ČESKÝ NÁRODNÍ KORPUS

FILOZOFICKÁ FAKULTA
UNIVERZITY KARLOVY
V PRAZE

Ústav Českého národního korpusu
doc. Mgr. Václav Cvrček, Ph.D.
nám. Jana Palacha 2, Praha 1, 116 38
+420 221 619 837 | vaclav.cvrcek@ff.cuni.cz

VÝZKUMNÉ ZAMĚŘENÍ A POSLÁNÍ

Výzkumná infrastruktura zajišťovaná Ústavem Českého národního korpusu FF UK a Ústavem teoretické a komputační lingvistiky FF UK

Projekt Český národní korpus (ČNK), založený v roce 1994, se zaměřuje na kontinuální mapování češtiny ve všech dostupných dimenzích (časové, regionální i žánrové).

ČNK vytváří a zpřístupňuje rozsáhlé elektronické sbírky textů (jazykové korpusy), které slouží jako báze pro jakýkoli jazykově orientovaný výzkum.

Vedle toho se věnuje rozvoji metodologie empirického lingvistického výzkumu a vývoji nástrojů na vytěžování jazykových korpusů.

ČNK od roku 2012 slouží jako výzkumná infrastruktura pro mnoho oblastí společenských a humanitních věd (zejm. pro lingvistiku, psychologii, sociologii, historii, počítačové zpracování přirozeného jazyka ad.).

Pro svoji rozsáhlou a kvalitně budovanou základnu je ČNK vyhledávaným partnerem pro zahraniční vědeckou spolupráci.

Vedle toho se zaměřuje i na poskytování poradenství, vytváření analýz pro vědecké i popularizační účely, poskytování dat pro výzkum češtiny i dalších jazyků pro srovnání s ní a automatické zpracování jazykových dat.

CO UMÍME

ČNK se soustředí na sběr textů zejména v následujících dimenzích:

- **současný psaný jazyk** (beletrie, odborná literatura, publicistika)
- **současný neformální mluvený jazyk** (komunikace v rodině nebo mezi přáteli), včetně propojení přepisu se zvukem
- **čeština v kontrastu s jinými jazyky** (český text propojený s jeho překlady do různých jazyků)
- **starší vývojové fáze jazyka** (od 14. století po současnost)

Jazyková data jsou v rámci ČNK dále podrobně zpracovávána; texty jsou opatřeny metadaty, procházejí sérií datových konverzí a konečně také lingvistickým značkováním (lemmatizací a tagováním). V případě paralelních korpusů (obsahujících texty spolu s jejich překlady z nebo do několika jazyků) je zpracování navíc obohaceno o metody poloautomatického zarovnání (nacházení ekvivalentů na rovině vět i slov). Jazykové zdroje jsou následně zpřístupňovány pomocí speciálních aplikací (viz portál www.korpus.cz), na jejichž vývoji se ČNK podílí.

CO NABÍZÍME

ČNK vyvíjí pro svoji činnost nástroje pro zpracování přirozeného jazyka, aplikace pro jazykovědný výzkum a provádí analýzy textů. Konkrétně jde o:

- **automatickou lemmatizaci textů,** tj. přiřazení slovníkového tvaru ke slovům v textu (např. *pes* k tvarům *pes*, *psa*, *psům* apod.)

https://www.korpus.cz
https://ucnk.ff.cuni.cz

- **automatické morfologické značkování,** tj. přiřazení informací o gramatických vlastnostech každému slovu v textu (např. slovní druh, pád, číslo, rod, osoba, způsob apod.)
- **lingvistickou analýzu textů** (např. frekvenční zastoupení jednotlivých slov spolu s jejich srovnáním s referenčním korpusem češtiny, analýza spojitelnosti zadaných slov atp.)
- **vytváření seznamů lexikálních jednotek češtiny** (na základě frekvence a/nebo v kombinaci s různými gramatickými či stylovými charakteristikami)
- **zkoumání běžných překladových řešení a automaticky generované seznamy pravděpodobných ekvivalentů v různých jazycích**

POPTÁVKA

Služby ČNK můžou být zajímavé pro různá odvětví, která pracují s jazykem:

- **média a PR agentury** (analýzy textů, projevů, veřejného diskurzu)
- **ICT vývojáři** (jazykové technologie vyžadující trénovací nebo testovací data – psaný jazyk, mluvený jazyk, překladové páry)
- **vzdělávací zařízení** (podklady pro výuku, výukové materiály a aplikace)
- **překladatelé a tlumočníci** (možnosti hledání překladových ekvivalentů)

NEJVĚTŠÍ ŘEŠENÉ PROJEKTY

Projekt Český národní korpus poskytuje svoje data a nástroje všem registrovaným uživatelům prostřednictvím portálu www.korpus.cz. Kromě obecných nástrojů na práci s rozsáhlými jazykovými korpusy (KonText) zde zájemce najde i nástroje specializované, např. SyD pro analýzu variant, Treq pro vyhledávání překladových ekvivalentů nebo KWords pro zjišťování klíčových (prominentních) slov v textech.

Český národní korpus úzce spolupracuje např. s nakladatelstvím Fraus na vytváření pomůcek pro výuku češtiny a pravidelně organizuje praktické semináře a workshopy zaměřené na práci s korpusy pro veřejnost i odborníky z řad učitelů, překladatelů nebo redaktorů.

NEJVĚTŠÍ DOSAŽENÉ ÚSPĚCHY

- Vytvoření a zveřejnění **největší a nejpestřejší veřejně přístupné databáze českých textů**, která je volně dostupná pro badatelské a pedagogické účely.
- Vytvoření **sady nástrojů pro práci s rozsáhlými korpusy** (KonText, SyD, KWords, Treq).
- Vývoj **unikátních nástrojů pro počítačové zpracování českých textů.**
- Zařazení projektu ČNK na cestovní **mapu výzkumných infrastruktur na období 2016–2022.**

PUBLIKACE

- Bartoň, T. a kol.: *Statistiky češtiny.* Nakladatelství Lidové noviny, Praha 2009.
- Cvrček, V. a kol.: *Mluvnice současné češtiny.* Nakladatelství Karolinum, Praha 2010.
- Čermák, F. – Křen, M. (eds.): *A Frequency Dictionary of Czech: Core Vocabulary for Learners.* Routledge, London 2011.
- Čermák, F. – Rosen, A.: *The case of InterCorp, a multilingual parallel corpus.* International Journal of Corpus Linguistics, 2012, 17(3), 411–427.
- Čermák, F. – Křen, M. (eds.): *Frekvenční slovník češtiny.* Nakladatelství Lidové noviny, Praha 2004.
- Čermák, F. – Cvrček, V.: *Slovník Bohumila Hrabala.* Nakladatelství Lidové noviny, Praha 2009.

https://www.korpus.cz
https://ucnk.ff.cuni.cz