Internetová jazyková příručka

Karel Pala, Pavel Šmerk, FI MU

1 Úvod

Čeština má, jako ostatně i jiné evropské jazyky, více podob. Při běžném styku mezi sebou, obvykle neformálním, používáme zpravidla nějakou její nespisovnou podobu. I v mluvené komunikaci však existuje řada situací, v nichž se používá v zásadě spisovná čeština, např. v rozhlase nebo v televizi. Pokud chceme přejít od mluvené podoby jazyka k psané, musíme mít k dispozici to, čemu lingvisté říkají pravopisný systém, tj. soubor pravidel a konvencí, podle nichž se mluvená podoba jazyka převádí na podobu psanou. Ve škole se učíme pravopisným pravidlům odpovídajícím spisovné podobě jazyka, která slouží k vytváření naší kulturní paměti, k zaznamenávání našich znalostí a také k oficiální komunikaci, při níž se snažíme dodržovat pravidla jazykové správnosti. Je celkem přirozené, že i přes soustavnou a dlouholetou školní výuku spisovného jazyka si čeští mluvčí při snaze o kultivovaný projev nemusejí být vždy jisti, co je vlastně správně.

Ještě donedávna mohli v takovém případě lidé správnou odpověď na takovou otázku zjišťovat buď z jazykových příruček, tedy Pravidel českého pravopisu, slovníků, mluvnic ap., nebo dotazem do Jazykové poradny Ústavu pro jazyk český AV ČR v Praze nebo v Brně. Cílem Internetové jazykové příručky, o níž informuje tento článek, je takové zpřístupnění potřebných informací o spisovné podobě jazyka, které umožní uživatelům jazyka v co nejširším spektru případů samostatně a pouze prostřednictvím svého webového prohlížeče zjistit, jaké jazykové prostředky jsou pro jejich konkrétní situaci adekvátní. Výhodami takového řešení je mimo jiné i snížení zátěže Jazykové poradny[1] a zejména možnost průběžné aktualizace, zpřesňování informací a doplňování jazykových dat podle aktuální situace a odezvy uživatelů, což je u tištěných příruček pochopitelně nemožné.

[1] Například mezi lety 2002 a 2007 stoupl počet emailových dotazů do Jazykové poradny více než trojnásobně, zároveň ovšem rostl i počet telefonických dotazů, což nutně vedlo ke stručnějším odpovědím. Naopak pokud mohou uživatelé odpovědi na jednodušší dotazy získávat i samostatně, uvolňuje to pracovníkům Jazykové poradny kapacity pro detailnější zodpovídání nerutinních, komplikovanějších dotazů.

2 Internetová jazyková příručka

Internetová jazyková příručka (IJP) na adrese `http://prirucka.ujc.cas.cz` je výsledkem grantového projektu *Jazyková poradna na internetu*, na kterém se podíleli pracovníci Oddělení jazykové kultury Ústavu pro jazyk český AV ČR (vytvoření, opravy a doplňování jazykových dat a informací) a pracovníci Centra zpracování přirozeného jazyka Fakulty informatiky MU (technická realizace, údržba a další rozvoj[2]). IJP sestává ze dvou hlavních částí, slovníkové, která obsahuje téměř 62 tisíc hesel, a výkladové, zahrnující 158 kapitol, v nichž může uživatel nalézt obecnější popis a vysvětlení jednotlivých jazykových jevů.

Základem pro zpracování výkladové části byla aktuální Pravidla českého pravopisu a současné mluvnice. Oproti nim jsou ale výklady často podrobnější, ucelenější a zpřesňující, protože přidávají a shrnují i informace ze speciálních jazykových příruček, z odborných časopiseckých studií či z českých státních norem ap. Ve výkladech jsou uváděny i případné rozpory mezi jednotlivými zdroji informací, stejně jako rozdíly mezi kodifikací a spisovným územ. V takových situacích je připojen hodnotící komentář a doporučená řešení.

Slovníková část vychází z hesláře školního vydání Pravidel českého pravopisu, Slovníku spisovné češtiny, výběrově jsou do ní začleněna i hesla z Nového akademického slovníku cizích slov, ze slovníků neologismů Nová slova v češtině 1, 2 a výrazy z poradenské databáze. Jednotlivá hesla mohou obsahovat informaci o možném dělení slova a jeho výslovnosti, pravopisné či tvaroslovné varianty (*balon*/*balón*, *brambora*/*brambor*), nebo naopak slova formálně shodná či podobná, ale významově odlišná (*rys*-zvíře/*rys*-nákres, *sjednat*/*zjednat*), dále informaci o významu a etymologii, ustálené

[2] Server Internetové jazykové příručky je fyzicky umístěn na Fakultě informatiky MU.

vazby (*diskutovat něco*), odvozená slova, frazeologii, příklady použití (slovní spojení či celé věty) a další. Podstatným rozdílem proti tištěným příručkám jsou tabulky tvarů u podstatných jmen a sloves a dále u vybraných zájmen a číslovek.[3] Pokud je potřeba, jsou jednotlivé potenciálně problematické tvary okomentovány prostřednictvím poznámek. Slovníková část je pomocí hypertextových odkazů propojena s výkladovou částí, aby si uživatel mohl snadno zobrazit obecné popisy jednotlivých jazykových jevů, které se na vyhledaném slově projevují.

Internetová jazyková příručka poskytuje primárně informace o pravopisných jevech a není tedy soustavným a komplexním popisem současného gramatického systému češtiny, jejím cílem není nahradit existující mluvnice. Zde je třeba upozornit, že běžní uživatelé jazyka si často pletou pravopis s gramatikou — zdůrazňujeme, že Internetová příručka zahrnuje pravidla českého pravopisu a jazykové správnosti. Ve výkladové části jsou rozebírány především ty jevy, na které se uživatelé češtiny v jazykové poradně opakovaně dotazovali. Stejné kritérium hrálo svou roli také jak při výběru slov zařazených do slovníkové části, tak i při výběru a volbě míry podrobnosti informací (zejména příkladů a poznámek) uváděných u jednotlivých slov.

Primární cílovou skupinou Internetové jazykové příručky jsou samozřejmě rodilí mluvčí, případně šířeji ti, kdo už český jazyk ovládají. Projekt je ale oceňován i cizinci, kteří se češtinu teprve učí nebo s ní přicházejí do kontaktu jiným způsobem. Kromě jiného jim totiž umožňuje dohledávat základní tvary nepravidelných slov, kdy běžný překladový slovník neobsahuje všechny možné slovní tvary jako třeba *stojí* či *psovi*, přičemž ale cizinec může jen stěží uhodnout, že má ve svém slovníku hledat slova *stát* či *pes*. Pro tyto uživatele je k dispozici i anglické rozhraní, a třebaže jde jen o překlad názvů jednotlivých položek či kratších popisků, a nikoli článků o jazykových jevech či vysvětlujících poznámek u jednotlivých slov a podobně, podle ohlasů je i jen takovéto zpřístupnění jazykových dat pro cizince velmi cenné.

Po technické stránce aplikace vychází z lexikografické platformy DEB II (Dictionary Editing and Browsing) [2] vyvinuté v Centru zpracování přirozeného jazyka FI MU. Mimo IJP je DEB II využit například v nástroji Debdict[4], což je prohlížeč umožňující po registraci přístup k šesti hlavním českým slovníkům a některým dalším zdrojům. Tento nástroj využívá v současnosti skoro 700 uživatelů z ČR a celého světa. Serverová strana je realizována v programovacím jazyce Ruby, data jsou uložena v XML databázi Berkeley DB XML. Vedle „viditelné" části zpřístupňující data veřejnosti obsahuje IJP i neveřejnou část, která umožňuje editaci a správu dat. Za zmínku stojí, že při naplňování slovníku ušetřilo velké množství práce použití morfologického analyzátoru `ajka` [1][5], jehož pomocí byly vygenerovány tvary jednotlivých slov, takže je pak editoři nemuseli vepisovat ručně, ale mohli je pouze zkontrolovat, jestli neobsahují chyby.

3 Využití IJP uživateli

Internetová jazyková příručka byla veřejnosti v plném rozsahu zpřístupněna v polovině ledna roku 2009[6], lze tedy dnes v několika statistických údajích přiblížit první dva roky její existence. Poznamenejme úvodem, že agregované údaje z přístupových logů mohou být nejen zajímavé pro utvoření obecné představy o využití IJP, ale zejména jsou cenným zdrojem informací, jaká slova považují tazatelé za problematická, které jazykové jevy stojí v popředí zájmu veřejnosti a čemu by tedy měla být ze strany editorů dat věnována zvláštní pozornost. Snažíme se proto tyto přístupové logy pokud možno co nejvíce očistit od požadavků generovaných automaticky (vyhledávacími roboty ap.), což sice z principu nebude nikdy možné dokonale, nicméně následující čísla by už měla s mírnou tolerancí odpovídat pouze „klikání" reálných uživatelů.

[3] U ostatních číslovek a zájmen a u přídavných jmen jednotlivé tvary uváděny nejsou, protože jejich tvorba je pro rodilého mluvčího neproblematická, případné jednotlivé výjimky jsou ukázány v příkladech nebo vysvětleny v poznámkách k heslu. U přídavných jmen a příslovcí jsou uváděny tvary druhého a třetího stupně, pokud jsou doloženy v praxi.

[4] `http://deb.fi.muni.cz/debdict`

[5] `http://nlp.fi.muni.cz/projekty/wwwajka`

[6] Výkladová část byla přístupná už od začátku dubna 2008.

Od zveřejnění slovníkové části zaznamenala IJP přes 10 000 000 přístupů z celkem více než 480 000 různých IP adres. Za poznámku stojí, že — alespoň měřeno využitím IJP — „pracovním" minimem týdne není neděle, jak by napovídalo i její pojmenování, ale sobota. Naopak nejaktivnější jsou uživatelé v průměru v úterý a v pondělí, ve zbytku týdne využití klesá tak, že páteční zátěž se už od nedělní ani příliš neliší. Průměrný denní počet přístupů za celou dobu je zhruba 13 500, využití IJP ale postupně roste, takže v posledních měsících už průměr v pracovní dny přesahuje 20 000 požadavků.

Nejčastějšími dotazy do slovníkové části jsou *jenž* (cca 14 500 dotazů), *jež* (11 000) a *práce* (5 500), následované slovy *datum*, *mě*, *den*, *já*, *narozdíl*, *ona*, *on*, *zapomněl* (vše už okolo 5 000) atd. V průběhu času se tento pomyslný žebříček nijak zvlášť nemění, například první desítka nejčastějších dotazů za první rok provozu IJP je v podstatě stejná (i včetně pořadí) jako první desítka za druhý rok. Jedinou výjimku tvořil relativně krátký časový úsek po počátečním zveřejnění a medializaci IJP, kdy by běžný člověk v první stovce nejčastějších dotazů pravděpodobně našel naprostou většinu vulgárních slov, která zná, přičemž ta úplně „nejprofláklejší" držela se spolehlivým odstupem první tři místa.

Ve výkladové části uživatelé nejčastěji pokládají dotazy *pomlčka* (cca 3200 dotazů), *číslovky* (2 900), *nebo* (2 400), *uvozovky*, *zájmena*, *datum*, *spojovník*, *jak tak*, *než*, *jako* (vše už okolo 2 000) atd. Ke každému dotazu jsou nabídnuty vyhovující výklady, z nichž si uživatel může vybrat. Přestože mezi první desítkou nejčastějších dotazů není čárka, tři nejčastěji takto vybírané výklady jsou *Psaní čárky ve větě jednoduché*, *Psaní čárky v souvětí* a *Psaní čárky před spojkami a, i, ani*. Stejné výklady jsou nejvíce preferovány i při přímém výběru ze seznamu výkladů na úvodní stránce IJP. K výkladům se uživatel může dostat i prostřednictvím odkazů z jednotlivých hesel slovníkové části, v takovém případě jsou nejčastěji zobrazovány výklady *Dělení slov*, *Psaní předpon s-, z-* a *Vyjmenovaná slova*. Stejně jako u dotazů do slovníkové části, i zde jsou uživatelské preference z dlouhodobého pohledu vesměs stabilní.

Přestože pozorný čtenář denního tisku by o následujícím tvrzení mohl nezřídka zapochybovat, největšími uživateli IJP jsou média a státní či veřejné instituce. Nejvíce dotazů přichází ze strojů v doménách[7] patřících mediálním domům Mafra, Mladá fronta a Vltava-Labe-Press, dále to jsou UK a MU, Česká televize, nakladatelství Economia, Portál veřejné správy (gov.cz), Ministerstvo spravedlnosti a deník Právo. I za touto první desítkou výrazně převažují domény z mediální a státní/veřejné sféry.

4 Ocenění

Projekt Internetové jazykové příručky byl velmi úspěšný a jeho výsledky jsou standardně využívány v celé České republice. V roce 2009 byl kolektiv autorů oceněn ministrem školství nejvyšším resortním oceněním, Medailí Ministerstva školství, mládeže a tělovýchovy 1. stupně „za zlepšování podmínek pro výuku mateřského jazyka na všech typech škol".

Literatura

[1] R. Sedláček, P. Smrž. A New Czech Morphological Analyser `ajka`. In *Proc. TSD 2001. LNCS 2166*, pp. 100–107. Springer-Verlag 2001.

[2] A. Horák, K. Pala, A. Rambousek, P. Rychlý. New clients for dictionary writing on the DEB platform. In *DWS 2006: Proc. 4th International Workshop on Dictionary Writings Systems*, pp. 17–23. Lexical Computing Ltd. 2006. □

[7] Pochopitelně nejsou započítány domény firem zprostředkujících připojení k internetu jako O2, UPC a další.

Obrázek 1: Medaile MŠMT ČR