Koncepce rozvoje *Ústavu teoretické a komputační lingvistiky FF UK*
na období 1. 2. 2013 – 31. 1. 2016

1. Úvod

Ústav teoretické a komputační lingvistiky FF UK (dále ÚTKL) byl založen roku 1990 prof. PhDr. Petrem Sgallem, DrSc., jako sesterské pracoviště *Ústavu formální a aplikované lingvistiky MFF UK* (dále ÚFAL) a jeho obecné zaměření je zřejmé z názvu. Je to ústav zaměřený především vědecky, v oblasti pedagogické ústav zajišťuje (spolu s Ústavem Českého národního korpusu FF UK) výuku doktorského studia oboru *matematická lingvistika*. Ředitelem ústavu je od června roku 1994 doc. RNDr. Vladimír Petkevič, CSc., naposledy byl jmenován ředitelem dne 21. 1. 2013, a to na tříleté funkční období od 1. 2. 2013 do 31. 1. 2016.

2. Celkové obecné zaměření ÚTKL

Ústav pracuje v těchto lingvistických odvětvích a oborech:

1. matematická lingvistika
   - korpusová lingvistika
     - příprava rozsáhlých korpusů současné češtiny, a to zejména v rámci projektu velké infrastruktury *Český národní korpus* (hlavní řešitel prof. PhDr. František Čermák, DrSc., z Ústavu Českého národního korpusu FF UK, dále ÚČNK) a v různých obdobích i projektů Grantové agentury České republiky (GAČR), popř. Evropského sociálního fondu (ESF). V projektu infrastruktury ÚTKL konkrétně řeší tuto problematiku: (a) gramatické (morfologické a syntaktické) značkování korpusů (tokenizace vstupního textu, větná segmentace, morfologická analýza, morfologická disambiguace včetně disambiguace slovnědruhové a lemmatizace); (b) příprava různých typů korpusů, zejména cizojazyčných paralelních korpusů v rámci infrastrukturního podprojektu *Interkorp*; ÚTKL zde v součinnosti s ÚČNK konkrétně zajišťuje automatické zarovnávání (alignment) a dále gramatické značkování a úkoly související. Dále se ÚTKL zabývá budováním syntakticky anotovaného korpusu (treebank) umožňujícího mj. různé pohledy na táž jazyková data;
     - příprava žákovských korpusů češtiny nerodilých mluvčích (ve spolupráci s pracovníky Katedry českého jazyka a teorie komunikace FF UK, s pracovníky Technické univerzity Liberec a Matematicko-fyzikální fakulty UK)
   - formální lingvistika
     - formální a teoretický popis přirozeného jazyka (zejména češtiny);
     - syntax přirozeného jazyka a její teoretické a počítačové zpracování (automatická syntaktická analýza češtiny);
     - tvorba valenčního slovníku a slovníku kolokací;
     - gramatické formalismy (Funkční generativní popis jazyka, Head-Driven Phrase-Structure Grammar aj.);
     - příprava encyklopedických hesel oboru matematická lingvistika (včetně lingvistiky korpusové);
2. obecná lingvistika – vydávání (překladových) publikací souvisejících hlavně s Pražskou školou strukturní lingvistiky
3. teoretická lingvistika, a to zvláště v oblasti morfologie a syntaxe přirozeného jazyka, především češtiny.

Pracovníci ústavu vyučují v seminářích oboru matematická lingvistika v těchto 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
- gramatika češtiny
- korpusová lingvistika.

Pracovníci ÚTKL mimoto vyučují tyto předměty v oborech obecná lingvistika a jazykovědná bohemistika:

- základy jazykovědy a úvod do obecné jazykovědy
- gramatika češtiny (v rámci projektu ESF *Moderní mluvnice češtiny* 2010–2013).

I v letech 2013 až 2015 se ÚTKL soustředí na vědeckovýzkumnou a pedagogickou činnost ve výše uvedených odvětvích s tím, že bude zejména plnit úkoly stanovené ve zmíněném projektu velké infrastruktury *Český národní korpus* (2012–2016) a podle možností v projektu *PRVOUK*, a to ve spolupráci s ÚČNK. Mimoto bude řešit úkoly stanovené v jiných projektech (podrobněji v odst. 4).

3. Oblast personálního rozvoje

V ÚTKL pracuje v současnosti šest pracovníků, z toho pět na plný úvazek:

- **doc. RNDr. Vladimír Petkevič, CSc.**, ředitel – netermínovaná smlouva
- **ing. Alexandr Rosen, Ph.D.**, zástupce ředitele – smlouva do 30. 9. 2014
- **RNDr. Hana Skoumalová, Ph.D.**, tajemnice – smlouva do 30. 9. 2014
- **RNDr. Milena Hnátková, CSc.** – smlouva do 30. 9. 201
- **Mgr. Tomáš Jelínek, Ph.D.** – smlouva do 31. 12. 2014
- **Lenka Horčičková**, sekretářka – smlouva celkově na plný úvazek (půl úvazku v ÚTKL, půl úvazku v Ústavu Blízkého východu a Afriky FF UK) do 30. 9. 2014.

V ústavu je tedy pět vědeckých pracovníků na plný úvazek a sekretářka na úvazek poloviční. Pracovníci jsou v současnosti, tj. v roce 2013, placeni jednak z prostředků přidělených přímo ústavu v podobě tzv. mzdového „balíčku“, jednak z projektu velké infrastruktury *Český národní korpus*, jednak i z projektu GAČR *Treebank češtiny na základě gramatiky* (od 1. 2. 2013). Podrobněji o těchto projektech viz níže v odst. 4.

Práci na uvedených projektech považuje ústav za svou vědeckovýzkumnou prioritu. Na projektu infrastruktury a grantu GAČR se podílejí a budou podílet též (zejména) mladí externisté, kteří jsou zaměstnáváni na dohody o provedení práce a na dohody o pracovní činnosti. Zvládnout vědeckou práci na uvedených projektech je ovšem velmi náročné, a v tomto směru, proto plánujeme (pochopitelně v závislosti na finančních možnostech) takovéto personální posílení ústavu k řešení grantových úkolů:

- o programátora-počítačového lingvistu
- nejméně o jednoho lingvistu.

Ústav úzce spolupracuje především s ÚČNK a dále pak s Ústavem formální a aplikované lingvistiky MFF UK, zčásti i s Ústavem pro jazyk český AV ČR, Filozofickou fakultou Masarykovy univerzity v Brně a Filozofickou fakultou Univerzity Palackého v Olomouci a dalšími obdobně zaměřenými pracovišti.

Za jednu z priorit považuje ústav *kvalifikační postupy pracovníků* ÚTKL, tj. profesury, docentury a úspěšné absolvování doktorského studia doktorandů ať už oboru matematická

lingvistika, tak oborů jiných (pracovníci ústavu jsou školiteli i doktorandů jiných jazykových oborů).

4. Rozvoj vědecké činnosti

Hlavní náplní činnosti ÚTKL je vědeckovýzkumná práce v oboru *matematická lingvistika*. Během své existence obdržel ústav řadu českých a mezinárodních grantů, v nichž figuroval jako hlavní řešitelské nebo spoluřešitelské pracoviště, většinou s doc. Petkevičem jako hlavním odpovědným řešitelem.

Byly to zejména tyto granty:

Granty Grantové agentury České republiky (GAČR):

- **Počítačový korpus českých psaných textů (Počítačový fond češtiny)** (1993–1995) (ÚTKL byl hlavním řešitelským pracovištěm)
- **Počítačové nástroje pro komputační zpracování českých textů** (1995–1997) (ÚTKL byl hlavním řešitelským pracovištěm)
- Komplexní projekt GAČR **Čeština ve věku počítačů** (1996–2001) (hlavním řešitelským pracovištěm byl Ústav formální a aplikované lingvistiky MFF UK, ÚTKL byl spoluřešitelským pracovištěm)
- **Elektronizace postupů diachronní lexikografie** (1999–2001) (hlavním řešitelským pracovištěm byl Ústav pro jazyk český AV ČR, spoluřešitelským pracovištěm pak FF UK, konkrétně ÚTKL)
- **Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu** (2003–2005) (hlavním řešitelským pracovištěm byl Ústav pro jazyk český AV ČR, spoluřešitel pak FF UK, konkrétně ÚTKL)
- **Slovní poklad češtiny v informační společnosti** (2003–2005) (hlavním řešitelským pracovištěm byl Ústav pro jazyk český AV ČR, spoluřešitelským pracovištěm pak FF UK, konkrétně ÚTKL)
- **Velké jazykové korpusy a jejich automatická analýza** (2003–2005) (hlavním řešitelským pracovištěm byl Ústav formální a aplikované lingvistiky MFF UK, ÚTKL se podílel na plnění úkolů projektu)
- **Syntaktická analýza českých textů** (2010–2012) (ÚTKL byl hlavním řešitelským pracovištěm)

Grant MŠMT:
Příprava testovacích dat a nástrojů pro testování lingvistického software (2005–2007) (hlavním řešitelským pracovištěm byl Ústav pro jazyk český AV ČR, ÚTKL se podílel na plnění úkolů projektu).

Výzkumný záměr:

- **Český národní korpus a korpusy jiných jazyků** (2005–2011, reg. č. MSM0021620823, vedoucí záměru prof. PhDr. František Čermák, DrSc., ÚČNK) ve spolupráci s ÚČNK

Projekt typu ESF (Evropský sociální fond):

- **Inovace vzdělávání v oboru čeština jako druhý jazyk** (červen 2009 – květen 2012, reg. č. CZ.1.07/2.2.00/07.0259, *OP Vzdělávání pro konkurenceschopnost*, hlavní koordinátor prof. PhDr. Karel Šebesta, CSc.) ve spolupráci s Katedrou českého jazyka a teorie komunikace FF UK a ve spolupráci s Technickou univerzitou Liberec.

Z mezinárodních grantových projektů uvádíme tyto:

- Mezinárodní projekt **Language Technologies for Slavic Languages** (**LATESLAV**) – **PECO 2824** (1993–1995). ÚTKL se mimo Ústav formální a aplikované lingvistiky MFF UK podílel na plnění úkolů tohoto grantu za českou stranu.
- Mezinárodní projekt **MULTEXT–EAST. Multilingual Text Tools and Corpora for Central and Eastern European Languages** (1995–1997) (COP106). ÚTKL byl hlavním řešitelským pracovištěm za českou stranu.
- Mezinárodní projekt **TELRI (Trans–European Language Resources Initiative**) (1995–1997). ÚTKL se mimo další pracoviště v České republice podílel na plnění úkolů tohoto grantu za českou stranu.
- Mezinárodní projekt **CONCEDE. Consortium for Central European Dictionary Encoding** (1998–2000) (PL-1142). ÚTKL byl hlavním řešitelským pracovištěm za českou stranu.
- **Shared formal grammar of Czech and Polish** (program vědecko-technické spolupráce MŠMT KONTAKT 2004/23). ÚTKL byl spoluřešitelským pracovištěm.

V současnosti pracuje ÚTKL na těchto projektech:

- Velká infrastruktura **Český národní korpus** (2012–2016, reg. č. LM2011023, vedoucí prof. PhDr. František Čermák, DrSc., ÚČNK) ve spolupráci s ÚČNK.

Budoucí projekty:

- Grant GAČR **Treebank češtiny na základě gramatiky** (1. 2. 2013–31. 12. 2015, reg. č. P406/13-27184S). Hlavním řešitelem tohoto čerstvě přiděleného grantu bude doc. RNDr. Vladimír Petkevič, CSc., hlavním řešitelským pracovištěm bude ÚTKL.

Pracovníci ÚTKL se též jako externí spolupracovníci podílejí na řešení projektů:

- **Moderní mluvnice češtiny** (grant ESF, 2010 – 30. 6. 2013)
- **Nový encyklopedický slovník češtiny** (ESF, 2012–2014).

Vědecká činnost ÚTKL v nejbližším období bude probíhat především v rámci těchto projektů:

- Velká infrastruktura **Český národní korpus**. ÚTKL zde spolupracuje a bude spolupracovat s ÚČNK podle stanoveného rozvrhu do konce roku 2016 včetně;
- Grant GAČR **Treebank češtiny na základě gramatiky** (podrobnosti viz výše).

Mimo uvedené oblasti činnosti pracovníci ústavu příležitostně přednášejí/vyučují v zahraničí. Zúčastňují se vědeckých konferencí doma i v zahraničí (z posledních jmenujme účast s referáty na konferencích ve Varšavě, Řezně, Sankt-Petěrburku, Dubrovníku, Istanbulu, Göttingen).

Ústav spolupracuje s obdobně zaměřenými zahraničními lingvistickými pracovišti, například s Univerzitou v Řezně (spolupráce v oblasti syntaxe slovanských jazyků a paralelních korpusů) s Univerzitou v Torontu (dlouhodobá spolupráce na vývoji topologického parseru a formalismu pro adekvátní popis slovosledu), s univerzitou ve

Varšavě (gramatické formalismy). Ke kontaktům se zahraničím patří i návštěvy význačných zahraničních odborníků, kteří ústav navštívili v poslední době: byli to například prof. Zygmunt Saloni z Varšavské univerzity, dr. Ruprecht von Waldenfels z Univerzity v Bernu, prof. Leonid Iomdin z Ruské akademie věd, prof. Gerald Penn z Univerzity v Torontu, doc. Viktor Zacharov z Univerzity v Petrohradě.

Pracovníci ústavu odborně působí rovněž v oblasti obecné lingvistiky, zejména pečují o myšlenkové dědictví Pražské lingvistické školy. Roku 2005 vyšla česká verze publikace *Lingvistického slovníku Pražské školy* od Josefa Vachka, v říjnu roku 2011 byla vydána publikace *Jindřich Toman: Příběh jednoho moderního projektu. Pražský lingvistický kroužek 1926–1948* – na obou publikacích se podíleli pracovníci ÚTKL. Připravuje se též rozsáhlá publikace korespondence členů Pražské školy strukturní lingvistiky (ve spolupráci s dr. Marií Havránkovou z Ústavu pro českou literaturu AV ČR) i sborník českých překladů klíčových statí významných protagonistů Pražské školy *Prague School Reader in Linguistics*, jehož redaktorem byl prof. Josef Vachek.

4.1 Zhodnocení realizace cílů dosavadní koncepce vědecké činnosti

V oblasti vědecké činnosti byly v posledním období, tj. v letech 2010–2012 (tedy v předcházejícím funkčním období staronového ředitele ústavu) veškeré naplánované úkoly v rámci hlavního výzkumného projektu – výzkumného záměru **Český národní korpus a korpusy jiných jazyků –** splněny. Mimo plánovanou činnost se ÚTKL výrazně podílel i na projektu **Inovace vzdělávání v oboru čeština jako druhý jazyk**, který byl úspěšně dokončen v květnu roku 2012. Proběhly závěrečné práce na grantu GAČR **Syntaktická anotace českých korpusů** a v roce 2012 byly zahájeny práce na projektu infrastruktury **Český národní korpus**.

5. Rozvoj pedagogické činnosti

ÚTKL zajišťuje (ve spolupráci s ÚČNK) na FF UK studijní obor: *Filologie – matematická lingvistika* (doktorské studium), nabízí výuku v kursech *počítačové a formální lingvistiky*, *základů jazykovědy a obecné lingvistiky* v podobě povinně volitelných a výběrových přednášek a seminářů pro pregraduální i postgraduální studenty. Takto se také podílí na výuce pro tyto obory: *logika, srovnávací jazykověda, lingvistika a fonetika, český jazyk a literatura* na FF UK a konečně *počítačová a formální lingvistika* na MFF UK. Na téže fakultě vyučuje také zahraniční studenty v rámci magisterského oboru *jazyk a informační technologie* programu Erasmus Mundus. Mimo vlastní přednášky a semináře se ústav rovněž podílí na výuce v semináři *korpusové lingvistiky*, který je organizován ÚČNK. Přednášky a semináře navštěvují nejen studenti (hlavně doktorandi) FF UK, ale i studenti z MFF UK a dalších fakult UK. Pracovníci ústavu vedli a vedou také doktorandy v oborech *matematická lingvistika*, *český jazyk a literatura*, *obecná lingvistika* a *germanistika*.

Členové ústavu doc. Petkevič, dr. Skoumalová a dr. Rosen jsou členy oborové rady doktorského oboru *matematická lingvistika.* Doc. Petkevič je rovněž členem oborové rady oboru *logika* a *český jazyk a literatura* na FF UK*, obecný a indoevropský jazykozpyt* na Filozofické fakultě Masarykovy univerzity v Brně (FF MU) a připravuje se jeho jmenování členem oborové rady oboru *český jazyk* na Filozofické fakultě Univerzity Palackého v Olomouci (FF UP). Doc. Petkevič a dr. Rosen jsou též členy oborové rady oboru *matematická lingvistika* na MFF UK. Pracovníci ústavu bývají oponenty disertačních prací v oboru *matematická lingvistika* a dalších oborech (například *český jazyk a literatura*, *germanistika*), píší posudky na habilitační práce a zasedají v komisích pro státní doktorské zkoušky v uvedených oborech na FF UK, MFF UK, Fakultě informatiky Masarykovy univerzity (FI MU), FF MU, FF UP. Píší také posudky na projekty podávané u Grantové

agentury Univerzity Karlovy. Rovněž příležitostně přednášejí v semináři formální lingvistiky pořádaném Ústavem formální a aplikované lingvistiky MFF UK.

Doc. Petkevič a dr. Rosen se podílejí na výchově vědeckých pracovníků v rámci doktorského studia oboru *matematická lingvistika*, i oborů jiných (český jazyk a literatura, germanistika).

5.1 Zhodnocení realizace cílů dosavadní koncepce pedagogické činnosti

Cíle dosavadní koncepce byly splněny ve výše uvedeném smyslu. Z hlediska konkrétních úspěchů v pedagogické činnosti v poslední době oznamuji, že v červnu roku 2012 obhájil svou doktorskou práci Mgr. Tomáš Jelínek pod vedením doc. Petkeviče jako školitele. Jeho plný titul nyní zní Mgr. Tomáš Jelínek, Ph.D.

6. Výhled dalších oblastí rozvoje základní součásti

Mimo plány uvedené výše se ÚTKL bude v příštích letech věnovat obecně těmto aktivitám:

(a) širší spolupráci s dalšími oborovými pracovišti, zejména pak s těmito: s ÚČNK a dalšími lingvisticky zaměřenými ústavy a katedrami FF UK, katedrou logiky FF UK, ÚFAL MFF UK, ÚJČ AV ČR, FI MU, FF MU a FF UP;
(b) prohlubování spolupráce se zahraničními oborovými pracovišti, s nimiž už má ÚTKL dlouhodobé kontakty, a navazování odborných kontaktů s novými pracovišti.

7. Stručný profil kandidáta na ředitele doc. RNDr. Vladimíra Petkeviče, CSc.

Stručné odborné curriculum vitae

- Narodil se 2. 3. 1954 v Praze
- vystudoval Matematicko-fyzikální fakultu Univerzity Karlovy (MFF UK), obor *Matematické zabezpečení výpočetní techniky* (1974–1979)
- Doktorát z přírodních věd (RNDr.), obor: *matematická informatika a teoretická kybernetika* (1985)
- Kandidát věd (CSc.), obor: *matematická informatika a teoretická kybernetika* (1992)
- Zaměstnán ve *Výzkumném ústavu matematických strojů (VÚMS)*. Podílel se na vývoji operačních systémů pro sálové (mainframe) počítače, dále na vývoji překladačů programovacích jazyků a na vývoji databázových programů (1979–1992).
- Dne 1. ledna 1993 přijat jako samostatný vědecký pracovník do *Ústavu teoretické a komputační lingvistiky* FF UK v Praze.
- Od 10. 6. 1994 dosud ředitelem ÚTKL FF UK.
- Habilitace v oboru *matematická lingvistika* (1996), název práce: *Underlying Structure of Sentence Based on Dependency*
- Od roku 1996: vedoucí lingvistické sekce (pro synchronní jazyk) Ústavu Českého národního korpusu FF UK, a to do roku 2006
- Přednášky na zahraničních univerzitách: Düsseldorf, Heidelberg, Erlangen, Tübingen (SRN), Bratislava; výuka na letní škole (Sozopol 2002)

- Účast na mezinárodních konferencích (nejvýznamnější: COLING´88, ACL´93, EURALEX´94, COLING–ACL´98, EURALEX´2000, Corpus Linguistics 2005 Birmingham, Corpus Linguistics 2011 Sankt-Petěrburg, SlaviCorp 2011 Dubrovník)
- člen ACL (Association for Computational Linguistics)
- člen Jazykovědného sdružení České republiky
- člen Pražského lingvistického kroužku
- člen těchto oborových rad: pro obor *matematická lingvistika* (FF UK), *logika* (FF UK), *český jazyk a literatura* (FF UK), *matematická lingvistika* (MFF UK), *obecný a indoevropský jazykozpyt* (FF MU).

Odborná specializace:

matematická lingvistika a lingvistická bohemistika:

- korpusová lingvistika – lingvistické značkování jazykových korpusů (zejména korpusů synchronní češtiny v rámci projektu *Český národní korpus*)
- syntax a morfologie přirozeného jazyka a její formální a počítačové zpracování, automatická morfologická a syntaktická analýza češtiny, gramatické formalismy

obecná jazykověda
- pražský strukturalismus

Konkrétní pedagogická a odborná činnost v posledních pěti letech:

- výuka v oboru matematická lingvistika, konkrétně: základy oboru (formální popis přirozeného jazyka, korpusová lingvistika, matematické metody v lingvistice, gramatické formalismy); výuka v oboru matematika pro filology; výuka základů jazykovědy a úvodu do obecné jazykovědy; občasné přednášky na semináři pořádaném na MFF UK a na jiných fakultách (např. FF UP, FF MU, FI MU, Západočeská univerzita v Plzni)
- vedení devíti doktorandů v doktorském oboru *matematická lingvistika*, z nichž čtyři doktorandi v období prosinec 2007 – prosinec 2012 úspěšně absolvovali doktorské studium.
- jazykové značkování jazykových korpusů češtiny (zejm. morfosyntaktické)
- statistický výzkum současné češtiny
- příprava žákovského korpusu nerodilých mluvčích
- výzkum (morfo)syntaxe české věty na základě korpusů současné češtiny
- překlady v oboru lingvistika a matematická lingvistika
- redakční činnost při přípravě publikací v uvedených oborech.

V Praze dne 21. ledna 2013

doc. RNDr. Vladimír Petkevič, CSc., ředitel ÚTKL FF UK