.

KAZMÍŘ, Silvestr (2012). *Slovník valašského nářečí I–II*. Zlín: Alisa Group. ISBN 978-80-903965-3-1 (I. díl), 978-80-903965-4-8 (II. díl). Též online. Dostupné z: https://sites.google.com/site/silvestrkazmir/Home.

KLÍMOVÁ, Dagmar a OTČENÁŠEK, Jaroslav (2012). *Česká pohádka v 19. století*. Praha: Etnologický ústav AV ČR. ISBN 978-80-87112-50-2.

KLOFEROVÁ, Stanislava a ŠIPKOVÁ, Milena (eds., 2018). *Život ve slovech, slova v životě. Procházka labyrintem českých nářečí*. Praha: Nakladatelství Lidové noviny. ISBN 978-80-7422-657-1.

KLOFEROVÁ, Stanislava (2017). Dialekt. In: KARLÍK, Petr, NEKULA, Marek a PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Online. Dostupné z: https://www.czechency.org/slovnik/DIALEKT.

KLOFEROVÁ, Stanislava (2007). Dialektologie. In: PLESKALOVÁ, Jana a kol. (eds.). *Kapitoly z dějin české jazykovědné bohemistiky*. Praha: Nakladatelství Academia, s. 336–376. ISBN 978-80-200-1523-5.

KLOFEROVÁ, Stanislava (2000). *Mluva v severomoravském pohraničí*. Brno: Masarykova univerzita. ISBN 80-210-2470-4.

KOLAŘÍK, Josef (1998). *Současná běžná mluva obyvatel v Napajedlích*. Olomouc: Univerzita Palackého v Olomouci. ISBN 80-7067-921-2.

KOMRSKOVÁ, Zuzana, KOPŘIVOVÁ, Marie, LUKEŠ, David, POUKAROVÁ, Petra a GOLÁŇOVÁ, Hana (2017). New spoken corpora of Czech: Ortofon and Dialekt. *Jazykovedný časopis*, roč. 68, č. 2, s. 219–228. ISSN 0021-5597.

KOPEČNÝ, František (1957). *Nářečí Určic a okolí. Prostějovský úsek hanáckého nářečí centrálního*. Praha: Nakladatelství Československé akademie věd.

KOPŘIVOVÁ, Marie, KOMRSKOVÁ, Zuzana, POUKAROVÁ, Petra a LUKEŠ, David (2019). Relevant criteria for selection of spoken data: Theory meets practise. *Jazykovedný časopis*, roč. 70, č. 2, s. 324–335. ISSN 0021-5597.

KRAAK, Menno-Jan a ORMELING, Ferjan (2010). *Cartography: Visualization of Geospatial Data*. Boca Rabon: CRC Press. ISBN 978-1138613959.

KRČMOVÁ, Marie a CHLOUPEK, Jan (2017). Národní jazyk. In: KARLÍK, Petr, NEKULA, Marek a PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Online. Dostupné z: https://www.czechency.org/slovnik/NÁRODNÍ JAZYK.

KRČMOVÁ, Marie (2017). Transkripce. In: KARLÍK, Petr, NEKULA, Marek a PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Online. Dostupné z: https://www.czechency.org/slovnik/TRANSKRIPCE.

KRČMOVÁ, Marie (1997). Současná běžná mluva v českých zemích. In: DANEŠ, František a kol. *Český jazyk na přelomu tisíciletí*. Praha: Academia, s. 160–172. ISBN 80-200-0617-6.

KŘEMELA, Filip (1940). *Hanáckó kolajó*. 2. vydání. Brno: Moravské kolo spisovatelů.

KŘIŽKOVÁ, Eliška a kol. (2010). *O kopanickej reči*. Starý Hrozenkov: Informační středisko pro rozvoj Moravských Kopanic. ISBN 978-80-254-7891-2.

KUBÍN, Josef Štefan (1964–1971). *Lidové povídky z Podkrkonoší I–II*. Praha: Odeon.

KULDA, Beneš Method (1874–1894). *Moravské národní pohádky a pověsti I–IV*. Praha: I. L. Kober; V. Kotrba.

KYNČL, Rudolf (1943–1945). *Paměti uječka Matěja Škrobáka*. Vyškov: F. Obzina.

KYNČL, Rudolf (1928). *Uječek Matěj Škrobák dragúnem*. Brno: Moravské nakladatelství.

LABOV, William (1984). Field Methods of the Project on Linguistic Change and Variation. In: BAUGH, John a SHERZER, Joel (eds.). *Language in Use. Reading in Sociolinguistics*. Englewood Cliffs, N. J.: Prentice-Hall, s. 28–53.

LABOV, William (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

LAMPRECHT, Arnošt a kol. (1976). *České nářeční texty*. Praha: Státní pedagogické nakladatelství.

LEIX, Alicija (2003). K problematice transkriptu ve společenských vědách. *Biograf*, č. 31, s. 69–84. ISSN 1211-5770.

LINDBLOOM, Jana (2004). Preklopenia asymetrie po získaní údajov. Pomáha informovaný súhlas alebo škodí? *Biograf*, č. 35, s. 85–95. ISSN 1211-5770.

MALAČKA, Ondřej (2011–2020). *KdeJsme.cz*. Brno. Online. Dostupné z: https://www.kdejsme.cz/.

MARKL, Jaroslav (1987). *Nejstarší sbírky českých lidových písní*. Praha: Supraphon.

MARTÍNEK, Vojtěch (1977). *Kamenný řád I–III*. 4. vydání. Ostrava: Profil.

MATÚŠOVÁ, Jana (2015). *Německá vlastní jména v češtině*. Praha: Nakladatelství Lidové noviny. ISBN 978-80-7422-369-3.

MAZLOVÁ, Věra (1949). *Výslovnost na Zábřežsku. Fonetická studie z moravské dialektologie*. Praha: Filosofická fakulta University Karlovy.

MICHÁLKOVÁ, Věra (1971). *Studie o východomoravské nářeční větě*. Praha: Academia.

MILROYOVÁ, Lesley a GORDON, Matthew (2012). *Sociolingvistika: Metody a interpretace*. Přeložil Jan Chromý. Praha: Karolinum. ISBN 978-80-246-2125-8.

MIOVSKÝ, Michal, MIOVSKÁ, Lenka a GAJDOŠÍKOVÁ, Lenka (2004). Některé etické aspekty terénního výzkumu uživatelů nelegálních drog. *Biograf*, č. 35, s. 33–52. ISSN 1211-5770.

MINISTERSTVO VNITRA (1991). Oznámení č. 1/1991 Ú.v., Přehled změn v územní organizaci, v názvech obcí a jejich částí, ve střediskových obcích a v matričních obvodech s účinností ke dni 1. března 1990 ke dni voleb do obecních zastupitelstev v roce 1990. *Ústřední věstník ČR*, č. 1, s. 1.

MOVING IMAGE GATEWAY (2024). *British Library: Sounds Familiar? Accents and Dialects of the UK*. Online. Dostupné z: http://bufvc.ac.uk/gateway/index.php/site/1098.

MURRAY, Thomas E. a MURRAY, Carmin Ross (1992). *On the Legality and Ethics of Surreptitious Recording*. Publication of the American Dialect Society, č. 76, s. 15–75.

MY LEARNING (2024). *A Sound Map of Leeds*. Online. Dostupné z: https://www.mylearning.org/resources/a-sound-map-of-leeds.

NÁRODNÍ ÚŘAD PRO KYBERNETICKOU A INFORMAČNÍ BEZPEČNOST (2024). *Povolené datové formáty pro příjem dokumentů v elektronické podobě*. Online. Dostupné z: https://nukib.gov.cz/cs/kontakty/povolene-datove-formaty.

NEJEDLÝ, Petr (2021). Ke gramatickým charakteristikám folklorního písňového jazyka. *Bohemica Olomucensia. Linguistica*, roč. 13, č. 2, s. 10–35. ISSN 1803-876X. Též online. Dostupné z: https://kb.upol.cz/fileadmin/userdata/FF/katedry/kbh/veda_a_vyzkum/bohemica_olomucen sia/bo_2021_2_web.pdf.

NÉTEK, Rostislav, ŠTRUBL, Ondřej a STUPŇÁNEK, Bronislav (2022). *InteGra*. Specializovaná veřejná databáze. Olomouc: Univerzita Palackého v Olomouci. Online. Dostupné z: https://www.ceskanareci.cz/geoportal/integra/.

NÉTEK, Rostislav (2020). *Webová kartografie – specifika tvorby interaktivních map na webu*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-5827-4.

NOSKOVÁ, Jana (2014). *Biografická metoda a metoda orální historie*. Brno: Etnografický ústav AV ČR. ISBN 978-80-87112-84-7.

NOVOTNÁ, Hedvika (2014). Funkce a logika etických kodexů v sociálních vědách. *Národopisný věstník*, roč. 31 (73), č. 2, s. 7–17. ISSN 1211-8117.

ODEHNAL, Petr a SAINT-EXUPÉRY, Antoine de (2021). *Malučký princ*. Brno: Jota. ISBN 978-80-7565-746-6.

OPENDATA.CZ (2024). *Stupně otevřenosti otevřených dat a česká legislativa*. Praha: OPENDATA.CZ. Online. Dostupné z: https://opendata.gov.cz/informace:stupn%C4%9B-otev%C5%99enosti-datov%C3%BDch-sad.

PACHOLÍK, František Karel (2020). *Všelijaký poudání*. 2. vydání. Liberec: Bor. ISBN 978-80-88367-00-0.

PALÁTOVÁ, Dagmar (1958). *Povídky lidových vypravěčů z Čech, Moravy a Slezska zaznamenané v letech 1945–1954*. Praha: Orbis.

ParCzech 3.0 (2013–2017, 2017–2021). LINDAT/CLARIAH-CZ. Online. Dostupné z: http://lindat.mff.cuni.cz/services/teitok/parczech-3.0/.

POLÁČEK, Jan (2010–2011). *Lidové písně z Hané I–III*. Prostějov, Boskovice: Albert. ISBN 978-80-7326-200-6 (I. díl), 978-80-7326-201-3 (II. díl), 978-80-7326-189-4 (III. díl).

POPELKA, Pavel (2016–2018). *Na Slovácku po slovácku I–II*. Osvětimany: Pavel Popelka.

POPELKA, Pavel (2016). *Nimródi a lovy I–III*. Osvětimany: Pavel Popelka.

POSPÍŠILOVÁ, Jana (ed., 2016). *To sem čuła na vlastní oči… Tradiční vyprávění na nedašovském Závrší*. 2., opr. a dopl. vydání. Valašské Klobouky; Brno: Muzejní společnost ve Valašských Kloboukách ve spolupráci s Etnologickým ústavem AV ČR, v. v. i., pracoviště Brno. ISBN 978-80-905408-4-2 (Muzejní společnost ve Valašských Kloboukách), 978-80-88081-04-3 (Etnologický ústav AV ČR, Praha).

PREISSOVÁ, Gabriela (1910). *Gazdina roba. Drama o třech jednáních*. Praha: J. Otto.

PŘIKRYL, Ondřej (1943). *Haná a Romža*. 3. vydání. Brno: Moravské kolo spisovatelů.

PŘIKRYL, Ondřej (1929). *Komurka. Stréc Karásek*. Olomouc: R. Promberger.

PŘIKRYL, Ondřej (1928). *Dule z hroške. Obrázke z Hané okolo púlke 19. století*. Brno: Národní noviny.

PŘIKRYL, Ondřej (1925). *Ondřeji a inši obrázke z Hané*. Olomouc: R. Promberger.

PSJČ = *Příruční slovník jazyka českého* (1935–1957). Praha: Česká akademie věd a umění; Československá akademie věd.

QUINN, Sterling D. (2018). Web GIS. In: WILSON, John P. (ed). *The Geographic Information Science & Technology Body of Knowledge*. Online. Dostupné z: https://gistbok-topics.ucgis.org/CP-04-014.

RABANUS, Stefan (2020). Language Mapping Worldwide: Methods and Traditions. In: BRUNN, Stanley D. a KEHREIN, Ronald (eds.). *Handbook of the Changing World Language Map*. Cham: Springer. ISBN 978-3030024376. Též online. Dostupné z: https://doi.org/10.1007/978-3-030-02438-3_151.

RADFORD, Alec, KIM, Jong Wook, XU, Tao a kol. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. Online. Dostupné z: arXiv preprint arXiv:2212.04356.

RITCHIE, Donald A. (2015). *Doing Oral History*. Third edition. New York: Oxford University Press. ISBN 978-0-19-932933-5.

RŮŽKOVÁ, Jiřina a ŠKRABAL, Josef (2006). *Historický lexikon obcí České republiky 1869–2005*. Praha: Český statistický úřad. ISBN 80-250-1277-8.

RYBNIKÁŘ, Fanyn (2021). *Mudrosloví národa hluckého*. 2., rozšířené vydání. Hluk: Město Hluk. ISBN 978-80-11-01209-0.

RYCHLÍK, Bedřich (2001). *Pověsti, pohádky a vyprávění moravských Kopanic*. Brno: Doplněk. ISBN 978-80-7239-344-2.

RYCHTAŘÍKOVÁ, Jitka, VONDRÁKOVÁ, Alena, VOŽENÍLEK, Vít a PÁSZTO, Vít (2021). *Obyvatelstvo Česka v období 1995–2019: vitální index, věková struktura a index ekonomického zatížení*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-6063-5.

SATKE, Antonín (ed., 1958). *Hlučínský pohádkář Josef Smolka*. Ostrava: Krajské nakladatelství.

SCHUSTER, Mike a PALIWAL, Kuldip K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, roč. 45, č. 11, s. 2673–2681. ISSN 1053-587X.

SINGLETON Jr., Royce A. a STRAITS, Bruce C. (2012). Survey Interviewing. In: GUBRIUM, Jaber F., HOLSTEIN, James A., MARVASTI, Amir B. a McKINNEY, Karyn D. (eds.). *The SAGE Handbook of Interview Research. The Complexity of the Craft*. 2nd ed. Los Angeles: SAGE, s. 77–98. ISBN 978-1-4129-8164-4.

SLAVIČÍNSKÝ, Josef Misárek (1927). *Vlk Krampotů. Valašská povídka z roků čtyrycátých*. 3. vydání. Olomouc: R. Promberger.

SLAVÍK, Bedřich (1947). *Písemnictví na moravském Valašsku*. Olomouc: R. Promberger.

SLAVÍK, Bedřich (1940). *Hanácké písemnictví*. Olomouc: R. Promberger.

SNČJ = *Slovník nářečí českého jazyka* (2011–2024). Brno: Ústav pro jazyk český AV ČR. Online. Dostupné z: https://sncj.ujc.cas.cz/.

SPILKA, Josef (1956–1975). *Rčení a pranostiky*. Osobní fond Josef Spilka II. Rčení a pranostiky. Praha: Etnologický ústav AV ČR.

SSJČ = *Slovník spisovného jazyka českého* (1960–1971). Praha: Nakladatelství Československé akademie věd.

STARÁ, Alena (2008). *Lidová přísloví, přirovnání a pořekadla*. Jablunkov: Alena Stará.

STEINBECK, John (1941). *Hrozny hněvu*. Přeložil Vladimír Procházka. Praha: Evropský literární klub.

STEUER, Felix (1932). Branické podřečí. *Časopis Vlasteneckého spolku musejního v Olomouci*, roč. 45, s. 211–249.

STOLAŘÍK, Ivo a ŠTIKA, Jaroslav (eds., 1997–2001). *Těšínsko 1–3*. Šenov u Ostravy: Tillia. ISBN 80-86101-01-0 (soubor).

STUDNIČKA, Alois (1953). *Nářečí Vnorov a okolí*. Disertační práce. Brno: Masarykova universita.

STUPŇÁNEK, Bronislav a kol. (2022). *DiDa*. Databáze elektronických dokumentů. Brno: Ústav pro jazyk český AV ČR. Online. Dostupné z: https://www.ceskanareci.cz/geoportal/dida/dump.sql.

STUPŇÁNEK, Bronislav a VONDRÁKOVÁ, Alena (2022a). *Vybrané jevy hláskoslovného vývoje českých nářečí ve 20. století*. Olomouc: Univerzita Palackého v Olomouci ve spolupráci s Ústavem pro jazyk český AV ČR. ISBN 978-80-244-6157-1 (print), 978-80-244-6161-8 (iPDF). Též online. Dostupné z: https://www.ceskanareci.cz/geoportal/?sdm_process_download=1&download_id=143.

STUPŇÁNEK, Bronislav a VONDRÁKOVÁ, Alena (2022b). *Slezské asibilace v druhé polovině 20. století*. Olomouc: Univerzita Palackého v Olomouci ve spolupráci s Ústavem pro jazyk český AV ČR. ISBN 978-80-244-6158-8 (print), 978-80-244-6162-5 (iPDF). Též online. Dostupné z: https://ceskanareci.cz/dokumenty/MAPS_20_mapa.pdf.

STUPŇÁNEK, Bronislav a VONDRÁKOVÁ, Alena (2022c). *Výsledky vývoje tvrdého* y *ve 20. století*. Olomouc: Univerzita Palackého v Olomouci ve spolupráci s Ústavem pro jazyk český AV ČR. ISBN 978-80-244-6159-5 (print), 978-80-244-6163-2 (iPDF). Též online. Dostupné z: https://ceskanareci.cz/dokumenty/MAPS_21_mapa.pdf.

STUPŇÁNEK, Bronislav a VONDRÁKOVÁ, Alena (2022d). *Ústup dvojího* l *během 20. století*. Olomouc: Univerzita Palackého v Olomouci ve spolupráci s Ústavem pro jazyk český AV ČR. ISBN 978-80-244-6160-1 (print), 978-80-244-6164-9 (iPDF). Též online. Dostupné z: https://ceskanareci.cz/dokumenty/MAPS_22_mapa.pdf.

STUPŇÁNEK, Bronislav a IREINOVÁ, Martina (2020). Druhý život nářečí. Centrální středomoravské dialekty užívané ve veřejné komunikaci a tzv. hanácké obrození. *Národopisná revue*, roč. 30, č. 4, s. 317–325. ISSN 0862-8351 (print); ISSN 2570-9437 (online). Též online. Dostupné z: https://revue.nulk.cz/wp-content/uploads/2021/04/r4-2020.pdf.

SUŠIL, František (1998). *Moravské národní písně. S nápěvy do textu vřaděnými*. 5. vydání. Praha: Argo. ISBN 80-7203-096-5.

ŠIMEČKOVÁ, Marta, KARAFIÁT, Martin a PLCHOT, Oldřich (v tisku). Automatická detekce dialektů strojovým učením: nové možnosti české dialektologie. In: *Slovanské dialekty v době digitálních technologií. Nářeční prameny a jejich současné zpracování*. Praha: Slovanský ústav AV ČR.

ŠIMEČKOVÁ, Marta, KUBEČEK, Filip, ŽIŽKA, Josef a NÉTEK, Rostislav (2024*). Mapa ligy superdialektologů*. Verze 1. Online. Dostupné z: https://www.jamap.cz/supermapa.

ŠIMEČKOVÁ, Marta a kol. (2022–). *Ve slovech*. Online. Dostupné z: https://veslovech.cz/.

ŠIMEČKOVÁ, Marta (2024a). *Archiv zvukových záznamů nářečních promluv*. Praha: Academia. ISSN 2464-6245. Též online. Dostupné z: http://www.vedakolemnas.cz/sys/galerie-download/VKN-132_.pdf.

ŠIMEČKOVÁ, Marta (2024b). Zvukový archiv Veslovech.cz. *Jazykovědné aktuality*, roč. 61, č. 1, s. 60–63. ISSN 1212-5326. Též online. Dostupné z: https://www.jazykovednesdruzeni.cz/wp-content/uploads/2024/08/JA_1_24.pdf.

ŠIMEČKOVÁ, Marta (2024c). *Metoda rozhovoru (interview) ve společenských a humanitních vědách // The interview method in the social sciences and humanities*. Online. Dostupné z: https://forms.gle/avU6ZDnPCqV3HHHPA. [El. dotazník.]

ŠIMEČKOVÁ, Marta (2022). Jazyky a nářečí jako součást regionální identity a kulturní dědictví (na příkladu České republiky). *Národopisná revue*, č. 2, s. 114–124. ISSN 0862-8351. Též online. Dostupné z: https://revue.nulk.cz/wp-content/uploads/2022/07/r2-2022.pdf.

ŠIPKOVÁ, Milena (1992). Podmínkové věty v hanáckých nářečích. *SFFBU*, roč. 41, č. A 40, s. 55–59. ISSN ISSN 0231-7567.

ŠKARNÉTKA, Cyril, KUBÍKOVA, Francka a SAINT-EXUPÉRY, Antoine de (2024). *Malušenký princ*. Brno: Jota. ISBN 978-80-7689-376-4.

ŠTRUBL, Ondřej, NÉTEK, Rostislav a STUPŇÁNEK, Bronislav (2022). *DiaMa*. Geoportál dialektologických map. Olomouc: Univerzita Palackého v Olomouci. Online. Dostupné z: https://www.ceskanareci.cz/geoportal/diama/.

TAGLIAMONTE, Sali A. (2006). *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press. ISBN 978-0-521-77115-3.

TAGLIAMONTE, Sali A. (2005). *APPENDIX B. INTERVIEW SCHEDULE. Guideline Questions (Adapted from Labov 1973)*. Online. Dostupné z: https://www.cambridge.org/files/5213/6689/9619/2846_APPENDIX_B.pdf.

TAUSCH, Jaromír (2006). *Batelovsko. Kronika o lidové kultuře, tradicích a životě lidí*. Batelov: Obecní úřad v Batelově. ISBN 80-86391-21-3.

TED2020 (2020). Online. Dostupné z: https://opus.nlpl.eu/TED2020/cs&en/v1/TED2020.

The Survey of English Dialects /SED/ (2024). Online. Dostupné z: https://dialectandheritage.org.uk/about/the-survey-of-english-dialects/.

THOMA, Ludwig (1966). *Dopisy poslance bavorského zemského sněmu*. Přeložil Jaroslav Homola. Praha: Odeon.

TKÁČ, Filip (2023). *Metody využívané v dialektologii na příkladu výzkumu nářečí v Kozlovicích*. Diplomová práce. Brno: Masarykova univerzita.

Transkripce v korpusu DIALEKT (2018). Online. Dostupné z: https://wiki.korpus.cz/doku.php/cnk:dialekt:pravidla.

UTĚŠENÝ, Slavomír (1960). *Nářečí přechodného pásu česko-moravského*. Praha: Nakladatelství ČSAV.

VAJDOVÁ, Zdenka, ČERMÁK, Daniel a ILLNER, Michal (2006). *Autonomie a spolupráce: důsledky ustavení obecního zřízení v roce 1990*. Praha: Sociologický ústav Akademie věd České republiky. ISBN 80–7330–086-9.

VALIHRACHOVÁ, Zdenka (1971). *Nářeční promluvy Ludmily Sojkové ze Staroviček*. Diplomová práce. Brno: Universita Jana Evangelisty Purkyně.

VANĚK, Miroslav, MÜCKE, Pavel a PELIKÁNOVÁ, Hana (2007). *Naslouchat hlasům paměti. Teoretické a praktické aspekty orální historie*. Praha: Ústav pro soudobé dějiny AV ČR. ISBN 978-80-7285-089-1.

VASWANI, Ashish a kol. (2017). Attention is All you Need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*, s. 6000–6010.

VAŠÍČEK, Michal (2020). *Dynamika jihokarpatských nářečí*. Praha: Slovanský ústav AV ČR. ISBN 978-80-86420-74-5.

VÁŽNÝ, Václav (1955). K otázce jazykového atlasu zemí českých. *Slovo a slovesnost*, roč. 16, č. 3, s. 159–173.

VOŽENÍLEK, Vít, IREINOVÁ, Martina, VONDRÁKOVÁ, Alena a KONÍČEK, Jakub (2022). Mapping, synthesis and visualization of Czech dialects. *International Journal of Cartography*, roč. 8, č. 1, s. 148–163. ISSN 2372-9333. Též online. Dostupné z: https://doi.org/10.1080/23729333.2021.1978039.

VOŽENÍLEK, Vít, KAŇOK, Jaromír a kol. (2011). *Metody tematické kartografie – vizualizace prostorových jevů*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978–80–244–2790–4.

WEBDEVEL (2024). *Responsivní design*. Online. Dostupné z: https://www.webdevel.cz/responzivni-design/.

WELLS, John C. (1995). Computer-coding the IPA: a proposed extension of SAMPA. Online. Dostupné z: https://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf.

WIELING, Martijn, REBERNIK, Teja a JACOBI, Jidde (2023). SPRAAKLAB: a mobile laboratory for collecting speech production data. In: SKARNITZL, Radek a VOLÍN, Jan (eds.). *Proceedings of the 20th International Congress of Phonetic Sciences*. Praha: Guarant International, s. 2060–2064. ISBN 978-80-908 114-2-3.

WILSON, James (2017). Teorie akomodace. In: KARLÍK, Petr, NEKULA, Marek a PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Online. Dostupné z: https://www.czechency.org/slovnik/TEORIE AKOMODACE.

WODARZ, Jan (1955). Soustava melodických prostředků v nářečí západního Hlučínska. *Slezský sborník*, roč. 53, s. 527–543.

WOLFRAM, Walt (1993). Ethical Considerations in Language Awareness Programs. *Issues in Applied Linguistics*, roč. 4, č. 2, s. 225–255. ISSN 1050-4273.

WYDERKA, Bogusław (2014). O rozwoju polskich dialektów. *Poznańskie Studia Polonistyczne. Seria Językoznawcza*, roč. 21, č. 2, s. 103–113. ISSN 1233-8672.

YU, Dong a DENG, Li (2016). *Automatic Speech Recognition: A Deep Learning Approach*. London: Springer. ISBN 978-1-4471-6967-3.

ZÁBRANSKÝ, Alois (1919). *Břehule čili poslední hastrman na Hané. Trilogie*. Praha: Unie.

ZAORÁLEK, Jaroslav (1996). *Lidová rčení*. 3. vydání. Praha: Aurora. ISBN 80-85974-10-X.

ZÍBRT, Čeněk (1909–1911). *Veselé chvíle v životě lidu českého I–VIII*. Praha: Šimáček.

ZÍKOVÁ, Magdalena a KŘIVAN, Jan (2014). Nahrávání v terénním lingvistickém výzkumu: jak získat kvalitní záznam řeči? *Studie z aplikované lingvistiky*, roč. 5, č. 1, s. 65–82. ISSN 2336-6702.

ZHAO, Ding, SAINATH, Tara N., RYBACH, David a kol. (2019). Shallow-Fusion End-to-End Contextual Biasing. In: *Proc. Interspeech*, s. 1418–1422.

12

Seznam publikací, které předcházely metodice a byly publikovány

SEZNAM PUBLIKACÍ, KTERÉ PŘEDCHÁZELY METODICE A BYLY PUBLIKOVÁNY

Metodika je na počátku své implementace, z toho důvodu budou další publikační výstupy obsahující zkušenosti z jejího uplatnění teprve následovat. Veškeré výstupy budou k dispozici prostřednictvím informačního systému VaV, jejich seznam bude též průběžně aktualizován na projektovém webu **www.jamap.cz** (sekce Výstupy).

Publikované výstupy dedikované k projektu, jehož součástí je též předložená metodika:

MATĚJKA, Pavel, SILNOVA, Anna, SLAVÍČEK, Josef, MOŠNER, Ladislav, PLCHOT, Oldřich, KLČO, Michal, PENG, Junyi, STAFYLAKIS, Themos a BURGET, Lukáš (2023). Description and Analysis of ABC Submission to NIST LRE 2022. In: HARTE, Naomi, CARSON-BERNDSEN, Julie a JONES, Gareth (eds.). *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Dublin: International Speech Communication Association, s. 511–515. ISSN 2308457X.

Hlavní publikované výstupy, které sice nebyly dedikovány k projektu, jehož součástí je předložená metodika, avšak na nichž se podíleli autoři metodiky a jež obsahují postupy v této metodice rozvíjené:

KOCOUR, Martin, CÁMBARA, Guillermo, LUQUE, Jordi, BONET, David, FARRÚS, Mireia, KARAFIÁT, Martin, VESELÝ, Karel a ČERNOCKÝ, Jan (2021). BCN2BRNO: ASR System Fusion for Albayzin 2020 Speech to Text Challenge. In: *Proceedings of IberSPEECH 2021*. Vallaloid: International Speech Communication Association, s. 113–117. Též online. Dostupné z: https://www.fit.vut.cz/research/publication/12577/.

ŠIMEČKOVÁ, Marta (2024). *Archiv zvukových záznamů nářečních promluv*. Praha: Academia. ISSN 2464-6245. Též online. Dostupné z: http://www.vedakolemnas.cz/sys/galerie-download/VKN-132_.pdf.

VOŽENÍLEK, Vít, IREINOVÁ, Martina, VONDRÁKOVÁ, Alena a KONÍČEK, Jakub (2022). Mapping, synthesis and visualization of Czech dialects. *International Journal of Cartography*, roč. 8, č. 1, s. 148–163. ISSN 2372-9333. Též online. Dostupné z: https://doi.org/10.1080/23729333.2021.1978039.

VYDANA, Hari K., KARAFIÁT, Martin, ŽMOLÍKOVÁ, Kateřina, BURGET, Lukáš a ČERNOCKÝ, Jan (2021). Jointly Trained Transformers Models for Spoken Language Translation. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario: IEEE Signal Processing Society, s. 7513–7517. ISBN 978-1-7281-7605-5. Též online. Dostupné z: https://www.fit.vut.cz/research/publication/12522/.

PŘÍLOHY

PŘÍLOHA 1

Struktura obsahových metadat v *Databázi nářečních promluv pro odbornou veřejnost*

- **NAHRÁVKA**
  - jazyk
    - místní a pomístní názvy
    - nářečí
    - příjmení a přezdívky, posměšky
    - cizí jazyky
  - útvar
    - humorné vyprávění
    - historka
    - zpěv, recitace
    - pověst, pohádka, pověra
      - poklad
      - strašidla
        - bludičky, světýlka
        - čarodějnice
        - čert
        - divoženka
        - duch
        - hejkal
        - ohnivý muž
        - polednice
        - vodník
    - hádanka, vtip
    - rozhovor s dítětem
  - dotazník
    - dotazník ATLAS
    - dotazník JAMAP 1
    - dotazník JAMAP 2

PŘÍLOHA 1
Struktura obsahových metadat v *Databázi nářečních promluv pro odbornou veřejnost*

- **PŘÍRODA**
  - rostliny
    - plodiny, užitkové rostliny
      - obilí
      - kukuřice
      - konopí
      - len
      - chmel
      - luštěniny
      - zelenina
        - brambory
        - zelí
        - řepa
      - ovoce, ovocné stromy a keře
        - švestky
        - vinná réva
    - luční rostliny
      - seno
    - lesní rostliny a plody
      - dřevo, klestí
      - houby
    - léčivé byliny
    - nemoci rostlin
  - živočichové
    - hospodářská zvířata
      - drůbež
      - holubi
      - hovězí dobytek
      - husy
      - koně
      - kozy
      - králíci
      - ovce
      - prasata
      - včely
    - škůdci
    - domácí zvířata, mazlíčci
    - exotická zvířata

- lesní zvířata
- polní zvířata
- nemoci zvířat
- lidské tělo, zdraví a nemoci
  - lidské nemoci, úrazy
  - postižení
    - fyzické postižení
    - duševní postižení
  - pobyt v nemocnici
  - lidové léčitelství, bylinkářství
  - závislost
    - kouření, tabák
    - alkoholismus
- počasí
  - roční období
  - přírodní katastrofa
    - bouřka
    - povodeň
    - požár
    - mráz
    - sněhová vánice, závěj
    - sucho

KULTURA

- hmotná kultura
  - obydlí, lidová architektura
    - stavba a rekonstrukce
    - statek, hospodářství
    - střecha
    - svícení
    - pec, kamna
    - černá kuchyně
    - kuchyně
    - nábytek
    - kaplička, kostel (stavba)
    - stodola
    - studna
    - žudro

- pokrmy a nápoje
  - pokrmy, jídlo
    - chléb
    - pečivo
    - kaše
    - přílohy
    - knedlíky
    - sladké pokrmy
    - maso a masné výrobky
    - mléko a mléčné produkty
    - polévky
    - sádlo
  - nápoje, pití
    - pivo
    - víno
    - tvrdý alkohol, lihoviny
    - káva a její náhražky
  - zavařování
  - zabijačka
- odívání, vzhled
  - obuv
  - oděv
  - krojové odívání
  - šátek
  - účes

duchovní kultura

- zvyky, svátky
  - advent
  - dožínky
  - Dušičky
  - hody
  - poutě
  - vinobraní
  - jízda králů
  - masopust
  - Mikuláš
  - obchůzka lucek
  - obchůzka barbor

		průvod králek, královniček
		pálení čarodějnic
		smrtná neděle, vynášení smrtky
		stavění, kácení máje
		svatodušní svátky
		Tři králové
		Vánoce
		Velikonoce
	víra, náboženství	
		kostel, farnost
		modlení, mše
		křest
		svaté přijímání
		biřmování
		pouť, procesí
		svatí
		zázrak
		ministrování
	škola, vzdělávání	
		učitel
		výběr školy, budoucí studium
		budoucí povolání
		školní předměty
		zájmový kroužek
		školní výlet, tábor
		prázdniny
volný čas		
	četba	
	kino, film	
	rádio	
	televize	
	cestování	
	dovolená	
	chalupaření	
	koupání	
	kulturní, společenské akce	
		divadlo
		lidový tanec

- obec a okolí
  - vybavenost a infrastruktura
    - popis obce, proměna obce
    - budování obce
      - elektrifikace
      - plynofikace
      - obecní vodovod
    - pošta
    - telefonování
    - doprava
      - auto
      - autobus
      - vlak
    - hasiči
  - lidé
    - obyvatelé obce
    - mladá generace
    - příslušníci národnostních menšin
    - německé obyvatelstvo
    - polské obyvatelstvo
    - romské obyvatelstvo
    - židovské obyvatelstvo
    - cizinci
  - sport, sportovní akce
  - spolky
    - Skaut
    - Sokol
    - Orel
    - Junák
    - folklorní spolek, soubor
  - dětské hry, hračky
  - hudba, kapela
  - hospoda
    - maškarní, ples
    - taneční
    - taneční zábava

PŘÍLOHA 1
Struktura obsahových metadat v *Databázi nářečních promluv pro odbornou veřejnost*

- **PRÁCE**
  - dům a usedlost
    - pozemek
      - pole
      - louka
      - les
      - zahrada, sad
      - vinohrad
    - rostlinná výroba
      - pěstování
      - sklizeň, sběr
      - zpracování
    - živočišná výroba, lov
      - chov
        - pasení
        - dojení
      - myslivost
      - rybaření
      - pytláctví
    - domácí práce
      - draní peří
      - háčkování, pletení, šití, vyšívání
      - bělení, praní n. sušení prádla
      - předení
      - úklid
      - vaření, pečení
    - nářadí, nástroje a stroje
      - cep
      - nůše
      - trakař
      - pluh, ruchadlo
      - postroj
      - selský vůz a jeho součásti
      - kolovrátek
  - tradiční řemeslo a výroba
    - výroba došků
    - knoflíkářství

	košíkářství, ošatkářství	
	výroba mýdla	
	výroba povřísel	
	metlařství	
	klempířství	
	kovářství	
	krejčovství	
	mlynářství	
	pekařství	
	řezbářství	
	tkalcovství, plátenictví	
	ševcovství	
	tesařství	
	truhlářství	
	řeznictví	
	vinařství	
	výroba, pálení alkoholu	
	podomní obchodníci	
zaměstnání a provoz		
	přírodní zdroje	
		práce na pile
		práce v lese, lesnictví
		vorařství, plavení dřeva
		rybníkářství
		práce v lomu
		hornictví
		práce v drůbežárně
		práce v zemědělském družstvu (rostlinná výroba)
		práce v zemědělském družstvu (živočišná výroba)
	průmysl	
		práce v továrně
		obuvnictví
		sklářství

			zednictví
			cukrovarnictví (= cukrovar)
		služby	
			formanství
			obchodování
			trh, jarmark
			pašování
		zvláštní typy prací	
			robota
			služba
			dětská práce
			učňovství, vyučení
			brigáda
			nezaměstnanost
			žebráctví, vandráctví
UDÁLOSTI			
	historie, dějiny		
		válka	
			1. světová válka
			2. světová válka
			osvobození
			mobilizace
			totální nasazení
			partyzáni
			dezerce
			zajetí
			ukrývání
			válečné zranění
			rusko-ukrajinská válka
		meziválečné období	
		poválečné období (po r. 1945)	
			odsun Němců
		komunismus, totalita	
			kolektivizace, JZD
			rok 1968
			rok 1989
		politika	
		odchod do exilu	

	rozpad Československa
	slavné osobnosti
pobyt, místo	
	Praha
	Albánie
	Amerika
	Argentina
	Belgie
	Francie
	Holandsko
	Itálie
	Jugoslávie
	Maďarsko
	Německo
	Polsko
	Rakousko
	Rakousko-Uhersko
	Rumunsko
	Rusko
	Řecko
	Slovensko
	Sovětský svaz
	Srbsko
	Ukrajina
	Velká Británie
rodina	
	těhotenství, porod
	dětství
	péče o dítě, o člena rodiny
	seznámení, námluvy
	svatba
	výročí svatby
	nevěra, zálety
	rozvod
	důchod
	výměnkářství
	smrt
	pohřeb

	dědictví
	příbuzenské vztahy
	rodinné neshody
sociální a právní situace	
	vojna, vojenské cvičení
	chudoba
	nedostatek jídla
	dluhy
	soud
	vězení
	krádež, přepadení
	neštěstí
	sebevražda
	vražda

PŘÍLOHA 2

Leták s výzvou k nářečnímu výzkumu na jihovýchodní Moravě

MAPUJEME NÁŘEČÍ
PRO PŘÍŠTÍ GENERACE!

Co se chystá:
NÁŘEČNÍ VÝZKUM V OBCÍCH JIHOVÝCHODNÍ MORAVY

Kdy:
V PRŮBĚHU ROKU 2023

Proč:
CHCEME ZDOKUMENTOVAT JAZYKOVÉ DĚDICTVÍ NAŠICH PŘEDKŮ

Kdo přijede: **výzkumníci Akademie věd České republiky**

Koho hledáme: **nářeční mluvčí, ochotné popovídat si s námi**

Výzkum probíhá formou přátelského rozhovoru, není se čeho bát!

Mluvíte nářečím? Pak hledáme právě vás!
Zapojte se do výzkumu, a pomozte tak zdokumentovat jazykové dědictví svých předků.

Hledáme mluvčí starší 60 let,
nejlépe **starousedlíky a aktivní nositele dialektu**,
kteří by byli ochotni najít si čas popovídat si s námi.

Výzkum není žádné zkoušení,
probíhá formou rozhovoru na různá témata,
například jaké zvyky se ve vaší obci udržují,
co se dříve vařilo a peklo nebo jak se u vás říká pampelišce.

Je to u vás **pumpeliška, pampelica, pléška, plešek, nebo mlíčák?**

Chcete se zapojit do výzkumu? Pak se nám ozvěte.
Kontakt: Marta Šimečková, tel. +420 775 397 427, simeckova@ujc.cas.cz

Výzkumy jsou organizovány ve spolupráci s Muzejním spolkem Prušánky, z. s.

Pro více informací navštivte **www.jamap.cz**

PŘÍLOHA 3

Informovaný souhlas k pořízení zvukového záznamu, jeho archivaci a nakládání a ke zpracování a zpřístupnění osobních údajů

Já, níže podepsaný/á (dále jen informant) .....,

podle zákona č. 110/2019 Sb., o zpracování osobních údajů, zákona č. 89/2012 Sb., občanského zákoníku, ve znění pozdějších předpisů, a v souladu s nařízením (EU) 2016/679 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů (GDPR) a v souladu s Etickým kodexem Akademie věd ČR, tímto uděluji svůj výslovný a svobodný souhlas se zpracováním svých osobních údajů, které jsem poskytl/a během rozhovorů pořízených ve zvukové či audiovizuální podobě v rámci terénních výzkumů realizovaných pod záštitou projektu *Jazyková paměť regionů České republiky. Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka* (NAKI III, č. DH23P03OVV010) pro následující účely:

- vědecký výzkum v oblasti humanitních a přírodních věd;
- archivace ve veřejném zájmu;
- výstavní činnost (možnost využití nahrávek či jejich transkriptů na výstavě);
- vzdělávací činnost (možnost využití nahrávek či jejich transkriptů při výuce nebo na vzdělávacích či popularizačních akcích pro veřejnost);
- publikační činnost (možnost využití nahrávek či jejich transkriptů ve studiích, monografiích, slovnících, atlasech, mapách aj.);
- digitalizace a možnost zpřístupnění na veřejné webové platformě https://www.jamap.cz/.

Souhlasím s poskytnutím a zpřístupněním následujících osobních údajů:

- rok narození;
- pohlaví;
- geografické údaje týkající se rodiště a místa pobytu;
- geografické údaje týkající se rodiště a místa pobytu rodičů a prarodičů;
- informace o zaměstnání (a to i před nástupem do důchodu);
- informace o nejvyšším dosaženém vzdělání.

Souhlasím s poskytnutím následujících údajů pro interní potřeby a evidenci Ústavu pro jazyk český AV ČR, v. v. i.:

- rodné jméno a příjmení;
- korespondenční adresa;
- e-mail;
- telefonní číslo.

Jsem si vědom/a skutečnosti, že správcem těchto údajů se na dobu nezbytně nutnou pro využití dat k danému vědeckému výzkumu podpisem tohoto souhlasu stává Ústav pro jazyk český AV ČR, v. v. i., který údaje a nahrávky (vč. jejich transkriptů) archivuje ve sbírce *Databáze nářečních promluv pro odbornou veřejnost*. Také jsem si vědom/a toho, že svůj souhlas se zpracováním osobních údajů mohu kdykoli odvolat, a to písemně nebo elektronickou poštou na níže uvedených kontaktních adresách.

PŘÍLOHA 3
Informovaný souhlas k pořízení zvukového záznamu a nakládání a ke zpracování a zpřístupnění osobních údajů

Souhlasím s tím, aby texty a informace v nich obsažené byly v případě potřeby za souhlasu Etické komise Akademie věd ČR a v souladu s etickým kodexem Akademie věd ČR poskytnuty v písemné, elektronické či audiální podobě na základě podmínek, jež stanoví správce, také dalším subjektům (domácí či zahraniční akademická pracoviště a výzkumné instituce, badatelé z řad veřejnosti), a to pro účely vědeckého výzkumu a souvisejících výstav a pro účely žurnalistiky, které zahrnují možnost zveřejnění rozhovorů, či pro účely jiného bádání, které není v rozporu s mými zájmy (např. zájmové bádání laické veřejnosti).

Hodící se zaškrtněte: ☐ souhlasím ☐ nesouhlasím

Jsem si vědom/a, že jsem byl/a informován/a o tom, že podle výše uvedených právních předpisů o zpracování osobních údajů mám právo:

- kdykoli odvolat udělený souhlas, a to pro každý ze shora uvedených účelů samostatně;
- vyžádat si informaci o tom, jaké osobní údaje jsou o mně zpracovávány;
- vyžádat si opravu nebo doplnění svých osobních údajů;
- žádat okamžitý a bezpodmínečný výmaz osobních údajů;
- žádat o omezení zpracovávání údajů, které jsou nepřesné, neúplné nebo u nichž odpadl důvod jejich zpracování;
- žádat umožnění přenesení zpracovávaných údajů;
- dostat odpověď na svou žádost bez zbytečného odkladu, v každém případě do jednoho měsíce od obdržení žádosti správcem.

Pro kontaktování správce ve věci ochrany osobních údajů lze využít následující kontakty:

- správce *Databáze nářečních promluv pro odbornou veřejnost*: Mgr. Marta Šimečková, Ph.D., dialektologické oddělení ÚJČ AV ČR, v. v. i., Veveří 97, 602 00 Brno, e-mail: simeckova@ujc.cas.cz; tel.: +420 775 397 427
- Úřad pro ochranu osobních údajů; e-mail: posta@uoou.gov.cz; tel.: +420 234 665 111

Datum, místo podpisu: .....

Podpis: .....

METODIKA PRO PŘEVOD STRUKTUROVANÝCH ZNALOSTÍ Z OBORU DIALEKTOLOGIE DO STROJOVÉHO UČENÍ

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

Program na podporu aplikovaného výzkumu
v oblasti národní a kulturní identity 2023–2027

JaMap

PRO VÍCE INFORMACÍ SLEDUJTE

www.jamap.cz

Ústav pro jazyk český
Akademie věd České republiky

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

KATEDRA GEOINFORMATIKY
Univerzita Palackého v Olomouci