**Pojďte si hrát s korpusy a korpusovými nástroji (Treq a Skell – jak využít korpus při překladu a studiu lexika L2)[[1]](#footnote-1)**

***Klára Osolsobě***

**osolsobe@phil.muni.cz**

**Anotace**

Cílem přednášky bude prezentace dvou korpusových nástrojů, které nabízejí uživatelům podporu při překládání, popřípadě při pokročilém studiu jazyka. První z nich je založen na paralelních korpusech budovaných v rámci projektu Intercorp Ústavu Českého národního korpusu (<http://ucnk.korpus.cz/intercorp/>)[[2]](#footnote-2), druhý je produktem firmy Lexical Computing, rovněž využívá jazykové korpusy ke studiu L2 (<https://www.sketchengine.eu/skell/>).[[3]](#footnote-3)

**Úvod**

Studium/výuku L2 a překlad si v dnešní době lze jen těžko představit bez využití nejrůznějších počítačových nástrojů a podpor. Překladatelé mohou využívat a využívají nejen množství elektronických slovníků a podpory opírající se o strojový překlad, ale existují a budují se i tzv. paralelní korpusy ([http://wiki.korpus.cz/doku.php/pojmy:paralelni](http://wiki.korpus.cz/doku.php/pojmy%3Aparalelni)).

**Krátký historický exkurz**

Hovoříme-li o korpusové lingvistice, pak máme většinou na mysli poměrně mladé odvětví spojené s technickými možnostmi, které se objevují v 50. letech XX. století a rozšiřují se od 80. let téhož století.

Je ovšem také třeba si uvědomit, že myšlenka korpusu je mnohem starší. Již ve strukturalistických přístupech k jazyku, v myšlence toho, že systém (langue) poznáváme na základě pozorování jeho realizací na rovině parole (promluv, textů) je patrná idea korpusu (parolových dat) a procedury (pravidel, kterými jsou vydělitelné jednotky jako hláska, morf, ... a vztahy mezi nimi – např. dvojice lišící se distinktivními rysy).

Ještě starší jsou jazykové analýzy opřené o textové korpusy **(papírové korpusy)**. Jeden ze základních termínů korpusové lingvistiky – **konkordance** je termín převzatý z **biblické kritiky**.[[4]](#footnote-4) V korpusové lingvistice se termínem konkordance označuje **soupis všech výskytů klíčového slova** (může jít ovšem i o klíčové spojení více slov) **ve všech kontextech**, v nichž se v korpusu vyskytuje.

Typické bylo užití korpusů (sbírek textů) v **lexikografické praxi**. Podíváme-li se na slavného slovníkáře **Samuela Johnsona (1709-1784)**, tak je známo, že ve výkladu používal doklady z pečlivě sestaveného korpusu. Podobně když se podíváme na **Příruční slovník jazyka českého (PSJČ)**[[5]](#footnote-5), tak vidíme, že u řady dokladů jsou zkratky odkazující ke jménům klasiků české obrozenecké literatury. Slovníkářská praxe se dnes zcela běžně opírá o korpusy, existuje tzv. korpusová lexikografie.[[6]](#footnote-6)

**Korpus a internet**

V dnešní době představuje asi největší sbírku elektronicky uložených textů internet. V čem je jazykový korpus lepší?

Jednou z podmínek, kterou musí korpus jako zdroj dat pro studium jazyka splňovat, je v moderní době elektronické uložení a elektronická přístupnost jazykových dat, která korpus tvoří, dále vymezený obsah a rozsah jazykových dat, která korpus tvoří a standardní anotace, které data v korpusu popisují. Tyto podmínky (vymezený rozsah a obsah a standardní anotace) nesplňuje internet. Pro češtinu existuje v současnosti velké množství jazykových korpusů, které tuto podmínku splňují.

**Význam slov prozradí jejich okolí**

Nejobvyklejším základním nástrojem pro práci s jazykovým korpusem je korpusový vyhledávač (manažer), který v zásadě umožňuje klást dotazy na jazykové jednotky a zobrazovat jejich výskyt ve formátu KWIC (**k**ey **w**ord **i**n **c**ontext).

V jazyce platí podobné zákonitosti, jaké sleduje psychologie osobnosti: *„Řekni mi, kým se obklopuješ, a já ti řeknu, jaký jsi“*, **o významu slova ti leccos napoví jeho kontext**.

Víte například, co je to ***hořčák***? Zadáme-li lemma *hořčák* do vyhledávače google, pak na první stránce narazíme na obrázky, z nichž je patrné, že jde o druh houby, týž význam nabízí i česká Wikipedie.

Jak si ale poradíte s tím, když Vás někdo pozve na *hořčák*?

Na obrázku 1 vidíme, jak vypadá konkordance klíčového lemmatu *hořčák* z korpusu SYN2015[[7]](#footnote-7).

Obrázek 1



Prostým pročtením 12 kontextů můžeme odlišit tři různé významy:

1. *hořký alkoholický nápoj* (obrázek 2)

Obrázek 2



1. *hořká houba* (obrázek 3)

Obrázek 3



1. *hořkost* (obrázek 4)

Obrázek 4



Podobný postup můžeme zopakovat například se substantivem *svízel*. Na obrázku 5 vidíme část konkordance lemmatu *svízel* z korpusu SYN2015.

Pročítáme-li konkordance, narazíme na doklad, z jehož okolí se dozvíme, že slovo se v češtině vyskytuje ve dvou různých významech, které se navíc liší gramatickým rodem (obrázek 6). První význam ilustrují řádky 1, 3, 4, 5, 7, 8, 9, druhý význam řádky 2, 6, 10, 11 na obrázku 5.

Obrázek 5



Obrázek 6



***Treq* nástroj pro překladatele**

Kromě jednojazyčných korpusů se budují i korpusy vícejazyčné, tzv. paralelní korpusy. Ve Wiki ÚČNK ([https://wiki.korpus.cz/doku.php/pojmy:paralelni](https://wiki.korpus.cz/doku.php/pojmy%3Aparalelni)) je paralelní korpus definován jako „Jeden z [typů korpusů](https://wiki.korpus.cz/doku.php/pojmy%3Akorpus#typy_korpusu), jehož účelem je rozšířit metodologii, propracovávanou v rámci korpusové lingvistiky původně pouze pro jeden jazyk, také na výzkum kontrastivní (mezijazykové srovnání). Paralelní korpusy se dělí na srovnatelné (angl. comparable) a překladové (angl. translational).“ Překladové paralelní korpusy (dvou jazyků) nemusí obsahovat pouze originály a překlady, ale mohou obsahovat též dvojici překladů stejného textu ze třetího jazyka. Jako projekt ČNK vznikl vícejazyčný překladový paralelní korpus [InterCorp](https://wiki.korpus.cz/doku.php/cnk%3Aintercorp) ([https://wiki.korpus.cz/doku.php/cnk:intercorp](https://wiki.korpus.cz/doku.php/cnk%3Aintercorp)). Nad korpusem InterCorp byla automaticky vytvořena databáze česko-cizojazyčných a anglicko-cizojazyčných ekvivalentů *Treq* (<http://treq.korpus.cz/>)*.* Na obrázku 7 vidíme příklad anglických ekvivalentů slova *svízel*.

Obrázek 7



Na obrázku 8 vidíme naopak ekvivalenty anglického *trouble*. Nejobvyklejším ekvivalentem v příslušné databázi je slovo *problém.* Povšimněme si, že slovo *svízel* se mezi frekventovanými ekvivalenty nevyskytuje.

Obrázek 8



Můžeme se podívat i na kontexty, v nichž se jako ekvivalent anglického *trouble* (obrázek 9) slovo *svízel* (obrázek 10) objeví.

Obrázek 9



Obrázek 10



***Skell* (Sketch Engine for Language Learning) nástroj pro studenty L2**

*Skell* (https://csskell.sketchengine.co.uk/run.cgi/skell) je jednoduchý nástroj pro studenty a učitele cizích jazyků, prostřednictvím kterého mohou zjistit, zda se nějaké slovo, či fráze užívají a jak je užívají rodilí mluvčí. Umožňuje automatickou extrakci zadaných slov v kontextech, v nichž byla užita rodilými mluvčími. Automatickými metodami založenými na statistickém zpracování jazyka vybírá a nabízí uživateli podobná slova. Nástroj pracuje nad databází obsahující 1,4 miliardy slov ve více než 90 milionech vět. Texty pocházejí z webové domény archivované Českým Webarchivem, z článků a diskusí české Wikipedie a z webového korpusu novinek Institutu Jožefa Štefana v Lublani.[[8]](#footnote-8)

Význam adjektiva *jízlivý* pomocí vhodných příkladových vět vidíme na obrázku 11.

**Word sketch** = kolokační profil, tj. nejčastější vazby zachycující kolokační a gramatické chování slova *jízlivý* vidíme na obrázku 12.

Na obrázku 13 jsou blízká slova, tj. nejen synonyma, ale i slova vyskytující se v podobných kontextech jako adjektivum *jízlivý*.

Obrázek 11



Obrázek 12



Obrázek 13



**Závěr**

Naším cílem bylo ukázat, jak fungují v praxi dva nástroje založené na korpusech, které lze používat jako pomůcku při učení se češtině i při překládání do/z češtiny.

V rámci přednášky bude prostor na další praktické ukázky práce s oběma nástroji i na diskusi.

Ti, kteří se přednášky účastní, si mohou dopředu připravit jedno až tři slova, jejichž význam přibližně znají. Pokusí se popsat jeho význam, přičemž se zaměří na slova významově blízká.

Součástí přednášky pak bude ukázka analýzy korpusových konkordancí uvedených slov i užití nástrojů *Treq* a *Skell*.

**Bibliografie**

Cvrček, V. – Čermáková, A. – Křen, M. (2016): Nová koncepce synchronních korpusů psané češtiny. Slovo a slovesnost, 77 (2), 83–101.

Charciarek, A. (2018): Možnosti využití korpusu InterCorp v česko-polské překladové lexikografii. *Časopis pro moderní filologii*, 100 (2), 206-222.

Milena Hnátková, Michal Křen, Pavel Procházka, Hana Skoumalová (2014): The SYN-series corpora of written Czech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), s. 160–164. Reykjavík: ELRA. <http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf>

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: ten years on. Lexicography, 1, s. 7–36.

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A. (2016): SYN2015: Representative Corpus of Contemporary Written Czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2522–2528. Portorož: ELRA. http://www.lrec-conf.org/proceedings/lrec2016/pdf/186\_Paper.pdf

Škrabal, M. – Vavřín, M. (2017): Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii*, 99 (2), 245-260.

**Elektronické korpusy a korpusové nástroje**

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A.: SYN2015: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný z WWW: http://www.korpus.cz

Nástroj *KonTetxt* dostupný z: WWW: <https://kontext.korpus.cz/first_form>.

Nástroj SkELL dostupný z: WWW: <https://www.sketchengine.co.uk/skell/>.

# Příruční slovník jazyka českého (1935–1957) dostupný z: WWW: <https://psjc.ujc.cas.cz/>.

1. Text vznikl za podpory projektu **MUNI/A/1061/2018** Čeština v jednotě synchronie a diachronie – 2019. [↑](#footnote-ref-1)
2. „Paralelní korpus slouží jako zdroj dat pro teoretické studie, lexikografii, studentské práce, výuku, zejména výuku cizích jazyků, počítačové aplikace, překladatele i veřejnost. Korpus se buduje s účastí řešitelských pracovišť a je dostupný on-line pro registrované uživatele.“ Viz také podrobně [http://wiki.korpus.cz/doku.php/cnk:intercorp](http://wiki.korpus.cz/doku.php/cnk%3Aintercorp). [↑](#footnote-ref-2)
3. Sketch Engine je nejmodernější online nástroj pro budování, správu a prohledávání velkých textových kolekcí (tzv. korpusů) v desítkách jazyků. Tento nástroj využívají po celém světě jednotlivci i společnosti, kupříkladu Cambridge University Press, Oxford University Press nebo Macmillan. [↑](#footnote-ref-3)
4. Šlo o alfabetický soupis slov Bible s odkazy na všechna místa, kde se vyskytují. Proč? Např. tzv. synoptická evangelia – Mk., Mt., Lk., líčí paralelně události života Kristova, podobně existují odkazy na biblické texty SZ v NZ. První biblickou konkordanci sestavili dominikáni k latinské Vulgátě (Hugo de Sancto Charo, vyšlo v roce 1230, později vznikly konkordance pro hebrejskou Bibli a pro Septuagintu). Pro angličtinu vznikly první konkordance v 16. stol. České konkordance jsou až z 20. stol. V dnešní době existují počítačové programy pro tvorbu biblických konkordancí. [↑](#footnote-ref-4)
5. Viz <http://www.ujc.cas.cz/elektronicke-slovniky-a-zdroje/Prirucni_slovik_jazyka_ceskeho.html>. [↑](#footnote-ref-5)
6. ###  Ke konkrétním výstupům bylo zaměřeno několik projektů nakladatelských domů Longman a Collins Cobuild – *Longman Dictionary of Contemporary English* a *Collins Cobuild English Language Dictionary*). Za vrchol se pokládá *OED2*. V českém prostředí vyšla péčí Ústavu Českého národního korpusu cela řada slovníků (autorské slovníky Karla Čapka a Bohumila Hrabala, frekvenční slovníky psané i mluvené češtiny a řada dalších, např. *A Frequency Dictionary of Czech: Core Vocabulary for Learners*, viz <https://ucnk.ff.cuni.cz/cs/veda-a-vyzkum/publikace/>).

 [↑](#footnote-ref-6)
7. SYN2015 je stomilionový korpus synchronních psaných českých textů (viz [https://wiki.korpus.cz/doku.php/cnk:uvod](https://wiki.korpus.cz/doku.php/cnk%3Auvod)). [↑](#footnote-ref-7)
8. Prohledávané texty nemusely projít jazykovou korekturou, mohou tudíž obsahovat pravopisné a stylistické chyby, jichž se dopouštějí rodilí mluvčí (skýtají tak reálný obraz jazyka). [↑](#footnote-ref-8)