Korpus českého verše

Petr Plecháč
Ústav pro českou literaturu AV ČR, v.v.i.
Na Florenci 3/1420
110 00 Praha 1
plechac@ucl.cas.cz

Robert Kolár
Ústav pro českou literaturu AV ČR, v.v.i.
Na Florenci 3/1420
110 00 Praha 1
kolar@ucl.cas.cz

Abstract

In following we present the Corpus of Czech Verse (i.e. lemmatised, phonetically, morphologically, metrically and strophically annotated corpus of Czech poetry) and the online tools that give access to its data. The following online tools are described: Database of Czech metres, Gunstick, Hex, Euphonometer, and Babel. English presentation of these tools may be found in Plecháč, Kolár 2015 or at the website of Versification Research Group (http://versologie.cz/en/).

1 Úvod

Korpus českého verše (KČV) je lemmatizovaný, foneticky, morfologicky, metricky a stroficky anotovaný korpus české poezie 19. a počátku 20. století.[1] Na rozdíl od standardních jazykových korpusů je tedy ke každé jednotce připojeno nejen lemma a morfologická značka, ale i fonetická transkripce; dále jsou každému verši přiřazeny atributy „metrum“ (jamb, trochej...), „počet stop“, „klauzule“ (ženská, mužská...) a „metrický vzorec“. Na vyšších rovinách jsou pak anotovány rýmové dvojice (resp. *n*-tice) a tzv. pevné formy (sonet, rondel...). V současnosti KČV obsahuje:

- 1 689 básnických sbírek
- 76 699 básní
- 2 664 989 veršů
- 14 592 037 slov

Data obsažená v KČV jsou zpřístupněna pomocí on-line nástrojů dostupných na http://versologie.cz.

2 On-line nástroje

2.1 Databáze českých meter (DČM)

DČM zpřístupňuje metrickou a strofickou rovinu anotace KČV. S aplikací lze pracovat ve dvou základních módech: (1) Prohlížení databáze, (2) Vizualizace.

Mód prohlížení databáze umožňuje vyhledávat básnické sbírky i konkrétní básně na základě filtrů (1) „jméno autora“, (2) „rok vydání“, (3) „název“ sbírky a/nebo básně. U každého výsledku jsou uvedeny bibliografické údaje a odkaz na externí zdroj (plný text sbírky dostupný v České elektronické knihovně),[2] v případě jednotlivých básní i podrobné informace o užitých metrech, rýmových a strofických schématech a o tom, zda byla báseň anotována jako realizace některé z pevných forem.

[1] Lemmatizaci a morfologickou anotaci provedli pracovníci Ústavu teoretické a komputační lingvistiky FF UK ve spolupráci s pracovníky Ústavu formální a aplikované lingvistiky MFF UK. Fonetická, metrická a strofická anotace byla provedena pomocí počítačového programu Květa (Ibrahim, Plecháč 2011).

[2] http://www.ceska-poezie.cz/cek/

Sborník Semináře o digitálních zdrojích a službách ve společenských a humanitních vědách (WDH 2015), str. 74–77, Praha, 24. září 2015.

Mód vizualizace umožňuje výše uvedené informace agregovat a zobrazit v přehledných grafech. Vizualizace lze vytvářet jednak pomocí základního filtru (s předem nastavenými datovými řadami), jednak pomocí filtru pokročilého (datové řady definuje uživatel). Pro podrobnější informace o DČM viz Kolár, Plecháč 2015.

Figure 1. Databáze českých meter: výsledek dotazu. Relativní frekvence jambických a trochejských básní v jednotlivých letech.

2.2 Gunstick – databáze českých rýmů

Aplikace Gunstick slouží k výzkumu frekvence rýmových párů a jejich historického vývoje. Při práci s aplikací uživatel zadává slovo (token), pro nějž budou vyhledány všechny rýmové páry doložené v KČV před rokem 1920 (databáze obsahuje přes milion rýmových párů). Dotaz může uživatel omezit jen na určitý časový úsek nebo určitý typ klauzule (mužská, ženská, akatalektická, neurčená). Výsledkem takového dotazu je interaktivní kruhový diagram zobrazující četnost výskytů jednotlivých rýmových párů. Kliknutím na jednotlivé výseče může uživatel promítnout vybraná data (1) do plošného diagramu zobrazujícího výskyt vybraného rýmového páru v jednotlivých letech a (2) do tabulky obsahující detaily jednotlivých záznamů (mimo jiné text obou rýmujících se veršů a odkaz na plný text sbírky dostupný v České elektronické knihovně).

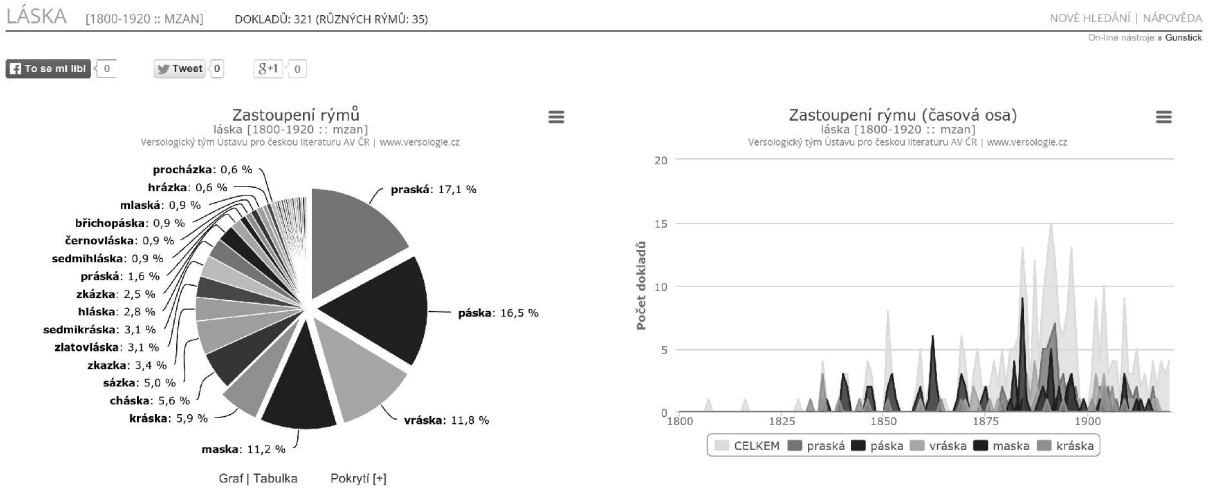

Figure 2. Aplikace Gunstick: výsledek dotazu. Relativní četnost rýmů na slovo „láska“ a absolutní četnost vybraných párů v jednotlivých letech.

2.3 Hex – klíčová slova v české poezii

Aplikace Hex umožňuje vyhledávat v KČV texty, které obsahují uživatelem specifikované klíčové slovo, nebo naopak u uživatelem specifikovaného okruhu textů zobrazit všechna klíčová slova v nich nalezená. V obou případech může uživatel vyhledávání omezit na určitý časový úsek a/nebo dílo jednoho či více autorů. Při prohledávání specifikovaného okruhu textů lze navíc užít filtry „název

sbírky“ a „název básně“. Jako klíčová slova jsou označena lemmata, jejichž frekvence v dané básni statisticky významně převyšuje jejich frekvenci v celém korpusu. Statistická významnost je ověřována zároveň testem χ^2 s Yatesovou korekcí a testem log-likelihood. Uživatel má možnost specifikovat, zda budou testy provedeny na hladině významnosti $\alpha = 0{,}001$ nebo $\alpha = 0{,}01$, a které slovní druhy mají být zařazeny do stop listu.

Vyhledává-li uživatel klíčové slovo, je výsledkem dotazu interaktivní diagram zobrazující četnost výskytů v jednotlivých letech, a to buď jako absolutní frekvenci nebo relativní frekvenci měřenou (a) počtem básní, (b) počtem veršů nebo (c) počtem slov. Dále je zobrazena tabulka obsahující pro každý záznam (báseň) mimo jiné bibliografické informace, odkaz na seznam všech klíčových slov, která byla za daných parametrů v básni nalezena, a odkaz na plný text sbírky dostupný v České elektronické knihovně.

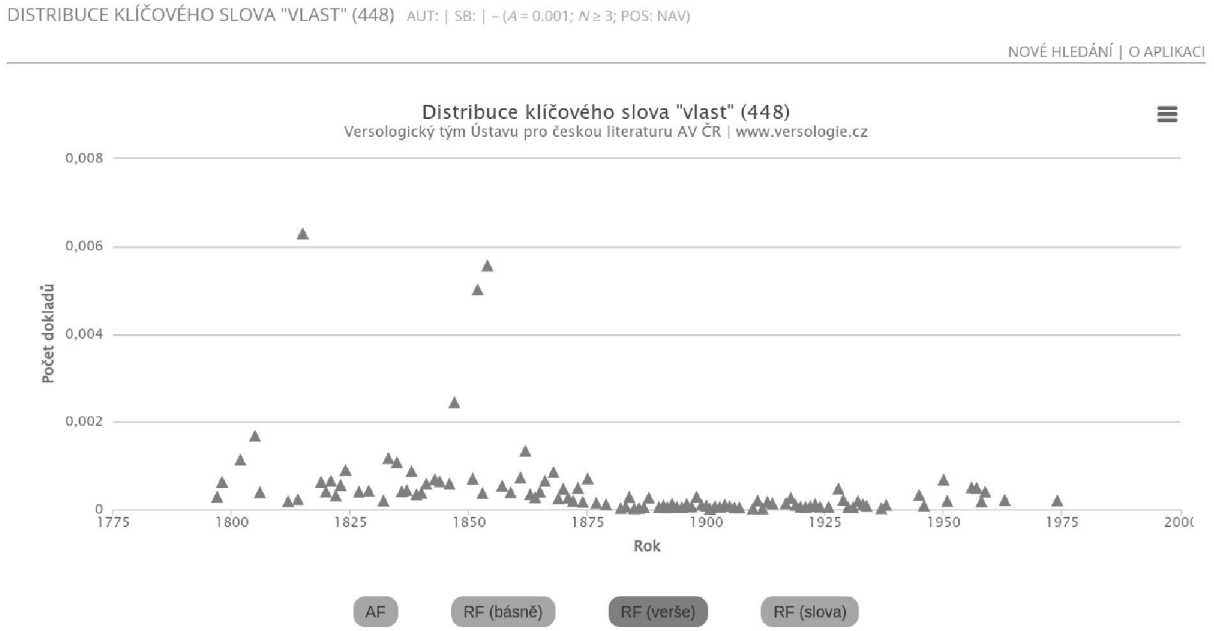

Figure 3. Aplikace Hex: výsledek dotazu. Relativní četnost básní obsahujících klíčové slovo „vlast“ v jednotlivých letech (měřeno počtem veršů); hladina významnosti $\alpha = 0{,}001$; minimální četnost lemmatu $n \geq 3$.

2.4 Eufonometr

Aplikace Eufonometr umožňuje na základě hodnot naměřených v KČV kvantifikovat míru nenáhodnosti hláskových opakování v libovolném uživatelem vloženém textu (tzv. eufonický koeficient). Aplikace vychází z binomického testu navrženého Gabrielem Altmannem (Altmann 1966; Čech et al. 2011) a jeho pozdějších úprav (Plecháč, Říha 2014). Výsledkem analýzy je hodnota eufonického koeficientu každého řádku vloženého textu a celkový (průměrný) eufonický koeficient, který je možné srovnat s hodnotami naměřenými v jednotlivých básních obsažených v KČV.

2.5 Babel

Konzolová aplikace Babel představuje nejmocnější nástroj pro práci s KČV. Oproti ostatním nástrojům sice vyžaduje alespoň elementární znalosti dotazovacího jazyka SQL, zato ale umožňuje uživateli pracovat (téměř) bez omezení se všemi rovinami anotace KČV. Aplikace odesílá dotaz zadaný uživatelem do databáze sqlite. Kvůli co možná nejsnadnějšímu dotazování jsou data rozdělena do malého počtu tabulek. Aplikace je rozšířena knihovnou sqlite3-pcre, která zajišťuje podporu regulárních výrazů (REGEXP) ve formátu programovacího jazyka Perl. Dotazy do databáze jsou kladeny asynchronně. Aplikace tedy nečeká na odpověď serveru a při zpracovávání dotazu s ní lze dále pracovat.

3 Závěr

Mezi automaticky anotovanými veršovými korpusy[3] patří KČV množstvím zpracovaných textů i množstvím anotovaných jevů k nejobjemnějším na světě. Domníváme se, že díky volně přístupným a do značné míry intuitivně ovladatelným on-line nástrojům se může KČV stát cenným zdrojem dat nejen pro specialisty-versology (nebo šířeji literární vědce), ale mimo jiné i pro lingvisty,[4] či pedagogy a studenty.[5]

Reference

Gabriel Altmann. 1966. The Measurement of Euphony. In *Teorie verše I*, 263–264. UJEP, Brno.

Klemens Bobenhausen. 2011. The Metricalizer – Automated Metrical Markup of German Poetry. In *Current Trends in Metrical Analysis*, 119–132. Peter Lang, Frankfurt am Main et al.

Radek Čech, Ioan-Iovitz Popescu, Gabriel Altmann. 2011. Euphony in Slovak Lyric Poetry. *Glottometrics*, 22: 5–16.

Daniele Fusi. 2009. An Expert System for the Classical Languages: Metrical Analysis Components. *Lexis*, 27: 25–46.

Robert Ibrahim, Petr Plecháč. 2011. Toward Automatic Analysis of Czech Verse. In *Formal Methods in Poetics*, 295–305. RAM, Lüdenscheid.

Robert Kolár, Petr Plecháč. 2015. Databáze českých meter a výzkum českého verše 19. století. *Česká literatura*, 63(2): 236–246.

Igor Pilshchikov, Anatoli Starostin. 2011. Automated Analysis of Poetic Texts and the Problem of Verse Meter. In *Current Trends in Metrical Analysis*, 133–140. Peter Lang, Frankfurt am Main et al.

Petr Plecháč, Jakub Říha. 2014. Measuring Euphony. In *Methodology and Practices of Russian Formalism*, 194–199. Azbukovnik, Moskva.

Petr Plecháč, Robert Kolár. 2015. The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1): 107–118.

Thomas Rainsford, Olga Scrivner. 2014. Metrical Annotation for a Verse Treebank. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, 149–159. Universität Tübingen.

[3] Srov. např. Fusi 2009; Bobenhausen 2011; Pilshchikov, Starostin 2011.
[4] Srov. Rainsford, Scrivner 2014
[5] Srov. Jan Bouchner: Korpus českého verše a jeho využití ve výuce; `http://spomocnik.rvp.cz/clanek/19305/korpus-ceskeho-verse-a-jeho-vyuziti-ve-vyuce.html`