MLUVENÉ SLOVO V POŘADECH ČESKÉHO ROZHLASU: ZPRÁVA O VÝZKUMNÉM PROJEKTU ZPŘÍSTUPNĚNÍ ARCHIVU ČESKÉHO ROZHLASU PRO SOFISTIKOVANÉ VYHLEDÁVÁNÍ

Mluvené slovo se v poslední době dostává stále častěji do středu pozornosti různých vědních oborů. Obvyklý problém pro práci s audiodaty představuje paralelnost ubíhání textu s reálným časem, a tedy velmi obtížná zachytitelnost konkrétních zkoumaných jevů, možnost vytváření jejich databáze a následná práce s těmito daty. Ruční přepisy mluveného slova pro potřeby výzkumu jsou natolik časově i fyzicky náročné, že není možné je využít pro sestavování statisticky validních databází, které by pro výzkum poskytly potřebné množství vstupů.

Jakkoli se dnes díky počítačovým technologiím stává možné analyticky pracovat přímo se zvukovými nahrávkami, je zvukový záznam mluveného slova vždy těžko uchopitelný. Z tohoto důvodu pořizování textových přepisů zvukových verbálních nahrávek je, a zřejmě i nadále bude, standardní součástí zpracování dat, zvláště kvůli možnostem jejich analýzy a interpretace.

Následující text informuje o projektu, jehož cílem je zpřístupnění dat archivu vysílání Českého rozhlasu pro sofistikované vyhledávání,[1] tj. zpřístupnění velkého množství dat primárně akustického kódu paralelně s automatickým převodem do kódu grafického.

Výzkumný projekt, který chceme představit, se zaměřuje na automatický přepis mluveného slova pořadů Českého rozhlasu od jeho založení (1923) až po současnost.

Projekt je realizován od roku 2010 a potrvá do roku 2015, kdy budou jeho výsledky zpřístupněny pro veřejnost. Řešitelským pracovištěm je Fakulta mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci, na projektu spolupracuje Katedra českého jazyka a literatury Fakulty přírodovědně-humanitní a pedagogické téže univerzity, celý výzkumný tým pracuje pod vedením prof. Ing. Jana Nouzy, CSc., z Ústavu informačních technologií a informatiky Fakulty mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci.

Pořady Českého rozhlasu představují kulturní dědictví inspirativní pro mnohá vědní odvětví; obsahují stovky hodin mluveného slova, které není možné vzhledem k rozsahu převádět do psané textové podoby manuálním způsobem, ale je třeba využít přepisu automatického. Ústav informačních technologií a elektroniky Fakulty mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci dlouhodobě pracuje na systému pro rozpoznávání mluvené češtiny. Systém použitý pro transkripci archivních dokumentů, který je na Technické univerzitě v Liberci vyvíjen a stále zdokonalován od roku 2003, je založen na rozpoznávání souvislé řeči za použití velkého slovníku. Tento systém je uzpůsoben pro potřeby vysoce flektivního jazyka, který kvůli zachycení všech gramatických podob jednotlivých slov potřebuje enormně velký slovník (Nouza et al., 2012).

1 Projekt Zpřístupnění archivu Českého rozhlasu pro sofistikované vyhledávání (č. DF 11P010 VV013) se uskutečňuje v rámci Programu aplikovaného výzkumu a vývoje národní a kulturní identity vyhlášeného Ministerstvem kultury ČR.

K automatické transkripci systém využívá akustický model, jazykový model a slovník. Akustický model pro tento systém je složen z kontextově závislých fonémových jednotek (trigramů). Současný slovník pro češtinu využívaný daným systémem obsahuje 490 000 vstupů/položek s 540 000 přiřazenými výslovnostmi. Jazykový model je založen na bigramech a je trénován na korpusu českých textů rozličných žánrů. Kromě standardní transkripce ortografické, zachovávající současnou pravopisnou normu češtiny, systém separátně uchovává výslovnost každého zachyceného slova, časy začátku a konce slov, ale i neřečové události jako ticho, hluk velkého rozsahu (např. znělka v úvodu rozhlasových novin), vokální hluk, hezitační zvuky atd. Každý dokument je také rozčleněn na dílčí textové segmenty volené převážně s ohledem na střídání mluvčích v promluvách. Jednotlivé segmenty obsahují dílčí metajazykové deskriptory jako např. označení mluvčího (tj. jeho jméno, pokud je známo, pokud ne, pak alespoň označení žena/muž). Všechny tyto údaje je možno využít jednak pro vyhledávání, jednak pro statistické a analytické účely.

Zvláště pro jazykový výzkum opřený o tento automatický přepisovač[2] je třeba dodat, že daný transkripční systém nepracuje zcela bezchybně, tj. není zaručena 100% bezchybnost automatického přepisu. Autoři systému však neustále pracují na jeho vylaďování a dochází k nepřetržité evaluaci stávajících přepisů, které jsou používány pro následné trénování a korekci. Experimentální data ukazují, že v současné době je chybovost v přepisu slov (*word-error-rate*) okolo 11 % u vysílání současných pořadů a okolo 14 % u pořadů z let 1969 až 1989. Na první pohled se toto procento může zdát příliš vysoké, avšak většina těchto chyb jednak pokrývá vynechání rozsahově krátkých slov (tj. slov sestávajících z jednoho písmene nebo dvoupísmenné prepozice a konjunkce), jednak dochází k chybnému rozpoznání finálních morfémů slov, které jsou si akusticky blízké (podrobněji viz Boháč — Nouza — Blavka, 2012). Pokud však není databáze použita pro zkoumání flektivní morfologie češtiny, může být dopad těchto chyb v tak velkém statistickém vzorku považován za nesignifikantní.

Pro tento automatický přepisovač byl vyvinut editor transkriptu nahrávek (Seps, v tisku). Editor je určen k tomu, aby se texty získané automatickým přepisem daly dodatečně opravit či upravit. Velkou výhodou tohoto vyvíjeného programu je možnost paralelního sledování mluveného slova společně s grafickým přepisem. Díky editoru je tak možné získat stoprocentně správné přepisy, jestliže je to pro dané účely nutné. Na korekturách spolupracují studenti Katedry českého jazyka a literatury Fakulty přírodovědně-humanitní a pedagogické Technické univerzity v Liberci; každý semestr opraví 100 hodin zvukových nahrávek. Nahrávky jsou nejprve přepsány automaticky a následně ručně opravovány ve výše zmíněném editoru. I tak zpracování jedné hodiny přepisu zabere cca tři hodiny lidské práce. Tyto opravené přepisy jsou používány jak pro vlastní indexaci pořadů pro účely vyhledávání, tak například i pro účely měření úspěšnosti automatického rozpoznávání, kdy je přepis z editoru považován za referenční a jeho porovnáním s automaticky přepsaným textem lze vyhodnotit procentní míru přesnosti.

2 V roce 2012 bylo těchto dat využito pro zkoumání hranic větných celků mluveného textu (viz Škodová — Kuchařová — Šeps, 2012; Plocová, 2012) a také pro studie onomastické (viz Lábus, 2012a; 2012b).

Jak bylo řečeno výše, program využívá tzv. akustického modelu pro rozpoznávání řeči. Jedná se o statistický model, který reprezentuje akustický projev všech hlásek češtiny a dalších zvuků v nahrávkách. Pro účely akustického modelu bylo postupně nahráno a foneticky anotováno 80 hodin mikrofonní řeči a cca 240 hodin rozhlasové řeči, tj. čtené rozhlasové zprávy a publicistické pořady. Pro představu o náročnosti této práce lze uvést, že přesná fonetická anotace jedné hodiny nahrávky trvá s využitím vyvinutých nástrojů osm hodin v závislosti na charakteru promluvy a kvalitě výslovnosti mluvčího.

Důležitou součástí automatického rozpoznávání češtiny je slovník. Tento slovník je specifický tím, že ke každému slovu může být přiřazeno více výslovnostních variant. Analýzou cca 10 GB samotných přepisů, ale i novinových textů, které svým charakterem slovní zásoby odpovídají potřebám přepisu lexika rádiových pořadů, byl vytvořen zásobník obsahující cca 500–600 tisíc slov. Do procesu frekvenční analýzy je zapojen i kontrolor českého pravopisu. I tato součást přepisovače vyžaduje velký podíl lidské manuální práce. Velká část slov přidávaných do slovníku musí být ručně kontrolována, neboť se většinou jedná o slova a jména cizího původu, názvy firem, organizací apod., u nichž není jednotná forma přepisu a u nichž musí být spárování výslovnosti a grafické podoby vytvořeno ručně. Nejnovější verze slovníku nyní obsahuje slova, která se vyskytla alespoň 20krát ve všech dostupných textech.

LITERATURA:

Boháč, Marek — Nouza, Jan — Blavka, Karel (2012): Investigation on most frequent errors in large-scale speech recognition applications. In: Petr Sojka — Aleš Horák — Ivan Kopeček — Karel Pala (eds.), *Text, Speech and Dialogue: 15th International Conference, TSD 2012: Brno, Czech Republic, September 3–7, 2012: Proceedings*. Heidelberg: Springer, s. 520–527.

Nouza, Jan et al. (2012): Making Czech historical radio archive accessible and searchable for wide public. *Journal of Multimedia*, 7(2), s. 159–169.

Lábus, Václav (2012a): Atyp v cihle aneb O jednom progresivním způsobu neologizace. *Naše řeč*, 95(4), s. 187–197.

Lábus, Václav (2012b): Nisa, nebo Nysa? *Acta onomastica*, 53, s. 207–218.

Plocová, Hana (2012): *Porovnání hranic větných celků v závislosti na realizační formě textu* [nepublikovaná bakalářská práce]. Liberec: Technická univerzita v Liberci, Fakulta přírodovědně-humanitní a pedagogická.

Seps, Ladislav (v tisku): NanoTrans editor for orthographic and phonetic Transcriptions. *2013 36th International Conference on Telecommunications and Signal Processing: (TSP 2013): Rome, Italy: 2–4 July 2013*. Rome.

Škodová, Svatava — Kuchařová, Michaela — Šeps, Ladislav (2012): Discretion of speech units for the text post-processing phase of automatic transcription (in the Czech language). In: Petr Sojka — Aleš Horák — Ivan Kopeček — Karel Pala (eds.), *Text, Speech and Dialogue: 15th International Conference, TSD 2012: Brno, Czech Republic, September 3–7, 2012: Proceedings*. Heidelberg: Springer, s. 446–455.

Svatava Škodová a Michaela Kuchařová