Český národní korpus

 # Korpus, jazykový korpus
 # Text, analýza textu
 # Zpracování přirozeného jazyka
 # NLP, lingvistika

Co nabízíme

- Nástroje pro zpracování přirozeného jazyka, aplikace pro jazykovědný výzkum, analýzy textů
- Automatická lemmatizace textů, tj. přiřazení slovníkového tvaru ke slovům v textu
- Automatické morfologické značkování, tj. přiřazení informací o gramatických vlastnostech každému slovu v textu
- Lingvistická analýza textů (např. frekvenční zastoupení jednotlivých slov spolu s jejich srovnáním s referenčním korpusem češtiny, analýza spojitelnosti zadaných slov atp.)
- Vytváření seznamů lexikálních jednotek češtiny
- Zkoumání běžných překladových řešení a automaticky generované seznamy pravděpodobných ekvivalentů v různých jazycích

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

- Média a PR agentury
- ICT vývojáři (jazykové technologie vyžadující trénovací nebo testovací data)
- Vzdělávací zařízení
- Překladatelé a tlumočníci

Co umíme

ČNK se soustředí na sběr textů zejména v těchto dimenzích:

- Současný psaný jazyk (beletrie, odborná literatura, publicistika)
- Současný neformální mluvený jazyk 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 lingvistickým značkováním (lemmatizací a tagováním)
- V případě paralelních korpusů (obsahujících texty spolu s jejich překlady z nebo do několika jazyků) je zpracování obohaceno o metody poloautomatického zarovnání (nacházení ekvivalentů na rovině vět i slov)
- Jazykové zdroje jsou zpřístupňovány pomocí speciálních aplikací (viz portál www.korpus.cz), na jejichž vývoji se ČNK podílí

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.
- ČNK 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.

Největší řešené projekty

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

Zajímá vás tato expertíza?

Kontaktujte CPPT UK
Web: www.cppt.cuni.cz/
Mail: transfer@cuni.cz
Tel.: +420 224 491 255

Naši experti a jejich pracoviště

Mgr. Michal Křen, Ph.D.
Filozofická fakulta UK
Web: www.ff.cuni.cz