Koncepce rozvoje Ústavu Českého národního korpusu v období 2013-2015
Václav Cvrček, Ph.D.

Ústav Českého národního korpusu (ÚČNK) byl založen před téměř 20 lety a od té doby se stal centrem oboru korpusová lingvistika a hlavní datovou základnou pro jazykovědný výzkum u nás. Za svoji úspěšnou historii vděčí především zakladatelům (zejm. prof. F. Čermákovi), kteří celý projekt vybudovali z ničeho, navíc v situaci, kdy korpusový přístup nebyl v české jazykovědě uznáván. Projekt Českého národního korpusu (ČNK) se nyní dostává do situace, kdy zakladatelská generace pomalu odchází, a všichni následovníci se musí snažit obstát ve srovnání s jejich úsilím i výsledky.

Vědomí tradice ovšem nesmí být svazující, ale inspirující k dalšímu rozvíjení toho, co se osvědčilo, k podpoře toho, čemu dosud nebyla věnována pozornost, a ke změně tam, kde se stávající praxe ukazuje jako neefektivní nebo zastaralá. Stejně jako ve všech oborech, i v korpusové lingvistice a v činnostech s ní souvisejících jde vývoj ve světě rychle kupředu, a je proto třeba tyto trendy sledovat a snažit se o jejich aplikaci na češtinu.

ÚČNK je jako základní součást FF UK specifický minimálně ze dvou hledisek: 1) jedná se o pracoviště s většinovým podílem vědecko-technické práce a 2) jeho financování je kryto především z prostředků mimo FF UK. Z toho plyne, že i koncepce jeho rozvoje bude (na rozdíl od ostatních pracovišť fakulty) akcentovat jiné problémy a úkoly.

Právě oblast financování, v níž se ÚČNK dlouhodobě spoléhá na externí zdroje, musí představovat rámec všech úvah o jeho budoucnosti. V současnosti je jeho provoz z 95 % hrazen z projektu „Velké infrastruktury pro výzkum, vývoj a inovace“, který umožňuje jeho rozvoj především po stránce datové, technické, metodologické a popularizační. Vedle toho byla ÚČNK přiznána i podpora v rámci UK (Prvouk/VVZ), která umožňuje realizovat tu část výzkumu, který se přímo netýká budování korpusů (tedy jeho designu, metod vytěžování apod.). Oba zdroje financování předpokládají splnění závazných výstupů, jejichž realizace nesmí být ohrožena, a které je tedy třeba považovat za *implicite* obsažené v této koncepci.

Jako obecný a kontinuální úkol rozvoje ÚČNK je ve světle těchto specifik třeba postulovat hledání rovnováhy mezi činností vědecko-pedagogickou a servisní (zejm. budování a poskytování korpusů). Obě složky jsou nezbytné pro zdravé fungování projektu ČNK, jedna je na druhé životně závislá, jsou si vzájemnou inspirací, obě jsou výsledkem nepřetržité a dlouhodobé činnosti, kterou nelze (v případě potřeby po případném vynuceném útlumu) rychle a snadno resuscitovat. V praktické rovině se hledání této rovnováhy projevuje snahou o získávání finančních zdrojů pro obě hlavní složky činnosti ÚČNK zaručující jejich soustavné rozvíjení.

Oblast personálního rozvoje

Historickým vývojem a v důsledku nutné specializace činností při budování korpusů byly vytvořeny (a infrastrukturou de facto potvrzeny) pracovní sekce: lingvistická, komputační, diachronní, sekce mluvených korpusů a sekce paralelních korpusů (anotační sekce je zajišťována partnerským ÚTKL). Autonomie těchto celků se ukazuje být jako funkční, výhledově je ovšem třeba zajistit účinnou kontrolu výstupů a koordinaci činností (vhodnou platformou je tradiční interní seminář, jehož potenciál ovšem není plně využit).

V současnosti v ÚČNK pracuje 22 zaměstnanců:

- 2 akademičtí pracovníci
- 13 vědecko-pedagogických pracovníků (z toho 2 na rodičovské dovolené)
- 7 technicko-administrativních pracovníků (z toho 1 na rodičovské dovolené)

Z těchto pracovníků je 7 zaměstnáno v částečném pracovním poměru (většinou jsou to vědecko-pedagogičtí pracovníci). V tomto akademickém roce by měl být kolektiv technicko-administrativních pracovníků rozšířen ještě o další dva půlúvazky. Tři vědecko-pedagogičtí pracovníci jsou v důchodovém věku.

Ačkoli je nízký věkový průměr celého týmu nesporným pozitivem pro budování mladého a dynamického oboru, jako kritický se do budoucna ukazuje nedostatek pracovníků s vyššími

akademickými tituly, což je podmínka nutná zejména pro udržení akreditace doktorského studia. Obor v tomto smyslu dosud zaštiťují dva profesoři (oba ovšem v důchodovém věku) a jeden docent (z ÚTKL), ostatní vědecko-pedagogičtí pracovníci mají titul Ph.D./CSc. nebo jsou v závěrečné fázi doktorského studia; v roce 2012 bylo zahájeno jedno habilitační řízení.

Pro zachování akademického rázu celého pracoviště a zejména akreditace oboru doktorského studia (jejíž platnost vyprší na konci roku 2017) je třeba klást důraz na zvyšování kvalifikace všech vědecko-pedagogických pracovníků, předpokladem čehož je činnost vědecká i pedagogická. Nutnost profesního postupu se týká všech akademických pracovníků – příslušnost k sekci by neměla být omezujícím faktorem kariérního růstu; je třeba vytvořit takové podmínky, aby každý měl možnost profesně postupovat (při současném splnění všech úkolů, které jsou danou sekcí zajišťovány). Toho je možné dosáhnout kombinací několika koncepčních změn:

1. větším zapojením doktorandů do činností souvisejících s budováním korpusů a do servisní činnosti v rámci infrastruktury obecně; pro ÚČNK je takový postup výhodnější než krátkodobé najímání externistů (zvlášť při řešení úkolů vyžadujících delší zaškolení), pro studenty je přímá účast na budování korpusů nejen vhodnou odbornou praxí, ale zároveň i možným dodatečným zdrojem příjmu, který je možné částečně krýt ze stipendijního fondu.
2. snížením nároků na objem dat sesbíraných v každém roce (samozřejmě při dodržení kvót, k nimž se projekt zavázal v rámci grantů); v minulosti byla preferována kvantita (nad plán), což bylo pochopitelné vzhledem k celkovému nedostatku dat, dnes je ovšem v tomto smyslu situace stabilizovaná, bylo by proto krátkozraké věnovat růstu korpusů veškeré úsilí na úkor kariérního postupu pracovníků.
3. obnovením studijních dnů, v rámci nichž má každý pracovník možnost věnovat část svého úvazku badatelské činnosti; podmínkou úspěšného znovuzavedení tohoto opatření je pravidelná kontrola vykonaných vědeckých pokroků ze strany vedoucího příslušné sekce.
4. kontinuálně je třeba vyvíjet tlak na uznávání vytvořených korpusů jako výsledků vědecké činnosti (např. pro RIV apod.); hlavní koordinátoři přípravy korpusů by měli být u těchto vědeckých výstupů jmenovitě uvedeni (a citováni) jako vedoucí autoři.

Z dlouhodobého hlediska je třeba také postupovat koncepčněji při přijímání nových pracovníků do všech sekcí. Vhodné spolupracovníky není obtížné sehnat jenom v technické oblasti, kde ÚČNK čelí silné konkurenci na trhu práce IT specialistů. Fluktuace zaměstnanců přitom není výhodná v žádném oboru, při budování kontinuálního projektu však může být velmi komplikujícím faktorem.

Personální politika ÚČNK by proto obecně měla víc využívat toho, že v rámci doktorského programu jsou vychováváni fundovaní a talentovaní odborníci. Pokud to tedy situace dovolí, měli by být při přijímání preferováni kandidáti, kteří mají potenciál a úmysl zapojit se do korpusového výzkumu. Právě realizace oboru korpusová lingvistika – který je svojí povahou zajímavý jak pro lingvisty, tak pro matematiky nebo informatiky – by se měla stát zdrojem potenciálních nových spolupracovníků (mnozí ze současných zaměstnanců jsou absolventi oboru).

Rozvoj pedagogické činnosti

Na ÚČNK se dnes realizuje svojí větší částí doktorský studijní program matematická lingvistika. Ten je zajišťován především pedagogickými pracovníky ÚČNK a ÚTKL. V současnosti program studuje 16 aktivních doktorandů, z toho 4 v prezenčním studiu.

ÚČNK také nabízí úvodní semináře korpusové lingvistiky hlavním filologickým oborům vyučovaným na FF UK (bohemistika, anglistika, germanistika, obecná lingvistika, překladatelství). Je třeba nabídku kurzů dále rozšiřovat (např. obnovením spolupráce s Pedagogickou fakultou), protože při absenci samostatného bakalářského a magisterského studia plní tyto semináře důležitou propagační roli. Vedle toho se od roku 2012/2013 začínají realizovat specializační semináře pro pregraduální studenty filologických oborů (statistika, programování, mluvené korpusy, korpusová lexikologie), na jejichž výuce se můžou podílet pracovníci z různých sekcí.

Je třeba dále podporovat současnou iniciativu propagace korpusových metod mezi vysokoškolskými a středoškolskými pedagogy (např. formou školení, víkendových workshopů nebo

letní školy). Nejen, že tato činnost zvyšuje celostátní prestiž pracoviště, je zároveň jednou z klíčových úloh, kterou by servisní pracoviště (ve smyslu infrastruktury vědy) mělo plnit.

Pro další personální rozvoj pracoviště (výchova nových spolupracovníků i získávání pedagogické praxe nutné pro zahájení habilitačního a profesorského řízení) i pro disciplínu samotnou je proto zásadní upřít snahu ke zkvalitnění a přeměně doktorského studia v obor otevřený a atraktivní. Zásadní je v této souvislosti zvýšení požadavků na odbornou přípravu absolventů a dále větší kompetitivnost oboru. Odborná a servisní spolupráce s MFF UK a ČVUT dává předpoklady přilákat ke studiu matematické lingvistiky nejen absolventy FF UK, ale i zájemce z řad techničtějších oborů, kteří můžou být pro pracoviště velkým přínosem. Obor musí samozřejmě zůstat atraktivní i pro hlavní skupinu uchazečů, kterými jsou absolventi filologických oborů. To souvisí především s odbornou úrovní ústavu, který může – pouze pokud realizuje zajímavé vědecké projekty – přilákat kvalitní uchazeče o studium.

Reorganizace (a pozdější reakreditace) doktorského studia by měla zahrnovat zejména následující změny:

- větší zapojení doktorských studentů do úkolů řešených na ÚČNK (ať už v oblasti technické, výzkumné nebo budování korpusů, viz výše)
- změna osnov hlavního doktorského semináře – je třeba dbát jak na teoretické proškolení studentů, tak na jejich praktické schopnosti zpracovávat textová data
- klást větší důraz na samostatné práce jako formu atestace (namísto ústní zkoušky)

Studium přitom musí být podřízeno profilu absolventa. Ten by měl být mj. schopen:

- detailně ovládat nástroje užívané pro analýzu korpusů, vymýšlet a aplikovat nové způsoby vytěžování dat
- analyzovat jazykový problém a mít přehled o jazykovědě a možnostech jejího matematického (a speciálně počítačového) zpracování
- detailně popsat vnitřní uspořádání a strukturu jazykových dat v korpusech
- vyznat se v základních problémových okruzích budování jazykových korpusů

Absolvent by měl nacházet uplatnění v jakémkoli oboru, který zpracovává a analyzuje jazyková data (především kvantitativně), a to jak po stránce lingvistické, tak komputační/statistické.

Rozvoj vědecké činnosti

Stěžejním vědeckým úkolem ÚČNK, který vyžaduje neustálé expertní a odborné vedení, je budování korpusů; mělo by proto být jako vědecká činnost uznáváno a vykazováno (viz výše). Situaci v oblasti rozvoje datové základny je možné považovat za stabilizovanou, ačkoli je třeba neustále zefektivňovat postupy získávání dat. Podporovat je třeba i inovativní metody jejich zpracovávání (viz např. započaté práce na lemmatizaci diachronního korpusu či víceúrovňový přepis mluvených dat apod.).

Kvalitní badatelská činnost je základním předpokladem personálního i vědeckého rozvoje celého projektu. Pouze pracoviště s inovativním výzkumem může přilákat kvalitní zájemce o postgraduální studium (a tedy i potenciální spolupracovníky), umožňovat profesní růst, získávat externí granty, na nichž je ÚČNK od svého založení existenčně závislý, a neustále zlepšovat možnosti popisu jazykové materie na základě nových metod a aplikací pro práci s korpusy. Za hlavní a dlouhodobý výzkumný úkol je tak možné označit zkoumání možností výstavby korpusů a jejich vytěžování. S tím souvisí minimálně dva okruhy témat:

1. Teoretická příprava a následná realizace nově pojaté reprezentativnosti zejm. psaných korpusů, která musí zohledňovat nové textové typy objevující se v češtině (internetová komunikace) i změny ve vymezení synchronie.
2. Návrh, realizace a testování nových nástrojů pro vytěžování korpusů. Nezastupitelnou roli zde hraje kontinuální badatelská činnost, která se bude systematicky věnovat korpusovému pohledu i na ty oblasti jazykovědy, které dosud nebyly korpusově zkoumány (hlavní pozornost byla v současnosti věnována lexikologii a gramatice). Nové nástroje můžou vzniknout pouze na základě inovativního bádání a objevování nových metod vytěžování;

popularizace těchto aplikací je smysluplně možná pouze realizací a publikováním vlastního (třeba parciálního) výzkumu s jejich pomocí.

Vedle toho je třeba umožnit pracovišti další rozvoj v oblastech, které byly tradičně korpusovými metodami pokrývané. Vedle řady menších a individuálních výzkumů jde především o započetí prací na koncepci kolokačního slovníku a výzkum gramatických kategorií. Obě oblasti zájmu je třeba za stávající situace (alespoň částečně) financovat z jiných zdrojů, než je infrastruktura, obě v budoucnu předpokládají získání externích finančních prostředků, což by zaručilo větší zajištění celému projektu. Jejich přínos pro další rozvoj oboru je přitom zjevný – realizace kolokačního slovníku bude vyžadovat vytvoření adekvátního korpusového nástroje (dosud používaná aplikace Word Sketch Engine se jeví jako postačující pouze z části), výzkum gramatických kategorií slibuje přinést nový pohled do značkování korpusů (lemmatizace a tagování).

Z větších projektů, které se již začaly realizovat a jejichž přínos je pro projekt ČNK i obor rovněž zásadní, je možné jmenovat výzkum specifik překladového jazyka (jehož výsledky najdou svoje uplatnění jak v otázce reprezentativnosti, tak případně i v další strategii budování paralelního korpusu InterCorp) a analýzu textů pomocí klíčových slov s přesahem do corpus-assisted discourse studies, což je oblast, která byla v české lingvistice (na rozdíl od vývoje ve světě) poněkud zanedbávána. Postupným zvětšováním časového záběru synchronních korpusů se před korpusovou lingvistikou otvírá i velké pole možností výzkumu a zpracování diachronních témat mapující nedávnou jazykovou historii (modern diachrony).

Výhled dalších oblastí rozvoje ÚČNK

Ústav ČNK, jako pracoviště svým složením i zaměřením na FF UK specifické, má na starosti – vedle rozvíjení oboru a jeho metodologie – shromažďování a poskytování jazykových dat (vlastních i hostovaných) pro lingvistický výzkum. V době vzniku ÚČNK byl nedostatek rozsáhlých korpusů hlavní překážkou empirického jazykového výzkumu. Od té doby se situace nejen v češtině změnila; ačkoli – a to je třeba obzvlášť zdůraznit – potřeba kontinuálně mapovat jazykový vývoj češtiny ve všech jeho aspektech stále trvá, hlad po velkých korpusech (i v souvislosti s rozvojem korpusů webových) byl částečně uspokojen, i když pouze v oblasti současného a psaného jazyka. Je třeba na nastalou situaci reagovat, a to především změnou preferencí: od prosazované kvantity dat zejména ke kvalitě jejich zpracování a vytěžování. Není tedy nutné zvyšovat tempo růstu datové základny, rozšířením celého pracoviště je třeba sledovat spíš větší kvalitu poskytovaných dat (při zachování stejného objemu nových textů do korpusů vstupujících) a větší pestrost v možnostech jejich vytěžování.

Soudím proto, že je třeba zahájit práce na dlouhodobé koncepci rozvoje ČNK, která by měla obsahovat cíle, jejichž naplnění je daleko za horizontem jednoho funkčního období. Příprava koncepce by měla probíhat nejprve v rámci vedení a vedoucích sekcí, poté by měla být podrobena několika kolům připomínkování (od všech zaměstnanců a příp. i externích hodnotitelů). Tento dlouhodobý výhled by měl mj. specifikovat stanovisko k následujícím výzvám, které obor i projekt ČNK čekají (některé z nich už byly zmíněny výše):

1. otázky technické podpory – zejm. problematika jednotného rozhraní pro korpusové aplikace, dále pak otázky spojené se serverovou částí korpusového manažeru (Manatee, CWB ad.)
2. otázky anotování korpusů – do jaké míry a v jakých oblastech chce být projekt ČNK v dlouhodobém horizontu závislý na nástrojích vyvíjených mimo ÚČNK a ÚTKL
3. otázky reprezentativnosti korpusů psaného jazyka – stanovení obecného rámce, na jehož základě bude (nebo nebude) reprezentativnost korpusů aktualizována (v reakci na neustálý vývoj jazyka)
4. otázky reprezentativnosti korpusů mluveného jazyka – dosud je mapována (zcela unikátně) neformální a nepřipravená komunikace, je třeba proto rozhodnout, zda a jak datově podchytit oblasti mluveného jazyka, které jsou z hlediska mluvenosti spíše na periférii (např. projevy formálnější, veřejné apod.)
5. otázky zveřejňování a skladby korpusů – s jakou periodou korpusy zveřejňovat a v jaké

podobě (ne/referenční, různé typy reprezentativnosti atp.), jaké specifické (sub)korpusy by česká lingvistická komunita měla mít k dispozici
6. důkladné zvážení přínosů webových korpusů a nákladů na jejich budování
7. celkové zvýšení prestiže pracoviště – např. výhledovou akreditací habilitačního, příp. i profesorského oboru korpusová lingvistika, pravidelnou sebeprezentací v oborových a popularizačních periodikách, jednotným a profesionálním grafickým stylem pracoviště pro prezentaci na veřejnosti apod.
8. mezinárodní vztahy – ÚČNK může stavět na relativně hojných kontaktech ve světě, je však třeba systematicky hledat cesty, jak se dostat do celosvětového povědomí, např. nabídkou nástrojů pro práci s hlavními světovými jazyky, popularizací InterCorpu, reakreditací doktorského oboru v cizím jazyce apod.

Tato dlouhodobá koncepce by pak měla sloužit jako ukazatel směru či seznam desiderát, od něhož by se měla odvíjet další badatelská i technická činnost. Je zjevné, že v průběhu funkčního období nebude možné uskutečnit vše, identifikace dlouhodobého horizontu by ovšem měla zajistit minimalizaci nekoncepčních *ad hoc* řešení, zvláště v situacích, které nevyžadují bezprostřední zásah. Zároveň by měla být návodem pro včasnou identifikaci výzkumných problémů, pro jejichž realizaci je nezbytné usilovat o externí financování.

V Praze, 5. listopadu 2012