V. Popis projektu

(Text max. 10 běžných stran formátu A4)

Povinná osnova popisu projektu (nutno vyplnit všechny body)

1. Vymezit konkrétní cíl(e) projektu v souladu s jedním či více specifickými cíli globálního cíle programu a způsob jejich naplnění.

Po stránce společenského přínosu je hlavním cílem projektu zpřístupnění jedinečného jazykového materiálu z dotazů položených jazykové poradně Ústavu pro jazyk český AV ČR, v. v. i., nejširší veřejnosti ve zcela nové, uživatelsky přívětivé softwarové databázi jazykové problematiky řešené v jazykové poradně. Nejrůznější uživatelské vrstvy tím získají přístup k jazykovému materiálu, který zpravidla nelze nalézt v žádném jiném zdroji poučení o jazyce, a pokud ano, nebývá o něm v takovém zdroji často pojednáno z perspektiv, které tazatel očekává, tj. tazateli se může dostat obecného poučení, avšak aplikace na jeho konkrétní případ může být obtížná, může být vyložen jen dílčí aspekt problematiky, tazatel potřebuje ujištění o správném chápání kodifikačního doporučení apod.

Kromě tohoto praktického cíle – poučit o možnostech řešení konkrétního obtížného jazykového problému – bude však mít vytvoření databáze pro české národní jazykové společenství i pro zahraniční zájemce o češtinu daleko hlubší smysl. Databáze, jež by měla vzejít z tohoto projektu, není navrhována jako nástroj ke zveřejnění uzavřeného okruhu dotazů bez dalších perspektiv, nýbrž hlavně jako otevřený a flexibilní nástroj k trvalému doplňování a rozšiřování, a proto bude sloužit jako jakási kolektivní historická paměť jednak tvorby české normativní mluvnice a pravopisných pravidel, jednak jako co možná nejspolehlivější vědecky zpracovaný záznam o vývoji českého národního jazyka – vývoji nahlíženém nikoli z perspektivy tvůrců jazykových příruček, ale samotných uživatelů jazyka.

Po technické stránce je hlavním cílem projektu tvorba specifického databázového softwaru a webového rozhraní pro vytvoření lingvisticky strukturované softwarové databáze dotazů (LSSDD) položených jazykové poradně Ústavu pro jazyk český AV ČR; dále zdokonalení softwaru pro převod mluvené řeči na psaný text tak, aby co nejlépe vyhovoval potřebám přepisu telefonických hovorů v jazykové poradně coby podkladů pro databázové záznamy o jazykové problematice řešené v konkrétním telefonickém rozhovoru tazatele a pracovníka jazykové poradny. Tyto telefonické hovory se předpokládají jako primární zdroj dat pro LSSDD.

Vytvoření LSSDD sleduje tyto společenské a lingvistické cíle:

1) Zachytit a zpřístupnit veřejnosti zcela nové jazykové jevy bezprostředně po jejich vzniku a umožnit sledování průběhu existence daného jazykového prostředku v češtině. Dotazový materiál dobře odráží nástup, průběh existence, aktuálnost, popř. i zánik jazykových jevů. Např. první záznam o neživotně pojímaných počítačových červech (počítačoví červi, nebo počítačové červy?) se v dotazech poprvé objevuje v roce 1998, od té doby zaznamenáváme po určitou dobu relativně stabilní množství dotazů na týž typ a následný pokles; dotazy na pravopis slova tsunami/cunami přicházejí zprudka a ve velkém množství v závěru roku 2004 jako důsledek událostí ve světě a ustávají v průběhu roku následujícího. Strukturovaná

archivace dotazů má tedy zásadní význam v dlouhodobé perspektivě. Kromě korpusů, jež však nejsou ze své podstaty zaměřeny na klasifikaci a popis jednotlivých jazykových jevů, česká lingvistika nemá k dispozici jiný nástroj, který by umožňoval s obdobnou (tedy alespoň přibližnou) spolehlivostí určit nástup a průběh existence některých jazykových jevů; přitom právě nové, inovační a (dosud) nepopsané jevy bývají častým předmětem dotazů, protože se ještě nedostaly do jazykových příruček, které veřejnost běžně užívá (slovníky, Pravidla českého pravopisu). Tyto životní cykly jazykových jevů bude pomocí LSSDD možné snadno sledovat, protože se počítá s opatřováním databázových záznamů přesnými daty zachycení. Sledování nástupu a průběhu existence jazykových jevů je cenným záznamem dokumentujícím vývojové tendence současné i budoucí češtiny.

2) Umožnit uživatelům češtiny praktickou orientaci v tom, které jazykové jevy jsou kodifikované a které nikoli, jak aktuální a vyčerpávající je kodifikační popis a v čem jsou případné nedostatky kodifikačního popisu vyváženy doporučeními jazykové poradny. Tyto informace jsou velmi žádané nejen žáky, učiteli, jazykovými redaktory a korektory, ale často také samotnými lingvisty. Kromě toho, že jde o žádaný typ služby veřejnosti, bude mít LSSDD v tomto ohledu zásadní význam jako systematická a systematizovaná tvorba podkladů pro kodifikační činnost či obecněji pro jazykovou deskripci. Záznamy o stavu kodifikace je z výše uvedených důvodů třeba (vůbec poprvé v historii české kodifikace) důsledně formalizovat, systematizovat a strukturovat. LSSDD bude pro tento konkrétní účel navržena tak, aby umožňovala zachytit následující okruh základních údajů.

U kodifikovaných, resp. popsaných jevů:

a) ve které kodifikační příručce je jev zachycen, tzn. jakou kodifikační váhu má doporučované řešení (jinak posuzujeme řešení zachycené v Pravidlech českého pravopisu, jinak řešení v popularizačních pracích);

b) rozpory v kodifikaci, tj. rozdílná řešení v různých příručkách, např. Indián/indián – rozdílně řešeno v Pravidlech českého pravopisu a v některých vydáních Slovníku spisovné češtiny pro školu a veřejnost;

c) řešení či doporučení, která se v praxi přežila či se zhusta nedodržují, např. skloňování (ten) datum – datumu atd. užívané v běžné praxi a odporující kodifikovanému (to) datum – data atd.;

d) nedostatečný popis jevu v příručkách, např. skloňování složených číslovkových výrazů.

U jevů nekodifikovaných, resp. nepopsaných:

Poradenská činnost vytváří soubor doporučení, která respektují jazykové zákonitosti i potřeby uživatelů, a proto mají potenciál stát se racionální a přijímanou součástí budoucích kodifikačních příruček.

3) Zpřístupnit veřejnosti co nejširší okruh problematických jevů příslušných typů („zexplicitnit kodifikaci“). Mezi časté dotazy patří např. psaní velkých písmen v nejrůznějších typech názvů. Obecná pravidla pro řešení lze vyhledat v základních příručkách, chybí však pokud možno úplné výčty, často žádané: „Nemáte nějaký seznam

(problematických) názvů ulic/hradů a zámků/institucí/dokumentů apod.? A není někde na webu?“ LSSDD by tento problém v relativně krátké době pomohla účinně řešit, protože se bude postupně plnit příslušnými konkrétními výrazy i s potřebnými výklady u každého z nich.

4) Zajištění jednotnosti v poradenské činnosti jakožto službě veřejnosti. Jednotnost v odpovědích se při současném vytížení poradny sice daří udržovat, avšak jen se značným úsilím a s rizikem pochybení úměrným objemu zpracovávaného materiálu. Je třeba vytvořit nástroj zjednodušující dodržování jednotných postupů v každodenní poradenské činnosti, a to zejména u jevů nových, dosud nezachycených, a u jevů nekodifikovaných (jednoznačně). Jednotnost je v poradenské praxi nezbytná, tazateli jazykové poradny bývají instituce, úřady a média, jejichž jazykové chování je vysoce normotvorné – má celostátní, veřejnou a oficiální působnost. Rozdílná řešení určitých jazykových problémů mohou mít pro jazykovou praxi závažné důsledky. Zveřejnění dat jazykového poradenství posílí společenskou odpovědnost pracoviště, protože s LSSDD společnost získá přehled o všech aspektech jazykověporadenské práce.

5) Využití materiálu z LSSDD pro popularizační činnost, přípravu učebnic, skript i jiných učebních textů (poskytnutí konkrétního jazykového materiálu tříděného podle mluvnických kategorií) atd., a to jak pracovníky poradny, tak tvůrci učebních materiálů z řad veřejnosti (učitelé všech stupňů škol, pracovníci nakladatelství – tvůrci učebnic).

6) Využití pro další elektronické zpracování, např. jako podklad pro tvorbu počítačových aplikací zaměřených na automatické odpovídání.

7) Informace získané při vyplňování LSSDD v průběhu řešení projektu budou využity jako referenční expertní znalost pro algoritmy strojového učení pro systém poloautomatického zpracování dat pro vkládání do této databáze. Výsledná LSSDD bude obsahovat jedinečné informace o jazyce získané od uživatelů poradny doplněné o odborné znalosti pracovníků poradny. Tyto informace bude možné použít jako zdroj referenčních dat (informací od učitele) pro vývoj algoritmů strojové klasifikace obsahu dotazů a porozumění přirozenému jazyku.

Takto koncipovaná softwarová databáze bude potřebnou „kolektivní pamětí“ kodifikace a lingvistiky vůbec a nástrojem, který umožní jednak překlenovat mezidobí do vydání nové kodifikace, jednak zaplnit mezery v existující kodifikaci, protože dokáže zprostředkovat odborné poučení o jevech (dosud) nezachycených v kodifikaci, popř. jevech, které v kodifikaci nejsou popsané dostatečně.

2. **Uvést zda byl nebo je předmět výzkumu v minulosti řešen v rámci jiné výzkumné aktivity podporované z veřejných zdrojů a pokud ano, uvést její identifikaci.**

Předkládaný návrh inovované strukturace a zveřejnění dat získaných z dotazového materiálu jazykové poradny (JP) Ústavu pro jazyk český ideově čerpá z projektu (IAA961409 – Archivace lingvistických dat/dokumentů na počítači a jejich

sociolingvistická analýza; 1994–1996, řešitelka PhDr. Ludmila Uhlířová, CSc.). Tímto projektem byl položen základ formálního zpracování lingvistických dat z dotazového materiálu završený využitím získaných dat při tvorbě Internetové jazykové příručky (projekt 1ET 200610406 – Jazyková poradna na internetu; 2004–2008, řešitelka PhDr. Ivana Svobodová).

Předkládaný projekt je však zcela odlišný – jeho primárním cílem není získat neveřejné podklady k vytvoření jiného lingvistického zdroje, nýbrž zpřístupnit jedinečný jazykový materiál obsažený v dotazech adresovaných jazykové poradně doplněný o instruktivní výklady a strukturované vyhledávání podle mluvnických kritérií ve zcela novém, veřejně přístupném zdroji dat o národním jazyce.

Při řešení úlohy automatického přepisu nahrávek bude využito zkušeností řešitelského týmu získaných při řešení projektu MK NAKI (DF12P01OVV022 – Zpřístupnění rozsáhlého videoarchivu kulturního dědictví pomocí metod automatického rozpoznávání mluvené řeči a strojového překladu, 2012–2015, řešitel doc. Ing. Luděk Müller, Ph.D.), v němž byl vytvořen akustický a jazykový model systému automatického rozpoznávání orálních výpovědí svědků holokaustu zaznamenávaných kamerovou technikou. Pro dosažení maximální přesnosti automatického přepisu audionahrávek je třeba systém automatického rozpoznávání mluvené řeči vždy vybavit akustickým a jazykovým modelem specifickým pro danou řešenou doménu; v případě předkládaného projektu půjde o akustický a jazykový model telefonních dotazů a odpovědí z jazykové poradny ÚJČ. Úlohy vytváření scénářů pro vedení dialogu, segmentace nahrávek a kategorizace segmentů pak v projektu nebyly řešeny vůbec.

3. Rozbor stavu řešení problému v ČR a v zahraničí s odpovídajícími citacemi v odborné literatuře.

Soustavné jazykové poradenství je akademickými institucemi poskytováno i v některých dalších evropských zemích v blízkém okolí ČR: Slovensko – Jazykovedný ústav Ľudovíta Štúra SAV; Bulharsko – Ústav pro bulharský jazyk Bulharské akademie věd; Polsko – poradna nakladatelství Polské akademie věd; v ostatních jazykové poradenství poskytují instituce jiného typu: Německo – různé jazykové poradny univerzit i nakladatelství (nejznámější je poradna nakl. Duden). Slovenská jazyková poradna zveřejňuje vybrané jazykové dotazy s odpověďmi na webové stránce uvedené výše, nejde však o softwarovou databázi s lingvistickou strukturací. Polská jazyková poradna příležitostně zveřejňuje vybrané dotazy a odpovědi. Poradna nakladatelství Duden uveřejnila omezený okruh nejčastějších dotazů. Výsledky poradenské činnosti se u nás i v zahraniční obvykle publikují nesoustavně, v dílčích lingvistických studiích, v popularizačních textech nebo bývají předávány formou přednášek pro veřejnost apod. Není nám známo žádné evropské jazykověporadenské pracoviště, jež by disponovalo veřejně přístupným zdrojem, který by svým charakterem a strukturou symetricky odrážel službu veřejnosti poskytovanou jazykovými poradnami, tj. pokud možno důsledně zaznamenával, formalizoval a zpřístupňoval prakticky veškeré vstupy a výstupy jazykového poradenství. Softwarová databáze navrhovaná v rámci tohoto projektu tedy bude patrně v evropském kontextu zcela ojedinělým nástrojem pro archivaci a zpřístupnění kulturního dědictví národního jazyka, jeho dějin a dějin jeho kodifikace z perspektivy jazykového poradenství.

Problematika jazykového poradenství a s ním souvisejících otázek je u nás zpracovávána zatím nesoustavně, formou dílčích studií, z poslední doby viz např.:

Beneš, M. – Prošek, M. (2011). Ke konceptu minimální intervence. Slovo a slovesnost, 72, s. 39–55.

Beneš, M. – Prošek, M. – Smejkalová, K. (2014). Kodifikace a její role v současné společnosti. In: O. Uličný (ed.), Studie k moderní mluvnici češtiny 2 – komunikační situace a styl. UPOL, Olomouc, s. 15–24.

Pravdová, M. (2012). Elektronizace jazykových dat – nové výzvy pro jazykovou kulturu. In: S. Čmejrková, J. Hoffmannová, J. Klímová, Čeština v pohledu synchronním a diachronním. Karolinum, Praha 2012, s. 819–823.

Prošek, M. (2007). Konstanty a proměnné morfologických dotazů v jazykové poradně. Naše řeč, 90, s. 174–194.Prošek, M. – Smejkalová, K (2011). Kodifikace - právo, nebo pravomoc? Naše řeč, 94, s. 231–241.

Svobodová, I. (2013). Názory respondentů v dotazníku na psaní velkých písmen. Naše řeč, 96, s. 17–35.

Štěpánová, V. (2013a). Fonetická problematika v jazykové poradně. Naše řeč, 96, s. 61–77.

Štěpánová, V. (2013b). Výslovnost cizích slov, vlastních jmen, zkratek a některé další fonetické dotazy v jazykové poradně. Naše řeč, 96, s. 117–140.

Úlohou automatického rozpoznávání mluvené řeči se zabývá mnoho světových pracovišť, v ČR pak kromě ZČU zejména pracoviště na Technické univerzitě v Liberci a na Vysokém učení technickém v Brně [1]. V případě rozpoznávání spontánní, předem nepřipravené a často běžně mluvené „telefonní“ řeči jsou výsledky na úrovni 70 % WER [2], naopak při kvalitních nahrávkách s adaptací na konkrétního řečníka lze dosáhnout výsledků lepších. Například v projektu řešeném ZČU s Českou televizí a společností SpeechTech, TAČR 01011264, Eliminace jazykových bariér hendikepovaných diváků České televize je v současnosti dosahováno přesnosti přepisu originální zvukové stopy na skryté titulky až 90 % WER.

V oblasti segmentace nahrávek dialogů se v současné době využívá kombinace přístupů detekce latentních témat – v rámci stejného dotazu se předpokládá velká podobnost jednotlivých krátkých úseků nahrávky a klasifikace na základě předpokládané velké rozdílnosti na hranici dvou dotazů [3]. V České republice se tímto problémem zabývá například Ústav formální a aplikované lingvistiky na Matematicko-fyzikální fakultě UK v Praze [2]. V oblasti následné klasifikace segmentovaných dotazů budeme navazovat na předchozí výzkum v oblasti klasifikace témat novinových článků [5]. Automatické zpracování a porozumění textu je v současnosti stále nedořešený problém.

[1] Karafiát, M.; Veselý, K.; Szöke, I.; Burget, L.; Grézl, F.; Hannemann, M.; Černocký, J.: BUT ASR System for BABEL Surprise Evaluation 2014. In: Proceedings of 2014 Spoken Language Technology Workshop. South Lake Tahoe, Nevada: IEEE Signal Processing Society, 2014, ISBN 978-1-4799-7129-9, p. 501–506.

[2] Psutka, J. et al: System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. EURASIP Journal on Audio, Speech and Music Processing, 2011, 2011(10), p. 1-19

[3] Arguello, J.; Rosé, C.: Topic segmentation of dialogue. In Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech (ACTS '09). Association for Computational Linguistics, Stroudsburg, PA, USA, p. 42–49.

[4] Galuščáková, P.; Pecina, P.: Experiments with segmentation strategies for passage retrieval in audio-visual documents. In Proceedings of International Conference on Multimedia Retrieval, April 1–4, Glasgow, UK, ACM, 2014, p. 217.

[5] Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J.: Automatic Topic Identification for Large Scale Language Modeling Data Filtering . Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 6836, Springer, Heidelberg, 2011, p. 64-71.

4. Uvést metodiku řešení projektu.

V rámci projektu bude vytvořena databáze s nejpodrobnější možnou lingvistickou strukturací – LSSDD. Základní lingvistické kritérium, kterým budou řízeny všechny lingvistické práce na struktuře softwarové databáze, je dáno charakterem lingvistické práce poradny. Poradna řeší dotazy ze všech rovin jazyka i ze všech oblastí pravopisu, a proto i struktura LSSDD musí být schopna uspokojivě pojmout veškerou dosavadní jazykovou problematiku, kterou jazyková poradna řešila a o které má záznamy, a zároveň musí být dostatečně flexibilní na to, aby bylo strukturu LSSDD v budoucnu možno modifikovat v případě potřeby zachytit dosud opomenuté nebo zcela nové jazykové jevy.

Data budou do LSSDD vkládat samotní pracovníci jazykové poradny na základě automaticky přepsaných telefonních hovorů v jazykové poradně. Písemný podklad je pro záznam nezbytný, protože tazatelé mívají více dotazů v jednom telefonátu, často popisují důležité a pozoruhodné okolnosti vzniklého jazykového problému, prezentují cenné argumenty, vedou diskuse s pracovníky poradny a i k řešení obě strany často dospívají v netriviální diskusi. Bez písemných podkladů nelze zcela přesně zpětně rekonstruovat průběh diskuse a argumentace obou stran a služba veřejnosti zamýšlená v tomto p

rojektu by nebyla úplná a dostatečná. K náležitě podrobné analýze nelze často uspokojivě dospět ani opakovaným poslechem zvukového záznamu, zejména např. v diskusích s korektory, redaktory, pracovníky médií, ale i s laiky s živým zájmem o jazyk a přirozenou jazykovou erudicí.

Pro tyto účely bude třeba zdokonalit technické řešení nahrávání hovorů v jazykové poradně (oddělené kanály – rozlišení řeči pracovníka poradny a tazatele pro odstranění překryvů ztěžujících automatické rozpoznávání; vyšší kvalita nahrávky), dále přizpůsobit systém převodu mluvené řeči do textu konkrétní úloze a dostupným datům a navrhnout adekvátní strukturu databáze na základě lingvistické strukturace jazykových dat. Pro usnadnění práce odborných pracovníků poradny při zpracování dotazů bude navržen systém pro poloautomatickou analýzu dat (SADA – SemiAutomatic Data Analysis) pro jejich vložení do LSSDD. Systém SADA bude uživateli pomáhat s anotací nahraných dotazů: poskytne návrh segmentace nahrávky na jednotlivé dotazy, návrh tématu usnadňující uživateli správné zařazení dotazu do databáze a automatický přepis nahrávky spolu s možností orientace ve zvukové nahrávce: zadavatel dat bude mít možnost kliknutím na slovo v automatickém přepisu spustit přehrávání příslušného úseku nahrávky telefonátu. LSSDD bude mít veřejnou a neveřejnou část. Neveřejná část bude přístupna pouze pracovníkům poradny a jejím účelem bude formalizace práce poradny – vkládání dat za účelem dosažení

výše uvedených cílů projektu; účelem veřejné části bude zpřístupnit uživatelsky přívětivou formou informace o jazyce v souladu s cíli projektu.

LSSDD – Koncepce a lingvistická struktura

Pro databázi bude vyvinut co nejpodrobnější systém klasifikace („tagování, indexování“) dotazů podle navržené struktury lingvistických identifikátorů. Podle těchto identifikátorů se dotazy budou ukládat a následně vyhledávat a třídit. Indexování dotazů bude navrženo tak, aby co nejlépe vyhovovalo potřebám uživatelů služeb jazykové poradny a LSSDD. Na nejobecnější rovině je třeba připravit databázi pro vstup následujících typů dotazů:

1. Dotazy na jazykové jevy nepřesahující rozsahem 1 slovo, tj. týkající se hláskové, morfematické, tvaroslovné, lexikálněsémantické, stylistické nebo etymologické problematiky, při níž není třeba přihlížet k syntaktickému okolí (výslovnost slova koncert, morfematický rozbor slova písek, psaní slova ližiny, původ slova medvěd, volba tvaru okresech/okresích, volba podoby briefing/brífing/brífink apod.).
2. Dotazy přesahující rozsahem 1 slovo, tj. dotazy při nichž je třeba přihlížet k syntaktickému okolí (Je správně zákaz vstupu cestujících do kolejiště, nebo zákaz vstupu cestujícím do kolejiště? Píše se čárka před než? apod.).
3. Dotazy, které nelze vázat na žádné konkrétní „klíčové slovo v dotazu“ (Jak se správně píší čárky v infinitivních konstrukcích? Kdy vyšla první Pravidla českého pravopisu? Kolik má čeština slov? Znepokojuje mne pronikání anglicismů, lze proti tomu něco podnikat? apod.)

Databázový záznam pro každý jednotlivý dotaz bude komplexnější, než byl sám předmět dotazu. Např. záznamy dotazů z první a druhé skupiny budou do databáze uloženy podle klíčového slova dotazu (KSD – koncert, ližiny; vstup, popř. též zákaz, apod.), a to morfematicky rozčleněného s indexy popisujícími typy jednotlivých morfů, kategoriemi označujícími jednotlivé okruhy pravopisné nebo gramatické problematiky (pravopis → psaní i/y → v kořeni/kmeni/příponě/koncovce nebo morfologie → skloňování → podstatná jména obecná … atd.), původ slova (slova domácí; slova přejatá → anglicismy apod.), stylový/komunikační příznak tradiční či deskriptivní (spisovné neutrální; v komunikaci IT specialistů apod.); kodifikační doporučení (např. kodifikováno v PČP, vyd. 1993, str. 56), úplnost kodifikačního doporučení (např. kodifikováno dostatečně a jednoznačně); správné řešení, resp. poradenský výklad a doporučení (nebude-li zřejmé ze samotného zápisu KSD); dále údaje o datu vložení apod. Dotazy ze třetí skupiny budou řazeny pouze kategoriálně do okruhů typu „psaní čárky v infinitivních konstrukcích“ apod. U opakujících se dotazů na touž problematiku (např. psaní ližiny, nebo *lyžiny) nebude vytvořen duplicitní záznam, nýbrž bude pouze zvýšen statistický údaj o počtu vznesení tohoto dotazu. Důvodem je též přehlednost – uživatel zajímající se o správné psaní slova ližiny nebude nucen procházet stovky, ba tisíce záznamů téhož typu.

Podle všech uvedených kritérií či jejich smysluplných kombinací bude uživatel mít možnost dotazy též vyhledávat – může vyjít z klíčového slova ližiny a obdrží všechny dotazy na toto slovo kategorizované podle jednotlivých oblastí problematiky (bude-li to v daném případě relevantní): psaní i/y, skloňování, kolokace apod. Bude mít možnost též např. vyhledávat

všechna slova domácí s kořenem -ved- (a jeho případnými alomorfy), anglicismy zakončené na -ing (-ink), slova zakončená na -ště, spojení obsahující slovo zámek (zámek Horšovský Týn, zámek Karlova koruna) apod.

Navržení podrobného systému klasifikace jazykových jevů je nejvýznamnějším a nejnáročnějším lingvistickým výstupem projektu, který bude vyžadovat řadu netriviálních rozhodnutí i technických řešení umožňujících výstižně zachytit případnou kompromisní povahu kategorizace daného jevu. V zájmu zajištění co nejkvalitnějšího provedení tohoto výstupu se při rozpracování klasifikace podle jazykových kritérií předpokládá specializace jednotlivých pracovníků. Každý pracovník bude zodpovědný za adekvátní zpracování určitého okruhu dotazové problematiky, což bude obnášet prostudování dostupných pramenů, aplikaci na dotazový materiál a navržení takové strukturace, která umožní sledovat centrum daného jevu a jeho periferii, tzn. prvky nové (dosud nezaznamenané), odchylné, progresivní, ustupující atd.

Vzhledem ke komplexnosti navržené lingvistické strukturace softwarové databáze bude možno ji využít nejen jako nástroj pro ukládání dotazových dat, ale jako obecně použitelnou databázi pro excerpci, třídění a statistické zpracování dat. O využití v tomto směru předběžně projevila zájem domácí univerzitní pracoviště (viz přiložené doklady zájmu o výstupy projektu) se záměrem dát databázi k dispozici svým studentům a pracovníkům, kteří s její pomocí budou moci snadno a kvalitně zpracovávat a vyhodnocovat jazykový materiál pro své výzkumy a kvalifikační práce, a rovněž jazyková poradna Jazykovedného Ústavu Ľudovíta Štúra SAV.

LSSDD – statistika, správa uživatelů a zabezpečení

Kromě lingvistické strukturace bude zapotřebí důkladně propracovat též systém statistického vyhodnocení zaznamenávaných dotazů. Významnou výpovědní hodnotou o jazykovém či pravopisném systému a také o stavu jazyka jsou totiž také údaje o tom, na které jevy se uživatelé jazyka ptají častěji, na které méně často či zcela ojediněle, v kterém období přibývá dotazů na některý jev apod. Na rozdíl od vybrané jazykové problematiky z dotazů jazykové poradny je tato oblast dosud empiricky zcela nezpracovaná, šlo by o historicky první soustavné empirické zpracování statistiky dotazů, navíc zároveň o první veřejně přístupné zpracování.

Pro řízení chodu databáze je nezbytné, aby byla LSSDD opatřena též systémem zabezpečení (přístupová hesla) a správou uživatelů.

LSSDD – vkládání prvních dat a očekávané přírůstky

Nahrávání telefonických hovorů v jazykové poradně bylo uvedeno do provozu v prosinci 2013 po konzultaci s Úřadem pro ochranu osobních údajů. Volající jsou na zaznamenávání hovoru upozorněni před započetím hovoru. Data z hovorů jsou využívána pouze k vědeckým účelům, tzn. čerpá se z nich pouze jazyková problematika. Veškeré údaje, které by mohly vést (byť nepřímo) k identifikaci volajícího, nejsou předmětem zkoumání a nejsou žádnou formou zveřejňovány. Stejný přístup bude uplatněn v LSSDD. Do databáze bude z obsahu dotazu přenesena pouze jazyková problematika. Mimo záznam a zveřejnění zůstanou pouze dotazy takového typu, u nichž hrozí (byť nepřímo) možnost identifikace

volajícího a poškození jeho práv na ochranu osobních údajů, takových je však v poradenské praxi mizivé množství.

Od počátku prosince 2013 do 20. dubna 2015 bylo zaznamenáno celkem 4 600 telefonátů, tj. průměrně cca 290 telefonátů měsíčně. Při rozvaze počtu je nutno mít na zřeteli, že řada z nich obsahuje více než jeden jazykový dotaz, proto údaje počítající s telefonáty jsou vlastně minimální počty záznamů do databáze, dotazů bude vždy více než telefonátů. K 1. 3. 2016 (datum případného zahájení prací na projektu) by tedy mělo být pořízeno cca 7 800 telefonátů. Tyto telefonáty budou během řešení projektu též zaneseny do databáze. Do databáze budou dále postupně ukládána data z telefonických hovorů pořízených po dobu řešení projektu. Počet dotazů v jazykové poradně dlouhodobě nevykazuje klesající tendenci, a tak lze předpokládat průměrný roční počet telefonátů cca 3 500. Kromě zmíněných 7 800 telefonátů pořízených do doby před započetím projektu bude tedy databáze postupně ještě v průběhu projektu obohacována minimálně o cca 3 500 záznamů ročně a stejné množství bude každoročně přibývat i po skončení projektu. Po technické stránce je pamatováno na to, aby systém přepisu mluvené řeči vyvinutý pro poradnu i LSSDD byly i po skončení projektu udržitelné s minimálními nároky na údržbu.

Do LSSDD budou v průběhu trvání projektu uloženy též záznamy o jazykových jevech z archivní neveřejné databáze, která byla výstupem projektu L. Uhlířové z bodu 2 tohoto oddílu přihlášky. Jedná se celkem o cca 10 000 e-mailů s jazykovými dotazy. Nebude-li technicky možné tato data využít, použije se výběr 10 000 e-mailů se stále aktuální jazykovou problematikou z více než 70 000 dosud databázově nezpracovaných e-mailů, které jazyková poradna obdržela před rokem 2010.

Skladba a struktura dat v LSSDD bude podrobně popsána ve veřejném webovém rozhraní LSSDD, aby uživatel přesně věděl, v jakém materiálu vyhledává.

Strukturovaná softwarová databáze dotazů a software pro zpracování dat

V technickém řešení softwarové databáze LSSDD se předpokládá, že většina ukládaných dat bude využívat služeb relačního databázového systému. Tento systém mimo jiné automaticky zajišťuje integritu databáze a také nabízí možnost správy uživatelů databáze a jejich uživatelských rolí a práv. Toho lze využít k řízenému přístupu k datům, kdy vyhrazená data jsou dostupná pro čtení (i zápis) jen určeným uživatelům. Nelze vyloučit možnost, že některá data budou aplikací ukládána do jiné databáze, kterou bude spravovat tzv. nerelační databázový systém. Použité databázové systémy a přesný způsob ukládání dat bude cílem hlubší analýzy v první části řešení projektu. Rozhodně bude kladen důraz na použití produktů typu open source, které budou vyhovovat technickým požadavkům počítačového vybavení ÚJČ. Základní nástin aplikací přicházejících v úvahu je již předjednán se správou sítě ÚJČ. Chod aplikace nebude při předpokládaném objemu dat vyžadovat příliš nákladné inovace hardwaru. Práce s LSSDD bude možná i prostřednictvím internetu, pracovníci ÚJČ tedy budou moci s LSSDD pracovat i mimo budovu ÚJČ.

Vlastní návrh LSSDD bude vznikat v součinnosti pracovníků ZČU s pracovníky ÚJČ. V začátcích projektu vznikne prvotní databáze, která umožní ukládání nahrávek a informací k nim doplněných pracovníky poradny tak, aby tyto informace bylo následně možno využít k trénování a nastavení systému poloautomatického zpracování nahrávky (SADA). V průběhu projektu bude databáze rozšiřována dle vznikajících požadavků uživatelů a souběžně s ní navrhován systém SADA, který uživatele povede při ukládání nových dat do databáze.

Přístup k údajům v databázi bude prostřednictvím webového rozhraní pro registrované

uživatele ze strany ÚJČ (uživatel s právy čtení i zápisu), tak v pozdější části projektu i webové rozhraní pro obecného uživatele ze strany široké veřejnosti.

Automatický přepis nahrávek a metody automatického zpracování/porozumění textu

Existující systém pro automatický přepis bude nutné přizpůsobit konkrétní úloze, tedy natrénovat jazykový i akustický model a sestavit specifický slovník výrazů vyskytujících se v nahrávkách. Takováto úprava vyžaduje sběr dat pro metody strojového učení.

V začátcích projektu budou již existující nahrávky dotazů ručně anotovány a z nich bude natrénován specifický jazykový model a sestaven oborový slovník. Protože v nahrávkách se vyskytují specifické jazykové termíny, budou k trénování slovníku i jazykového modelu využity další dostupné textové zdroje, jako jsou například dotazy poradny pokládané v dřívější době formou e-mailů, Internetová jazyková příručka a další elektronické zdroje pojednávající o českém jazyce.

Dále budou zpracována nově nahraná data v požadované kvalitě (s oddělenými kanály a vyšší kvalitou nahrávek), s jejichž využitím bude existující akustický model adaptován na prostředí telefonní poradny. Díky malému množství mluvčích na straně poradny a odděleným kanálům v nahrávce bude možné systém adaptovat na řeč pracovníků poradny a tím zvýšit přesnost rozpoznávání odpovědi.

Protože se předpokládá zpracování nahrávek dotazů z poradny ex-post (například na konci dne) a protože v nahrávce jednoho hovoru je obvykle zaznamenáno více různých dotazů najednou, bude navržen systém pro poloautomatickou analýzu zaznamenaných dat pro usnadnění jejich zpracování (SADA) a pro následné vložení výsledků této analýzy do softwarové databáze LSSDD. Zpracování automatického přepisu nahrávek metodami strojového učení umožní automatický návrh možných témat v nahrávce, návrh rozdělení nahrávky na jednotlivé dotazy aj., s kterými bude moci uživatel dále pracovat.

Úspěšnost přepisů nahrávky je očekávána v rozmezí cca 70 % – měřeno procentem správně rozpoznaných slov, tzv. WER (Word Error Rate). Text bude přesto nápomocný uživateli databáze při orientaci v nahrávce k jejímu snadnějšímu zpracování. Přepisy bude možné zpětně fulltextově prohledávat pro lepší orientaci v záznamech.

Společně s pracovníky poradny vznikne návrh scénáře vedení dialogu a formulace klíčových frází (nyní zopakuji tu větu; shrnu řešení; problémem tedy je... apod.), pomocí kterých bude nahrávaný hovor usměrňován a které usnadní systému přepis nahrávky a jeho následné automatické zpracování pro pomoc se správným vyplněním informací do LSSDD.

Pro podporu poloautomatického zpracování zaznamenaných hovorů bude pro nahrávací systém vyvinut modul, který umožní v průběhu nahrávaného hovoru zaznamenávat (například pomocí klávesových zkratek zadávaných pracovníkem poradny) dodatečné informace k nahrávce týkající se jejího obsahu, které uživateli i systému SADA pomohou s jejím zpracováním.

Systém SADA bude umožňovat jak zpracování přepisů nahrávek z poradny, tak i obecných textových zdrojů (např. dotazů položených poradně formou e-mailů) z daného oboru, ve kterém bude systém natrénován.

5. **Stručně popsat vybavenost pracoviště – materiální, laboratorní, přístrojové, případně jiné vybavení řešitelského pracoviště nebo pracovišť, přístup k informačním zdrojům potřebným k řešení projektu.**

Oddělení jazykové kultury Ústavu pro jazyk český je vybaveno odpovídající výpočetní technikou potřebnou pro řešení projektu, zapotřebí bude jen průběžná obměna dožitých a technicky zastaralých (zastarávajících) počítačů, z nichž některé ještě využívají nepodporované operační systémy. Pracoviště má vlastní příruční knihovnu pro jazykověporadenskou činnost, využívá též služby ústavní knihovny i knihovny AV ČR, a má tudíž přístup k významným mezinárodním odborným zdrojům, tištěným i elektronickým. Pro účely zaznamenávání dat v požadované formě bude nutné stávající systém renovovat, aby se zvýšila kvalita zaznamenávaných dat pro následné automatické zpracování. Na straně pracoviště ZČU jsou k dispozici mimo jiné nové výpočetní stroje vybavené výkonnými GPU kartami pro náročné výpočty, pracoviště má přístup ke zdrojům Národní Gridové Infrastruktury MetaCentrum, jejíž členství umožňuje přístup k dalším masivním výpočetním prostředkům cloudového typu. Tyto prostředky budou využity k zpracování velkého množství dat pro vývoj a testování statistických modelů systému rozpoznávání řeči a zpracování textu.

6. **Specifikovat výsledky projektu (výčet všech očekávaných výsledků).** *Očekávané výsledky musí být rozděleny na výsledky **hlavní a vedlejší**. **Rozdělení jednotlivých druhů výsledků do skupin hlavní a vedlejší výsledky je uvedeno v zadávací dokumentaci, v části 5.4. Očekávané výsledky.** Výsledek musí být specifikován písmenem a textem uvedeným v zadávací dokumentaci (viz tabulka „Pomocné kritérium pro hodnocení poskytovatele z hlediska naplnění indikátorů programu NAKI II“). Povinnou součástí specifikace každého předpokládaného výsledku projektu je:*

písmeno označující druh výsledku (např. R, G, B atd.)	R (2x), J (2x), D (14x)
kategorie výsledku: hlavní/vedlejší (lze uvést v záhlaví pro celou skupinu, pokud od jedné kategorie bude více druhů výsledků a/nebo vícečetné výsledky jednoho druhu výsledku stejné kategorie)	hlavní, vedlejší, vedlejší
předpokládaný název výsledku	LSSDD - lingvisticky strukturovaná softwarová databáze dotazů SADA - SemiAutomatic Data Analysis Článek v odborném časopise - 2x Článek ve sborníku - 14x
krátká charakteristika výsledku	LSSDD - lingvisticky strukturovaná softwarová databáze pro zaznamenávání a zveřejňování dotazů jazykové poradny ÚJČ SADA - systém pro poloautomatickou analýzu dat pro

	usnadnění předzpracování dat pro LSSDD Článek v odborném časopise - 2x Článek ve sborníku - 14x
předpokládaný rok uplatnění výsledku	2019
předpokládaní **budoucí uživatelé výsledku** (tento údaj o uživatelích výsledku nebude uváděn pouze u výsledků publikačních typu B, C, D a J; u ostatních druhů výsledků hlavních i vedlejších pro program NAKI II je nepominutelný). Doklad o zájmu budoucího možného uživatele o navrhovaný výsledek je možné přiložit k přihlášce projektu, pokud jej lze zajistit.	LSSDD - FF MUNI, PF MUNI, PF TUL, JÚLŠ SAV SADA - MFF UK (doklad o předběžném zájmu předložen)
dedikace výsledku - u vedlejších výsledků bude uvedeno, zda bude výsledek dedikován výlučně k projektu NAKI II. Pokud nebude výlučně vázaný na NAKI II (s výjimkou výsledku druhu B - odborná kniha, A - specializovaná veřejně přístupná databáze, kde je tento postup vyloučen - viz ZD), je nutné uvést všechny souvztažné výzkumné aktivity, z kterých bude výsledek rovněž podporován, instituce a autory, kteří se budou na výsledku rovněž spolupodílet.	všechny vedlejší výsledky budou dedikovány výlučně projektu NAKI II

V případě, že uchazeč předpokládá více jak jeden výsledek přísl. druhu výsledku, je nutné uvést jejich počet a specifikace u každého z očekávaných výsledků příslušného druhu výsledku.

U specifického výsledku pro program NAKI II E - uspořádání výstavy - je nutné dodržet podmínky uvedené v zadávací dokumentaci v části 5.4., včetně zveřejnění publikace typu B (která bude kritickým katalogem výstavy a která musí být v přihlášce projektu jednoznačně jako kritický katalog výstavy označena - v poli krátká charakteristika výsledku). U očekávaných a v přihlášce vymezených výsledků uvést případný mezinárodní přínos. Dále se doporučuje respektovat programem pro daný specifický cíl očekávané druhy výsledků případně další výsledky aplikovaného výzkumu a experimentálního vývoje definované v platné Metodice hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů. Při hodnocení projektu nebude brán zřetel na uvedené očekávané výsledky, které neodpovídají druhům výsledků uvedených ve struktuře RIV15 (např. rukopis, studie, abstrakt apod.).

Na závěr bodu 6. bude povinně vyplněna níže uvedená přehledová tabulka počtu předpokládaných výsledků projektu odpovídající komentáři v bodě č. 6. *Definice druhů výsledků jsou uvedeny v platném znění Metodiky hodnocení výsledků výzkumných organizací a výsledků ukončených programů.*

předpokládané výsledky projektu	počet
Hlavní výsledky	
F_{uzit} - užitný vzor	
F_{prum} - průmyslový vzor	
G_{prot} - prototyp	
G_{funk} - funkční vzorek	
N_{met} - certifikovaná metodika	
N_{pam} - památkový postup	

předpokládané výsledky projektu	počet
N_{map} - specializovaná mapa s odborným obsahem	
P - patent	
- "evropský“ patent (EPO), patent USA (USPTO) a Japonska	
- český nebo národní patent (s výjimkou patentu USA a Japonska), který je využíván na základě platné licenční smlouvy	
- ostatní patenty[6]	
R - software	2
Z_{polop} - poloprovoz	
Z_{tech} - ověřená technologie	
H_{leg} - výsledky promítnuté do právních předpisů a norem	
H_{neleg} - výsledky promítnuté do směrnic a předpisů nelegislativní povahy závazných v rámci kompetence příslušného poskytovatele	
E - uspořádání výstavy - **specifický výsledek programu NAKI** Jedná se o nejméně dva měsíce trvající veřejnou prezentaci kulturních či kulturně historických hodnot s minimální návštěvností 1000 návštěvníků za dobu trvání výstavy, která je výlučně výsledkem výzkumných projektů v rámci Programu aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI), a její součástí je kritický katalog s řádně přiděleným ISBN, jehož obsah prošel recenzním řízením. O případné výnosy ze vstupného musí být poníženy způsobilé náklady projektu.	
Vedlejší výsledky	
A - audiovizuální tvorba, elektronické dokumenty	
B - odborná kniha	
C - kapitola v odborné knize	
D - článek ve sborníku (z konference)	14
J - recenzovaný odborný článek	2
M - uspořádání konference	
W - uspořádání workshopu	

7. Poslání a očekávané přínosy projektu ve vazbě na očekávané přínosy programu NAKI *(část 2.3 programu), včetně zdůvodnění potřebnosti projektu pro naplnění cílů programu NAKI.*

Cílem projektu je všobecná prezentace a zpřístupnění aspektů týkajících se utváření klíčové hodnoty národní identity – národního jazyka –, a to konkrétně těch aspektů, které nelze a nebude lze zachytit v jiných existujících zdrojích poučení o jazyce, protože jejich zaměření je jiné a protože jsou prostorově omezené. Z hlediska uživatele se LSSDD bude jevit patrně především jako užitečný nástroj k řešení praktických problémů s jazykem, z hlediska koncepčního však půjde hlavně o dlouhodobě a trvale doplňovaný informační zdroj, ve kterém budou uchovány jazykové prostředky obohacené o komentáře sociokulturní povahy, komentáře objasňující příčiny (ne)kodifikovanosti a také (ne)kodifikovatelnosti daného jevu, to vše ve vývojové perspektivě, s přesným časovým určením, statistickými

[6] Český nebo jiný národní patent udělený, doposud nevyužívaný nebo využívaný vlastníkem patentu.

informacemi a možností vyhledávat příbuzné jevy, analogické jevy, jevy náležející do stejných kategorií a podobně. Navrhovaná struktura LSSDD jako specializované veřejné softwarové databáze bude významně sloužit také vzdělávacím účelům – studentům a učitelům se dostane strukturovaného poučení o tom, že ne všechny jevy patří do tradiční kodifikace, budou objasněny příčiny tohoto stavu u každého konkrétního jazykového prostředku a bude názorně ilustrován fakt, že ve fázích mezi zveřejněním nové kodifikace jazyk prochází živým vývojem, který je třeba průběžně reflektovat. Všechny uvedené funkce nemůže poskytnout žádný existující lingvistický zdroj. Svrchovaně důležitá je též ta skutečnost, že vytvoření LSSDD nevychází z potřeby lingvistů, ale samotné společnosti, která potřebnost takového zdroje často spontánně vyjadřuje v telefonátech do jazykové poradny.

8. Kritické předpoklady dosažení cíle projektu, popis rizik projektu.

Kromě obecných kritických předpokladů platících pro obdobné projekty (odchody kvalifikované pracovní síly, reorganizace apod.) jsou konkrétní rizika pro tento projekt dané jeho technologiemi založenými na strojovém učení z rozsáhlých dat:

- nedostatek nahrávek a textových dat pro vývoj statistických modulů a
- malá informační hodnota v získaných datech pro kvalitní klasifikaci textu pro předem neznámé a řídké problémy jazyka.

Pracoviště ZČU se problematkou databázových systémů a systémů rozpoznávání a porozumění řeči dlouhodobě zabývá a s rezervou disponuje potřebnými prostředky i kvalifikovanou pracovní sílou, čímž snižuje riziko nedosažení cíle projektu z důvodu odchodu jednotlivců pracujících na projektu.

Zmírnění dalších uvedených rizik bude dosaženo provedením manuální transkripce řečových dat, dispozicí rozsáhlými korpusy pro jazykové modelování češtiny a možností obohatit nahrávaná data o apriorní informaci formou předpřipravených scénářů vedení dialogu. důležité je podotknout, že systém SADA bude možno učit v průběhu projektu, a to na základě oprav prováděných jeho uživatelem. Odborný pracovník vkládající data do LSSDD je garantem správnosti dat.

9. Etapy projektu

Pro každou etapu projektu je nutné uvést (etapy na sebe musí časově a věcně navazovat, popř. se mohou částečně překrývat, ale musí být uvedeny a nesmí být všechny plánovány na celou dobu řešení):

a) Číslo, název a cíl etapy

1. Nová strukturace poradenských hovorů – návrh automatické segmentace

Součinnost ZČU a ÚJČ: Návrh klíčových orientačních výrazů ve vedení dialogu pracovníků poradny s tazateli pro podporu automatické segmentace a klasifikace částí hovoru, algoritmy porozumění jazyku, zpracování modelů pro segmentaci dotazu, zapracování do modulu porozumění jazyku (součást systému SADA).

b) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-03-01

c) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-12-31

d) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

e) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

D – článek ve sborníku (plán TSD 2017/2018)

f) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN, předání v rámci průběžné zprávy projektu.

g) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2018-12-31

h) Číslo, název a cíl etapy

2. Kategorizace dat

ÚJČ: Tvorba kategorií dotazů, vytvoření trénovacích dat, zpětná vazba k poloautomatické kategorizaci.

ZČU: Zpracování modelů pro kategorizaci dotazu – úprava algoritmů pro identifikaci tématu v dotazu, zapracování do modulu porozumění jazyku (součást systému SADA).

i) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-01-01

j) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2019-12-31

k) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

l) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

D - článek ve sborníku (plán Interspeech 2018/2019)

J - článek v časopisu (plán Naše řeč 2018)

m) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN, předání v rámci průběžné zprávy projektu.

n) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2019-12-31

o) Číslo, název a cíl etapy

3. Tvorba databáze

ÚJČ: Vytvořit základní návrh struktury databáze – veřejné i neveřejné části – a propracovávat jej, a to tak aby co nejlépe vyhovovala potřebám uživatelů (bude odvozeno od toho, na jaké jevy se tazatelé ptají a co konkrétně o nich chtějí vědět), charakteru a objemu dat a způsobům práce zadavatelů dat.

ZČU: Vytvořit programový návrh databáze LSSDD.

p) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-03-01

q) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-12-31

r) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

s) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

D - článek ve sborníku (plán Language Management Symposium 2017)

D - článek ve sborníku (plán Language Management Symposium 2019)

D - článek ve sborníku (plán 2 příspěvky Konference Lingvistika Praha 2016)

D - článek ve sborníku (plán 2 příspěvky Konference Lingvistika Praha 2017)

D - článek ve sborníku (plán 2 příspěvky Konference Lingvistika Praha 2018)

D - článek ve sborníku (plán 2 příspěvky Konference Lingvistika Praha 2019)

D - článek ve sborníku (plán databázová konference 2019)

t) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN, předání v rámci průběžné zprávy projektu.

u) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2019-12-31

v) Číslo, název a cíl etapy

4. Průběžné zadávání dat do databáze.

ÚJČ: V průběhu celé etapy bude probíhat zadávání dotazů (archivní dotazy i dotazy pořízené v průběhu trvání projektu) do LSSDD.

w) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2017-01-01

x) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2019-12-31

y) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

z) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

J - článek v časopisu (plán Naše řeč 2019)

aa) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN, předání v rámci průběžné zprávy projektu.

bb) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2019-12-31

cc) Číslo, název a cíl etapy

5. Tvorba jazykového modelu ZČU: Příprava dat pro jazykový model – přepis nahrávek poradny pořízených v době před započetím prací na projektu a jejich úprava pro trénování jazykového modelu systému automatického přepisu řeči, zapracování do modulu automatického rozpoznávání řeči (součást systému SADA). ÚJČ: Připraví podkladová data.

dd) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-03-01

ee) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-12-31

ff) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

gg) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

-

hh) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

-

ii) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

-

jj) Číslo, název a cíl etapy

6. Tvorba nového akustického modelu ZČU: Příprava dat pro akustický model – z nahrávek poradenských hovorů pořízených v průběhu trvání projektu natrénovat nový akustický model, zapracování do modulu automatického rozpoznávání řeči (součást systému SADA). ÚJČ: Připraví podkladová data.

kk) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2017-01-01

ll) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2019-12-31

mm) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

nn) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

-

oo) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

-

pp) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

-

qq) Číslo, název a cíl etapy

7. Programový návrh systému SADA
ZČU: Vytvořit programový návrh systému SADA.
ÚJČ: Součinnost při návrhu rozhraní systému SADA.

rr) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-03-01

ss) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2019-12-31

tt) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

uu) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

D – článek ve sborníku (plán TSD 2019)

R – SW systém SADA pro poloautomatické zpracování dat pro vložení do databáze

vv) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.

Popisu funkčnosti výsledku typu R společně s licenčními podmínkami pro využití, SW dostupný přes web, předání výsledků v rámci závěrečné zprávy projektu.

ww) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2019-12-31

xx) Číslo, název a cíl etapy

8. Webové rozhraní přístupu do databáze LSSDD

ZČU: Vytvořit webové rozhraní přístupu do databáze LSSDD.

ÚJČ: Součinnost při návrhu rozhraní databáze LSSDD, rozdělení privátní a veřejné části databáze.

yy) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2017-03-01

zz) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2019-12-31

aaa) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

bbb) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

R – softwarová databáze LSSDD s rozhraním pro přístup k datům (čtení /zápis/ různá úroveň práv)

ccc) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.

Popisu funkčnosti výsledku typu R společně s licenčními podmínkami pro využití, SW

dostupný přes web, předání výsledků v rámci závěrečné zprávy projektu.

ddd) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2019-12-31

10. Uvedení oponentů projektu, se kterými uchazeč nesouhlasí z důvodů možné podjatosti při hodnocení předloženého projektu *(lze uvést max. 3 osoby nebo pracoviště)*.

prof. Ing. Jan Nouza, CSc., Ústav informačních technologií a elektroniky, Technická univerzita v Liberci doc. PhDr. Karel Pala, CSc., Centrum zpracování přirozeného jazyka, Fakulta informatiky, Masarykova univerzita