Mgr. Tomáš Jelínek, Ph.D.

Koncepce rozvoje Ústavu teoretické a komputační lingvistiky FF UK

na období 1. 7. 2018 – 1. 7. 2021, s výhledem na další roky

Ústav teoretické a komputační lingvistiky (dále ÚTKL) působí na Filozofické fakultě Univerzity Karlovy od svého založení v roce 1990 v oblasti počítačového zpracování přirozeného jazyka (především češtiny) a formálního popisu jazyka. Ústav je zaměřen vědecky, pracovní náplní ústavu je z větší části řešení výzkumných projektů. V pedagogické oblasti pak zajišťuje spolu s Ústavem Českého národního korpusu (dále ÚČNK) výuku doktorského studia v oboru *matematická lingvistika*. Ředitelem ústavu je od roku 1994 doc. Vladimír Petkevič, CSc., který na svou funkci rezignoval ke 30. červnu 2018.

ÚTKL se podílí spolu s ÚČNK na plnění mnoha společných úkolů, především na vytváření a správě Českého národního korpusu. Od založení ústavu úzce spolupracuje s Ústavem formální a aplikované lingvistiky MFF UK (ÚFAL) a s Ústavem pro jazyk český AV ČR. V rámci FF UK má ÚTKL (kromě ÚČNK) nejvíce společných zájmů s Ústavem českého jazyka a teorie komunikace (ÚČJTK) a s Ústavem bohemistických studií (ÚBS).

Koncepce rozvoje ÚTKL představí současný stav, plánovaný rozvoj v nejbližších třech letech a výhled na další roky ve třech oblastech: ve vědeckém výzkumu, v pedagogické činnosti a v oblasti personální.

I. Vědecký výzkum

1. Vědecké zaměření ústavu

ÚTKL se zaměřuje na počítačové zpracování češtiny, lingvistický výzkum češtiny a na formální popis jazyka a obecnou lingvistiku. Ve všech těchto oblastech je nutné sladit vlastní výzkumný zájem s možnostmi financování výzkumu vzhledem k tomu, že příspěvek („balíček“) na vědu poskytovaný fakultou je v rozpočtu ÚTKL zanedbatelný.

1.1 Počítačové zpracování češtiny

V oblasti počítačového zpracování češtiny se ÚTKL v současné době věnuje především automatické morfologické anotaci a lemmatizaci češtiny, pro tyto účely dlouhodobě vyvíjí desambiguační systém LanGr založený na lingvistických pravidlech, věnuje se úpravám morfologické analýzy a tokenizace.

Dále se zabývá syntaktickou anotací češtiny i dalších jazyků (ve standardu Pražského závislostního korpusu, nově i ve standardu Universal Dependencies), automatickou anotací frazémů a jiných víceslovných lexikálních jednotek, chybovou anotací nestandardního jazyka (čeština nerodilých mluvčích) aj. Po několik let také ÚTKL vede kolektiv lingvistů, kteří spolu vytvářejí návrh nového systému morfologické anotace češtiny NovaMorf.

V příštích třech letech by se v oblasti počítačového zpracování češtiny měl ÚTKL soustředit na zdokonalování a rozvoj počítačových nástrojů, které již vyvíjí, tj. především na zkvalitňování automatické morfologické a syntaktické anotace textů, dále na práci s víceslovnými lexikálními jednotkami a s nestandardním jazykem. ÚTKL nebude rozšiřovat záběr na zcela nové výzkumné oblasti. ÚTKL se tedy zaměří na vylepšení automatické morfologické anotace pomocí systému LanGr, především na vyhledání a odstranění desambiguačních chyb, které LanGr v určitých kontextech způsobuje. Více než dosud se bude ÚTKL zabývat lemmatizací a morfologickou anotací nestandardních textů (textů mluvených, webových, textů nerodilých mluvčích), což dosud nástroje vyvinuté v ÚTKL zvládají jen v omezené míře. Požadavky na kvalitní (v rámci možností) značkování nestandardních textů přibývají jak pro účely infrastruktury ČNK (jako hlavního zadavatele výzkumné práce pro ÚTKL), tak pro další uživatele.

Pokud projekt NovaMorf postoupí do fáze implementace (což mj. záleží na vedení infrastruktury ČNK), bude nutné nemalou část kapacit ústavu věnovat úpravě desambiguačního systému LanGr a dalším souvisejícím činnostem.

I po dokončení grantu zaměřeného na značkování víceslovných lexikálních jednotek bude ÚTKL (v rámci infrastruktury ČNK) pracovat na vývoji databáze těchto jednotek a na automatickém značkování textů s využitím této databáze.

Pracovníci ÚTKL budou nadále ve spolupráci s ÚČNK pracovat na vytváření paralelních korpusů a zlepšování jejich automatické lemmatizace, morfologické (a výhledově i syntaktické) anotace.

V oblasti automatické syntaktické anotace bude ÚTKL pracovat na zlepšení syntaktické anotace češtiny podle Pražského závislostního korpusu (mimo jiné s cílem lepší syntaktické anotace korpusu syn2020) a na syntaktické anotaci paralelních korpusů ve formátu Universal Dependencies.

ÚTKL bude také pokračovat ve vývoji metod a nástrojů pro zpracování a chybové značkování češtiny nerodilých mluvčích, a to jednak v návaznosti na dříve řešené projekty, jednak podle potřeb nových projektů ÚBS a ÚČJTK.

Z dlouhodobého hlediska je třeba zaměření výzkumu v oblasti počítačového zpracování přirozeného jazyka přizpůsobovat jednak možnostem financování daného výzkumu, jednak širšímu vývoji v oboru, kvůli němuž některé v současnosti používané metody mohou být za několik let již nekonkurenceschopné. ÚTKL se při zvažování účasti v nových projektech nevzdá své výzkumné priority, jíž je studium české morfologie a syntaxe s využitím počítačů a vývoj počítačových nástrojů pro tyto účely.

ÚTKL se bude nadále věnovat především automatické morfologické a syntaktické anotaci češtiny, ovšem jaký systém bude používat, nelze v delším časovém horizontu rozhodnout. Financování ÚTKL po roce 2022 závisí z větší části na pokračování infrastruktury ČNK, ÚTKL by tedy měl vyvinout maximální úsilí při spolupráci na předpokládaném podání přihlášky do nové výzvy (v roce 2020 či 2021).

1.2 Výzkum českého jazykového systému

Zároveň s vývojem nástrojů pro automatické zpracování češtiny probíhá na ÚTKL i výzkum českého jazykového systému, který se opírá o rozsáhlého textové korpusy češtiny, korpusy paralelní, žákovské aj. Údaje zjištěné pro účely například morfologické desambiguace jsou většinou zároveň poznatky o české morfologii, syntaxi nebo lexikologii. Tyto poznatky ÚTKL systematicky shromažďuje a (podle možností) publikuje.

V této vědecké činnosti budou pracovníci ÚTKL pokračovat i v příštích letech. Pokud to dovolí jiné úkoly, mohou publikační činnost v této oblasti ještě zintenzivnit. Jedním z dlouhodobých cílů ÚTKL (jehož plnění ovšem nebylo ještě započato) je vydat publikaci o českém jazykovém systému z pohledu automatické morfologické desambiguace.

1.3 Formální popis jazyka a obecná lingvistika

Pracovníci ústavu se věnují i formálnímu a teoretickému popisu přirozených jazyků (zvláště češtiny), rozvíjejí formální gramatiky a gramatické formalismy, zvláště HPSG (Head-Driven Phrase-Structure Grammar). Kromě teoretického výzkumu a vývoje formálních gramatik po teoretické stránce se věnují i jejich počítačové implementaci, např. v prostředí Trale.

Pracovníci ÚTKL se také věnují výzkumu v oblasti obecné lingvistiky, navazují zejména na myšlenky Pražské lingvistické školy. V roce 2011 byla například vydána publikace *Jindřich Toman: Příběh jednoho moderního projektu. Pražský lingvistický kroužek 1926–1948* (knihu přeložil Vladimír Petkevič), na počátku roku 2015 byla vydána publikace *Pražská škola v korespondenci. Dopisy z let 1924–1989* (knihu připravila Marie Havránková z Ústavu pro českou literaturu AV ČR spolu s Vladimírem Petkevičem). Připravuje se též sborník českých překladů klíčových statí významných protagonistů Pražské školy *Prague School Reader in Linguistics*, jehož redaktorem byl prof. Josef Vachek.

V této činnosti budou pracovníci ÚTKL pokračovat i v následujících letech, nakolik to jejich pracovní povinnosti v ostatních výzkumných směrech ústavu umožní (na tuto vědeckou činnost v současnosti nemá ÚTKL zajištěn žádný zdroj financování a není ani pravděpodobné, že by takový zdroj získal v dohledné době).

2. Vědecké projekty

ÚTKL se v současné době podílí na následujících vědeckých projektech:

- Český národní korpus, projekt velké infrastruktury pro VaVaI, LM2015044
- Český národní korpus, program Progres Q08 FF UK
- Jazyková variabilita v CNC, projekt OP VVV CZ.02.1.01/0.0/0.0/16_013/0001758
- Mezi slovníkem a gramatikou, grant GAČR 16-07473S
- Čeština nerodilých mluvčích z pohledu teoretického a komputačního, grant GAČR 16-10185S

V rámci infrastruktury **Český národní korpus** pracuje ÚTKL na automatické lemmatizaci, morfologické a syntaktické anotaci korpusů, na anotaci frazémů a jiných víceslovných lexikálních jednotek. Spolupracuje také na tvorbě a anotaci paralelních (tj. vícejazyčných) korpusů.

Infrastruktura ČNK je schválena na roky 2016–2019, ale díky úspěšné průběžné evaluaci z roku 2017 se dostala na tzv. cestovní mapu velkých infrastruktur (Roadmap of Large Infrastructures for Research, Experimental Development and Innovation of the Czech Republic for the years 2016–2022), jež by měla zajišťovat financování do konce roku 2022.

V programu Progres **Český národní korpus** se pracovníci ÚTKL věnují obecnějším, teoretičtějším otázkám a publikační činnosti v oblasti lingvistiky (obecné, komputační, korpusové), které nelze realizovat v rámci infrastruktury, jež je zaměřena na aplikovaný výzkum. Program je schválen na období 2017–2018.

Cílem projektu OP VVV **Jazyková variabilita v CNC** je shromáždění a zpřístupnění informací o variabilitě jazykových jednotek (slov, víceslovných lexikálních jednotek) v češtině, primárně na základě synchronních korpusů ČNK (psaných i mluvených). Období realizace: 7/2017 – 6/2020. Po ukončení tohoto projektu OP VVV lze předpokládat vyhlášení dalšího projektu OP VVV pro doplnění financování velkých infrastruktur.

V grantu GAČR **Mezi slovníkem a gramatikou** zkoumá ÚTKL ve spolupráci s některými pracovníky ÚČNK a externisty víceslovné lexikální jednotky (frazémy, idiomy, přísloví a úsloví aj.) a buduje jejich databázi. Grant byl přidělen na roky 2016–2018.

Grant GAČR **Čeština nerodilých mluvčích z pohledu teoretického a komputačního**, na němž ÚTKL spolupracuje s pracovníky ÚBS, ÚJOP UK a ÚFAL MFF UK, je zaměřen na vývoj nového systému chybové anotace češtiny nerodilých mluvčích a na vývoj nových metod automatické jazykové analýzy nestandardního jazyka. Grant byl přidělen na roky 2016–2018.

V letech 2019-2021 by měla být spolupráce s ÚČNK na velké infrastruktuře **Český národní korpus** a na společném grantu OP VVV **Jazyková varibilita v CNC** hlavní činností ústavu. Tento projekt ústavu zajišťuje finanční stabilitu a zároveň umožňuje věnovat se dlouhodobým výzkumným zájmům ÚTKL.

V roce 2018 ÚTKL nové granty (které by nahradily současné dva granty **Mezi slovníkem a gramatikou** a **Čeština nerodilých mluvčích z pohledu teoretického a komputačního** končící v roce 2018) na GAČR nepodává, především kvůli nutnosti většího zapojení do řešení úkolů v rámci infrastruktury ČNK. Jednotliví pracovníci ÚTKL budou (v případě úspěšného podání) zapojeni do nově podávaných grantů vedených jinými ústavy (ÚČJTK, ÚBS) a do projektu **KREAS**. V dalších letech bude ÚTKL samostatně nové granty podávat, pouze bude-li jasné, že stávající granty nevyužijí plně kapacitu pracovníků.

II. Pedagogická činnost

Ústav teoretické a komputační lingvistiky zajišťuje spolu s ÚČNK výuku doktorského studia v oboru matematická lingvistika. Pracovníci ÚTKL vedou semináře v následujících oblastech:

- formální zpracování přirozeného jazyka včetně matematické teorie formálních jazyků a automatů
- teoretická lingvistika, s důrazem na deklarativní (netransformační) teorie
- gramatické formalismy a jejich aplikace na popis přirozeného jazyka
- korpusová lingvistika.

Kromě toho vyučují pracovníci ústavu v oborech obecná lingvistika a jazykovědná bohemistika předmět základy jazykovědy a úvod do obecné jazykovědy. V rámci volitelných seminářů ve filologických oborech magisterského studia vedou také semináře v oboru korpusové lingvistiky.

V příštích třech letech by výuka v doktorském studiu měla pokračovat ve stejné míře jako dosud. V bakalářském a magisterském studiu je třeba vzhledem k plánovanému zařazení korpusové lingvistiky do společného základu filologických oborů počítat s požadavkem na navýšení počtu hodin v seminářích korpusové lingvistiky pro bakaláře a magistry. Lze také očekávat nárůst zájmu o volitelné (specializovanější) semináře v tomto oboru. Na tento zájem bude ÚTKL muset adekvátně reagovat (i s cílem vyvolat zájem o studium matematické / korpusové lingvistiky v doktorském studiu u vyššího počtu studentů ve studiu magisterském a bakalářském).

III. Personální oblast

Na Ústavu teoretické a komputační lingvistiky pracuje k 1. květnu 2018 pět akademických pracovníků na plný úvazek a jedna pracovnice neakademická (administrativní) na poloviční úvazek:

- **Doc. RNDr. Vladimír Petkevič, CSc.**, současný ředitel – netermínovaná smlouva
- **Ing. Alexandr Rosen, Ph.D.**, zástupce ředitele – smlouva do 30. 9. 2019
- **RNDr. Hana Skoumalová, Ph.D.**, tajemnice – smlouva do 30. 9. 2019
- **RNDr. Milena Hnátková, CSc.** – smlouva do 30. 9. 2019
- **Mgr. Tomáš Jelínek, Ph.D.** – smlouva do 31. 12. 2019
- **Lenka Horčičková**, sekretářka – smlouva celkově na plný úvazek (půl úvazku v ÚTKL, půl úvazku v Ústavu Blízkého východu a Afriky FF UK) na dobu neurčitou.

Většina vědeckých pracovníků má smlouvy jen do roku 2019, nicméně vzhledem k jejich zapojení do infrastruktury ČNK a do projektu OP VVV (viz výše) je financování jejich mezd zaručeno i v delším časovém horizontu, smlouvy tedy bude možné prodloužit.

Věkový průměr ústavu je poměrně vysoký, 56 let, je tedy důležité včas zajistit generační obnovu ústavu tak, aby bylo možné znalosti a dovednosti starších kolegů předat novým pracovníkům. První krok k této obnově byl nejspíš učiněn vyhlášením výběrového řízení (VŘ) na nového pracovníka ÚTKL (s nástupem od září 2018), jehož práce bude financována z infrastruktury ČNK (ve výběrovém řízení ovšem věk nebude hrát žádnou roli, rozhodne jen kvalifikace uchazeče).

V následujících třech letech bude mít tedy ústav šest akademických pracovníků na plný úvazek. Pokud se v těchto třech letech nebo v letech následujících některý z kolegů rozhodne snížit svůj úvazek, bude vyhlášeno VŘ, aby byly kapacity ústavu odpovídajícím způsobem doplněny. ÚTKL však nebude usilovat o rozšíření nad šest AP na plný úvazek vzhledem k prostorovým i finančním omezením. Zároveň se ÚTKL bude více než dosud snažit zapojovat do své práce studenty v doktorském studiu tak, aby mohli postupně přebírat úkoly i zkušenosti od starších kolegů, kteří budou v následujících letech snižovat své úvazky kvůli důchodovému věku.