Přihláška programu Progres

1) Kód programu (nevyplňujte – vyplní rektorát UK):

2) Název programu v českém jazyce:

Český národní korpus

Název programu v anglickém jazyce:

Czech National Corpus

3) Vědní oblast v českém jazyce:

AI - jazykověda

Vědní oblast v anglickém jazyce:

AI - linguistics

4) Výčet fakult a vysokoškolských ústavů UK, na kterých má být program uskutečňován:

FF UK

5) Stručná anotace programu v českém jazyce:

Program Český národní korpus navazuje na stejnojmenný program PRVOUK P11, který se zaměřoval na rozvoj a popularizaci korpusového přístupu při zkoumání jazyka a komunikace. Český národní korpus (ČNK) funguje od roku 2012 jako národní výzkumná infrastruktura financovaná MŠMT (LM2015044), jejíž náplní je sběr, zpracování a zpřístupňování rozsáhlých jazykových dat pro jazykově orientovaný empirický výzkum v oblasti společenských a humanitních věd. Podpora z programu Progres by měla sloužit jako komplementární zdroj prostředků pro projekt ČNK, které umožní vedle infrastrukturní činnosti financovat i stěžejní aktivity vědeckého charakteru, které nelze hradit z dotace pro výzkumnou infrastrukturu, ačkoli jsou pro její rozvoj a úspěšnost klíčové.

Základní výzkumnou orientací programu by měla být metodologie využívání korpusu zejm. v oblastech gramatického popisu češtiny, korpusové lexikologie a lexikografie, korpusové analýzy diskurzu, diachronního a kontrastivního výzkumu.

Stručná anotace programu v anglickém jazyce:

The Czech National Corpus (CNC) program is a follow-up to the PRVOUK P11 project of the same name which focused on the development and popularization of corpus-based approaches to research on language and communication. Since 2012, the CNC operates as a national research infrastructure financed by the MEYS (LM2015044). Its primary goal is to collect, process and provide access to a large amount of textual data for language-oriented empirical research in social sciences and humanities. Participation in the Progres scheme should provide the CNC project with a complementary source of funding for core scientific research activities which cannot be financed using the MEYS subsidy, even though they are key to the infrastructure's continued success and development.

The main focus area of the program is the methodology of using corpora, esp. in the fields of grammatical description of Czech, corpus-based lexicology and lexicography, corpus-assisted discourse studies, and diachronic and contrastive corpus related research.

6) Údaje o koordinátorovi

Jméno, příjmení, tituly: Doc. Mgr. Václav Cvrček, Ph.D.	
Fakulta (VŠ ústav) UK: FF UK	
Telefon: 221 619 846	**E-mailová adresa:** vaclav.cvrcek@ff.cuni.cz

Stručný životopis koordinátora

Jméno, příjmení, tituly: Doc. Mgr. Václav Cvrček, Ph.D.

Fakulta, VŠ ústav: FF UK	**E-mailová adresa:** vaclav.cvrcek@ff.cuni.cz

Vědecký profil:

Vzdělání:

11/2013	Doc., Český jazyk, FF UK
2004–2008	Ph.D., Filologie – Matematická lingvistika, FF UK
1999–2004	Mgr., Český jazyk a literatura, Lingvistika a fonetika, FF UK

Zaměstnání a zkušenosti:

od 02/2016	zástupce ředitele ÚČNK FF UK
02/2013–02/2016	ředitel ÚČNK FF UK
od 2013	člen oborové rady pro obor Matematická lingvistika, FF UK
06/2012–02/2013	vedoucí lingvistické sekce ÚČNK FF UK
08/2011–06/2012	Stipendijní pobyt na Brown University (RI, USA)
08/2006–	vědecký pracovník, ÚČNK FF UK

Badatelské zájmy:
korpusová lingvistika, kolokace, jazyková variabilita, analýza diskurzu, kvantitativní metody, morfologie češtiny, jazyková regulace

Publikační činnost za posledních 10 let:

Shrnutí:
Autor nebo spoluautor 9 monografií a slovníků (z toho v 6 publikacích jako autor jediný nebo hlavní), 14 statí v recenzovaných časopisech (z toho 3 zahraniční) a dalších 16 studií v domácích i zahraničních sbornících.

5 nejvýznamnějších publikací:

- Cvrček, V. 2008. *Regulace jazyka a Koncept minimální intervence*. Praha: Nakladatelství Lidové noviny.
- Cvrček, V. et al. 2010. *Mluvnice současné češtiny*. Praha: Karolinum.
- Cvrček, V. 2013. *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny.
- Cvrček, V. & Chlumská, L. 2015. Simplification in translated Czech: a new approach to type-token ratio. *Russian linguistics* 39/3. 309–325.
- Fidler, M. & Cvrček, V. 2015. A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis. *Journal of Slavic Linguistics* 23(2). 197–239.

7) Údaje o dalších navrhovaných členech rady programu

Jméno, příjmení, tituly	**Fakulta (VŠ ústav) UK**
Doc. Mgr. Václav Cvrček, Ph.D.	FF UK
Prof. PhDr. Karel Kučera, CSc.	FF UK
Doc. RNDr. Vladimír Petkevič, CSc.	FF UK
Mgr. Michal Křen, Ph.D.	FF UK
Mgr. Anna Čermáková, Ph.D.	FF UK

Stručné životopisy navrhovaných členů rady

Jméno, příjmení, tituly: Prof. PhDr. Karel Kučera, CSc.

Fakulta, VŠ ústav: FF UK | **E-mailová adresa:** karel.kucera@ff.cuni.cz

Vědecký profil:

1966–1971 – studium: Filozofická fakulta UK (čeština – angličtina), 1971 PhDr., 1976 CSc., 1986 docent, 1999 profesor, vše FF UK v Praze (lingvistika – český jazyk)

Zaměstnání a zkušenosti:

1971–1973 asistent, katedra neslovanských jazyků, Vysoká škola zemědělská Praha-Suchdol

1973–1986 odborný asistent, katedra českého jazyka FF UK v Praze

1986–1999 docent, katedra českého jazyka FF UK v Praze

1999–2009 profesor, Ústav českého jazyka a teorie komunikace FF UK v Praze

2009–dosud profesor, Ústav Českého národního korpusu FF UK v Praze

1995–dosud řešitel 2 grantů MŠMT, 2 velkých infrastruktur MŠMT, projektu IMPACT (EU) a projektu *Nástroje pro zpřístupnění tištěných textů 19. století a první poloviny 20. století* (program NAKI, MKČR)

Badatelské zájmy:

korpusová lingvistika, historická jazykověda, vývoj českého jazyka, paremiologie

Působení v zahraničí:

1985 přednáškový pobyt, univerzita Uppsala

1985–1988 lektor, University of Nebraska Omaha, USA; výzkum historie češtiny v USA

1992 přednáškový pobyt, univerzita Uppsala

1993 lektor, univerzita Amsterdam

2009, 2010, 2011 přednáškové pobyty, univerzita Regensburg

Publikační činnost za posledních 10 let:

Shrnutí:

Publikační činnost zaměřena zejména na vývoj češtiny a slovanských jazyků a jeho korpusový výzkum. Autor 2 statí v encyklopedické publikaci *Die slavischen Sprachen / The Slavic Languages* (De Gruyter Mouton), 8 statí v zahraničních recenzovaných sbornících, 3 vědeckých statí v češtině, spoluautor 2 slovníků (*Slovník Karla Čapka* a *Slovník Bohumila Hrabala*).

Nejvýznamnějších publikace:

- The Orthographic Principles in the Slavic Languages: Phonetic/Phonological. In: Kempgen, S. – Kosta, P. – Berger, T. – Gutschmidt, K. (eds.) *Die slavischen Sprachen / The Slavic Languages, Halbband 1,* De Gruyter Mouton, Berlin – New York 2009, s. 70–75.
- History and Development of the Latin Writing System in the Slavic Languages. In: *Die Slavischen Sprachen – The Slavic Languages. An International Handbook of their Structure, their History and their Investigation. Volume 2.* Karl Gutschmidt, Sebastian Kempgen, Tilman Berger, Peter Kosta (eds.). De Gruyter Mouton, Berlin 2014, s. 1514–1525.
- Diachronic Corpora: Seeing Histories of Words from Another Angle. In: Considine, J. (ed.) *Webs of Words. New Studies in Historical Lexicology*. Cambridge Scholars Publishing, Newcastle upon Tyne 2010, s. 1–6. ISBN 1-4438-1952-2 (spoluautor Martin Stluka).
- Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora. In: Levická, J., Garabík, R. (eds.): *Computer Treatment of Slavic and East European Languages*. Tribun, Bratislava 2007, s. 121–125. ISBN 978-80-87139-05-9.

Jméno, příjmení, tituly: **doc. RNDr. Vladimír Petkevič, CSc**

Fakulta, VŠ ústav: FF UK	**E-mailová adresa:** **Vladimir.Petkevic@ff.cuni.cz**

Vědecký profil:

Zaměstnání a zkušenosti:

1979	absolvent MFF UK
1985	RNDr., na MFF UK
1992	CSc., na MFF UK

1994–dosud	ředitel Ústavu teoretické a komputační lingvistiky, FF UK
1996	habilitace v oboru matematická lingvistika, Ústav teoretické a komputační lingvistiky, FF UK

Pedagogická praxe: 1993–dosud

Granty: V. Petkevič řídil 10 grantů z oblasti počítačového zpracování jazyka, v dalších byl buď spoluřešitelem, nebo spolupracovníkem. Podílel se na řešení úkolů dvou mezinárodních projektů jako zástupce České republiky.

Badatelské zájmy:

matematické a počítačové zpracování přirozeného jazyka; korpusová lingvistika; formální lingvistika, zvláště morfologie a syntax; obecná lingvistika; dědictví Pražské školy strukturní lingvistiky

Výběr ze zahraničních konferencí a působení v zahraničí:

1994	stáž v Düsseldorfu, SRN
2014	přednáškový cyklus v Sankt-Petěrburku, Ruská federace

Publikační činnost za posledních 10 let:

Shrnutí:
Publikační činnost zaměřena zejména na problematiku počítačového zpracování morfologie a syntaxe češtiny. Jedna monografie, dvě rozsáhlé překladové a ediční publikace, 23 statí v různých časopisech a knihách.

5 nejvýznamnějších publikací:

- **Petkevič, V**.: *Morfologická homonymie v současné češtině*, Nakladatelství Lidové noviny / Ústav českého národního korpusu, Praha 2014
- Bartoň, T.– Cvrček, V. – Čermák, F. – Jelínek, T., – **Petkevič, V.**: *Statistiky češtiny*. Nakladatelství Lidové noviny / Ústav českého národního korpusu, Praha 2009.
- Havránková, M. – **Petkevič, V.** (eds.): *Pražská škola v korespondenci. Dopisy z let 1924–1989*. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha 2014.
- **Petkevič, V.**: Kontrola české gramatiky (český grammar checker). *Studie z aplikované lingvistiky (SALi)* 2/2014, Filozofická fakulta Univerzity Karlovy v Praze, Praha 2014, s. 48–86.
- **Petkevič, V**.: L'accord en tchèque : le centre et la périphérie. In: Radimský, J. (ed.): *Écho des Études Romanes*, *vol. VI / Num. 1–2*. Université de Bohême du Sud, České Budějovice 2010, s. 143–160.

Jméno, příjmení, tituly: **Mgr. Michal Křen, Ph.D.**

Fakulta, VŠ ústav: FF UK	**E-mailová adresa:** michal.kren@ff.cuni.cz

Vědecký profil:

Vzdělání:

1991–1996	Mgr., Informatika, MFF UK

2005–2012	Ph.D., Filologie – Matematická lingvistika, FF UK

Zaměstnání a zkušenosti:

1999	zahraniční stáž, Institut für Deutsche Sprache, Mannheim
2003–2012	vedoucí odboru komputační techniky, ÚČNK FF UK
2013–2016	zástupce ředitele, ÚČNK FF UK
od 2013	člen oborové rady pro obor Matematická lingvistika na FF UK
od 2016	ředitel, ÚČNK FF UK

Badatelské zájmy:
korpusová lingvistika, složení jazykových korpusů, kolokace, jazyková variabilita a změny v jazyce

Publikační činnost za posledních 10 let:

Shrnutí:
Autor nebo spoluautor 2 monografií/slovníků (z toho 1 zahraniční), 1 kapitoly v zahraniční monografii, 3 statí v recenzovaných časopisech (z toho 1 zahraniční) a dalších 18 článků ve sbornících (z toho 15 zahraničních, 3 v Thomson Reuters CPCI).

5 nejvýznamnějších publikací:

- Křen, M. (et al.): SYN2015: Representative Corpus of Contemporary Written Czech. *Proceedings of LREC2016*, ELRA 2016.
- Křen, M.: Czech conditional conjunctions: a corpus-based study of development tendencies. *Prace Filologiczne* 67, 2015, s. 215–232.
- Benešová, L. – Waclawičová, M. – Křen, M.: Building a Data Repository of Spontaneous Spoken Czech. *Best Practices for Spoken Corpora in Linguistic Research*, Cambridge Scholars Publishing, Newcastle Upon Tyne 2014, s. 128–141.
- Čermák, F. – Křen, M. (et al.): *Frequency Dictionary of Czech: Core Vocabulary for Learners*. Routledge, London 2011.
- Křen, M.: Compilation of the Dictionary of Karel Čapek. *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. Peter Lang, Frankfurt am Main 2008, s. 469–481.

Jméno, příjmení, tituly: **Mgr. Anna Čermáková, Ph.D.**

Fakulta, VŠ ústav: FF UK	**E-mailová adresa:** anna.cermakova@ff.cuni.cz

Vědecký profil:

Zaměstnání a zkušenosti:

2001–2003	Centre for Corpus Linguistics, Department of English, University of Birmingham (research associate)
2003–2008	International Journal of Corpus Linguistics, John Benjamins, Amsterdam (výkonná redaktorka)
2008	Ph.D., Filologie – Matematická lingvistika, FF UK
2011–2013	Ústav Českého národního korpusu, FF UK (akademický pracovník)

2013–dosud Ústav Českého národního korpusu, FF UK (akademický pracovník, vedoucí lingvistické sekce)

Výběr ze zahraničních konferencí:

2016 Keynote speech "Current Trends in Corpus Linguistics", panel expertů organizovaný varšavskou univerzitou, Polsko

Causal Conjunctions in Spoken and Written Czech (s M. Kopřivovou), IVACS, Bath, Velká Británie

2015 *Denoting place in English and Czech: Methodological challenges of corpus-driven contrastive study of typologically different languages* (s L. Chlumskou), Language in Contrast. Diachronic, variationist, and cross-linguistic studies, GReG & LeCSeL, Paříž

Compiling Corpus for School Children to Support L1 Teaching: Case of Czech (s L. Chlumskou), Corpus Linguistics 2015, Lancaster, Velká Británie

Comparing the language of children literature (s L. Chlumskou), ICAME 36, Trier, Německo

Členství v mezinárodních radách:

2015–dosud Advisory Panel projektu *CliC Dickens,* http://clic.bham.ac.uk/

Publikační činnost za posledních 10 let:

5 nejvýznamnějších publikací:

Cvrček, V. – Čermáková, A. – Křen, M.: Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost* 77 (2), 2016, s. 83–101.

Čermáková, A.: Repetition in John Irving's novel "A Widow for One Year": A corpus stylistics approach to literary translation. *International Journal of Corpus Linguistic* 20 (3), 2015, s. 355–377.

Čermáková, A.: Jaký slovník uživatelé češtiny potřebují? O Slovníku současné češtiny nakladatelství Lingea. *Slovo a slovesnost* 74 (3), 2013, s. 195–210.

Čermáková, A.: *Valence českých substantiv*. NLN, Praha 2009.Teubert, W. – Čermáková, A.: *Corpus Linguistics: A Short Introduction.* Continuum, London 2007 (Chinese edition 2009).

8) Orientační údaje o počtech osob zúčastněných na programu

Fakulta (VŠ ústav)	Orientační počet			
	akademických a vědeckých pracovníků	studentů doktorských studijních programů	studentů magisterských studijních programů	studentů bakalářských studijních programů
FF UK	20	7	0	0

9) Popis programu, včetně návaznosti na dosavadní vědecké výsledky a programy; kritické zhodnocení postavení vědní oblasti na UK v národním a zejména mezinárodním kontextu; návrh klíčových kroků pro zlepšení tohoto postavení v horizontu doby uskutečňování programu, zdůvodnění a rámcový harmonogram těchto kroků, indikátory tohoto zlepšení

Kritické zhodnocení vědní oblasti na UK

Využití jazykových korpusů, tj. rozsáhlých elektronických sbírek autentických textů, představuje pro empirickou jazykovědu jeden z nejvlivnějších metodologických impulzů posledních dekád. Jako svébytné odvětví jazykovědy si korpusová lingvistika za dobu od svého vzniku v 60. letech vymezila specifické výzkumné okruhy (zejm. otázky související s frekvencí jevů, lexikální syntagmatikou, výzkumem kontextu apod.), vyvinula inovativní metody práce (využívající komputačních a statistických nástrojů) i vlastní terminologii. Vedle toho se ovšem korpusový přístup k popisu jazyka stále častěji stává nezbytností při řešení tradičních lingvistických úkolů (popis gramatiky a lexikonu) a nachází uplatnění i v řadě dalších jazykovědných disciplín (např. sociolingvistika, analýza diskurzu, kognitivní lingvistika, aplikovaná lingvistika), ba proniká i do zcela jiných oborů (literární věda, historie, psychologie, sociologie). Schopnost zacházet s jazykovými korpusy se tak v současnosti stává nezbytnou součástí profesní výbavy každého empirického jazykovědce.

Korpusová lingvistika, na jejíž rozvoj se předkládaný program Progres zaměřuje, nemá v českém prostředí příliš dlouhou tradici. Důvodem je především fakt, že pro systematické uplatňování kvantitativních empirických postupů v jazykovědné praxi, které korpusová lingvistika akcentuje, je nezbytná rozsáhlá a nákladná infrastruktura. Ta začala pro češtinu vznikat až v druhé polovině 90. let, kdy už v anglické či německé jazykovědě první rozsáhlé korpusy existovaly.

Počátky korpusové lingvistiky na UK (a tedy i v ČR) souvisejí se založením projektu Český národní korpus (ČNK) v roce 1994, který se od té doby systematicky věnuje kontinuálnímu a všestrannému datovému mapování češtiny (a dalších jazyků pro kontrastivní účely). Jeho primárním cílem je vytvářet rozsáhlé a kvalitně zpracované elektronické databáze textů – jazykové korpusy –, které slouží zejména lingvistům pro empirický popis jazyka. První reprezentativní stomilionový korpus současné češtiny, dostupný on-line, byl zveřejněn v roce 2000.

Od té doby lze sledovat pozvolný rozvoj uplatnění korpusových metod v jazykovědě – na základě korpusů vznikají gramatiky (první z nich – V. Cvrček a kol.: *Mluvnice současné češtiny*. Karolinum. 2010 – dokonce na půdě ČNK), slovníky (např. Čermák, F. – Křen, M. (eds.): *Frekvenční slovník češtiny*. Nakladatelství Lidové noviny, Praha 2004), lingvistické studie a další jazykové příručky; zvýšenou měrou se korpusový přístup začíná uplatňovat v počítačovém zpracování češtiny (morfologické a syntaktické analyzátory), ale i v jiných oborech společenských a humanitních věd (frekvenční údaje slouží jako podklad pro psychologické a neurologické testy zkoumající mentální procesy související s jazykem, korpusový přístup k literárním textům dává nové impulzy literární vědě atp.).

V oblasti množství dat a kvality jejich zpracování patří ČNK ke světové špičce; objem synchronních korpusů psané češtiny (tištěné texty, tj. bez webového obsahu) letos přesáhla hranici 4 miliard slov, což vztaženo k počtu mluvčích češtiny představuje patrně vůbec nejrozsáhlejší datovou základnu na světě. Tyto zdroje využívá přes 4800 registrovaných uživatelů, kteří v průměru položí 1500 dotazů denně (údaje z roku 2015).

Jazykové korpusy budované v rámci ČNK lze rozdělit do 4 hlavních skupin (z důvodů přehlednosti ponecháváme stranou korpusy specializované a korpusy hostované, tj. vytvořené třetí stranou a ve spolupráci s ČNK pouze finalizované a zveřejňované):

1. Korpusy synchronní psané češtiny (řada SYN) – v současnosti obsahuje přes 4 mld. textových slov a zahrnuje široké spektrum textových typů (od beletrie přes publicistiku až po texty oborové a odborné); vedle objemu textu je při tvorbě korpusů v této řadě kladen důraz také na jejich reprezentativní složení (stomilionové korpusy SYN2000, SYN2005, SYN2010 a SYN2015).
2. Korpusy mluvené češtiny (řada ORAL) a korpus nářeční (DIALEKT) – v oblasti mluveného jazyka se ČNK soustředí na vytváření datové základny mapující především prototypicky mluvený jazyk (neformální nepřipravené dialogické promluvy); vzhledem k náročnosti sběru a zpracování (nahrávky, transkripce, systém kontrol) má řada ORAL podstatně menší objem (cca 5 mil. textových slov), představuje nicméně velmi cenný výzkumný materiál, který dosud v takovémto objemu a pestrosti nebyl dostupný.
3. Korpusy diachronní (řada DIA) – korpusy starší češtiny si kladou za cíl poskytovat datovou základnu pro výzkum zaměřující se na vývoj češtiny od 14. století po rok 1945; aktuální objem zveřejněných textů v této řadě je 3,4 mil. textových slov; v současnosti se hlavní pozornost soustředí na důkladné datové pokrývání jazyka 2. pol. 19. století tak, aby bylo možné po propojení s řadou SYN vytvářet v dlouhodobém horizontu tzv. monitorovací korpus mapující období od roku 1850 až do současnosti po úsecích, v nichž budou mít data srovnatelné žánrové složení.
4. Korpusy paralelní (InterCorp) – vícejazyčný korpus InterCorp, který vzniká za přispění mnoha spolupracovníků (zejm. z FF UK), zahrnuje české texty spolu s jejich překlady z/do 39 jazyků s aktuálním objemem 187 mil. textových slov v české části korpusu (celkový objem korpusu, tj. po započítání cizojazyčných částí, činí 1,6 mld. slov); paralelní korpus je nedocenitelnou pomůckou při kontrastivním a translatologickém studiu jazyků a jako takový má významný potenciál přilákat k projektu ČNK zahraniční badatele.

Rozsáhlá textová data shromážděná v korpusech nelze využívat bez specializovaných nástrojů. Jejich vývoj ovšem předpokládá nejen dovednosti programátorské a detailní znalost odborných potřeb uživatelů, ale zejména výzkum v oblasti metodologie korpusového bádání. Rovněž na tuto oblast se ČNK dlouhodobě soustředí, výsledkem čehož je řada aplikací, s jejichž pomocí jsou korpusová data v přehledné podobě zpřístupňována uživatelům a které umožňují inovativní koncový výzkum (KonText, SyD, Morfio, KWords, Treq – vše viz www.korpus.cz).

K rozvoji oboru korpusová lingvistika přispívá ČNK rovněž samostatnou vědeckou činností: publikováním monografií, článků v recenzovaných časopisech a sbornících či příspěvky na domácích i zahraničních konferencích. Vedle textů metodologického zaměření se jedná zejména o případové studie, které inspirují domácí i zahraniční badatele k využití nejnovějších postupů při práci s korpusy ČNK. Klíčovou roli v této oblasti hraje ediční řada *Studie z korpusové lingvistiky*, kterou ČNK vydává ve spolupráci s Nakladatelstvím Lidové noviny od roku 2006 a v níž už vyšlo 22 svazků.

Pravidelná setkávání s výzkumnou komunitou jsou zajišťována prostřednictvím cyklu konferencí a kolokvií *Korpusová lingvistika Praha* (v roce 2016 se uskutečnil 5. běh) a dále pak pravidelně organizovanými workshopy pro zájemce o práci s korpusem z řad odborné veřejnosti, studentů, učitelů a jazykových profesionálů (překladatelé, redaktoři apod.).

Ačkoli tedy lze shrnout, že korpusová lingvistika na UK (a potažmo v celé ČR) se za posledních 20 let zvláště díky rozvoji infrastruktury jednoznačně etablovala, je třeba zároveň konstatovat, že metodologická úroveň některých vědeckých publikací stále zaostává za světovou produkcí.

Důvodem je zaprvé to, že postupy korpusové lingvistky operující nad velkými objemy textů vyžadují po uživatelích nezanedbatelnou průpravu v oblasti počítačového a statistického zpracování dat, tedy v oborech, které se mnohdy nacházejí za obzorem vědeckého zájmu badatelů v oblasti společenských a humanitních věd a které teprve postupně pronikají do kurikulárních materiálů těchto oborů. Druhým aspektem, jenž znemožňuje dosažení excelentních výsledků srovnatelných se světovou produkcí, je nedostatečná obeznámenost s principy korpusové lingvistiky a s širokými možnostmi využití korpusů pro inovativní výzkum. Obě tato desiderata současné empirické jazykovědy hodlá navrhovaný program Progres řešit.

Obecný popis projektu

Od svého založení byl projekt ČNK koncipován nejen jako infrastruktura pro badatele, ale také jako vědecké centrum. Infrastrukturní složka jeho činnosti je od roku 2012 jako národní výzkumná infrastruktura (projekt LM2011023) financovaná MŠMT. Tento status si ČNK udržel i v následné evaluaci, která proběhla v roce 2014 za účasti mezinárodních hodnotitelů a v níž uspěl s nejlepším možným hodnocením, což vyústilo v jeho další podporu na období 2016–2019 s výhledem do roku 2022 (LM2015044). Podle dohody s poskytovatelem dotace se hostitelská instituce na chodu infrastruktury podílí jednak přímo formou finanční spoluúčasti, jednak nepřímo formou podpory výzkumných aktivit na infrastruktuře prováděných.

Navrhovaný projekt v rámci programu Progres navazuje na PRVOUK P11 Český národní korpus, který sloužil právě jako komplementární podpora pro aktivity vědeckého a pedagogického charakteru, jež nelze hradit z dotace pro výzkumnou infrastrukturu, ačkoli jsou tyto aktivity poskytovatelem dotace předpokládané a pro rozvoj infrastruktury a její celkovou úspěšnost klíčové. Zatímco podpora ze strany MŠMT se zaměřuje především na budování vědecké infrastruktury, která slouží výzkumné komunitě v ČR i v zahraničí, a nepodporuje vlastní výzkum na infrastruktuře prováděný, program PRVOUK umožňoval rozvíjet vědecký potenciál zúčastněných pracovišť (ÚČNK FF UK a ÚTKL FF UK) a potažmo celého oboru korpusová lingvistika (ČNK je v současnosti jediné pracoviště v ČR, které věnuje systematickou pozornost rozvoji tohoto oboru a zejména jeho metodologie). Vedle odborného růstu pracovníků zapojených do projektu PRVOUK bylo nezanedbatelným přínosem jeho výstupů také zvýšení prestiže hostitelské instituce na poli současné korpusové lingvistiky (zvané přednášky významných zahraničních hostů z oboru), propagace korpusových metod mezi českými i zahraničními badateli (např. formou konferenčních příspěvků a case-studies) a rozvoj pedagogické složky projektu (ČNK dnes zajišťuje vedle vlastního doktorského programu se specializací na korpusovou lingvistiku také výuku řady seminářů zaměřených na práci s korpusy pro mnoho oborů pěstovaných na UK).

Navrhovaný projekt v rámci programu Progres by měl navazovat na zmíněné aktivity PRVOUK P11 a dál je rozvíjet s akcentem na výzkum v oblasti metodologie oboru. Je nesporné, že vzhledem k množství vědeckých výstupů, které jsou spjaty se zdroji ČNK (viz volně přístupná databáze http://biblio.korpus.cz, která eviduje přes 1000 výsledků vytvořených s pomocí dat a nástrojů ČNK v posledních 5 letech), je pro další rozvoj této oblasti výzkumu potřebná kvalitní metodologická podpora. ČNK se chce proto v rámci programu Progres věnovat rozvoji metodologie těchto oblastí korpusového výzkumu (podrobnější popis s konkrétními výzkumnými tématy spadajícími pod jednotlivé linie výzkumu viz níže):

- **korpusová lexikologie a lexikografie** – otázky spjaté zejména s lexikální syntagmatikou, kolokacemi a frazeologií, což jsou disciplíny, jimž se v ČNK věnuje pozornost dlouhodobě

- **gramatický popis češtiny** – zvláště s ohledem na automatické metody morfologického a syntaktického značkování korpusů rozvíjené v rámci ČNK
- na korpusu založená **analýza diskurzu** – otázky související jak s designem výzkumných dat, tak s jejich statistickým zpracováním a lingvistickou interpretací: extrakce klíčových slov, identifikace prominentních jednotek různé úrovně apod.; v této oblasti rozvíjí ČNK dlouhodobou spolupráci s Brown University (USA)
- **diachronní výzkum** – zejména v souvislosti s metodologií vytěžování připravovaného monitorovacího korpusu a s přesahem do oblasti historie
- **kontrastivní a translatologický výzkum**, které jsou realizovány obzvláště na základě paralelního vícejazyčného korpusu InterCorp (jde především o otázky korespondence, překladové univerzálie, design výzkumných dat zajišťující srovnatelnost výsledků, kontrastivní studium stylu a specifických typů textů)
- **aplikovaná lingvistika** – využití korpusů pro výzkum akvizice jazyka a ve výuce češtiny (pro rodilé i nerodilé mluvčí)
- **metodologie korpusové lingvistiky** – obecně metodologická témata spjatá s vytvářením korpusů a jejich využitím při popisu jazyka; inovativní postupy při vytěžování korpusových dat

Předpokládanými výstupy vědeckých aktivit v naznačených tematických okruzích jsou především odborné publikace a konferenční příspěvky, které jednak umožňují efektivní propagaci zdrojů ČNK a korpusového přístupu obecně a jednak (např. pomocí případových studií) ukazují další možnosti inovativního využití empirických dat, čímž inspirují badatele k jejich implementaci a dalšímu rozvíjení. Publikační činnost se orientuje na prestižní zahraniční i domácí časopisy; využívá také vlastní, dnes už etablované ediční řady *Studie z korpusové lingvistiky*, kterou ČNK vydává ve spolupráci s Nakladatelstvím Lidové noviny. Efektivní způsob zprostředkování metodologických zásad práce s korpusy a publikační výstup sui generis představuje také internetová *Příručka uživatele ČNK* (viz http://wiki.korpus.cz), která byla založena na sklonku roku 2013 a slouží jako dokumentace projektu ČNK (infrastrukturní část), ale rovněž jako neustále doplňovaná encyklopedie korpusové lingvistiky, on-line kurz a manuál (v lednu 2016 dosahovala průměrná návštěvnost 100 přístupů denně, celkový rozsah příručky odpovídá zhruba 300 stranám textu).

Dalším typem aktivit rozvíjejících obor jsou konference, kolokvia a workshopy, které ČNK organizoval díky podpoře v rámci PRVOUK a na něž by rád navázal. V dvouletém intervalu pořádá odborná setkání s názvem *Korpusová lingvistika Praha*, představující přirozenou platformu pro výměnu zkušeností domácích i zahraničních badatelů.

Rozklad dílčích témat (kroky ke zlepšení postavení)

Vzhledem k bezprecedentním podmínkám daným rozvinutou infrastrukturou, která byla vytvořena v minulých letech a v současnosti je dále rozvíjena z podpory MŠMT, lze výzkumný potenciál české korpusové lingvistiky zúročit na domácím i zahraničním poli bez nutnosti pořizovat nákladnější zařízení. Primárním cílem navrhovaného programu Progres je proto pomocí investic do lidských zdrojů zvýšit kvalitu korpusového výzkumu na UK na úroveň srovnatelnou s mezinárodními institucemi při zachování vůdčího postavení mezi institucemi domácími. Dlouhodobým cílem je pak zvýšit prestiž hostitelské instituce tak, aby představovala přirozené a respektované centrum korpusově orientovaného výzkumu pro střední Evropu, resp. pro celý slovanský areál.

Předpokladem pozvednutí kvality korpusového výzkumu je kromě infrastrukturního zabezpečení (viz výše) zejména personálně posílená badatelská aktivita v oblasti inovativního

využívání korpusových dat, která je základem pro vývoj nových aplikací; dále pak vytváření vlastního lingvistického výzkumu, který bude zaměřen primárně na implementaci metodologických inovací do popisu jazyka.

Výstupem těchto činností by měly být zahraniční publikace (zejména články v impaktovaných časopisech a příspěvky na mezinárodních konferencích), které spolu s neustále dobudovávanou infrastrukturou (zvláště vícejazyčný paralelní korpus InterCorp) pomůžou přilákat další spolupracovníky ze zahraničí.

Specificky by se pak členové pracovního týmu měli v rámci programu Progres věnovat následujícím výzkumným tématům (uspořádání témat kopíruje výše naznačené rozdělení na hlavní výzkumné linie; popis témat, jimž je v ČNK věnována soustavnější pozornost, jsou rovněž doprovázena odkazy na dosud publikovanou odbornou literaturu členů PRVOUK):

- Korpusová lexikologie a lexikografie
    - Původní motivací pro budování jazykových korpusů byla snaha poskytnout dostatek autentických dat pro tvorbu slovníků. Kromě toho, že ČNK slouží jako primární zdroj dat pro tvorbu nového výkladového slovníku češtiny vznikajícího v AV ČR, se ČNK podílel na vytvoření několika vlastních slovníkových publikací, konkrétně na tvorbě slovníků frekvenčních (např. Čermák, F. & Křen, M. (eds.): *Frekvenční slovník češtiny*. NLN, Praha 2004; Čermák, F. & Křen, M. (eds.): *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. Routledge, London 2011) a autorských či dobových (např. Čermák, F. & Cvrček, V.: *Slovník Bohumila Hrabala*. NLN, Praha 2009; Čermák, F. & Cvrček, V. & Schmiedtová, V. (eds.): *Slovník komunistické totality*. NLN, Praha 2010).
    - Trvalý badatelský zájem je v ČNK věnován problematice víceslovných jednotek, konkrétně kolokací a frazémů (např. Čermák, F. & Šulc, M. (eds.): *Kolokace*. NLN, Praha 2006; Čermáková, A.: *Valence českých substantiv*. NLN, Praha 2009). Vzhledem k absenci soustavného popisu lexikální syntagmatiky češtiny je tento směr bádání nejen zajímavý výzkumně, ale vytváří též předpoklady pro další aplikace a implementace (zlepšení lemmatizace a morfologické anotace využitím informací o ustálených spojeních, kolokacích, frazémech, víceslovných názvech apod.).
    - Výzkumná témata: aspekty frekvenčních charakteristik lexikálních jednotek, kolokabilita a valence (zejm. s ohledem na lexikografickou praxi), kvantitativní výzkum kontextu, sémantická prozódie.
- Gramatický popis češtiny
    - Autentická jazyková data jsou nezbytným předpokladem popisu reálného úzu nejen v oblasti lexikonu, ale také v gramatice. Na půdě ČNK vznikla vůbec první na korpusu založená mluvnice (Cvrček, V. a kol.: *Mluvnice současné češtiny*. Nakladatelství Karolinum, Praha 2010) a řada dalších dílčích studií publikovaných časopisecky nebo ve sbornících z konferencí.
    - Korpusový výzkum v oblasti gramatiky je rovněž předpokladem pro kvalitní automatickou morfologickou a syntaktickou analýzu. Unikátní nástroje vyvíjené v rámci ČNK na automatickou disambiguaci pomocí pravidel čerpají převážně z takto koncipovaného výzkumu, který nelze bez rozsáhlých korpusů a ručně označkovaných etalonových korpusů vůbec provádět (viz např. Petkevič, V.: *Morfologická homonymie v češtině*. NLN, Praha 2014; Jelínek, T.: *Skladební funkce a pád v korpusu: Frekvenční analýza*. NLN, Praha 2015). Jejich základem je soubor mnoha formálních pravidel, pomáhajících identifikovat ty interpretace, které neodpovídají struktuře české gramatiky, a postupně izolovat ty, které jsou v souladu

s konsenzuální představou o morfologické či syntaktické charakteristice daného jevu.
    - Výzkumná témata: útvarová/registrová příslušnost a variabilita gramatických jednotek, vzájemné vztahy gramatiky a lexikonu, povrchová syntax češtiny na základě automatické syntaktické analýzy, kvantitativní analýza gramatických kategorií.
- Korpusová analýza diskurzu
    - Ve spolupráci s Brownovou univerzitou vznikla vůbec první aplikace umožňující kvantitativní analýzu diskurzu – KWords. Na jejím základě pak vzniká řada studií zaměřených na analýzu politického diskurzu publikovaných v zahraničních odborných časopisech.
    - Oblast korpusového výzkumu mluveného jazyka je (nejen pro češtinu) dosud velmi málo metodologicky zpracovaná (viz Kopřivová, M. & Waclawičová, M. (eds.): *Čeština v mluveném korpusu*. NLN, Praha 2008), je proto žádoucí, aby badatelská pozornost byla napřena tímto směrem; jde mj. o otázky související s variabilitou mluveného jazyka, ale také s problematikou základních jednotek (slovo, věta), které v mluveném diskurzu nelze vymezovat totožně jako v psaných textech.
    - Výzkumná témata: role gramatických kategorií při analýze diskurzu, identifikace prominentních jednotek vyššího řádu (n-gramy), principy transkripce zvukového záznamu (a jejich důsledky pro popis), jednotky mluveného jazyka.
- Diachronní korpusový výzkum
    - Pro oblast diachronního bádání bude zásadní (vedle zveřejnění monitorovacích korpusů) rozvoj metodologie jejich využívání. Jde jednak o efektivní práci s časovou osou v rámci specializovaných aplikací, ale obzvláště o metody evaluace a validace výsledků získaných na specifickém vzorku dat monitorovacím korpusu (k tomu viz Křen, M.: *Odraz jazykových změn v synchronních korpusech*. NLN, Praha 2013).
    - Výzkumná témata: jazyková změna prizmatem korpusových dat, proměny kolokability, lemmatizace a morfologická anotace diachronních korpusů, registrová variabilita v diachronní perspektivě.
- Kontrastivní a translatologický výzkum
    - Kontrastivní popis jazyků založený na korpusu předpokládá využití zejména paralelního korpusu InterCorp (viz Čermák, F. et al. (eds.): *InterCorp: Exploring a Multilingual Corpus*. NLN, Praha 2010), a to jak v oblasti gramatiky, tak popisu lexikonu. Metodologický zřetel se uplatňuje mj. i v důrazu na aspekt translatologický, zohledňující překladovost a směr překladu jako potenciální faktor ovlivňující empirické bádání.
    - Z jiného hlediska s kontrastivním pohledem úzce souvisí problematika jednotné anotace různojazyčných textů (např. Universal Dependencies), jejíž teoretické a metodologické aspekty pro humanitní výzkum češtiny nejsou zatím dostatečně probádány.
    - Výzkumná témata: anotace paralelních korpusů jako metodologický problém, extrakce lexikálních ekvivalentů víceslovných jednotek, komparativní frazeologie a idiomatika.
- Aplikovaná lingvistika
    - Na poli aplikované lingvistiky by se ČNK v rámci programu Progres chtěl angažovat zejména v problematice využití korpusů při jazykové výuce ve školách. Toto nanejvýš aktuální téma, zasahující do výuky rodilých mluvčích i cizinců, přitom vyžaduje jak specifické nástroje (šité na míru specifickým potřebám), tak zcela inovativní pohled na jazyková data a proces jazykového vzdělávání. ČNK

v této souvislosti hodlá navázat na svoji dlouhodobou spolupráci s nakladatelstvím Fraus, pro které vytvářel výukové materiály založené na korpusu.
    - Perspektivně by se ČNK měl zaměřit i na dynamicky se rozvíjející oblast korpusové stylistiky, která se (zatím spíše jenom ve světě) zaměřuje na analýzu literárních děl. Podpora korpusových metod v literární vědě začíná mít první příznivce i v českém prostředí – důkazem je počínající spolupráce ČNK a ÚČL AV ČR na vytváření korpusu současné české poezie.
    - Výzkumná témata: data-driven metody učení, výzkum lexikonu a gramatiky publikací určených dětem, problém identifikace metafory, specifika básnického jazyka.
- Metodologie korpusové lingvistiky
    - Empirické zaměření korpusového výzkumu s sebou přináší nutnost statistického zpracování dat. V současnosti u nás etablovaný výzkum má povahu převážně deskriptivní, což je částečně způsobeno neznalostí metod a postupů inferenční a exploratorní statistiky a částečně nedostupností uživatelsky přívětivých nástrojů vhodných pro tento typ analýz. Právě rozvoj metodologie efektivního a korektního zpracování statistických údajů z korpusu spolu s návodnými ukázkami interpretace výsledků by proto měly hrát klíčovou roli v rozvoji vědní oblasti nejen na UK, ale v rámci celé lingvistické obce v ČR.
    - Dlouhodobým tématem korpusových studií ve světě je výzkum variability jazyka, kterému se koncentrovaně hodlá věnovat připravovaný projekt v rámci OP VVV – jeho cílem je pomocí multidimenzionální analýzy zmapovat variabilitu současného jazyka, a to jak na úrovni jednotlivých jevů (a jejich variant), tak na rovině textů (stylová rozrůzněnost, resp. variabilita registrů). Na základě výsledků tohoto projektu bude možné provádět dodatečná zkoumání specifičtějších periferních oblastí češtiny a vztahovat je k jádru jazyka.
    - Metodologie související s designem korpusů: odhlédneme-li od ad hoc sestavovaných oportunistických korpusů, které jsou využitelné jen v některých výzkumných oblastech, je předpokladem sestavení jakéhokoli korpusu koncept reprezentativnosti (viz Cvrček, V. & Čermáková, A., Křen, M.: Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost* 77, 2016, s. 83–101). Ta se odvíjí od populace textů, kterou má daný korpus reprezentovat, a vedle obecných metodologických otázek reprezentativnosti je třeba v této souvislosti zvažovat i specifika daného jazyka a komunity jeho mluvčích (např. absence či kritický nedostatek česky psaných odborných textů v určitých vědních oborech, vysoká proporce překladů mezi nově publikovanými texty apod.). Zatímco v případě psaného jazyka je možné se alespoň částečně inspirovat příklady ze zahraničí, unikátní projekty – jako jsou korpusy spontánního mluveného jazyka či korpusy monitorovací – narážejí na nedostatek vhodných vzorů mnohem palčivěji; zejména v těchto oblastech je třeba vlastní soustavný výzkum, na němž se ČNK podílí.
    - Výzkumná témata: signifikance a relevance rozdílu frekvenčních charakteristik, využití metod exploratorní statistiky (ANOVA, faktorová analýza apod.), žánrová/registrová klasifikace textů na internetu.

Harmonogram a výstupy

Navrhovaný program Progres je koncipován jako pokračování PRVOUK P11. Badatelský tým je tedy rámcově sestaven (v minulém období proběhla nutná generační obměna) a bude i nadále fungovat na půdorysu 6 sekcí (schéma kopíruje organizační strukturu projektu ČNK v rámci jeho infrastrukturní činnosti):

- lingvistická sekce (vedoucí: Mgr. Anna Čermáková, Ph.D.) se soustředí na otázky spjaté s metodologií vytěžování korpusů, návrhy inovativních aplikací, lexikální syntagmatiku, kontrastivní a translatologické aspekty a aplikovaný výzkum (využití korpusu ve školách);
- sekce mluvených a nářečních korpusů (vedoucí: PhDr. Marie Kopřivová, Ph.D.) se zaměřuje na problematiku metodologie vytěžování korpusů mluveného jazyka a rozvoj korpusové dialektologie;
- sekce diachronních korpusů (vedoucí: prof. PhDr. Karel Kučera, CSc.) rozvíjí metody práce s diachronními a monitorovacími korpusy (návrh aplikací pro jejich využívání, anotace starších textů);
- sekce lingvistické analýzy a anotace (vedoucí: Mgr. Tomáš Jelínek, Ph.D.) má na starosti automatickou morfologickou a syntaktickou anotaci; za tímto účelem se vědecky soustředí na výzkum v oblasti gramatiky a frazeologie;
- sekce paralelních korpusů (vedoucí: ing. Alexandr Rosen, Ph.D.) se podílí na kontrastivním výzkumu a rozvoji jeho metodologie;
- sekce komputační (vedoucí: Mgr. Pavel Vondřička, Ph.D.) má na starosti technickou podporu celému projektu, vývoj aplikací a technickou přípravu dat.

Součástí pracovního týmu jsou doktorandi oboru Matematická lingvistika (korpusová větev), kteří se podílejí jak na výzkumné, tak i propagační činnosti související s programem. Předpokládáme zapojení zhruba 7 doktorandů.

Výstupy z realizace programu budou dvojího druhu. Všechny výstupy mají přitom komplementární povahu vzhledem k financování výzkumné infrastruktury ČNK (LM2015044), která realizaci těchto aktivit neumožňuje vůbec nebo pouze v omezené míře:

1) vědecké výstupy
2) popularizační a propagační výstupy

Údaje o počtu jednotlivých výstupů v následujících sekcích mají indikativní charakter a jsou odvozeny od finanční podpory ve výši srovnatelné se současným stavem v rámci PRVOUK P11.

Ad 1) Vědecké výstupy

Publikační aktivita badatelů v rámci navrhovaného programu Progres se bude soustředit na články v zahraničních recenzovaných časopisech (v databázích WoS, Scopus nebo ERIH+). Rozsáhlejší výzkum, určený zejména českému publiku, je v humanitních vědách standardem publikovat ve formě monografie (zde se nabízí pokračování v již etablované řadě *Studie z korpusové lingvistiky*). Nezanedbatelný přínos pro sdílení informací a rovněž pro zvyšování vědecké prestiže hostitelské instituce představují konferenční příspěvky, případně příspěvky v konferenčních sbornících. Indikativní průměrné počty jednotlivých druhů výsledků za 5 let na 1 úvazek (FTE) shrnuje následující tabulka:

	Počet výstupu za 5 let na 1 FTE
Monografie	1
Články v recenzovaných časopisech	15
Konferenční příspěvky / příspěvky ve sbornících	20

Ad 2) Popularizační a propagační výstupy

Vzhledem k infrastrukturní povaze projektu ČNK je vedle samotného výzkumu třeba dbát na jeho diseminaci. K tomu slouží principiálně tři kanály:

a) odborná setkávání *Korpusová lingvistika Praha* – v průběhu realizace programu Progres by se měly realizovat minimálně dvě akce, které navážou na předchozí konference a kolokvia konané pod tímto jménem (předběžně v roce 2018 a 2020);
b) workshopy pro práci s korpusem – tyto aktivity organizuje ČNK v rámci své infrastrukturní činnosti zhruba v půlročních intervalech a účastní se jich zájemci z řad akademické obce, studentů, učitelů i jazykových profesionálů;
c) *Internetový manuál a příručka uživatele ČNK* (viz http://wiki.korpus.cz), který lze rozdělit na složku infrastrukturní (dokumentace projektu ČNK, popis korpusů a nástrojů) a složku vědeckou, hrazenou z prostředků Prvouk/Progres: on-line encyklopedie korpusové lingvistiky, kurz práce s korpusem, slovník pojmů korpusové lingvistiky. Obě složky jsou pravidelně aktualizovány a neustále rozšiřovány.

Indikátory zlepšení

Indikátory úspěšné realizace programu se odvíjejí od harmonogramu a navrhovaných výstupů (obojí viz výše). Předpokládané cílové hodnoty jednotlivých indikátorů se mohou měnit v závislosti na finančních možnostech projektu.

- Počet knižních monografií (předpoklad: jedna za 5 let na 1 FTE)
- Počet článků v recenzovaných časopisech, indexovaných ve WoS, SCOPUS nebo ERIH+ (předpoklad: 3 za rok na 1 FTE)
- Počet účastí na konferencích a oborových setkáních (předpoklad: 4 za rok na 1 FTE)
- Počet uspořádaných oborových setkání (předpoklad: jedno za 2 roky)
- Počet návštěv významného zahraničního odborníka (předpoklad: minimálně jeden za rok)

10) Podklady pro stanovení výše finanční bonifikace programu

Bonifikace mezinárodní spolupráce

ČNK vznikl v roce 1994 po vzoru zahraničních institucí či konsorcií usilujících o soustavné a kontinuální mapování jazykových dat pomocí elektronických korpusů pro účely jazykově orientovaného výzkumu. Svým zaměřením na kontinuitu sběru se ČNK podobá Národní knihovně (schraňující slovesné bohatství) či Národnímu archivu. V současnosti je ČNK jedinou institucí v ČR, která se soustředí na širokospektrální sběr autentických jazykových dat pro popis češtiny, a jako taková je i vyhledávaným partnerem zahraniční spolupráce zahrnující výzkum češtiny. Přes svoje bytostně národní zaměření na češtinu se ČNK angažuje v mnoha aktivitách, které mají mezinárodní přesah.

Od svého založení spolupracuje ČNK s řadou významných zahraničních institucí, které se soustředí na výzkum v oblasti korpusové lingvistiky. Předmětem těchto spoluprací je výměna dat, nástrojů a expertíz. S následujícími institucemi probíhá spolupráce na smluvním základě:

- Association des Originaires des Pays Tcheques et Slovaque, Paris, France
- Brown University (Department of Slavic Languages), Providence, USA
- Bulgarian Academy of Sciences (Institute of Parallel Processing, Linguistic Modelling Department), Sofia, Bulgaria
- "Orientale" University of Naples, Naples, Italy

- Polish Academy of Sciences (Institute of Computer Science, Department of Artificial Intelligence), Warsaw, Poland
- Saint-Petersburg State University (Faculty of Philology and Arts), Saint-Petersburg, Russia
- Slovak Academy of Sciences (Ľ. Štúr Institute of Linguistics), Bratislava, Slovak Republic
- Sorbian Institute (Department of Linguistics), Bautzen, Germany
- The Institute of German Language, Mannheim, Germany
- Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia
- Tübingen University (Faculty of Humanities, Slavisches Seminar), Tübingen, Germany
- University of Amsterdam, Amsterdam, Netherlands
- University of Bern (Faculty of Philosophy and History, Institute of Slavic Languages and Literature), Bern, Switzerland
- University of Bologna (The Department of Classic Philology and Italian Studies), Bologna, Italy
- University of Granada, Granada, Spain
- University of Latvia (Institute of Mathematics and Computer Science), Riga, Latvia
- University of Regensburg (Faculty of Philosophy IV, Slavonic Department), Regensburg, Germany
- University of Warsaw (Institute of Western and Southern Slavic Studies), Warsaw, Poland

Dlouhodobou vědeckou spolupráci realizuje ČNK především s Brownovou univerzitou (USA). Spolupráce je zaměřena na korpusový výzkum politického diskurzu a byla impulzem také pro vznik výše zmíněné aplikace KWords. Vedle publikačních výstupů se v rámci této spolupráce podařilo zorganizovat workshop pro zájemce o korpusovou lingvistiku z řad amerických slavistů s názvem *Quantitative Text Analysis for the Humanities and Social Sciences* (Brown University, Providence, RI, USA; 8.–9. dubna 2016.).

Níže uvádíme přehled nejvýznamnějších publikačních výstupů vytvořených členy týmu ČNK ve spolupráci se zahraničními badateli:

- Fidler, M. & Cvrček, V. (2015): A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis. *Journal of Slavic Linguistics* 23(2), s. 197–239.
- Kaczmarska, E. & Rosen, A. et al. (2015): A syntactico-semantic analysis of arguments as a method for establishing equivalents of Czech and Polish verbs expressing mental states. *Prace Filologiczne* LXVII, s. 135–158.
- Kaczmarska, E. & Rosen, A. (2013): Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego, *Studia z Filologii Polskiej i Słowiańskiej* 48, Warszawa 2013, s. 103–121.
- Piao, S. & Rayson, P. & Archer, D. & Bianchi, F. & Dayrell, C. & El-Haj, M. & Jiménez, R.-M. & Knight, D. & Křen, M. & Löfberg, L. & Nawab, R. M. A. & Shafi, J. & Teh, P. L. & Mudraya, O. (2016): Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. *Proceedings of LREC2016*, ELRA, s. 2614–2619.
- Richterová, O. & Stöckeler J. (2013): Partizipialadjektive im Deutschen und Tschechischen: Nichts- oder vielsagend über das Wesen der Wortklassen? *Studia Germanistica* 13/2013.

Od roku 2016 se ČNK v rámci programu Erasmus+ účastní projektu s názvem *DigiLing: Trans-European e-Learning Hub for Digital Linguistics*, jehož koordinátorem je univerzita v Lublani.

Cílem projektu je vytvořit mezinárodní odborný tým, který se bude soustředit na rozvoj oblasti digital humanities ve výuce lingvistických oborů.

11) Předpokládaný procentuální podíl finančních prostředků alokovaných jednotlivými participujícími fakultami (VŠ ústavy) v prvním roce uskutečňování Programu a přibližný výhled na další čtyři roky

FF UK 100%

12) Údaje o projednání přihlášky programu vědeckými radami všech fakult (VŠ ústavů) UK, na nichž má být program uskutečňován

Fakulta (VŠ ústav)	Datum projednání přihlášky vědeckou radou fakulty (VŠ ústavu)	Výsledek projednání přihlášky vědeckou radou fakulty (VŠ ústavu)
FF UK	29. 9. 2016	

13) Datum a podpis koordinátora: .....

14) Podpisy děkanů (ředitelů) všech fakult (jiných součástí) UK, na kterých má být program uskutečňován

Fakulta (VŠ ústav)	Jméno, příjmení, tituly děkana/ředitele	Datum a podpis děkana/ředitele
FF UK	Doc. Mirjam Fried	