Korpusové nástroje pro studenty a lektory jazyků (Miroslav Pošta)

Zdroj: <https://www.jazykovy-koutek.cz/?p=12272>

Jako překladatel, titulkář a redaktor už si neumím představit profesní život bez jazykových technologií. Některé z nich používám od 90. let, kdy ještě byly v plenkách, a v posledních letech s radostí sleduji, jak například s jazykovými korpusy pracuje čím dál tím víc překladatelů. Zdá se mi ale, že korpusy a další užitečné „vynálezy“ zatím moc nevyužívají studenti a lektoři jazyka. Pokusím se jim ukázat aspoň některé z nástrojů, které jim studium či výuku může usnadnit.

Řeč bude hlavně o korpusech, tedy o velkých elektronických archivech autentických textů, které bývají speciálně zpracované, aby v nich mohli lingvisté a zájemci o jazyk snadno hledat nejrůznější slova a jevy. Díky tomu můžeme například zjistit, která slova jsou v jazyce nejčastější, jestli je v češtině běžnější spojení *z třetí*, nebo *ze třetí*, s jakými přídavnými jmény se pojí anglické slovo *example*, můžete si vyjet věty se španělským slovem *ser*v různých tvarech (těch tvarů je přes 60 – je to vůbec možné?) nebo výskyty chorvatského slova *stvar*(věc) v 2. pádě množného čísla (protože si tímto tvarem v danou chvíli vůbec nejste jisti).

Tímto způsobem můžeme v „profesionálních“ korpusech vyhledávat právě proto, že jsou speciálně zpracované: v jejich útrobách je u každého slova uvedeno, jaký je jeho základní „slovníkový tvar“ (tzv. lemma) a jaké je gramatické zařazení daného tvaru – to je obsaženo ve značce, angl. tagu. Třeba tvar *věcí*ve 2. pádě množného čísla má v českém korpusu Syn2015 značku „NNFP2—–A—–“. Pojďme si ji rozluštit: N na první pozici znamená, že to je podstatné jméno, F na třetí pozici značí ženský rod, P na čtvrté pozici množné číslo, 2 na páté znamená 2. pád. (Na stránkách ČNK samozřejmě ke značkám najdete legendu.) A díky tomu, že je v útrobách korpusu u každého slova lemma a tag, je můžeme zadávat při hledání. Když zadáme lemma *věc*, najdeme nejen tvar 1. pádu jednotného čísla, ale i všechny ostatní tvary: *věci, věcí, věcmi*atd. Stejně tak můžeme hledat třeba všechna podstatná jména ženského rodu v 2. pádě množného čísla. To všechno navíc můžeme kombinovat, takže můžeme hledat třeba všechna přídavná jména zakončená v základním tvaru na *–ičí*.

Ukážu vám několik „kouzel“, a to hlavně v Českém národním korpusu, který se skládá z mnoha dílčích korpusů nejen češtiny, ale i řady dalších jazyků.

**Seznamy slov**

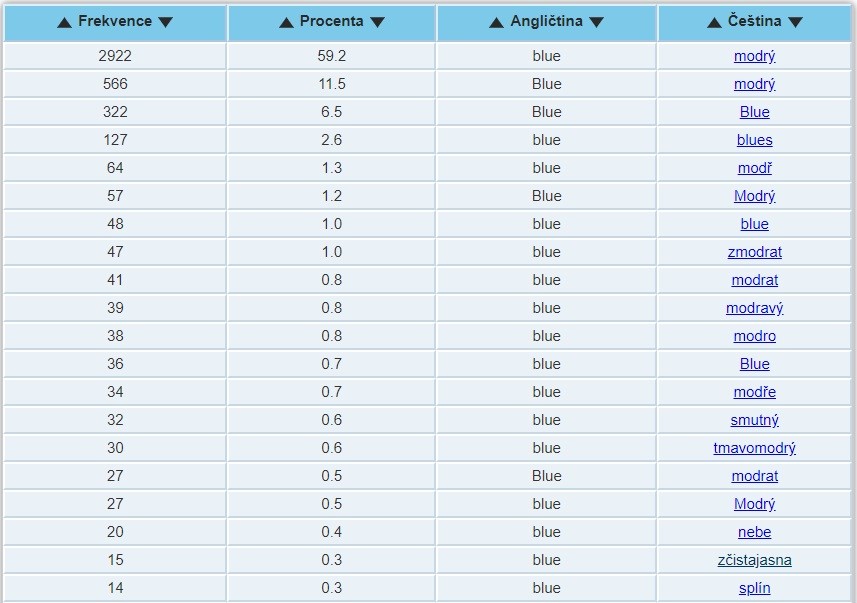
Jednou ze základních funkcí korpusů je tvorba frekvenčního seznamu slov. Na jeho základě si můžete udělat „inventuru“ slovní zásoby nebo se cíleně učit frekventovaná slova. Musíte ale samozřejmě zvolit správný korpus: když třeba zvolíte korpus textů EU, mezi nejčastější slova nejspíš bude patřit *nařízení* a *směrnice*, která vás možná úplně nezajímají.



***Obrázek.****Začátek seznamu slov extrahovaný ze španělského korpusu Araneum Hispanicum Maius.****[[i]](https://www.jazykovy-koutek.cz/?p=12272" \l "_edn1)****Na prvních místech jsou logicky samá „gramatická“ slova.*

**Překladové ekvivalenty**

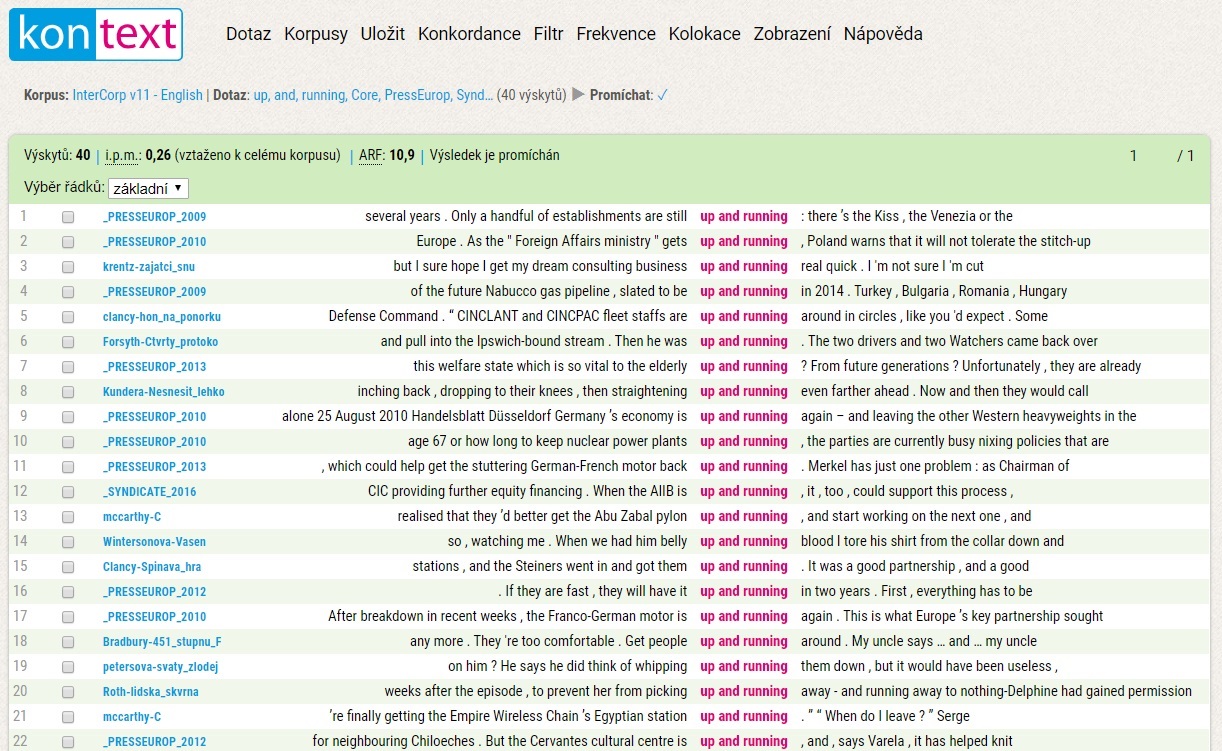
Mezi mé nejoblíbenější nástroje patří Treq, což je nástroj vytvořený na základě paralelních, tedy vícejazyčných korpusů. Je to vlastně překladový slovník, ve kterém můžete hledat nejenom ekvivalenty lemmat (slovníkových tvarů), ale i konkrétních tvarů. Takže když např. v chorvatštině zadáte *radio*, získáte jako ekvivalenty nejen *rádio*, ale i *dělal, pracoval*(*radio*je tvar slovesa *raditi*, tedy pracovat, dělat, fungovat). A často v ekvivalentu najdete i součást idiomu. Třeba jako jeden z ekvivalentů anglického *blue*získáte *zčistajasna,*a když si výsledek rozkliknete (čímž se dostanete do korpusu samotného), zjistíte, že jde o případy, kdy se slovo *blue*vyskytuje ve spojení *out of the blue*. A hledat koneckonců můžete i víceslovná spojení, tedy i *out of the blue*. Ekvivalenty jsou automaticky extrahované z korpusu na základě pokročilých algoritmů, takže klasický slovník plně nenahradí, ale i tak je Treq velmi užitečný, zvlášť u jazyků, pro které elektronický překladový slovník neexistuje a pro které je v ČNK dostatečně velký korpus.



***Obrázek.****Treq:****[[ii]](https://www.jazykovy-koutek.cz/?p=12272" \l "_edn2)****české ekvivalenty anglického slova*blue. *Povšimněte si mj. slovesa*zmodrat*, ale i slov*smutný, zčistajasna *a*splín.

**Hledání v korpusu: příkladové věty aj.**

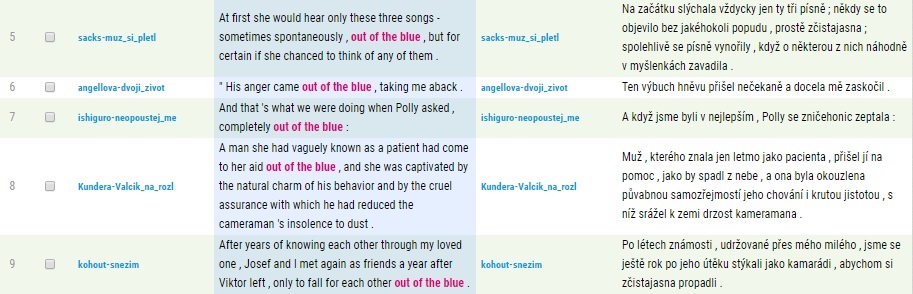
Konečně se dostávám k něčemu, co spousta uživatelů v korpusu oceňuje nejvíc, a to k vlastnímu vyhledávání v korpusu. Můžeme v něm totiž velmi rychle najít všechny výskyty libovolného slova, slovního tvaru, části slova a gramatické kategorie. A samozřejmě jejich kombinací. Mluvíme o takzvaném konkordování: nalezené doklady totiž program zobrazí v takzvané konkordanci. Klasická konkordance vypadá tak, že na každém řádku je zobrazen jeden výskyt hledaného jevu. Vidíme jej vždy uprostřed a nalevo a napravo je zobrazen nejbližší kontext.



***Obrázek.****Část konkordance vytvořené v (anglické složce) korpusu InterCorp.****[[iii]](https://www.jazykovy-koutek.cz/?p=12272" \l "_edn3)***

Tímto způsobem můžeme najít věty obsahující hledané slovo, sousloví nebo zkrátka jev, když si například chceme určité slovo zapsat do slovníčku s typickým kontextem, abychom se ho neučili osamoceně. Nebo jak už jsem řekl, můžeme srovnávat dva jevy, abychom zjistili, který je běžnější.

Podobným způsobem můžete vyhledávat nejen v jednojazyčném korpusu, ale i v korpusu paralelním. Pak vidíme i odpovídající větu v druhém jazyce.



***Obrázek.****Část paralelní konkordance anglicko-česká z korpusu InterCorp.****[[iv]](https://www.jazykovy-koutek.cz/?p=12272" \l "_edn4)***

**Kolokace**

Velmi užitečné je vyhledávání kolokátů, tedy slov, která s určitým slovem mají s ním nějaký nadstandardně silný vztah. Právě zvládnutí kolokací je nezbytné k tomu, abychom dosáhli opravdu vysoké úrovně znalosti jazyka.

V ČNK si nejprve necháme zobrazit konkordanci, potom klikneme na Kolokace > Vlastní…, vybereme rozmezí, ve kterém nás kolokace zajímají (např. -3 až 3 znamená od třetího slova vlevo po třetí slovo vpravo) a necháme vytvořit seznam kolokátů. Když nás například bude zajímat, jaká slova často stojí vlevo od anglického slova *example*, získáme v anglickém korpusu Araneum Anglicum Maius[[v]](https://www.jazykovy-koutek.cz/?p=12272" \l "_edn5) takovýto výsledek.

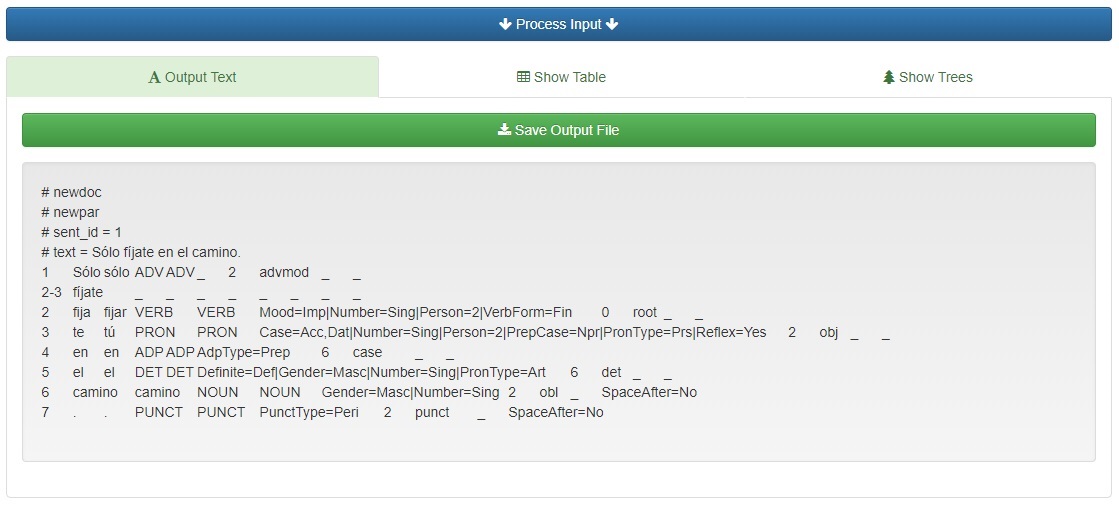


***Obrázek.****Kolokáty slova*example*v korpusu Araneum Anglicum Maius.*

**Nechte si označkovat větu**

Teď opustíme Český národní korpus a ukážu vážným zájemcům nástroj, který vlastně funguje úplně obráceně. Řekněme, že začínám se španělštinou, mám ve větě slovíčko, které za boha nemůžu najít v žádném slovníku a usoudím, že to bude nějaký tvar slovesa. Konkrétně mám větu *Sólo fíjate en el camino* a tím problémovým slovem je *fíjate.*

Větu můžu zadat do [nástroje UDPipe](http://lindat.mff.cuni.cz/services/udpipe/), vybrat španělštinu a kliknout na Process input. UDPipe mi ke každému slovu přiřadí lemma a značku, takže během okamžiku vím, že jde o tvar slovesa *fijar*spojený s tvarem zájmena *tú*.



***Obrázek.*** *Španělská věta analyzovaná pomocí nástroje UDPipe.*

Musím dodat, že lemmatizace a značkování je automatický proces, který není vždy bezchybný, a proto není stoprocentně spolehlivý. To se týká i všech korpusů, kde některá slova a jevy nejsou správně označkovány, většinou se ale jedná o zanedbatelné procento.

**Zkuste si to sami**

V tomto článku jsem mohl nabídnout jen drobnou ochutnávku jednotlivých nástrojů. Pokud budete chtít využít celý potenciál korpusů, budete se muset naučit klást pokročilé dotazy. Není to složité, ale vím, že mnoha humanitně orientovaným uživatelům se do toho příliš nechce. Proto vězte, že korpusy vám přinesou mnoho užitku, i když budete zadávat jen základní dotazy typu slovní tvar, fráze a lemma. Zkrátka to zkuste. [Český národní korpus](https://korpus.cz/) je uživatelsky příjemný a je zdarma. K využívání nástroje Treq se ani nemusíte registrovat, k plnému přístupu k vlastním korpusům je pak nutná registrace.

Český národní korpus samozřejmě není jediný korpus (nebo rodinka korpusů) na světě. Hledat můžete korpusy „svého“ jazyka v příslušných zemích. Některé jsou volně přístupné, jiné placené. Několik anglických korpusů najdete například na stránkách <https://www.english-corpora.org/>, několik španělských je tady: <https://www.corpusdelespanol.org/>. Doporučuji i [korpus SkELL, určený přímo pro studium jazyka](https://skell.sketchengine.co.uk/run.cgi/skell), (na stránce najdete i odkazy na korpus ruský, německý, italský, estonský a český) a pro velmi vážné zájemce i placené korpusy na [portálu Sketch Engine](https://www.sketchengine.eu/) (a vězte, že studenti mnoha univerzit do něj mají přístup zdarma).

Ale zpět k Českému národnímu korpusu. Na základní úrovni se jej naučíte ovládat velmi rychle. Doporučuji [wiki ČNK](https://wiki.korpus.cz/doku.php), kde je nejen přehled jednotlivých korpusů, ale i kurz práce s korpusem v sedmi lekcích. Pro vážné zájemce Ústav Českého národního korpusu pořádá i workshopy.

Anebo můžete zkusit mé články, knížku a přednášky, které jsou šité na míru překladatelům. Stejné postupy ale vlastně můžou využít i studenti a učitelé jazyků, jen jim technologie poslouží k trochu jinému účelu. To nejdůležitější jsem se snažil shrnout v několika [článcích na Translatoblogu](https://www.translatoblog.cz/tag/korpusy/), v první části knihy [*Technologie ve službách překladatele*](http://eshop.jazykovy-koutek.cz/odborne-knihy-pro-prekladatele/technologie-ve-sluzbach-prekladatele/) a ve dvou [přednáškách](http://www.translatoblog.cz/vyuka/). Najdete tam i informace o tom, jak si vytvořit vlastní malý korpus nebo jak si vyrobit bilingvní text zarovnaný po větách – bilingvní „zrcadlové“ publikace přece jen vycházejí jen pro největší jazyky a třeba už u španělštiny je nabídka velmi skromná.

Miroslav Pošta

**Miroslav Pošta** je titulkář, překladatel, redaktor a nakladatel. Je autorem praktických knih pro překladatele *Titulkujeme profesionálně,*[*Technologie ve službách překladatele*](http://eshop.jazykovy-koutek.cz/odborne-knihy-pro-prekladatele/technologie-ve-sluzbach-prekladatele/) a [*Titulkujeme: Audiovizuální překlad v otázkách a odpovědích*](http://eshop.jazykovy-koutek.cz/odborne-knihy-pro-prekladatele/titulkujeme--audiovizualni-preklad-v-otazkach-a-odpovedich/)*.*Píše blog pro překladatele [TranslatoBlog](http://www.translatoblog.cz/). O praktických aspektech jazykových technologií a tvorby titulků také přednáší.

[[i]](https://www.jazykovy-koutek.cz/?p=12272" \l "_ednref1) Benko, V.: *Araneum Hispanicum Maius, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný z WWW: http://www.korpus.cz

[[ii]](https://www.jazykovy-koutek.cz/?p=12272" \l "_ednref2) Vavřín, M. – Rosen, A.: Treq. FF UK. Praha 2015. Dostupný z WWW: *„http://treq.korpus.cz“*.

[[iii]](https://www.jazykovy-koutek.cz/?p=12272" \l "_ednref3) Klégr, A. – Kubánek, M. – Malá, M. – Rohrauer, L. – Šaldová, P. – Vavřín, M.: *Korpus InterCorp – angličtina, verze 11 z 19. 10. 2018*. Ústav Českého národního korpusu FF UK, Praha 2018. Dostupný z WWW: http://www.korpus.cz

[[iv]](https://www.jazykovy-koutek.cz/?p=12272" \l "_ednref4) Klégr, A. – Kubánek, M. – Malá, M. – Rohrauer, L. – Šaldová, P. – Vavřín, M.: *Korpus InterCorp – angličtina, verze 11 z 19. 10. 2018*. Ústav Českého národního korpusu FF UK, Praha 2018. Dostupný z WWW: http://www.korpus.cz

[[v]](https://www.jazykovy-koutek.cz/?p=12272" \l "_ednref5) Benko, V.: *Araneum Anglicum Maius, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný z WWW: http://www.korpus.cz