Hornjoserbski tekstowy korpus w nowej formje

Sonja Wölkowa

Wot 15. měrca 2013 je so změnił přistup k hornjoserbskemu tekstowemu korpusej na stronach Serbskeho instituta. Při tutej składnosći chcemy wo tutym projekće informować a někotre pokiwy za jeho wužiwanje dać.

Tekstowe korpusy su digitalne zběrki tekstow, kotrež su w dźensnišim času zwjetša w interneće přistupne a kotrež służa prěnjotnje jako zakład za přepytowanje rěče. W nich je zhromadźeny awtentiski rěčny material, kotryž dowoli objektiwniše posudźowanje rěčnych zjawow a je njeparujomny zakład za moderne linguistiske slědźenje a tohorunja za wudźěłanje rěčnych přiručkow, słownikow abo wučbnicow. Tola też za literaturowědnikow, bibliotekarow abo stawiznarjow móže tajki korpus při wotpowědnym wuběrje tekstow wunošne žórło być, a lektorojo abo redaktorojo móža so při rěčnych korekturach na njón zepěrać. Tekstowe korpusy eksistuja hižo za tójšto europskich rěčow, mjez nimi za nas wobdawacu němčinu[1] a za našej susodnej słowjanskej rěči pólšćinu[2] a čěšćinu[3]. A też za delnjo- a hornjoserbšćinu móžemy so na tajku informacisku resursu zepěrać. Wobaj korpusaj staj so wuwiłoj a hladatej a rozšěrjatej so w zamołwitosći Serbskeho instituta. Wo „DOlnoserbski Tekstowy KOrpus“ (skrótšeny jako DOTKO)[4] stara so wotrjad za delnjoserbske slědźenja w Choćebuzu, mjeztym zo je za HOrnjoserbski Tekstowy KOrpus (skrótka HOTKO)[5] zamołwity rěčespytny wotrjad Serbskeho instituta w Budyšinje w kooperaciji ze Serbskej centralnej biblioteku. Tutón naposledk mjenowany je tema předležaceho přinoška.

1. Nastaće hornjoserbskeho tekstoweho korpusa

Praktiske dźěło na hornjoserbskim tekstowym korpusu je so započało w lěće 1996, jeho zakłady je koncipował a stworił tehdyši sobudźěłaćer Serbskeho instituta Edward Wornar, nětko profesor na Lipšćanskej uniwersiće. Po jeho wotchadźe w lěće 2003 přewza zamołwitosć za korpus wjednica rěčespytneho wotrjada Serbskeho instituta Sonja Wölkowa. Pjeć lět po zahajenju dźěła na korpusu spřistupni so prěnja wersija na internetowej stronje Serbskeho instituta. Tehdyši staw wuwića kompjuteroweho koděrowanja słowjanskich pismikow po wšelakorych zasadach bě při tym wosebite wužadanje. Zo by korpus zajimcam po cyłym swěće njewotwisnje wot wužiwaneho koděrowanskeho standarda přistupny był, wuwi jeho założer a wobdźěłar Edward Wornar za specifisce serbske pismiki z diakritiskimi znamješkami transkripciju, kotraž wuńdźe ze 128 znamješkami tak mjenowaneho ASCII-koda.[6] Pismiki z diakritiku rozpušćichu so na dwě znamješce – zakładny pismik z předchadźacym symbolom za hóčku (*č = ^c*), smužku (*ć = /c*) resp. nakósnu smužku pola *ł/Ł* (*ł = _l*). Tak wupadachu drje serbske teksty chětro njezwučene, za to pak na wšěch kompjuterach jenak.[7]

Kompjuterowa technika pak je so dale wuwiwała, w koděrowanju je so mjezynarodnje přesadźił standard Unicode, kotryž też wosebite pismiki słowjanskich rěčow wobsahuje. Tohodla dotalny přistup k hornjoserbskemu tekstowemu korpusej nowym móžnosćam poněčim wjace njewotpowědowaše. K tomu přińdźe fakt, zo je wotrjad za delnjoserbske slědźenja kónc lěta 2010 spřistupnił delnjoserbski tekstowy korpus w spodobnej a za wužiwarja přijomnej formje na stronje www.dolnoserbski.de[8] a zdobom na zakładźe kooperacije z Institutom za Češki narodny korpus (ÚČNK) na tamnej stronje www.korpus.cz z přidatnymi móžnosćemi přepytowanja a rešeršowanja. Zo móžemy nětko tohorunja hornjoserbski tekstowy korpus na tutej stronje a ze samsnymi móžnosćemi wužiwać[9], za to mamy so kooperaciskim poćaham a sobudźěłu wjacorych partnerow dźakować. Kolegaj z Choćebuskeho wotrjada Serbskeho instituta Fabian Kaulfürst a Marcin Szczepański sposrědkowaštaj kontakt a zhromadne dźěło z Praskim ÚČNK a pomhaštaj formu podawanja tekstow a třebnych informacijow wo nich zjednotnić. Michal Křen z Praskeho instituta postara so mjez druhim wo segmentowanje běžnych tekstow na sady a wo jich přihotowanje za wšelake přepytowanske móžnosće, na př. za analyzowanje słownych skupin (kolokacijow).

2. Tuchwilny staw Hornjoserbskeho rěčneho korpusa

W běhu lět je Hornjoserbski tekstowy korpus dosć nahladnje rozrostł: Tuchwilu je w nim 384 datajow wšelakeho razu a wšelakeje wulkosće, dohromady je to wokoło 44 milionow tak mjenowanych „tokenow“ – z tym su měnjene jednotliwe separatne słowne formy.[10] Jako tokeny liča też w teksće wustupowace ličby a nimo toho kóžde interpunkciske znamješko, tak zo je w korpusu trochu mjenje słowow hač tokenow (ca. 36 milionow). Teksty hornjoserbskeho korpusa słušeja do wšelakich kategorijow (hlej wobr. 1): Najwjetši je podźěl publicistiki (57 %) a beletristiki (23 %), mjenje je nabožnych (4 %) a wědomostnych (4 %) tekstow, k tomu přińdu

někotre słowniki a rjad terminologijow za jednotliwe šulske předmjety (12%).[11] Najstarše teksty sahaja hač do prěnjeje połojcy 19. lětstotka, su to spisy Handrija Zejlerja a basnje Rudolfa Mjenja w nowowudaćach z 20. lětstotka, najmłódše su Serbske Nowiny a Katolski Posoł z lěta 2010. Podźěl jednotliwych časowych dobow je wšelaki: Najwjace mamy modernych tekstow z časa po přewróće 1989/90 (54%). Z 19. lětstotka pochadźa 10 %, z prěnjeje połojcy 20. lětstotka hač do 1945 je jich 18% a z lět mjez 1945 a 1990 19% – tak reprezentuje hornjoserbski tekstowy korpus z nimale třomi štwórćinami tekstow z časa po Druhej swětowej wójnje předewšěm rěč přitomnosće.[12]

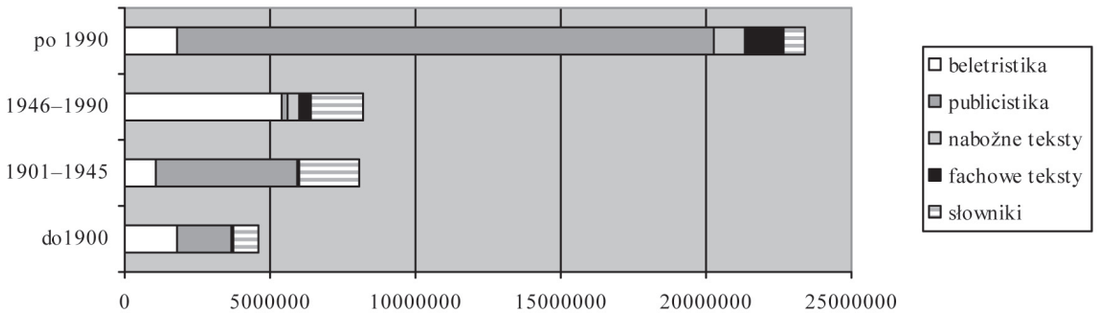
Dźakowano zrěčenjam z Ludowym nakładnistwom Domowina a Rěčnym centrumom WITAJ smy móhli wobšěrnu zběrku aktualnych tekstow za rešeršowanje spřistupnić, při čimž ma so wězo wobkedźbowanje awtorskich prawow zaručić.[13] Najwjetši dźěl tekstow pak je so zaskenował a potom z programami za spóznawaće teksta (t. mj. OCR-programy) wobdźěłał. Njedosahaceho personala dla njejsu so začitane teksty bohužel w dosahacej měrje skorigować móhli, tak zo njemóžemy wuzamknyć, zo jewja so zmylki. Za serbskeho wužiwarja mjenje problematiski drje je fakt, zo njejsu citowane pasaže w druhich rěčach, mjez druhim w delnjoserbšćinje, wosebje markěrowane – wšak je za njeho lochko spóznać, hač jedna so wo hornjoserbšćinu abo hinašu rěč. Mjeńši dźěl tekstow, předewšěm z 19. lětstotka, předleži w historiskim prawopisu (na př. pjeć lětnikow časopisa „Serbski hospodar“), pisanje kh za dźensniše ch na spočatku morfemow a mjechke ŕ pak namakatej so tež hišće w tekstach, nastatych do lěta 1945. Na internetnej stronje Serbskeho instituta smy wo tym zaměstnili nadrobnu informaciju (hlej wobr. 2, dypk: prawopisne warianty), kotraž móže wužiwarjej pomhać wšitke přikłady za swoje naprašowanje namakać, hačrunjež jewja so prawopisne warianty.

3. Wužiwanje hornjoserbskeho tekstoweho korpusa

Hornjoserbski tekstowy korpus namakaće na stronje Serbskeho instituta pod menijowym dypkom *online-publikacije* (http://www.serbski-institut.de/cms/os/48/hornjoserbski, hlej wobr. 2). Za wužiwarjow smy přihotowali nadrobnu dokumentaciju a pomocne informacije, kotrež namakaće, hdyž nakliknjeće jednotliwe dypki pod krótkim zawodnym tekstom.

Dokelž zmóžnja so přepytowanje hornjoserbskeho tekstoweho korpusa přez kooperaciju z Institutom za Češki narodny korpus (www.korpus.cz), dyrbja sej wužiwarjo na jeho stronach wosobinske wužiwarske konto założić. Za tutu proceduru smy na horjeka mjenowanej stronje spřihotowali wotpowědne wujasnjenja (wobr. 2, dypk: přistup ke korpusej).

wobraz 1



wobraz 2

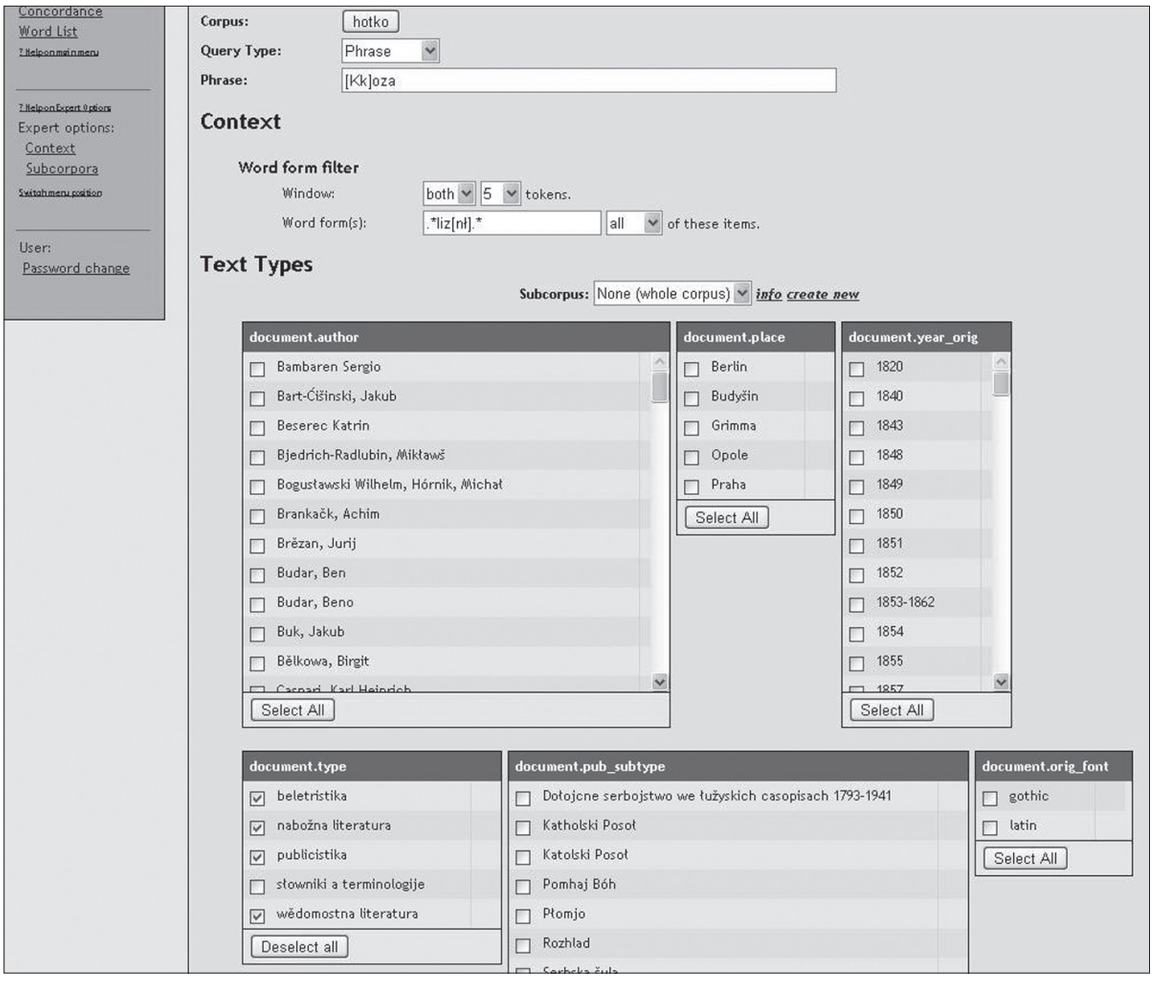
Štóž je sej na te wašnje přistup zarjadował, móže z pomocu wotprašowanskich programow[14] za wustupowanjom wěstych słowow w zapřijatych tekstach pytać – při tym namakaja so awtomatisce wobšěrne lisćiny přikładow (tak mjenowane konkordancy) za słowa, jich formy abo kombinacije, hromadźe z dokumentaciju, z kotrych žórłow pochadźeja. W tutych lisćinach hodźi so zaso přepytować, kak husto wěste słowa wustupuja a w kajkim susodstwje so wosebje jewja, móže so tež zwěsćić, hač su za wěstych awtorow typiske, na wěste družiny tekstow abo snano na wěstu časowu dobu wobmjezowane.

Hornjoserbski korpus njeje hišće lematizowany a gramatisce anotěrowany – to rěka, zo njejsu wšelake słowne formy jednotnemu hesłu přirjadowane kaž w słowniku (na př. *ruku, ruce, rukow, rukomaj* k hesłu *ruka*) a njejsu po słownych družinach abo gramatiskich formach analyzowane. To wobmjezuje pytanske móžnosće wosebje za rěčespytne prašenja z wobłuka syntaksy a morfologije. Za namakanje po móžnosći wšěch formow někajkeho pytaneho słowa pak móža sej wužiwarjo pomhać z tak mjenowanymi regularnymi wurazami (jendź. regular expressions). To su znamješka, kotrež funguja jako naměstniki za wšelake warianty, takrjec kaž jokery w kartowej hrě. Tak steji kombinacija .* za 0 abo wjace pismikow abo cyfrow. Hdyž zapodamy potajkim naprašowanje *měsac.**, namaka pytanski program tež přikłady za wotchilne formy tutoho słowa, na př. *měsaca*, *měsacej*, *měsacy*, *měsacow*, *měsacami.* Naprašowanje *.*spěwam* přinjese přikłady za *spěwam*, *N/njespěwam*, *Z/zaspěwam*, *W/wuspěwam*, *D/dospěwam* atd. Nadrobne serbske wujasnjenja k tutym znamješkam namakaće na stronje Serbskeho instituta (wobr. 2, dypk: regularne wurazy).

Nimo ryzy wědomostnych zaměrow hodźi so tekstowy korpus tohorunja za serbskorěčnu praksu zwužitkować. Awtorojo wučbnych materialijow namakaja z jeho pomocu zlochka mnóstwo awtentiskich přikładow za słowa abo słowne skupiny. Tute hodźa so na př. wužiwać za zestajanje lisćinow sadow z dźěrami k wupjelnjenju za leksikaliske zwučowanja. Też za rěčnymi wobrotami hodźi so derje pytać, hdyž kombinuje so naprašowanje we hłownym woknješku z kontekstowym filtrom, kaž sćěhowacy přikład pokazuje (wobr. 3).

Hdyž zapodamy na př. we hłownym woknješku sekwencu *[Kk]oza*[15] a w kontekstowym woknješku (w meniju wubjerje so *context*) zapisamy .***liz[nł].** a wuzamknjemy pod menijowym dypkom *subcorpora* jako žórło

wobraz 3



wobraz 4

słowniki, namaka program 72 přikładowych sadow z frazeologizmom *někoho koza liznje* we wšelakich tempusowych formach, kaž pokazuje sćěhowacy wurězk (wobr. 4).

Wězo njesmědźa so při wužiwanju ženje zabyć wěste wobmjezowanja za wuslědki rešeršow. Prěnje tajke wobmjezowanje rezultuje z toho, zo nimamy korpus wšěch hornjoserbskich tekstow. Hdyž potajkim někajku formu njenamakamy, rěka to jenož, zo njeje wona w korpusu. Dopokaz, zo wona njeeksistuje, to hišće njeje. Druhe wobmjezowanje leži w tym, zo móža teksty korpusa rozšěrjene rěčne zmylki dokumentować – hranica mjez zmylkom a nowej, so šěrjacej formu tež za rěčespytnika druhdy cyle wótra njeje. A třeće wobmjezowanje zależi na wužiwarju samym – kwalita wuslědkow korpusoweho pytanja wotwisuje wot kwality prašenjow, kotrež so jemu stajeja. A tu móže so kóždy najlěpje přez swójske nazhonjenja wukmanić. Nadźijamy so, zo namaka hornjoserbski tekstowy korpus tež w serbskoreěčnej praksy swojich zajimcow a wužiwarjow.

1 http://www.ids-mannheim.de/kl/projekte/korpora/ kaž tež http://wortschatz.uni-leipzig.de/
2 http://nkjp.pl/
3 www.korpus.cz
4 Přistup přez http://www.dolnoserbski.de/korpus/ abo přez http://www.korpus.cz/dotko.php.
5 Přistup pod http://www.serbski-institut.de/cms/os/48/hornjoserbski abo direktnje přez http://www.korpus.cz/hotko.php
6 Skrótšenka za: American Standard Code for Information Interchange.
7 K znazornjenju a wujasnjenju poda so w interneće přikład za transkripciju: Rjana _Lu^zica, sprawna, p^re/celna, mojich serbskich w/otcow kraj, mojich zb/o^znych sonow raj. Swjate su mi twoje hona.
8 Hlej Nowy Casnik 28.12.2010, str. 2.
9 Zjednorjeny přistup kaž pod www.dolnoserbski.de za delnjoserbski korpus za hornjoserbski tuchwilu hišće móžny njeje, za to dyrbja so hišće pjenjezy a programowar namakać.
10 Zestajane časowe formy kaž na př. perfekt *sym spěwał* liča jako dwě słowje.
11 Procentualne podźěle su so wuličili na zakładźe ličby tokenow (hlej horjeka).
12 Při dalšim rozšěrjenju korpusa budźe so wosebje na wurunanje disproporcijow mjez wšelakimi tekstowymi typami a časowymi dobami dźiwać dyrbjeć.
13 Tohodla je maksimalna wulkosć pokazowaneho konteksta pytaneho słowa wobmjezowana na 100 tokenow.
14 Za wužiwanje wšěch pytanskich a přepytowanskich opcijow je třěbna znajomosć jendźelšćiny.
15 Róžkate spinki signalizuja naprašowanskemu programej, zo maja so w nich stejace znamješka jako warianty rozumić.