Z Wikipedie, otevřené encyklopedie

Jazykový korpus je (většinou rozsáhlý) soubor textů určitého jazyka. Jedná se o „vnitřně strukturovaný, unifikovaný a obvykle i o indexovaný a ucelený rozsáhlý soubor elektronicky uložených a zpracovaných jazykových dat většinou v textové podobě, organizovaný se zřetelem k využití pro určitý cíl.“[1] Korpus slouží jednak pro lingvistický výzkum jazykové praxe, jednak jako datová základna pro tvorbu slovníků, korektorů, překladačů apod. Tvorbou korpusů se zabývá obor korpusová lingvistika.[2]

V současnosti mají korpusy digitální podobu, což výrazně usnadňuje sběr dat i jejich zpracování: speciální programy umožňují vyhledávání slov a slovních spojení v kontextu, zjištění frekvence výskytu v korpusu i zjištění původního zdroje textu.[3]

Popis

Korpusy slouží zejména jako lexikologický a lexikografický nástroj a stávají se mj. zdrojem pro zpracování jednojazyčných výkladových slovníků a automatických korektorů nebo vícejazyčných překladových slovníků a automatických překladačů. Kromě lingvistů korpusy stále častěji využívají i redaktoři, překladatelé a další tvůrci textů, učitelé a studenti cizích jazyků.

Texty jsou v různé míře opatřeny metajazykovými značkami vypovídajícími o samotném textu (autor, rok vydání, žánr apod.), o zařazení jednotlivých slov do kategorie slovních druhů, o frekvenci slova v korpusu, případně dalších lingvistických a frekvenčních aspektech. Pro formátování textů a vkládání značek se používá zejména standardizovaného jazyka XML, případně staršího SGML.

Rozdělení korpusů

Referenční korpus je stálý, takže opakované dotazy dávají vždy stejné výsledky. Naproti tomu *nereferenční* korpus je průběžně aktualizován, obvykle jednou ročně.

Některé korpusy jsou budovány jako *vyvážené*, což znamená, že by měly obsahovat vyvážený podíl textů tříděných podle žánrovosti, doby vzniku, případně dalších hledisek (mluvenost, psanost, regionálnost, užívanost apod.).

Synchronní korpusy jsou budované jako reprezentativní a vyvážené otisky jazyka v určitém relativně krátkém časovém období, během něhož lze považovat jazyk za neměnný systém. Většinou se jedná o korpusy současného jazyka. *Diachronní* korpusy zachycují jazyk v různých vývojových fázích a obsahují tudíž texty z rozsáhlejších období.

Podle dalšího kritéria rozlišujeme také korpusy jednojazyčné a vícejazyčné. Vícejazyčný korpus se také nazývá *paralelní* korpus a obsahuje stejné texty v různých jazycích zobrazené vedle sebe.

Korpusy češtiny

Budováním korpusů českého jazyka se v České republice zabývá od roku 1994[4] Ústav Českého národního korpusu (ÚČNK) při Filozofické fakultě UK[5], který založil lingvista František Čermák. Od ledna 2014 je možné vyhledávat v korpusech ÚČNK pomocí korpusového manažeru KonText.[6] Z českých korpusů mezi ně patří jak synchronní korpusy psané češtiny, tak synchronní mluvené korpusy a diachronní korpusy:

Korpusy řady SYN – všechny synchronní psané korpusy této řady jsou spojeny ve stejnojmenný korpus SYN. Dohromady obsahují přes dva tisíce testových slov (tokenů). Korpus není reprezentativní – sice jsou v něm spojeny reprezentativní korpusy, ale po sloučení převažují texty publicistické (SYN2006PUB, SYN2009PUB a SYN2013PUB). Publicistická část zahrnuje texty nejznámějších celostátních deníků a nespecializovaných časopisů. Ostatní korpusy (SYN2000, SYN2010, SYN2005) jsou žánrově vyvážené, obsahují tedy jak texty z publicistického prostředí, tak z beletrie nebo odborné literatury.

Korpus soukromé korespondence (KSK-dopisy) obsahuje 2000 elektronických přepisů ručně psané korespondence z let 1990–2004. Pisatelé pocházejí z různých částí České republiky, a tak jsou v dopisech dobře zachovány idiolekty. Všechny shromážděné texty obsahují sociologické charakteristiky pisatelů. Korpus byl vytvořen v Ústavu českého jazyka

na Filozofické fakultě Masarykovy univerzity v Brně pod vedením Zdeňky Hladké.

Pražský mluvený korpus (PMK) je první korpus mluvené češtiny. Zachycuje autentickou mluvenou češtinu z oblasti Prahy a okolí. Anonymních 304 nahrávek pochází z let 1988–1996.

Brněnský mluvený korpus (BMK) je v rámci ČNK prvním korpusem mluvené češtiny z oblasti Moravy. Zaznamenává autentickou tematicky nespecializovanou mluvu města Brna. BMK je elektronickým přepisem 250 anonymních magnetofonových nahrávek (tj. 596 009 pozic) z let 1994-1999 zachycujících 294 mluvčích.

řada ORAL zachycuje spontánní konverzaci výhradně v neformálních komunikačních situacích. Korpus ORAL2013 již pokrývá celé území České republiky a přepis je propojen se zvukovou stopou.

SCHOLA2010 obsahuje mluvu učitelů i žáků zaznamenanou ve vyučovacích hodinách o velikosti kolem milionu pozic. Je to proto sociolingvisticky nevyjádřený korpus formální i poloformální mluvené češtiny. Díky své interdisciplinárnosti může být využit nejen pro lingvistické, ale i pro psychologické, sociologické a další bádání.

CzeSL-plain (Czech as a Second Language) – neanotovaný žákovský korpus, který obsahuje tři subkorpusy – **ciz** (písemné práce nerodilých mluvčích různé úrovně), **kval** (odborné práce nerodilých mluvčích) a **rom** (písemné práce romských žáků z oblastí ohrožených sociálním vyloučením).

DIAKORP je diachronní korpus obsahující texty od 13. století, nejmladší pak pocházejí z roku 1989. Kvůli velkému časovému rozpětí musel být korpus vytvořen trochu odlišně od výše zmíněných. Texty jsou transkribovány a značkování zachycuje i podstatnou část lingvistických informací, která může transkripcí zaniknout.

Také další univerzitní pracoviště budují vlastní korpusy:

Korpus českého verše Ústavu pro českou literaturu AV ČR. Je lemmatizovaný, foneticky, morfologicky, metricky a stroficky anotovaný. Jeho obsah je čerpán z české poezie 19. a počátku 20. století. Na webových stránkách je k dispozici velká řada on-line nástrojů Archivováno 16. 6. 2016

na Wayback Machine., které umožňují s texty pracovat či si procvičovat rozpoznávání meter.

czTenTen12 – stejně jako to platí u ostatních korpusů řady TenTen, i zde jsou texty postahovány z internetu (pomocí nástroje Spiderling a Heritrix). Obsahuje přes 5,4 miliard tokenů, což z něj dělá největší český textový korpus. Spolu s korpusem **CzechParl** (Korpus stenografických záznamů českého parlamentu) a **Desam** (morfologicky označkovaný korpus českých textů) je dostupný přes korpusový manažer Sketch Engine, který vyvíjí společnost Lexical Computing Ltd. ve spolupráci s Centrem zpracování přirozeného jazyka při Fakultě informatiky Masarykovy univerzity.

Olomoucký mluvený korpus Od roku 2002 vzniká na Univerzitě Palackého v Olomouci pod vedením dr. P. Pořízky. Jeho základem jsou foneticky přepsané nahrávky.

Korpusy angličtiny

Brown Corpus

Brown University Standard Corpus of Present-Day American English neboli zkráceně **Brown Corpus** je dílem dvou autorů – Henryho Kučery (původem Čech, studoval na Univerzitě Karlově) a W. N. Francise, kteří tou dobou působili na Brownově univerzitě. Jedná se o korpus, který vznikal v letech 1963–1964, přičemž obsahuje texty z roku 1961 ve snaze zachytit jazyk v určitém období (trend, který se u tvorby korpusů volí i dnes). Cílem zkoumání je psaná americká angličtina rodilých mluvčích. V korpusu se neuchovávají celé texty, ale pouze vzorky, a to z toho důvodu, aby byl korpus vyvážený. Celkově je v něm využito 15 kategorií, mezi které patří časopisy, noviny, odborná literatura i beletrie. Z každého textu je vybrán vzorek 2 000 slov a celkový počet vzorků se rovnal 500. Celkový rozsah byl kolem jednoho milionu slov. Korpus Brown je morfologicky označkován, využito je 80 kategorií a značkovala se například interpunkce i speciální znaky.

Korpus Brown velice ovlivnil další generace lingvistů a je svým rozvržením vzorem mnoha dalších korpusů. Zajímavostí je, že v 80. letech vznikl **The Freiburg-LOB corpus of American English (Frown)**, který byl obdobou korpusu Brown. Vznikl na Freiburské univerzitě v Německu. Měl úplně stejnou

strukturu a snažil se zachytit britskou angličtinu z roku 1991. O značkování byl doplněn v roce 2007.

Na korpusu Brown je založeno několik publikací. Nejznámější z děl založených na korpusu vůbec je *Computional Analysis of Present-Day American English* od autorů korpusu (Kučera, Francis). Jedná se o statistickou studii, ve které se kombinuje lingvistika, psychologie a statistika.

Lancaster-Oslo/Bergen Corpus (LOB)

Tento korpus, který vznikal v letech 1970–1978, je britským protějškem ke korpusu Brown. Má stejnou strukturu (1 milion slov, 500 vzorků po 2 000 slovech, 15 žánrů) a snaží se taktéž zachytit jazyk v roce 1961, tentokrát se však jedná o britskou angličtinu. Jedním ze spoluautorů je i Geoffrey Neil Leech.

I k tomuto korpusu vznikla v Německu obdoba označená jako **The Freiburg-LOB Corpus of British English (FLOB)**. Zveřejněn byl v roce 1999 a zachycuje britskou angličtinu v roce 1991.

British National Corpus

British National Corpus (zkráceně **BNC**) je korpus, který vznikal ve spolupráci tří nakladatelů (Oxford University Press, Longman a W. & R. Chambers), dvou univerzit (Oxfordské univerzity a univerzity v Lancasteru) a britské národní knihovny v letech 1991–1994, přičemž zveřejněn byl v roce 1994. Jedná se o korpus se 100 miliony slov a rozvětvenou strukturou. Tento korpus je vyvážený, zahrnuje jak časopisy, noviny, tak odbornou literaturu i beletrii. V korpusu nalezneme vzorky jednotlivých textů, od jednoho autora maximálně 45 tisíc slov. Zachycuje britskou angličtinu (z let, ve kterých vznikal, tedy 1991–1994) a kromě psaných textů v něm nalezneme i mluvenou angličtinu v poměru 9 : 1 (psaná : mluvená). Mluvená angličtina byla do korpusu převedena pomocí ortografické transkripce. BNC je zafixovaný a nic se do něj nepřidává; jediné, co se mění, je značkování.

V rámci BNC jsou dva menší subkorpusy, které jsou určeny pro zkoumání jazyka. Prvním z nich je **BNC Sampler**, který obsahuje jeden milion mluvené angličtiny a jeden milion psaných textů. Druhým je **BNC Baby**, do nějž jsou zahrnuty čtyři milionové vzorky ze čtyř různých žánrů.

Na korpusu je založeno několik publikací, například článek *100 Million Words of English*, který napsal G. N. Leech, článek *Corpus Design Criteria* (1992), jehož autory jsou S. Atkins, J. Clear a N. Ostler a který popisuje strukturu BNC, a také kniha *Corpus: An Introduction* (2001), kterou napsali T. McEnery a A. Wilson.

Odkazy

Reference

1. ↑ ČERMÁK, František. 1995. Jazykový korpus : Prostředek a zdroj poznání. *Slovo a slovenost*. 1995, roč. 56, č. 2, s. 119 - 140. ISSN 0037-7031
2. ↑ https://www.ikaros.cz/jazykove-korpusy
3. ↑ Co je korpus? | Ústav Českého národního korpusu. *ucnk.ff.cuni.cz* [online]. [cit. 2022-01-13]. Dostupné online.
4. ↑ start - Příručka ČNK. *wiki.korpus.cz* [online]. [cit. 2022-01-13]. Dostupné online.
5. ↑ Portál | Český národní korpus. *www.korpus.cz* [online]. [cit. 2022-01-13]. Dostupné online.
6. ↑ KonText - query form. *www.korpus.cz* [online]. [cit. 2022-01-13]. Dostupné online.

Související články

- Korpusová lingvistika
- Paralelní korpus

Externí odkazy

- Obrázky, zvuky či videa k tématu jazykový korpus na Wikimedia Commons
- Jazykový korpus v České terminologické databázi knihovnictví a informační vědy (TDKIV)
- Ústav Českého národního korpusu
- Ústav formální a aplikované lingvistiky
- Centrum zpracování přirozeného jazyka
- Korpusový manažer Sketch Engine

- OPUS, mnohojazyčný paralelní korpus Archivováno 14. 12. 2013 na Wayback Machine.
- Britský národní korpus
- Cosmas II, *Deutsches Referenzkorpus*, DeReKo pro němčinu
- Hrvatski nacionalni korpus
- Konkordanční nástroj Korp pro vyhledávání ve švédských korpusech
- Narodowy Korpus Języka Polskiego
- Russian National Corpus (Национальный корпус русского языка)
- článek 100 Million Words of English od G. N. Leeche

Portály: Jazyk

Kategorie:

- Lexikografie
- Korpusy

Skryté kategorie:

- Monitoring:Články s odkazem na TDKIV
- Monitoring:Články s identifikátorem NKC
- Monitoring:Články s identifikátorem TDKIV
- Monitoring:Články s identifikátorem BNF
- Monitoring:Články s identifikátorem GND
- Monitoring:Články s identifikátorem LCCN
- Monitoring:Články s identifikátorem LNB
- Monitoring:Články s identifikátorem NLI
- Portál Jazyk/Zapojené články

Jazykový korpus

50 jazyků Přidat téma