K možnostem počítačového zpracování literárního textu

— Petr Pořízka —

Jednu z možností, jak dnes efektivně analyzovat literární text s použitím počítačů, nabízí korpusová lingvistika – obor, jehož hlavním předmětem zájmu jsou tzv. korpusy. Ty lze definovat jako metodologicky jednotné soubory textů (nejčastěji v elektronické verzi), sloužící k analýze prostřednictvím speciálních softwarových programů, tzv. konkordančních programů či korpusových manažerů.[1] Možnosti softwarových analýz jsou přitom určeny či podmíněny nejen počítačovými programy, ale (a to je třeba zdůraznit) i způsobem zpracování samotného textu. V zásadě lze pracovat s textem prostým (s původním

1 Jako korpusové manažery označujeme softwarové nástroje, jež umožňují komplexní práci s korpusem. Text je prostřednictvím tohoto nástroje zpracován a načten do implicitní databáze programu. Texty lze poté podle různých kritérií (filtrováním) prohledávat a analyzovat. Vyhledávat lze jednotlivé výrazy či slovní spojení, jež jsou zobrazeny i v omezeném kontextu (výsledkem jsou tzv. konkordance), vyhledávat lze také podle provedené anotace (viz pozn. 2) nebo s pomocí tzv. masek – speciálních metaznaků dotazovacího jazyka korpusového programu: blíže viz např. http://korpus.cz/bonito/regular.php. Provádět lze i statistické analýzy, vytvářet frekvenční či abecední slovníky ad. Standardně tyto nástroje nabízejí i možnost uložit výsledky analýz do externího samostatného souboru v některém z běžných textových formátů (*.txt*, *.rtf*, *.doc*).

textem, do nějž nejsou tvůrcem korpusu vnášeny dodatečné metatextové informace) nebo anotovaným (originální text je obohacen o metainformace vnětextového, ale i vnitrotextového charakteru).[2] Poměrně efektivně lze pracovat i s neanotovaným korpusem, provedená anotace ale může výrazným způsobem rozšířit možnosti analýz.

Během více než patnáctileté historie korpusové lingvistiky v České republice vznikla řada především jazykových/jazykovědných korpusů.[3] Postupně se ale metody a nástroje korpusové lingvistiky začínají aplikovat i na literárněvědné projekty. Nejznámějším a dosud bezesporu nejvýznamnějším literárněvědným korpusovým projektem je *Česká elektronická knihovna* (ČEK), fulltextová databáze české poezie 19. a počátku 20. století.[4]

Právě projekt ČEK nejlépe dokumentuje možnosti, jež počítačové zpracování textů a softwarové nástroje korpusové lingvistiky nabízejí. Uživatel může nejen vyhledávat výrazy prostým vepsáním do dotazového řádku webového rozhraní, ale lze vytvářet strukturované dotazy, tj. vyhledávat jak ve vybraných sbírkách, tak v jejich strukturách: básních, strofách, verších **[1]**.

Aplikace, jež spravuje textovou databázi, provádí i statistické analýzy – nabízí údaje o minimální, průměrné a maximální délce slova, verše či strofy **[2]**, umožňuje generovat frekvenční a abecední seznamy **[3]** a vyhledávat vazby slov a motivů.

Kromě toho lze volit mezi diplomatickou a ediční verzí textu nebo vpisovat uživatelské poznámky. Každý dokument je doplněn o ediční poznámky a ilustrace, jež jsou součástí knižního vydání. K textům je možno přistupovat skrze webové rozhraní s implementovaným korpusovým manažerem od firmy inSophy **[4]**.[5]

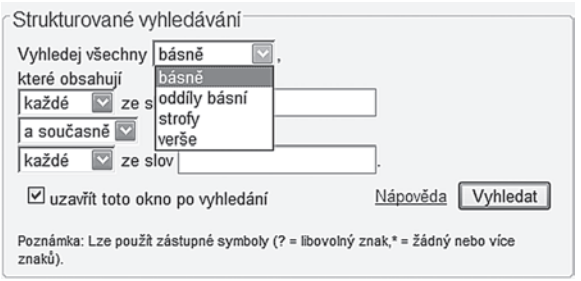
Jedním z projektů, jež by mohly být zajímavé pro literární vědce, je *Intercorp* – databáze paralelních korpusů obsahující literární

2 Anotací, její charakteristikou, možnostmi a typy jsme se blíže zabývali ve studii Pořízka – Schäfer 2010b.

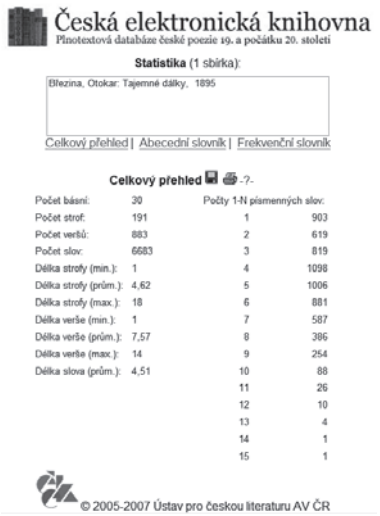
3 V roce 1994 byl založen Ústav Českého národního korpusu (ÚČNK), ale aktivity v oblasti počítačového zpracování jazyka vznikaly již řadu let předtím – mezi iniciátory patřili např. prof. František Čermák, doc. Karel Pala, doc. Vladimír Petkevič ad. Informace o stávajících korpusech ÚČNK lze nalézt na http://www.korpus.cz/struktura.php.

4 Iniciátory projektu byli Vladimír Macura a Pavel Janoušek; bližší charakteristika a popis projektu viz http://www.ceska-poezie.cz/cek/. Nutno ocenit, že stejně jako v případě korpusů ÚČNK je i k *České elektronické knihovně* poskytován bezplatný přístup. Databáze ČEK je o to cennější, že se dostaneme k plné verzi textů, na rozdíl od praxe ÚČNK, kdy je možno zobrazit v korpusech prostřednictvím korpusového manažeru jen hledaný výraz s omezeným kontextem.

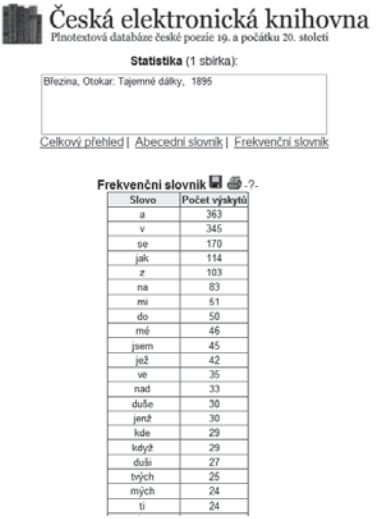
5 Http://www.insophy.cz/.



[1] Strukturované vyhledávání *České elektronické knihovny*



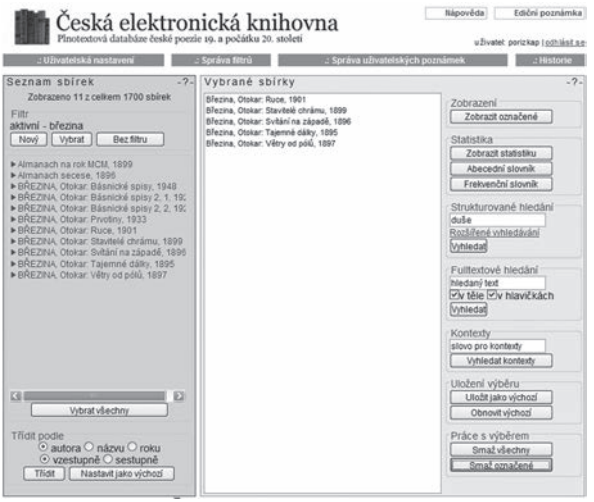
[2] Statistika *Tajemných dálek* Otokara Březiny



[3] Frekvenční slovník téže sbírky

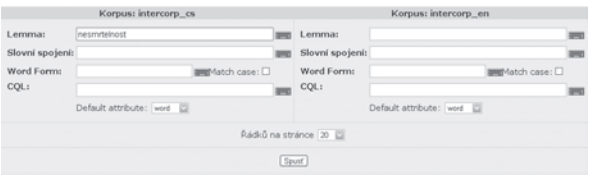
(beletristické) texty.[6] *Intercorp* vzniká v ÚČNK jako součást *Českého národního korpusu*. V současné době obsahuje texty z 22 jazyků, přičemž čeština má v korpusu pozici tzv. pivotu – česká verze (originál nebo překlad) je vztažena k jedné nebo více verzím cizojazyčným **[5]**.

6 Detailnější informace o projektu *Intercorp*, jeho struktuře, koncepci, užitých aplikacích, tvůrcích apod. lze nalézt na http://korpus.cz/intercorp-info.php, příp. http://www.korpus.cz/intercorp/. Na tvorbě tohoto paralelního korpusu se velkou měrou podíleli krom spolupracovníků ÚČNK pedagogové a studenti FF UK Praha.

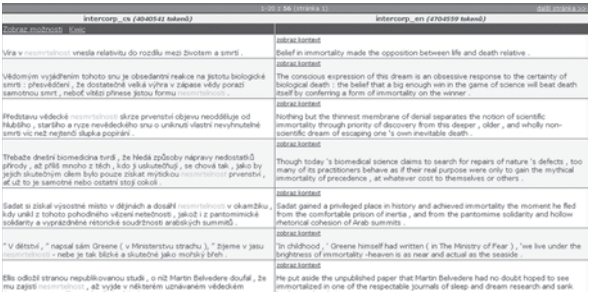


[4] Webové rozhraní (pracovní prostředí) *České elektronické knihovny*

[5] *Intercorp*: V levém sloupci je možno filtrem zvolit subkorpusy (zde český a anglický). V pravém bloku jsou texty dostupné v obou jazycích, v nichž je možno paralelně vyhledávat.



[6] Pracovní prostředí *Intercorpu* – dotazovací okno webového rozhraní *Park*



[7] Výsledek vyhledávání v *Intercorpu*: paralelně zarovnané konkordance českého a anglického subkorpusu.

Korpus je přístupný přes webové rozhraní *Park* (autor Michal Šťourač) **[6]**, jež je nadstavbou nejužívanějšího českého korpusového manažeru *Manatee*.[7] *Manatee* je komplexním korpusovým nástrojem a vždy záleží na tvůrcích daného korpusu, které možnosti programu využijí. Projekt *Intercorp* je zpracován způsobem, který umožňuje využít všechny základní funkce manažeru: je možné specifikovat prohledávané části korpusu – jazyky i konkrétní texty (použitím filtrů), vyhledávat podle řady kritérií – v jednom či více jazycích současně, podle slovního tvaru, frází či posloupností tvarů, podle dotazovacího jazyka programu *Manatee*,[8] podle lemmatu a morfosyntaktické značky (tagu).[9] Výsledky vyhledávání (konkordance) jsou zobrazeny jako paralelně zarovnané úseky textu ve zvolených jazycích. **[7]**

Statistické analýzy, mezi něž patří například výpočet absolutní a relativní frekvence výrazů, rozložení hledaného výrazu v korpusu apod., lze v programu *Manatee* aplikovat na jakýkoli text – anotovaný či prostý (neanotovaný), neboť jde o implicitní funkce, jejichž využití není přímo podmíněno mírou dodatečného zpracování textu.[10]

ÚČNK vydal i dvě lexikograficky zaměřené monotematické monografie – slovníky Karla Čapka a Bohumila Hrabala (Čermák 2007, Čermák – Cvrček 2009). Oba vyšly knižně s přiloženým CD, jež je de facto elektronickou verzí tištěné knihy, CD-ROM dokonce ve srovnání s knihou obsahuje korpus doplněný o lemmatizaci a morfologickou anotaci.[11] V obou případech se jedná v podstatě o abecedně uspořádaný frekvenční slovník. Charakteristiku obou slovníků (metodologicky se od sebe neliší) podává Štíchova recenze slovníku Karla Čapka

---

7 Autorem programu *Manatee* je Pavel Rychlý. Až na výjimky používají korpusový manažer *Manatee* všechny české korpusy, jde tedy o jakýsi český softwarový standard. Ke korpusům uloženým pod systém *Manatee* lze přistupovat dvěma způsoby: přes grafické uživatelské rozhraní *Bonito* (http://www.textforge.cz/products), nebo skrze novější verzi, webové rozhraní *Word Sketch Engine* (http://www.sketchengine.co.uk/, autoři Pavel Rychlý, Adam Kilgarriff a Jan Pomikálek). Zatímco systém *Manatee/Bonito* je poskytován zdarma, užití aplikace *Word Sketch Engine* je zpoplatněno. To je důvod, proč se i nadále ve většině korpusových projektů používá vývojově starší verze *Manatee/Bonito*.

8 Jde o tzv. *Corpus Query Language* (CQL), dotazovací jazyk vyvinutý na univerzitě ve Stuttgartu při práci na konkordančním programu *Xkwic*. Tento nástroj se stal základem systému *Manatee*. Vývoj projektu *Xkwic* v současné době pokračuje pod názvem *The IMC Open Corpus Workbench* – viz http://cwb.sourceforge.net/.

9 Otázkám lemmatizace a morfologického značkování jsme se věnovali ve studii Pořízka – Schäfer 2010a.

10 Zevrubnější popis základních statistických funkcí *Manatee/Bonita* viz http://korpus.cz/bonito/stat.php.

11 Nabízí se tedy i otázka ekonomického charakteru, neboť by slovníky mohly být vydány pouze na CD-ROMu, což by nepochybně přineslo velkou úsporu nákladů.

uveřejněná v *Naší řeči*: „Hlavní částí *Slovníku Karla Čapka* je abecedně uspořádaný slovník všech slov, která Čapek užil ve svém publikovaném díle literárním (próza, drama, poezie), odborném, v rozsáhlé publicistice i ve vydané soukromé korespondenci. U každého z těchto slov, řazených pod sebou ve dvou sloupcích na stránce, je v sedmi sloupcích vedle sebe uvedeno, kolikrát Čapek daného slova užil, a to nejdřív celkem a pak v šesti hlavních žánrech (próza, drama, publicistika, poezie, odborná literatura a korespondence)“ (Štícha 2009: 38–39).

K oběma slovníkům dodejme, že lze poměrně jednoduchým a nepracným způsobem vytvořit stejný abecední či frekvenční slovník libovolného autora s použitím volně dostupných konkordančních nástrojů (viz níže program *AntConc*).

Od roku 2008 se systematicky věnujeme možnostem tvorby malých, specializovaných korpusů, výsledkem těchto aktivit jsou i dílčí korpusové projekty, jež vznikají ve spolupráci se studenty na Katedře bohemistiky Filozofické fakulty Univerzity Palackého v Olomouci.[12] Jako první byl v roce 2008 sestaven korpus esejů Otokara Březiny,[13] v roce 2008–2009 vznikl korpus esejů Ladislava Klímy[14] a v roce 2010 jsme začali pracovat na výukovém korpusu Karla Čapka.[15] Všechny korpusy jsou stále ve fázi budování a každý sloužil či slouží jinému účelu.[16] Cílem, který spojuje všechny tři korpusy, bylo vypracovat efektivní postup tvorby malých autorských korpusů pro jazykovědné a literárněvědné účely. Budování korpusů ovšem zahrnuje několik etap a vyžaduje jak znalosti filologické, tak technické (kódování znaků, formát dat apod.). Jednotlivé korpusy tedy vznikaly jako pilotní projekty, jež měly prověřit možnosti zpracování textů korpusovými nástroji pro každou z klíčových oblastí tvorby textových korpusů.

---

12 Průběžné výsledky těchto aktivit jsme publikovali v následujících dvou studiích: Pořízka – Schäfer 2010a a Pořízka – Schäfer 2010b.

13 Zpracována byla prozatím první kniha esejů *Hudba pramenů* (1903). Korpus byl prezentován na mezinárodním sympoziu Otokar Březina 2008 v Jaroměřicích nad Rokytnou (Pořízka – Schäfer 2010a).

14 Tento korpus obsahuje text knihy *Svět jako vědomí a nic* (1904), prezentován byl na vědecké konferenci Ladislav Klíma konané na FF UPOL v Olomouci (Pořízka – Schäfer 2010b).

15 Vznik čapkovského korpusu je motivován čistě didaktickými účely, zatímco předchozí byly zaměřeny metodologicky. Vzniká v experimentálním semináři, v němž si studenti osvojují potřebné know-how, jak korpusy vytvořit. Obsahuje jen studentské seminární práce – dílčí texty Karla Čapka s různou formou zpracování. Materiálovým zdrojem jsou digitalizované texty e-knihovny Městské knihovny v Praze; srov. http://www.mlp.cz/karelcapek/ a rovněž http://www.mlp.cz/knihovna_on-line.htm.

16 Tyto dílčí korpusové projekty budou po dokončení zveřejněny na korpusovém portálu http://corpus.upol.cz.

Březinovský korpus byl zaměřen na lingvistickou anotaci: byla provedena lemmatizace (přiřazení reprezentativní formy slovních tvarů) a jednoduchá morfologická anotace přiřazením slovnědruhové interpretace. Použita přitom byla celá řada nástrojů urychlujících a automatizujících práci (počítačové skripty *fsm Tokenize*, *Annot1.pl*, *Annot1_to_Annot2.pl*) (Pořízka – Schäfer 2010a).

Klímovský korpus prověřoval možnosti typograficko-ediční a strukturně-obsahové anotace s pomocí značkovacího (meta)jazyka *XML* (*eXtensible Markup Language*), který uživateli umožňuje definovat si vlastní sadu značek se speciálním významem. V *XML* tak byla anotována struktura textu a jeho hierarchizace (kniha a její části: kapitoly, oddíly, odstavce, věty, slova…) a editorské/typografické jevy: u esejů Ladislava Klímy například a) grecismy, latinismy; b) korektury: překlepy, chybějící uvozovky; c) řezy písma: kurziva, bold; d) uvozovky, interpunkce apod. (Pořízka – Schäfer 2010b).

Zároveň byly během sestavování obou korpusů vypracovány a algoritmizovány jednotlivé kroky pro přípravu a importování souboru textů do aplikace *Manatee/Bonito*, včetně přesných formátů zápisu pro použití počítačových skriptů. Tento software totiž pracuje s tzv. binárními texty, jež je třeba s pomocí speciálních nástrojů (skriptů) zkonvertovat z prostého textu (ve formátu *.txt*) do požadovaného formátu – především nástrojem *encodevert.exe* (součást aplikace *Manatee*). Až po této konverzi a řadě dalších úkonů (zápisy do registrů, deklarace kódování znaků, korpusu a jeho atributů apod.) je možno s korpusem pracovat – přitom je nutné, aby byl text připraven v tzv. vertikále (jedno slovo na jeden řádek) a ve formátu *.txt*.

Tyto postupy a procesy při přípravě a zpracování textů mohou být pro začínajícího či méně zkušeného uživatele velmi komplikované a i sebemenší chyba během technického zpracování znamená nefunkčnost celého korpusu. Naštěstí pro běžného (i začínajícího) uživatele není nutno znát zevrubně všechny technické aspekty počítačového zpracování korpusů – pro základní práci s textovými daty postačí prostý (původní, tj. neanotovaný) text a jednoduchý konkordanční software.

Jedním z takových programů je *AntConc* (© Laurence Anthony),[17] aplikace umožňující pracovat s textem podobně jako *Manatee/Bonito* (v mnohém snese srovnání), ale tvorba korpusu je značně jednodušší – v podstatě triviální, neboť stačí text prostě jen importovat do aplikace.

17 Jedná se o výborný freewarový program, který je volně ke stažení ve verzi *AntConc3.2.1w* na webové adrese http://www.antlab.sci.waseda.ac.jp/software.html.

AntConc přitom pracuje s celou řadou formátů: s.*txt*,.*html*,.*xml*, importovat lze dokonce i wordovský.*doc* nebo.*rtf*, u nich ale nastanou s největší pravděpodobností problémy s kódováním (tj. se zobrazováním) znaků. *AntConc* umožňuje uživateli velmi snadno vyhledávat v textu jak slovní výrazy v kontextu (konkordance), tak s pomocí masek (speciálních metaznaků), provádět statistické výpočty, generovat abecední a frekvenční slovníky a další funkce.

Pro prezentaci základních možností tohoto konkordančního programu jsme zvolili Čapkovy *Povídky z jedné kapsy*.[18] Text lze načíst do programu standardní cestou: dokument lze otevřít v menu *File* → *Open File(s)* pro jednotlivé texty, nebo lze zvolit cestu *File* → *Open Directory* a importovat do aplikace ze zvolené složky počítače všechny dokumenty najednou. Načtené soubory se objeví v levém sloupci (*Corpus Files*) programu *AntConc*. Aby se texty zobrazovaly korektně, je nutno zkontrolovat či nastavit tzv. kódování znaků: v menu *Global Settings* → *Language Encodings* → *Edit* lze zvolit jedno ze tří kódování, která se dnes pro české texty užívají.[19] Poté již lze pracovat s textem a využívat všechny funkce, jež *AntConc* nabízí. Jak již bylo zmíněno, tou základní je možnost vyhledávat jednotlivé výrazy či slovní spojení (případně je kombinovat se speciálními zástupnými metaznaky). **[8]**

Obr. [8] zobrazuje kromě grafického prostředí aplikace jednotlivé konkordance (záložka *Concordance*) – výsledek hledaného výrazu „metod." (tečka představuje zástupný metasymbol s významem „jakýkoli znak"). Dotazový řádek se nachází v dolním panelu (*Search Term*), kde se také zobrazuje počet nalezených výskytů (*Concordance Hits*). Klíčové slovo je od kontextu barevně odlišeno.

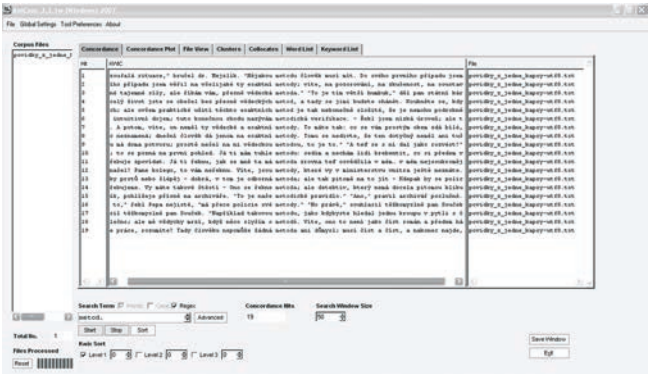
Velkou výhodou oproti systému *Manatee/Bonito* je možnost zobrazit celý text v úplnosti (záložka *File View*), nejen klíčové slovo s omezeným kontextem. **[9]**

Zajímavou funkci nabízí záložka *Concordance Plot*, jež zobrazuje rozložení hledaného výrazu, tj. jeho pozici v celém textu či korpusu – uživatel tak vidí, ve které části se daný výraz vyskytuje nejčastěji. **[10]**

V souvislosti se slovníky Karla Čapka a Bohumila Hrabala jsme

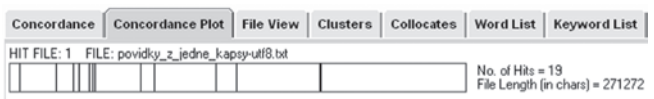
18 Zdroj: http://www.mlp.cz/koweb/00/03/34/75/70/povidky_z_jedne_kapsy.txt.

19 Pro češtinu lze zvolit 1. univerzální kódování *Unicode* (*utf8*), 2. *ISO Central Europe* (*iso-8859–2*) nebo 3. windowsovské *WinLatin2* (*cp-1250*). Při špatném kódování se znaky buď nezobrazují vůbec, nebo jsou zobrazeny nesprávně. Pokud uživatel neví, v jakém kódování je text zpracován (lze to jednoduše zjistit v některém z textových editorů), je třeba vyzkoušet postupně všechna tři kódování, až se všechny znaky zobrazí korektně.



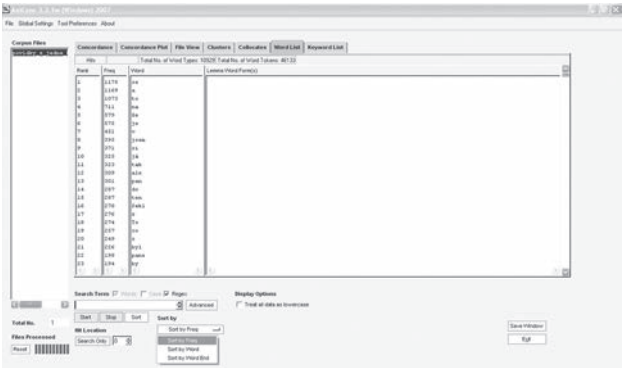
[8] Grafické uživatelské rozhraní programu *AntConc* s konkordancemi

[9] Funkce *File View*, jež zobrazuje barevně odlišená klíčová slova v celém dokumentu, nikoli jako konkordance



[10] Rozložení výrazu „metod.“ v korpusu s pomocí funkce *Concordance Plot*

uvedli, že existuje jednoduchý způsob, jak z libovolného textu vytvořit abecední či frekvenční slovník. Tímto způsobem je tak možno získat slovník kteréhokoli literárního autora, máme-li k dispozici elektronickou verzi příslušných (literárních) textů. Oba typy slovníků, ale i slovník retrográdní lze v programu *AntConc* vygenerovat v záložce *Word List*, přičemž jednotlivé typy výstupů se nastavují v dolní části panelu funkcí *Sort by Freq* (frekvenční slovník), *Sort by Word* (abecední slovník) a *Sort by Word End* (retrográdní slovník). **[11]**



[11] Tvorba frekvenčního slovníku, jejž lze z programu *AntConc* exportovat v textovém souboru

Slovníky, ale i všechny výsledky vyhledávání lze exportovat do textového souboru a uložit do počítače (menu *File* → *Save Output to Text File*). S výsledným textovým dokumentem může poté uživatel dále pracovat mimo program *AntConc*, například v běžném textovém editoru či procesoru.

Software *AntConc* nabízí řadu dalších funkcí či možností, například vyhledávání kolokací či klastrů, plné využití všech funkcí programu je však vázáno na anotované texty. Je možno například pracovat s lemmatizovaným textem či s dokumentem označkovaným v (meta)jazycích *XML* nebo *HTML*. Podobných programů jako *AntConc* existuje více, málokterý z nich – bereme-li v úvahu nekomerční, volně dostupné programy – ale nabízí takovou uživatelskou jednoduchost, spolehlivost a komplexnost.

Literatura

ČERMÁK, František (ed.)
2007 *Slovník Karla Čapka* (Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu)

ČERMÁK, František – CVRČEK, Václav (eds.)
2009 *Slovník Bohumila Hrabala* (Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu)

POŘÍZKA, Petr – SCHÄFER, František
2010a „Korpus esejů Otokara Březiny (fragment korpusu české esejistiky přelomu 19.–20. století)“, in Petr Holman (ed.): *Otokar Březina 2008* (Vsetín: Dalibor Malina), s. 95–105

2010b „Svět jako vědomí a nic Ladislava Klímy v olomouckém korpusu české esejistiky přelomu 19. a 20. století“, *Aluze* (v tisku)

ŠTÍCHA, František
2009 „Nad slovníkem Karla Čapka“, *Naše řeč* 92, č. 1, s. 38–39

Computer processing of literary texts: the opportunities involved
This study adopts an interdisciplinary approach towards text and deals with the technical options involved in text processing, which enables us by means of software tools to provide data retrieval, and to perform statistical analysis and other processes in accordance with preselected criteria and on the basis of an annotation text. The first part is devoted to the most important corpus projects focusing on literary texts: *Czech Electronic Library*, the *Intercorp* corpus and lexicographical dictionaries of Karel Čapek and Bohumil Hrabal. The second part presents the basic possibilities for the creation of small corpora, demonstrated on corpora of Otokar Březina and Ladislav Klíma, and primarily the usage of a corpus concordancer called *AntConc* during analysis of (literary) texts: data retrieval, creating alphabetical and frequency dictionaries, etc.

Keywords
corpus linguistics, text processing, data retrieval, corpus concordances, creation of corpora, Czech corpora of literary texts