Tým z FIT VUT pomáhá budovat jedinečnou mapu nářečí

7.2.2024 - | Vysoké učení technické v Brně

Na unikátním projektu mapování nářečí se podílí tým z Fakulty informačních technologií VUT pod vedením Martina Karafiáta. Ve spolupráci s Akademií věd ČR a Univerzitou Palackého v Olomouci vytváří webové stránky, na kterých si bude možné zvolit oblast České republiky a poslechnout si dialekty charakteristické pro dané místo. Projektoý tým navíc nahrávky, které jdou zpět až do 50. let minulého století, kategorizuje podle různých kritérií, například podle témat vyprávění.

Výzkumníci z dialektologického oddělení Ústavu pro jazyk český Akademie věd ČR se dlouhodobě snaží o zmapování a uchování nejrůznějších nářečí napříč Českem. V roce 2023 si na pomoc přizvali i odborníky ze skupiny Speech@FIT, kteří momentálně pracují na tvorbě systému, který by byl schopen dialekt identifikovat. A také vytvořit automatický přepis nahrávek. „Naše řečová skupina má velké úspěchy v oblasti identifikace jazyka, mluvčího a přepisu řeči. Primární myšlenka je tedy dát tyto oblasti dohromady, pracovat s unikátními daty a vytvořit systém, který bude schopen zvuková data automaticky přepisovat, což bude pro výzkumníky z Akademie věd ČR obrovská pomoc. Zejména proto, že jsou data specifická a klasické přepisovače od Googlu či Microsoftu selhávají," vysvětluje Martin Karafiát z FIT VUT. To potvrzuje i hlavní řešitelka projektu z Ústavu pro jazyk český Akademie věd ČR Marta Šimečková. „Naší snahou je vytvořit sadu nástrojů, které by nám, dialektologům, usnadňovaly práci. Jednak je to software na automatické rozpoznávání konkrétního nářečí na základě audionahrávky, jednak software, který by za nás pořizoval přepis nářečních promluv. Jde přitom o přepisy ve speciální dialektologické transkripci, která se v mnohém liší od spisovného zápisu," přibližuje Šimečková.

Archiv nářečních nahrávek vzniká už od 50. let minulého století a data pořád přibývají. „Kdysi byl v dialektologickém oddělení jeden kotoučový magnetofon. Navíc byly drahé pásky, takže se šetřilo a nahrávaly se jen malé úseky. Dnes se ale nechá nahrávání běžet i několik hodin. Data uložená na starých zvukových nosičích se ve spolupráci s Českým rozhlasem digitalizovala, následně se anotovala a katalogizovala. Systém katalogizace je ale dnes nevyhovující, a tak se přistoupilo k hloubkové revizi nahrávek a k vytvoření nového, moderního katalogu, ve kterém jsou k datům pořizovány popisky jednotným způsobem. Mimo jiné také informace o jejich obsahu," popisuje Martin Karafiát.

Archiv nahrávek nářečí vzniká už od 50. let minulého století | Autor: archiv AV ČR

Do budoucna by pak měli být zájemci schopni jednoduše v nahrávkách vyhledávat podle vybraného nářečí i tématu. „Chceme, aby si člověk mohl říct, že ho třeba zajímá, jak zní, když někdo povídá v hanáčtině o pečení chleba. A systém mu obratem takovou nahrávku nabídne," říká Karafiát. Podle Marty Šimečkové už je většina tradičních nářečí zmapovaná. „Zejména díky sběrům, které proběhly v 60. a 70. letech 20. století. Nahrávky z této doby tvoří jádro našeho zvukového archivu. Jedinými bílými místy je pohraničí, což je oblast nářečně nepůvodní, a tak se tu dříve spíše nezkoumalo. Naší snahou bude hlavně doplnit záznamy z tradičně nářečních oblastí, díky čemuž bude možné sledovat některé posuny dialektů v čase," dodává.

Podle Marty Šimečkové už je většina tradičních nářečí zmapovaná | Autor: archiv AV ČR Ačkoliv se na první pohled může zdát, že projekt pro výzkumníky z FIT VUT není žádnou obtížnou výzvou,

Martin Karafiát upozorňuje na složitost přepisu nářečí. „Je to podobné například vietnamštině. Ta také používá k zápisu latinku, ale pomáhá si i sadou pomocných symbolů, které určují, jak se má konkrétní znak vyslovit," vysvětluje Karafiát s tím, že budou muset systém naučit zaznamenávat například takzvané obalované l, které je typické pro některá nářečí na jihu Moravy a pro Jablunkovsko.

Tým už na podzim loňského roku vytvořil první verzi systému pro identifikaci dialektu. „Ten je schopen zhruba na 90 procent rozlišit čtyři hlavní nářeční skupiny. Když je ale rozdělíme na 13 podskupin, tak už je úspěšnost jen okolo 60 procent. Bude se to v čase zdokonalovat, protože systém ještě nebyl trénovaný na datech od Akademie věd ČR. Naše neuronová síť je trénovaná na 106 cizích jazycích, ale zatím neviděla dialektickou češtinu," upozorňuje Martin Karafiát.

On sám se soustřeďuje v projektu na přepis textu. Další členové týmu pak na automatickou identifikaci dialektu a tvorbu webového rozhraní. Ve spolupráci s Univerzitou Palackého v Olomouci totiž řeší i tvorbu dialektické mapy a vizuální zpracování dat. Spuštění této online mapy je naplánováno na rok 2027. „Půjde o první mapu svého druhu u nás. Uživatelé si budou moci přehrávat nářeční ukázky z různých regionů a zároveň procházet jejich přepisy. Mapa umožní zobrazení dat na různých podkladech a také vyhledávání v nahrávkách i přepisech podle různých parametrů, třeba podle témat vyprávění," říká Marta Šimečková.

Ve vyhledávání pamětníků, kteří ještě dobová nářečí ovládají, pomáhají výzkumníkům obecní úřady, školy i folklórní spolky. „Většinou se nám podaří v každé obci nahrát alespoň čtyři mluvčí, odmítnuti jsme jen výjimečně. Dokumentujeme zejména promluvy seniorů, kteří jsou často rádi, že je někdo ochotný jejich vyprávění naslouchat. A navíc tím pomůžou dobré věci, totiž uchování jazykového kulturního dědictví příštím generacím," přibližuje Marta Šimečková.

Pamětníky ovládající nářečí pomáhají vyhledávat obecní úřady i folklórní spolky | Autor: archiv AV ČR Tým Martina Karafiáta prý v začátku projektu uvažoval i nad tvorbou chatbota, který by jednotlivá nářečí ovládal a v budoucnu mohl se zájemci komunikovat ve vybraném dialektu. „Ale udělat dobře přepis řeči a naučit systém spolehlivě identifikovat, že se jedná například o některý jihozápadočeský nebo východomoravský dialekt, bude samo o sobě dost práce," dodává na závěr se smíchem Martin Karafiát.

https://www.zvut.cz/tema/tema-f38144/tym-z-fit-vut-pomaha-budovat-jedinecnou-mapu-nareci-d252599