Masarykova univerzita
Fakulta informatiky

Český národní korpus

(https://www.korpus.cz/)

esej z předmětu PV070 Digitální knihovny

Filip Kubeček
UČO 456251
3. ročník FF B-FI PLIN

6. 11. 2018

Obsah

Úvod

Český národní korpus (dále jen ČNK) je akademický projekt Filozofické fakulty Univerzity Karlovy (dále jen FF UK), který od jeho vzniku v roce 1994 spravuje Ústav Českého národního korpusu (dále jen ÚČNK). Na jeho vývoji se podílí i řada dalších pracovišť, především Ústav teoretické a komputační lingvistiky, rovněž náležící pod FF UK, či Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK. Projekt je dostupný na webu https://www.korpus.cz/.

Jde o obsáhlou digitální databázi autentických textů (= korpusů), v níž je možno vyhledávat různé jazykové jevy. Jeho výstupy slouží jako podklad pro výzkumy jazykové empirie v rámci korpusové lingvistiky, základní funkce projektu jsou však dostupné i bez nutnosti registrace veškerým zájemcům o aplikaci češtiny.

Obrázek 1: logo Českého národního korpusu

1 Cíle

Hlavním cílem ČNK je vytváření, spravování a zpřístupňování jazykových dat pro lingvistický výzkum. Tato data jsou formou různých korpusů pečlivě tříděna s ohledem na druh výzkumu – ČNK zahrnuje různé typy korpusů:

- jednojazyčné (výzkum češtiny) a paralelní (srovnávání jazyků),
- tématem všeobecné (vyvážená skladba druhů textů) a specializované (výzkum konkrétní tematické oblasti),
- psané a mluvené (přepisy nahrávek řeči),
- synchronní (texty ze současného období) a diachronní (texty z dřívější doby, výzkum jazykového vývoje). (1)

Cílem projektu je rovněž vytváření, spravování a vylepšování nástrojů, které slouží ke konkrétní práci s korpusy. V současné době jsou dostupné a rozvíjené tyto nástroje:

- Kontext – základní vyhledávací rozhraní, skrz něj lze přistupovat ke všem korpusům (https://kontext.korpus.cz/);
- SyD – nástroj pro výzkum jazykových variant (https://syd.korpus.cz/);
- Morfio – aplikace k odhadování rozsahu a produktivity slovotvorných modelů (https://morfio.korpus.cz/);
- KWords – nástroj k identifikaci klíčových slov v textu (https://kwords.korpus.cz/);
- Treq – databáze překladových ekvivalentů (http://treq.korpus.cz/).

Obrázek 2: loga korpusových nástrojů

Vzhledem k autentickému charakteru veškerého materiálu je v každém případě umožněno zkoumání reálně užívaného jazyka. Provádějí se kvantitativní analýzy nejmenších jazykových jednotek (fonémů a grafémů), stejně tak výzkum jevů morfologických, syntaktických a především lexikálních. Díky tomu vznikají specializované slovníky (frekvenční, kolokační), jež podávají obraz jazyka, který by bez použití korpusů a jejich nástrojů nebyl k dispozici. Výsledky korpusového zkoumání se uplatňují taktéž v pedagogice (tvorba učebnic nebo výukových materiálů, včetně výuky pro cizince), dialektologii, sociolingvistice, psycholingvistice, forenzní lingvistice nebo studiu akvizice jazyka. (2) Repozitář vědeckých publikací založených na ČNK je k dispozici na https://www.korpus.cz/biblio.

2 Popis

2.1 Korpusy

Vlajkovými korpusy ČNK jsou stomilionové korpusy psané současné češtiny řady SYN. Vycházejí po pěti letech od roku 2000 a jejich kompilaci představuje vůbec největší korpus SYN (jeho poslední verze, vydaná v roce 2017, obsahuje přes čtyři miliardy slov). Korpusy řady SYN jsou žánrově vyvážené, obsahují tudíž stejné zastoupení beletristických, odborných i publicistických textů. (3) Použití těchto textů je vázáno dohodami s jejich poskytovateli (vydavatelstvími) a kvůli autorským právům není nikdy možné zobrazit texty v jejich kompletním znění (vždy jen nejbližší kontext hledaného slova).

Pokud se týká mluvených, diachronních a paralelních korpusů, jejich anotaci a koordinaci sběru dat mají na starost samostatné sekce ÚČNK. (4) K mluveným korpusům náleží především řada ORAL, obsahující neformální dialogickou češtinu včetně sociolingvistických údajů o mluvčích. Kromě něj je k dispozici třeba Pražský nebo Brněnský mluvený korpus z konce minulého století. Zástupcem diachronního korpusu je DIAKORP, jenž pokrývá sedm století vývoje češtiny; zástupcem paralelního korpusu zase InterCorp, který obsahuje texty v češtině a k nim odpovídající překlady až ve 31 jazycích.

Mezi významné specializované korpusy ČNK patří například korpus češtiny jakožto druhého jazyka (CzeSL), korpus soukromé korespondence (KSK) nebo text románu 1984 (ORWELL).

2.2 Lingvistická analýza

Aby bylo možné provádět na základě korpusů komplexní jazykový výzkum, musejí být texty anotovány, a to především morfologicky: každému slovu je přiřazen jeho základní tvar (lemma) a značka obsahující příslušné gramatické kategorie (slovní druh, rod, číslo, pád, osoba, čas, …). Je-li dané slovo homonymní s jiným, je ještě zapotřebí provést jeho zjednoznačnění (desambiguaci) – např. tvar „je“ může mít platnost zájmena i slovesa. Značkování má na starost sekce lingvistické analýzy ÚČNK a je zpravidla prováděno automaticky morfologickým analyzátorem. Objevila se již u prvního psaného korpusu češtiny (SYN2000), ale vzhledem k zastaralosti a nedokonalosti tehdejší automatické anotace byly výsledky mnohdy nespolehlivé. V současné době však úspěšnost dosahuje až 95 %.

Mimoto existuje anotace syntaktická, která označuje závislostní vztahy mezi slovy ve větě a skladebné funkce slov. Syntaktické značkování se poprvé objevilo v korpusu SYN2015.

Anotují se rovněž celé texty, a to pomocí strukturních atributů, které ohraničují například jednotlivé dokumenty, odstavce či věty. (5)

2.3 Nástroje

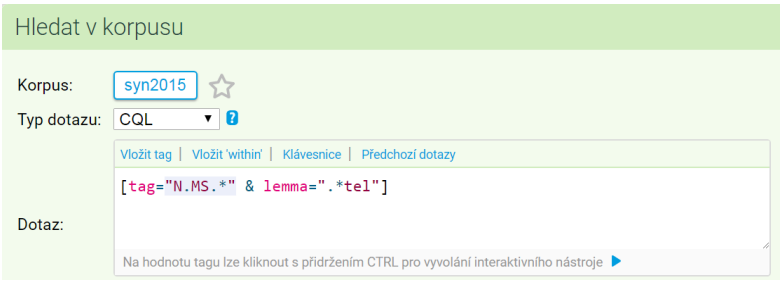
Jak již bylo zmíněno, ke korpusům ČNK se přistupuje primárně přes rozhraní Kontext. V rámci něj uživatel nejprve zvolí požadovaný korpus, ve kterém bude vyhledávání probíhat, a následně vybere typ dotazu – může jít o konkrétní slovní tvar, jeho část, frázi nebo lemma (typ lemma vyhledá všechny tvary daného slova, např. být – jsem, budu, bych, …).

Nejkomplexnější vyhledávání nabízí dotazovací jazyk CQL, pomocí nějž lze vyhledávat podle několika podmínek najednou, a to včetně morfologických či syntaktických kategorií u označkovaných korpusů. Výsledkem jsou slova zobrazená ve svém přirozeném kontextu a lze s nimi nadále

pracovat – zobrazovat jejich frekvenční seznamy podle různých kritérií (tvaru, lemmatu, morfologické značky, …), filtrovat a třídit je, hledat jejich nejčastější kolokace a výsledky ukládat v různých formátech (.csv, .xlxs, .xml, .txt).

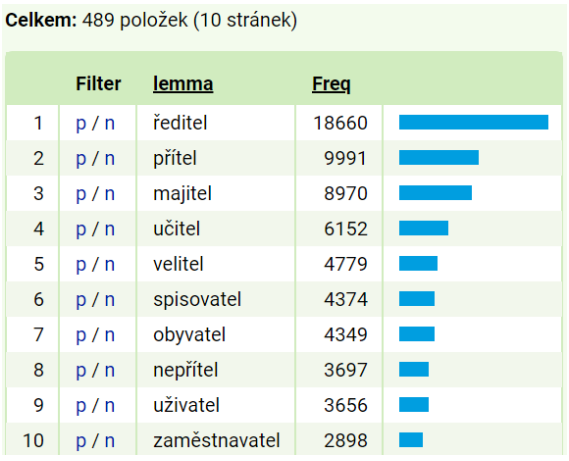
Podobně lze vyhledávat i v dalších nástrojích, zmíněných v předchozí kapitole. Vývoj těchto softwarů a správu IT vůbec má na starost komputační sekce ÚČNK. (4)

2.4 Příklad jednoduchého hledání



Obrázek 3: příklad hledání pomocí CQL dotazu v korpusu SYN2015 (substantiva rodu mužského životného v singuláru, jejichž základní tvar končí na sekvenci znaků tel*)*

Obrázek 4: ukázka části výsledku výše položeného dotazu



	Filter	lemma	Freq	
1	p / n	ředitel	18660	
2	p / n	přítel	9991	
3	p / n	majitel	8970	
4	p / n	učitel	6152	
5	p / n	velitel	4779	
6	p / n	spisovatel	4374	
7	p / n	obyvatel	4349	
8	p / n	nepřítel	3697	
9	p / n	uživatel	3656	
10	p / n	zaměstnavatel	2898	

Obrázek 5: ukázka části frekvenčního seznamu lemmat výše položeného dotazu

3 Aktuální stav

ČNK je projekt neustále modernizovaný a vylepšovaný. Průběžně se aktualizují různé jeho složky, ať už jde o tvorbu nových korpusů (v roce 2017 bylo zveřejněno 5 nových korpusů, celkově jich je 39 a obsahují téměř pět miliard slov), tak o vydávání nových verzí nástrojů – poslední aktualizace proběhla 30. října 2018, v rámci níž byla zveřejněná verze 0.12.0 rozhraní Kontext, která přinesla změny zefektivňující vyhledávání. (6) Jen o pár dní dříve vyšla 11. verze paralelního korpusu InterCorp, která tento korpus rozšířila a nově označkovala některé jazyky. Podobné aktualizace vycházejí několikrát do roka.

Samotný ÚČNK pravidelně vydává koncepce rozvoje a v té nejnovější (2016–2019) (7) hned v úvodu stanovuje, že jeho hlavním úkolem je péče o rozvoj ČNK, považuje jej za centrální bod svých aktivit. To konečně potvrzuje i snahou o popularizaci celého projektu – každým rokem se koná bezplatný korpusový workshop pro všechny zájemce; v roce 2018 proběhl 3. listopadu. Mimoto se může každý zájemce zúčastnit online kurzu práce s korpusem, který je v sedmi základních a čtyřech bonusových lekcích dostupný na http://wiki.korpus.cz/doku.php/kurz:uvod. Na téže doméně jsou rovněž k dispozici veškeré důležité a aktuální informace o službách ČNK, o korpusech jako takových a o vyhledávání v nich. K prostudování slouží i užitečný slovníčku pojmů.

Součástí projektu je také uživatelská podpora (https://podpora.korpus.cz/) otevřená všem uživatelům, která obsahuje poradnu umožňující pokládat jakékoliv dotazy týkající se práce s ČNK a také možnost hlášení chyb a/nebo návrhů na vylepšení.

Většina rozpočtu ČNK pochází v současné době z projektu Velkých infrastruktur pro výzkum, vývoj a inovace Ministerstva školství, mládeže a tělovýchovy (číslo projektu LM2011023). V rámci tohoto projektu získal ČNK velmi dobré hodnocení a splnil podmínku pro zajištění dalšího financování až do roku 2022. (7)

4 Vlastní hodnocení

V rámci svého studia českého jazyka se specializací počítačová lingvistika pracuji s ČNK, především s rozhraním Kontext, pravidelně. Absolvoval jsem několik předmětů zaměřených na využití korpusů, minulý rok jsem se zúčastnil workshopu zmiňovaného v předchozí kapitole a konkrétně na korpusech SYN2015 a SYN v4 je založena má právě vznikající bakalářská diplomová práce. A předpokládám, že i v budoucnu se bude oblast mého studia zaměřovat právě na korpusovou lingvistiku.

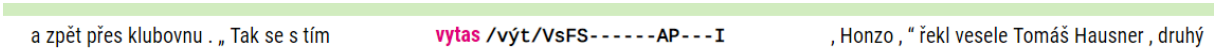
Kromě toho služby ČNK hojně využívám i v osobním životě, kupříkladu když se chci přesvědčit o frekvenci používání určitého slova či slovního spojení nebo když rozsuzuji diskuse svých známých o tom, který tvar slova je používanější (k tomu slouží nástroj SyD) nebo ve kterých typech textů se daný jev vyskytuje častěji. Ve většině případů najdu, co hledám, a zajímavé výsledky mě mnohdy překvapí.

Proto mohu konstatovat, že služby ČNK jsou na kvalitní úrovni a splňují svůj cíl. Přínos tohoto projektu pro lingvistiku je nepostradatelný a přínos pro laickou veřejnost značně obohacující.

Hlavní nedostatek spatřuji v automatické morfologické analýze, která v některých případech nedospěje ke správnému zařazení slova do všech kategorií, což může zkreslovat výsledky. U homonymních tvarů je to pochopitelně náročnější, ale někdy se chyba vyskytne i u na první pohled jasných záležitostí. Avšak s přihlédnutím k tomu, jakým tempem morfologická analýza pokročila za poslední roky, očekávám neustálé zdokonalování i nadále.

Rovněž bych uvítal uživatelsky přívětivější zobrazení nástrojů na mobilních zařízeních. Jak jsem se však dočetl v koncepci rozvoje ÚČNK, na tento problém je pamatováno a snad by taktéž mohl být brzy vyřešen.

Vzhledem k tomu, jak svědomitě je ČNK spravován nyní, věřím, že přicházet postupně budou i další vylepšení.



Obrázek 6: ukázka špatně označkovaného slova (místo náležitého lemmatu vytasit *bylo mylně určeno lemma* výt*)*

Bibliografie

1. Typy korpusů. *Wiki Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] https://wiki.korpus.cz/doku.php/pojmy:korpus#typy_korpusu.
2. Korpus a jeho využití. *Wiki Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] https://wiki.korpus.cz/doku.php/pojmy:korpus#korpus_a_jeho_vyuziti.
3. Korpus SYN. *Wiki Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] https://wiki.korpus.cz/doku.php/cnk:syn.
4. Profil ústavu. *Ústav Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] https://ucnk.ff.cuni.cz/cs/ustav/profil-ustavu/.
5. Anotace. *Wiki Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] https://wiki.korpus.cz/doku.php/pojmy:anotace.
6. Kontext verze 0.12.0. *Wiki Českého národního korpusu.* [Online] [Citace: 6. 11. 2018.] http://wiki.korpus.cz/doku.php/seznamy:kontext_verze#verze_0120.
7. Mgr. Michal Křen, Ph.D. Koncepce rozvoje Ústavu Českého národního korpusu. [Online] [Citace: 6. 11. 2018.] https://sites.ff.cuni.cz/ucnk/wp-content/uploads/sites/58/2016/10/kren_koncepce_rozvoje_2015-1.pdf.
8. Český národní korpus. [Online] [Citace: 6. 11. 2018.] https://korpus.cz/.

Seznam obrázků

Metadata v Dublin Core

```
<dc:title>Český národní korpus</dc:title>
<dc:creator>Filip Kubeček</dc:creator>
<dc:subject>databáze jazykových korpusů</dc:subject>
<dc:description>Esej představuje cíle, popis, aktuální stav a vlastní hodnocení Českého národního korpusu, digitální databáze korpusů, v níž je možno vyhledávat různé jazykové jevy a provádět výzkumy jazykové empirie.</dc:description>
<dc:date>2017-11-06</dc:date>
<dc:type>Text</dc:type>
<dc:format>public</dc:format>
<dc:identifier>http://korpus.cz</dc:identifier>
<dc:language>cz</dc:language>
```