Koncepce rozvoje Ústavu Českého národního korpusu FF UK

na období 1. února 2019 – 31. ledna 2022

Tato koncepce vychází v řadě ohledů z koncepce předcházející, kterou jsem jako kandidát na ředitele Ústavu Českého národního korpusu („ÚČNK“) předkládal před třemi lety. Sleduje tedy především zachování kontinuity v období, kdy je situace ústavu stabilní a kdy jsou předpoklady pro to, že se jeho institucionální zakotvení a finanční zajištění v nejbližších několika letech zásadně nezmění. Protože se však pochopitelně proměňuje okolní prostředí, a tedy i výhled do vzdálenější budoucnosti, přizpůsobuje jim tato koncepce konkrétní akcenty, které je na jednotlivé složky rozvoje ústavu potřeba klást.

Hlavním úkolem ÚČNK je bezesporu péče o kontinuální a všestranný rozvoj projektu Český národní korpus („ČNK“). Tento projekt je centrálním bodem aktivit ústavu, přesahuje však jeho rámec tím, že jsou do práce na něm zapojena také další pracoviště, zejména kolegové z Ústavu teoretické a komputační lingvistiky („ÚTKL“); na druhou stranu i aktivity ÚČNK jsou pochopitelně širší a neomezují se pouze na projekt ČNK.

ÚČNK je pracoviště, jehož charakter je ve srovnání s ostatními základními součástmi FF UK netypický, a tato netypičnost se pochopitelně projevuje také v předkládané koncepci. Jde především o vysoký podíl vědeckých a technických pracovníků, kteří zajišťují servisně-technickou složku činnosti ČNK (sběr, zpracování a zveřejňování dat, jejich anotace a uživatelská podpora), a také o způsob financování: existence ÚČNK totiž byla a je v rozhodující míře závislá na externích, mimofakultních zdrojích.

Pro ÚČNK má zásadní význam projekt *Český národní korpus* (angl. akronym „CNC“) v rámci programu Velkých infrastruktur pro výzkum, vývoj a inovace MŠMT (LM2015044, 2016–2019; „infrastruktura“), který zajišťuje především činnost servisně-technickou, a dále projekt *Jazyková variabilita v CNC* (07/2017–06/2020), který ÚČNK získal v rámci komplementární výzvy Výzkumné infrastruktury OP VVV. Tato výzva byla vyhlášena MŠMT jako součást nového vícezdrojového modelu financování Velkých infrastruktur, takže i tento projekt se odvíjí od infrastruktury CNC, přestože jeho náplní je kromě upgrade výpočetních kapacit především výzkumná složka. Oba tyto projekty společně v současné době pokrývají téměř 90 % rozpočtu ústavu, zbytek tvoří Progres, Kreas a fakultní balíček.

V roce 2017 MŠMT zorganizovalo interim hodnocení všech infrastruktur v ČR, v němž infrastruktura CNC získala na škále 0 (nejhorší) až 5 (nejlepší) celkovou známku 4. Na tomto základě by měla být zařazena na aktualizovanou Cestovní mapu ČR Velkých infrastruktur pro výzkum, experimentální vývoj a inovace, a tím splnit nutnou podmínku pro další financování v následujícím období, tj. 2020–2022. Přestože v tuto chvíli ještě není známo, jaké prostředky bude mít infrastruktura CNC po roce 2020 k dispozici, výhled je momentálně optimistický, podložený jednáními, která MŠMT vede a která by měla vyústit ve schválení jeho návrhů rozpočtů všech infrastruktur vládou ČR. Je však třeba konstatovat, že situace je nejistá po roce 2022, zejména s ohledem na předpokládaný výpadek strukturálních fondů EU používaných mj. ke kofinancování výzkumných

infrastruktur v ČR. V tuto chvíli tedy vidím jako hlavní mezník konec roku 2022 a tomu chci přizpůsobit naši činnost v nejbližších letech, kdy bychom bychom měli mít dostatek prostředků na rozvoj infrastruktury v oblastech, které považujeme za důležité, a kdy bychom se zároveň měli připravovat na různé varianty vývoje po roce 2022 tak, abychom na ně byli schopni pružně reagovat.

Personální oblast

Činnost projektu ČNK je rozdělena do šesti sekcí: sekce lingvistické, sekce komputační (technické), sekce mluvených korpusů, sekce diachronních korpusů, sekce paralelních korpusů a sekce lingvistické analýzy a anotace. Mezistupněm mezi jednotlivými pracovníky a vedením ČNK jsou vedoucí sekcí, kteří se na pravidelných schůzkách zabývají provozní agendou, koordinací činnosti sekcí a spolurozhodují také o prioritách dalšího směřování projektu. Řídící mechanismy jsou tedy nastaveny a jsou funkční, žádnou větší změnu neplánuji ani v nejbližší budoucnosti.

Pracovní kolektiv ÚČNK je převážně mladý, což je vzhledem k pracovnímu elánu a chuti inovovat nesporná výhoda, na druhou stranu však chybí starší generace a zejména pracovníci s vyššími akademickými tituly (členy ústavu jsou pouze jeden profesor a jeden docent). Situace v tomto ohledu rozhodně není uspokojivá a považuji ji za největší problém, který ÚČNK v současné době v personální oblasti má. Stejně jako v předchozím období proto budu podporovat kvalifikační růst akademických a vědeckých pracovníků, zejména ve směru k habilitacím (v následujících třech letech je realistické očekávat podání 1–2 habilitací). Na obecné rovině pak budu klást větší důraz na podporu osobního rozvoje a na opatření vedoucí k všestrannosti a flexibilitě, a to mj. jako přípravu na období po roce 2022; přitom však chci zachovat vysokou zainteresovanost na chodu projektu a udržet dobré mezilidské vztahy. Konkrétně půjde o následující kroky:

- podporu získávání a prohlubování zahraničních kontaktů akademických a vědeckých pracovníků s cílem snadnějšího zapojování do mezinárodních projektů;
- pokračování v nastoupeném trendu střídání ve vedení sekcí ČNK, které považuji za přirozené a prospěšné;
- v případě konkrétních úkolů vyžadujících zapojení více sekcí budu častěji využívat samostatné pracovní skupiny, které se podle mého názoru osvědčily: koordinátoři pracovních skupin tak získají zkušenosti s vedením menších týmů, zvýší se povědomí o celku i flexibilita všech zúčastněných.

Pedagogická činnost

Netypičnost postavení ÚČNK na fakultě se projevuje mj. v tom, že ÚČNK nemá vlastní studenty pregraduálního studia. Naší snahou proto je a nadále bude předávat znalosti pro práci s jazykovými korpusy především prostřednictvím volitelných seminářů pro studenty filologických oborů FF UK. Tato praxe se v posledních letech stabilizovala, nabídka i obsah seminářů se v souladu s možnostmi pracoviště a s obecnými trendy v oblasti korpusové

lingvistiky průběžně obměňují. Kromě toho bude ÚČNK od akademického roku 2019/2020 nabízet kurz „Úvod do práce s jazykovými korpusy“ jako součást společného základu, což by mělo zvýšit povědomí studentů FF o korpusech a prohloubit jejich dovednosti při práci s nimi. Z kapacitních důvodů bude předmět vyučován převážně formou e-kurzu, který v současné době připravujeme.

Co se týče doktorského studia, nedávno byla Radou pro vnitřní hodnocení UK schválena akreditace studijního programu „Korpusová a teoretická lingvistika“, který bude ÚČNK garantovat spolu s ÚTKL a který nahradí stávající „Matematickou lingvistiku“. Kromě obměny složení oborové rady je akreditace spojena s obsahovými změnami, zejména s důrazem na praktickou práci s jazykovými daty a na systematickou přípravu disertace.

Vědecká činnost

Vědeckou činností se v ÚČNK zabývají více či méně všechny sekce, hlavní díl práce a zodpovědnosti však leží na sekci lingvistické. V oblasti infrastrukturní vědecké činnosti, tedy takové, která vede k rozvoji a k vylepšování služeb infrastruktury ČNK, budu i nadále dbát na sledování vývoje v oblasti technických standardů a nástrojů tak, aby zpracování dat probíhalo co nejefektivněji. Všechny nástroje a jazykové zdroje vytvořené v rámci ČNK, které mají potenciál pro využití i mimo něj, by měly být koncipovány obecně tak, aby mohly být zveřejněny a používány co nejširší komunitou; to může přinést projektu mj. nové mezinárodní spolupráce.

Neméně důležitou oblastí je vědecká činnost neinfrastrukturní, která služeb infrastruktury ČNK pouze využívá, a musí být proto hrazena z jiných zdrojů (typicky Progres, Kreas nebo individuální granty). Tady je v současné době klíčový výzkum variability v jazyce, který je náplní projektu OP VVV; do nejbližší budoucnosti byly jako prioritní oblasti vybrány také kontrastivní korpusová lingvistika a kvantitativní metody. Ve všech těchto oblastech budu preferovat systematickou týmovou práci na dlouhodobějších projektech, které budou ÚČNK vědecky profilovat a které by v ideálním případě mohly poskytnout zpětnou vazbu pro rozvoj infrastruktury ČNK a/nebo pomoci výzkumnou činnost přetavit do dalšího veřejně dostupného nástroje či uživatelského rozhraní.

Kvalitní vědecká činnost v obou těchto oblastech je důležitá pro osobní rozvoj každého pracovníka i pro hodnocení pracoviště a je nezbytná také pro získávání dalších zdrojů financování, což může být v budoucnu stále důležitější. Ačkoli by se mohlo zdát, že ÚČNK není díky infrastruktuře na hodnocení vědy tolik závislý, jde pouze o přechodný stav, který je ovšem potřeba využít. Pro ústav, projekt i osobní rozvoj akademických a vědeckých pracovníků bude nezbytné stále více publikovat v renomovaných zahraničních časopisech. Chci proto k takovým publikacím motivovat všechny spolupracovníky, ať už se věnují vědecké činnosti směřující ke zlepšování infrastruktury ČNK nebo jde o výzkum neinfrastrukturní.

Další oblasti rozvoje a výhled do budoucna

Vzhledem k tomu, že rozhodující část rozpočtu ÚČNK pochází v současné době z infrastruktury, jsou priority činnosti ústavu v dalších letech do značné míry dány právě tímto projektem. Pro výhled do budoucna proto z návrhu projektu vybírám nejpodstatnější body a záměrně vynechávám běžnou agendu:

- rozšiřování sběru dat o formálnější podoby mluveného jazyka;
- dokončení prací na korpusu NET (specifický internetový jazyk);
- práce na monitorovacím korpusu, který by měl v dlouhodobější perspektivě pokrýt psaný jazyk od 2. pol. 19. století do současnosti a propojit tak diachronní korpusy se synchronními;
- rozvoj jednotné morfologické anotace pro standardní i nestandardní češtinu různého druhu (čeština 2. pol. 19. stol., neformální mluvený jazyk, nářečí, sociální sítě atd.), jejich slovníkové pokrytí a vytvoření zlatých standardů pro automatickou disambiguaci;
- zvyšování úspěšnosti syntaktické anotace;
- doplnění mezijazykově srovnatelné anotace podle Universal Dependencies (morfologie i syntax) do paralelního korpusu InterCorp;
- další rozvoj rozhraní KonText jako „vlajkové lodi“ nástrojů ČNK a podpora jeho nasazení jinými projekty;
- podrobnější rozpracování a implementace „word at a glance“, portálové stránky s přehledem základních charakteristik hledaného slova.

Lze myslím konstatovat, že ČNK je projektem, který se již na národní úrovni etabloval jako zdroj dat a nástrojů pro empirický výzkum češtiny. Do budoucna nás však stále čeká práce na jeho větším mezinárodním zviditelnění. V tomto kontextu bych chtěl zdůraznit řadu kroků, které již byly učiněny (viz „Zpráva o plnění koncepce rozvoje ÚČNK“) a jejichž přínos se projeví až za nějaký čas. Zejména chci nadále pokračovat v úzké spolupráci s infrastrukturou LINDAT/CLARIN (od příštího roku LINDAT/CLARIAH-CZ), v další integraci do ESFRI CLARIN, v zapojování do projektů z oblasti digital humanities a v neposlední řadě také v prohlubování spolupráce s projekty tvořícími národní korpusy jednotlivých jazyků. Přestože je ČNK projektem především národním, jsem přesvědčen, že máme co nabídnout i na mezinárodní úrovni.

V Praze dne 6. prosince 2018

Mgr. Michal Křen, Ph.D.