



## Univerzita Karlova v Praze, Filozofická fakulta Ústav Českého národního korpusu

nám. Jana Palacha 2, 116 36 Praha 1  
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

### Koncepce rozvoje *Ústavu Českého národního korpusu*

#### Preambule a nedávná minulost

Ústav Českého národního korpusu (ÚČNK), jemuž od jeho založení (r. 1994) stojím v čele, je pracoviště svou působností v celé zemi jedinečné a od ostatních pracovišť FFUK dosti odlišné. Vzniklo jako výsledek soustředěného úsilí řady institucí a kateder ze všech čelných univerzit, včetně ČVUT, i několika ústavů akademie věd; s nimi děkanové FF uzavřeli smlouvy o spolupráci s ÚČNK. Cílem jeho založení bylo a je budovat *Český národní korpus*, jedinečný, svého druhu jediný a nenahraditelný elektronický zdroj jazykových dat, a tedy i informace, a to nejen pro lingvisty, ale stejně tak pro odborníky řady dalších oborů (na základě prosté pravdy, že prakticky veškerá informace prochází jazykem) jako pro studenty a širokou veřejnost, včetně té mezinárodní. Jde dnes už o nepostradatelnou **informační základnu** jak **pro výzkum**, tak i pro ty nezákladnější aplikace, jako je budoucí nový slovník češtiny, jeho mluvnic a řada nezákladnějších příruček; jiná alternativa k němu dnes už není. Za oněch pár let si svými výsledky získal i značné mezinárodní renomé a respekt a jeho zkušenosti rádí dnes sdílí i jinde v zahraničí; přilákal už i mnohé výzvy ke spolupráci na evropské úrovni, nejčastěji v podobě společných evropských projektů. V současné době, v rozmezí několika let, nabídl odborné i široké veřejnosti zdarma k výzkumu dva stomilionové korpusy současné psané češtiny, první diachronní korpus a dva korpusy jazyka mluveného. Mám za to, že v takto dobře započaté a široce oceňované práci (stříbrná medaile rektora UK v r. 2001 a osobní cena ministryně školství za vědu v r. 2003) se musí pokračovat a jazykový vývoj a proměny mapovat kontinuálně dál do budoucnosti i minulosti, což se zatím díky poslednímu *Výzkumnému záměru (MSM 0021620823 na léta 2005-2011, Český národní korpus a korpusy dalších jazyků)* pod mým vedením daří. V poslední etapě se pokračuje dokonce v rozšířené podobě (subprojekt *InterCorp*).

Český národní korpus, který je vědecké a akademické povahy, byl a je podporován sérií celkem 8 udělených grantů a VZ (GAČR, GAUK, MŠMT), sponzemi (zvláště bank) a dokonce finanční participací Nakladatelství Lidových novin. Díky osvícenému vedení FFUK mohl být založen právě na ní (navzdory nemoudrému tehdejšímu postoji akademie věd) a dnes, vedle zmíněné mnohostranné spolupráce, jeho výsledků a služeb se využívá **celostátně** (právě pracovníci akademie, vysokých škol po celé ČR i některých škol středních) i **mezinárodně** (především zahraniční lektoráty, ale i spolupráce v projektech EU aj.).

Během krátké doby své existence musel tým ÚČNK začít řešit komplex zcela nových problémů technických a odborných a dokázal vyvinout vlastní postupy a metodologii, které jsou oceňovány i mezinárodně. Jejich odrazem je mj. vybudovaná větev **korpusové lingvistiky**, kterou lze dnes v doktorském studiu (matematická lingvistika) studovat právě jen na FFUK při ÚČNK, na druhé straně, vedle reprezentativních a široce využívaných korpusů, které stále rostou, dokázal tento tým ale také zpracovat a vydat mj. i oceňovaný velký *frekvenční slovník češtiny* apod. Tento rok byla založena celá nová ediční řada *Studie z korpusové lingvistiky*, vycházející v NLN, jejíž první 4 svazky jsou v tisku a mají do konce roku vyjít.

Je zřejmé, že náplň činnosti ÚČNK je primárně a většinou **vědecká**, podmíněná plněním udělených grantů (dnes VZ), sekundárně však i pedagogická, osvětová, servisní a další.

#### Můj osobní profil

Jsem profesorem českého jazyka a docentem obecné jazykovědy FFUK se specializací na bohemistiku, některé jazyky slovanské, germánské (zvl. holandštinu, angličtinu a jazyky skandinávské) a na jazyky ugrofinské (zvl. finštinu); specializuji se mj. na lexikologii, lexikografii, morfologii, frazeologii, typologii, obecnou jazykovědu, teorii jazyka a jeho metodologii, korpusovou lingvistiku aj.

Mé zatím publikované **dílo** zahrnuje (srov. též na webu ucnk.ff.cuni.cz/cermak.html):

Knihy (monografie): 7 (některé ve spoluautorství, některé v překladech, některé ve více vydáních)

Slovníky: 6 (obv. ve spoluautorství, některé ve více vydáních)

Skripta: 25 (některé ve spoluautorství, některé monografické, jiné učebnice, některé více vydání)



## Univerzita Karlova v Praze, Filozofická fakulta Ústav Českého národního korpusu

nám. Jana Palacha 2, 116 36 Praha 1  
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

Studie: 110 (časopisy, sborníky, ve slovnících, některé i v překladu do litev., ital, špan, a slovin.)  
Překlady knižní: 2 (z dánštiny a francouzštiny-italštiny, obě nejzákladnější práce moderní lingvistiky)  
Edice 6 (některé ve spoluautorství a spolueditorství)  
Recenze: cca 30 (některé ve spoluautorství)

Jsem **členem** redakčních rad odborných časopisů ve Slovinsku, Španělsku (dvou) a celosvětového *International Journal of Corpus Linguistics*.

Jsem členem dvou vědeckých rad (ÚJČ AVČR, FFUK), odborných sdružení našich (*Jazykovědné sdružení, Pražský lingvistický kroužek*) i mezinárodních (*Societas linguistica Europea, Europhras, Euralex*, u posledních dvou člen řídicího výboru), byl jsem členem několika projektů EU (zvl. *Telri 2*, věnovaného korpusům, kde jsem byl vědeckým koordinátorem).

Jsem členem **oborových rad** na FFUK: germanistika, ugrofenistika a matematická lingvistika (předseda posledních dvou); matematická lingvistika na MFF UK.

Jsem členem *Učené společnosti České republiky* a od r. 1994 jsem ředitelem ÚČNK.

### Personální situace ÚČNK

Původním vkladem FFUK do nově založeného ústavu byly v roce 1994 dva lidé, a tedy dva úvazky, jeden nový a jeden starý (můj, podotýkám ale, že svůj profesorský úvazek, vedle nového úvazku v ÚČNK, dosud v nezmenšené míře plním, paralelně mám de facto úvazky dva). Všichni ostatní pracovníci jsou **najímání** na práci na ČNK díky grantům a později VZ; v tomto se tedy naše pracoviště zásadně liší od všech ostatních (časem však do stálého kádru fakultních zaměstnanců byly převedeni další dva pracovníci). Dnešní stav, který je tak plánován v současném a už podruhé uděleném VZ *Český národní korpus a korpusy dalších jazyků*, jehož jsem nositelem, činí 15 interních pracovníků (většina je tedy do r. 2011 placena z tohoto VZ), čtyři interní doktorandi a řada externích; další jsou ovšem plánováni a je pravděpodobné, že na základě budoucího financování z projektu EU bude třeba najmout i nové lidi.

Vedle uvedených 15 interních pracovníků však náš VZ financuje na poloviční úvazky ještě 7 dalších participujících pracovníků z fakulty; v menší míře však financuje také na 20 koordinátorů jednotlivých jazyků z multilingválního subprojektu *Interkorp*, jehož činnost pod ÚČNK spadá. Konečně je třeba uvést, že pro ÚČNK pracuje velké množství lidí včetně studentů na specializovaných úkolech, především skenování a sběru a přepisu i mluvených textů (sem jde ročně na 2 miliony korun, formou OON).

**Odborně** je vlastní pracoviště relativně velmi mladé, a protože hledanou kvalifikaci nemohlo mít (takový obor předtím neexistoval), muselo se specializovat samo, především pod vedením starších a zkušenějších členů týmu. Jakkoliv jsou oborově různí (lingvisté, matematici, inženýři), je z nich dnes kompaktní a kompetentní tým, který mj. slouží i jako spolehlivá základna schopných konzultantů v řadě specializovaných otázek; nemluví přitom o jejich publikační činnosti, která je na webu a v bibliografiích, účasti na mezinárodních fórech aj. Tým se vzhledem k povaze práce nutně dále dělí na technické a akademické pracovníky. Přes relativní mládí z akademických už dva dosáhli hodnosti PhD a další dva jí v dohledné době dosáhnou. V **oborové radě** matematické lingvistiky, jejímž jsem předsedou, zasedají všichni relevantní specialisté u nás, a to nejen z FFUK.

Personální situace ÚČNK je tedy dobrá, mj. i díky od začátku rigorózně uplatňované praxi konkurzů, v němž se vybírají noví pracovníci jen na základě skutečných potřeb.

Já sám v současnosti vedu šest doktorandů, z toho jednu studentku z Lotyšska, jednu z Tchajvanu a jednu z Bulharska, ostatní jsou Češi a ještě jiní svou přípravu už skončili.

### Pedagogická činnost ÚČNK

Samo **studium korpusové lingvistiky** v rámci matematické lingvistiky je interdisciplinární, tj. zdaleka ne jen pro studenty z FFUK či České republiky (vedle zmíněných máme studenty např. ze Slovenska a Itálie). Vlastní



## Univerzita Karlova v Praze, Filozofická fakulta Ústav Českého národního korpusu

nám. Jana Palacha 2, 116 36 Praha 1  
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

studijní obor, patřící výhradně do 3. cyklu, je pochopitelně založený na studiu individuálním, resp. individuálních plánech doktorandů. Pedagogická činnost ÚČNK se soustřeďuje do centrálního korpusového semináře pro doktorandy a další zájemce, včetně studentů mimofakultních; přednáší a pracuje v něm řada specialistů, včetně mimofakultních (já ovšem také). Navíc někteří pracovníci vyhláší podle zájmu a potřeby různé další semináře pro studenty fakulty. Pedagogickou povahu má ovšem i řada seminářů pro učitele jiných fakult a univerzit, ba i středních škol v oboru vytěžování a studia korpusu.

Jak již bylo zmíněno výše, já sám plním vedle toho svůj profesorský úvazek (8 hodin) výukou českých a zahraničních studentů češtiny. To zahrnuje mj. i všechny doprovodné činnosti včetně vedení diplomových prací aj.

### Vědecká činnost ÚČNK

Vědecká činnost ÚČNK je jeho základní a vlastní náplní a odvíjí se od zadání VZ (blíže viz tam). Lze ji rozdělit do dvou oblastí, **výzkumu základního a aplikovaného**. Do toho **prvního** patří vlastní budování různých korpusů (projekt ČNK je komplexní, složený z mnoha částí), včetně organizace a realizace shromažďování a mnohastupňového zpracování dat, která nakonec jsou zveřejňována na Internetu ([ucnk.ff.cuni.cz](http://ucnk.ff.cuni.cz)) a hromadně využívána. Nedílnou součástí je i výzkum vedoucí k propracování a formulaci kritérií výstavby korpusu, precizace metodologie užívání korpusu, která je nutná, dosud nepoznaná a donedávna zcela chyběla (jde o kvalitativně i kvantitativně nové poměry spojující matematické metody s hromadnými jazykovými daty) aj. Do **druhého**, aplikovaného výzkumu patří tvorba vlastních výstupů, ať už v podobě individuálních studií, nebo výstupů týmových (jako je zmíněný *frekvenční slovník* či začínající *ediční řada* s prvními svazky už v tisku) či výstupů pedagogických týkajících se užívání korpusu (nedávná týmová kniha *Jak užívat ČNK*) a studia (*překladová čítanka* z korpusové lingvistiky) a další.

Sám se na všech typech vědecké činnosti podílím, většinu z nich koordinuju. Obraz o této mé činnosti podává i bibliografie zmíněná výše. Jen dodám, že jedna má knižní práce vyšla i v anglickém a litevském překladu, jiné (studie) byly přeloženy do dalších jazyků.

### Výhledy do budoucnosti

Vzhledem k nejistotě a vlastně i nedůstojnosti nutnosti stále znovu žádat o financování toho, co se již plně osvědčilo a čeho bude zapotřebí kontinuálně i do budoucna (namísto **trvalého zabezpečení** se tedy znovu a znovu periodicky dosud zdůvodňuje, co je zdůvodněno už dávno a musí vstupovat se do soutěže s odlišnými, většinou jasně krátkodobými projekty) lze výhled odhadovat jen obtížně. Optimisticky řečeno, tj. pokud financování projektu ČNK, o jehož potřebě nikdo nepochybuje, bude pokračovat, lze tu především vidět

- (a) prohloubené pokračování už existujících linií (výstavba korpusů synchronních, diachronního, zvláště potřebného mluveného a korpusů paralelních zatím mezi češtinou a cca dvacítkou dalších jazyků),
- (b) rozvoj metodologie vytěžování s postupnou realizací výsledků např. i v dalších seminářích,
- (c) větší propojení s jinými evropskými centry a korpusy,
- (d) výchova dalších specialistů, především v doktorském studiu,
- (e) další vlastní aplikované výstupy (pevně je např. naplánován už *slovník totality*, *dva autorské slovníky*, a to K. Čapka a B. Hrabala, a *kvantitativní popis češtiny*, *studie o mluveném jazyce*)
- (f) servisní činnost (údržba sítě, webového přístupu, administrativa s tím spojená aj.)
- (g) prohloubená kooperace s dalšími projekty (mj. s připravovaným velkým slovníkem současné češtiny)

15. října 2006

Prof. PhDr. František Čermák, DrSc.