Studia z Filologii Polskiej i Słowiańskiej

Studia z Filologii Polskiej
i Słowiańskiej, 49
SOW, Warszawa 2014

DOI: 10.11649/sfps.2014.008

Sonja Wölkowa
(Serbski institut, Budyšin)

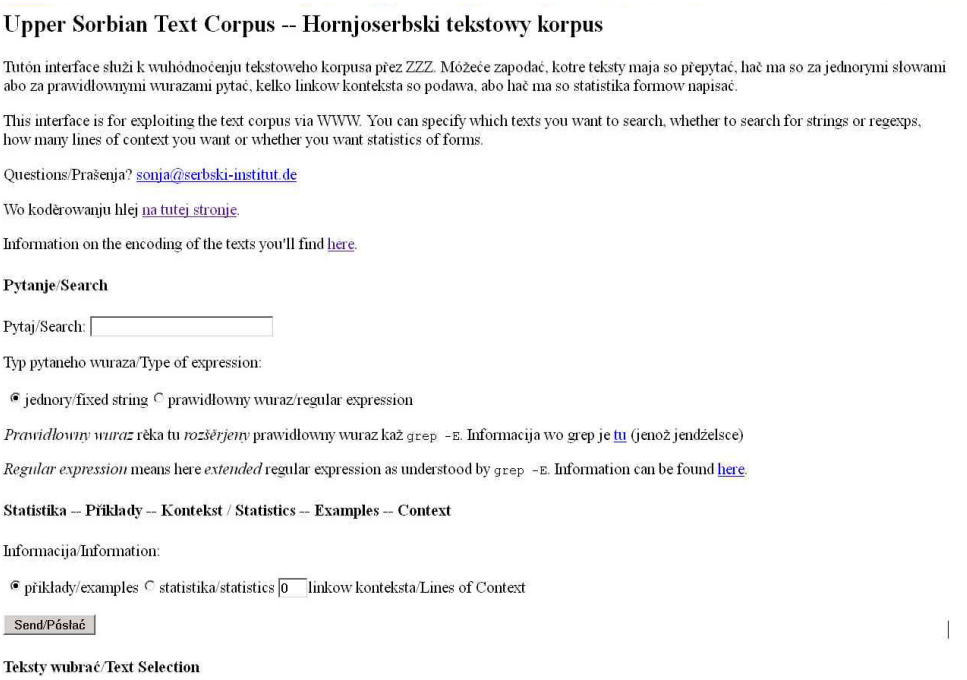
Tekstowy korpus a dalše informaciske srědki wo hornjoserbskej rěči w interneće

W dobje globalizacije a přiběraceje omniprezency interneta steji wopisanje hornjoserbšćiny před nowymi wužadanjemi – Franc Šěn je situaciju 2012 na konferency w Gdańsku jara markantnje charakterizował: „štož njeje w interneće, njeeksistuje" (Šěn, 2013, s. 131). A tak nastawa a rosće potrjeba za digitalnje a přez internet přistupnymi informaciskimi žórłami wo hornjoserbšćinje kaž za praksu tak tež za wědomostne slědźenje. Na aktualnosć tuteje problematiki pokazuje w lěće 2013 w nadawku Załožby za serbski lud założena dźěłowa skupina „Serbšćina w nowych medijach". Wona je jako tuchwilu najwažniši nadawk identifikowała a wobzamknyła wudźěłanje noweho němsko-hornjoserbskeho słownika w interneće, a tutomu nadawkej mamy so jako rěčespytnicy stajić. Jako palaca začuwa so tuta potrjeba wosebje tež hladajo na to, zo je so za delnjoserbšćinu w poslednim lětdźesatku w delnjoserbskim wotrjedźe Serbskeho instituta zrealizował a w syći spřistupnił cyły system informaciskich srědkow, ke kotremuž słuša nimo němsko-delnjoserbskeho słownika tež digitalny přistup k najwažnišim delnjoserbsko-němskim słownikam Zwahra, Muki, Šwjele a Starosty kaž tež delnjoserbski tekstowy korpus (Dolnoserbski tekstowy korpus, n.d.; www.dolnoserbski.de).

W swojim přednošku chcu tu předstajić hornjoserbski tekstowy korpus HoTKo w konteksće dalšich digitalnych projektow za hornjoserbšćinu,

Publisher: Institute of Slavic Studies, PAS & The Slavic Foundation
[Wydawca: Instytut Slawistyki PAN & Fundacja Slawistyczna]

realizowanych, přihotowanych a planowanych w Serbskim instituće. Wobšěrniši přehlad wo prezency serbskich wobsahow w interneće powšitkownje je podał w hižo mjenowanym přednošku w Gdańsku Franc Šěn, při tym wšak njebě městna na nadrobniše předstajenje jednotliwych poskitkow (Šěn, 2013). Prěnje plany stworjenja tajkich internetowych pomocnych srědkow za hornjoserbšćinu sahaja do prěnjeje połojcy 90tych lět 20. lětstotka. Tehdy naćisny so w Serbskim instituće projekt jednorečneho hornjoserbskeho słownika, kotryž měješe primarnje w digitalnej formje nastać (Šołćina & Wornar, 1996). Bohužel njeje so tutón projekt personalnych přičin dla realizować móhł. Tola w přihotowanskej fazy w zwisku z tutymi słownikarskimi planami započa so w lěće 1996 dźěło na digitalnym hornjoserbskim tekstowym korpusu, kotryž eksistuje hač do dźensnišeho a so dale wuwiwa (přir. Wölkowa, 2013). Jeho zakłady (runje tak kaž mjenowany słownikarski projekt) je koncipował a stworił Edward Wornar, tehdy sobudźěłaćer Serbskeho instituta, nětko profesor Instituta za sorabistiku na Lipšćanskej uniwersiće. Po jeho wotchadźe na profesuru w Lipsku w lěće 2003 je zamołwitosć za korpus přešła na awtorku předležaceho přinoška w kooperaciji z nawodu Serbskeje centralneje biblioteki Francom Šěnom.



Upper Sorbian Text Corpus -- Hornjoserbski tekstowy korpus

Tutón interface služi k wuhódnoćenju tekstoweho korpusa přez ZZZ. Možeće zapodać, kotre teksty maja so přepytać, hač ma so za jednorymi słowami abo za prawidłownymi wurazami pytać, kelko linkow konteksta so podawa, abo hač ma so statistika formow napisać.

This interface is for exploiting the text corpus via WWW. You can specify which texts you want to search, whether to search for strings or regexps, how many lines of context you want or whether you want statistics of forms.

Questions/Prašenja? sonja@serbski-institut.de

Wo koděrowanju hlej na tutej stronje.

Information on the encoding of the texts you'll find here.

Pytanje/Search

Pytaj/Search:

Typ pytaneho wuraza/Type of expression:

jednory/fixed string prawidłowny wuraz/regular expression

Prawidłowny wuraz rěka tu *rozšěrjeny* prawidłowny wuraz kaž `grep -E`. Informacija wo grep je tu (jenož jendźelsce)

Regular expression means here *extended* regular expression as understood by `grep -E`. Information can be found here.

Statistika -- Přikłady -- Kontekst / Statistics -- Examples -- Context

Informacija/Information:

přikłady/examples statistika/statistics 0 linkow konteksta/Lines of Context

Send/Pósłać

Teksty wubrać/Text Selection

Wobr. 1. Přistupowy formular za hornjoserbski tekstowy korpus 2001

Pjeć lět po zahajenju dźěła na korpusu, potajkim w lěće 2001, spřistupni so prěnja wersija zjawnje na internetowej stronje Serbskeho instituta.

Přistupowy formular bě dwuréčny serbsko-jendźelski a njebě tehdy ani estetisce wosebje naročny ani jara komfortabelny za wužiwanje. Tehdyši staw wuwića kompjuteroweho koděrowanja słowjanskich pismikow po wšelakorych zasadach a kodowych stronach bě při tym wosebite wužadanje. Zo by so korpus wot zajimcow po cyłym swěće njewotwisnje wot wužiwaneho koděrowanskeho standarda wužiwać hodźał, wuwi jeho założer a wobdźěłar Edward Wornar za specifisce serbske pismiki z diakritiskimi znamješkami transkripciju, kotraž wuńdźe ze 128 znamješkami koda ASCII. 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. K znazornjenju a wujasnjenju poda so w interneće přikład za transkripciju.

Example for Text Encoding -- Přiklad koděrowanja

Rjana _Lu^zica, sprawna, p^re/celna, mojich serbskich w/otcow kraj, mojich zb/o^znych sonow raj. Swjate su mi twoje hona.	Rjana Łužica, sprawna, přećelna, mojich serbskich wótcow kraj, mojich zbóžnych sonow raj. Swjate su mi twoje hona.

(_ for lslash/Lslash, ^ for caron, / for acute)

Kedźbu/Attention!

```
Hdyž jako "typ wuraza" wubjerjeće "prawidłowny wuraz", ma třěška ^ funkciju
wosebiteho znamješka. Tutu funkciju hasnje backslash před třěšku (\^):
Češki = \^c\^eski.
If you select "regular expression" from "type of expression", caron ^ gets
a special meaning. To turn off this meaning, type a backslash before caron (\^):
Češki = \^c\^eski.
```

Wobr. 2. Pokazka koděrowanja słowjanskich pismikow z diakritiskimi znamješkami

Kompjuterowa technika pak je so dale wuwiwała, w koděrowanju je so mjezynarodnje přesadźił standard Unicode, kotryž wobsahuje też wosebite pismiki słowjanskich rěčow. Tohodla dotalny přistup k hornjoserbskemu tekstowemu korpusej nowym móžnosćam po lětach wjace njewotpowědowaše. K tomu přińdźe fakt, zo je wotrjad za delnjoserbske slědźenja Serbskeho instituta delnjoserbski

digitalny korpus kónc lěta 2010 spřistupnił w spodobnej a za wužiwarja přijomnej formje na stronje www.dolnoserbski.de a zdobom na zakładźe kooperacije z Institutom za Češki narodny korpus (ÚČNK) na tamnišej stronje www.korpus.cz z přidatnymi móžnosćemi za přepytowanje a rešeršowanje. Zo móžemy nětko tohorunja hornjoserbski tekstowy korpus na tutej stronje a ze samsnymi móžnosćemi wužiwać, za to mamy so kooperaciskim poćaham a sobudźěłu wjacorych partnerow dźakować. Kolegaj z Choćebuskeho wotrjada Serbskeho instituta Fabian Kaulfürst a Marcin Szczepański posrědkowaštaj kontakt a zhromadne dźěło z Praskim ÚČNK a pomhaštaj formu podawanja tekstow a trěbnych 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). Zjednorjeny přistup za njelinguistow, kaž poskića so wón pod www.dolnoserbski.de za delnjoserbski korpus, tuchwilu za hornjoserbski hišće móžny njeje, za to dyrbja so hišće financielne srědki a programowar namakać.

Žórła korpusa a wosebitosće tekstow

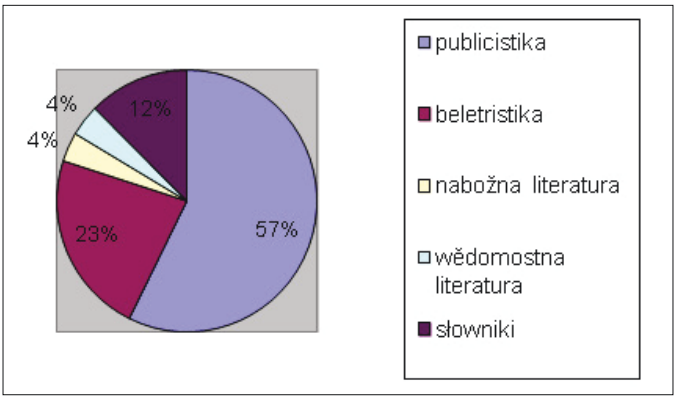
Dźakowano zrěčenjam z Ludowym nakładnistwom Domowina w Budyšinje, w kotrymž wuchadźa wjetšina publikacijow w hornjoserbskej rěči, a z Rěčnym centrumom WITAJ – to je wotrjad Domowiny, kotryž mjez druhim wudawa šulske wučbnicy a druhe wučbne srědki – smy móhli za rešeršowanje spřistupnić wobšěrnu zběrku aktualnych tekstow – nowinow, časopisow a knihow. Při tym ma so wězo wobkedźbowanje awtorskich prawow zaručić. Tohodla je maksimalna wulkosć pokazowaneho konteksta pytaneho słowa wobmjezowana na 100 tokenow. Najwjetši dźěl tekstow za korpus pak je so z pomocu wědomostno-techniskeho personala Serbskeho instituta zaskenował a wobdźěłał z pismo spóznawacymi OCR-programami. Dokelž njejsu so wuslědki tuteje procedury njedosahaceho personala dla w dosahacej měrje korigowali, njemóža so bohužel w tekstach korpusa zmylki wuzamknyć, za korektne citowanje je trjeba sahać po original. Tu čaka nas w přichodźe hišće wjele korekturneho dźěła, tola tuchwilu ma hišće prioritu kwantitatiwny přirost korpusa.

Wěsty problem za linguistiske korpusowe slědźenje je fakt, zo njejsu citowane pasaže w druhich rěčach, mjez druhim tež w delnjoserbšćinje a němčinje, wosebje markěrowane. Za serbskeho wužiwarja abo dobreho znajerja serbšćiny pak je poprawom 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") – dźensnišemu *š* wotpowěduje tu na př. *sch* (*mysch* město *myš*), za *ć/č* pisa so *cź/cž* (*cžakacź* město *čakać*), za *s/z* pisa so *ß/s* (*ßasy* město *sazy*). Pisanje *kh* za dźensniše *h* na spočatku morfemow a mjechke *ŕ* pak namakatej so tež hišće w tekstach nastatych w 20. lětstotku do lěta 1945. Na internetnej stronje Serbskeho instituta smy wo tym zaměstnili nadrobnu informaciju, kotraž móže wužiwarjej pomhać wšitke přikłady za swoje naprašowanje namakać, hačrunjež jewja so prawopisne warianty.

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 „tokenow", potajkim jednotliwych separatnych słownych formow[1]. Dokelž liča jako tokeny tež w teksće wustupowace ličby a nimo toho kóžde interpunkciske znamješko, je w korpusu trochu mjenje słowow hač tokenow (ca. 36 milionow).

Teksty hornjoserbskeho korpusa słušeja do wšelakich kategorijow: 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 %)[2].

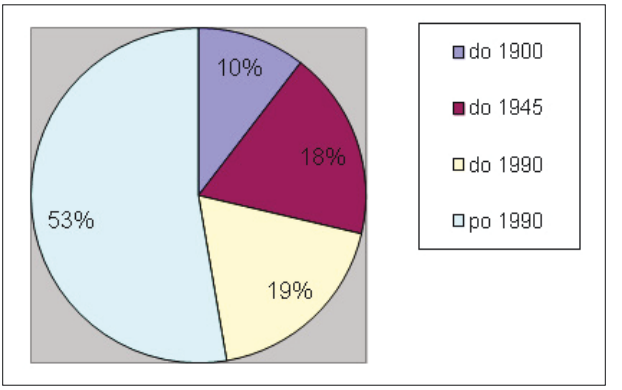


Wobr. 3. Podźěl tekstowych družin na korpusu

Najstarše teksty sahaja hač do prěnjeje połojcy 19. lětstotka, su to spisy Handrija Zejlerja a basnje Rudolfa Mjenja w nowowudaćach z 20. lětstotka a lětniki „Časopisa Maćicy Serbskeje" wot lěta 1848, najmłódše su „Serbske Nowiny"

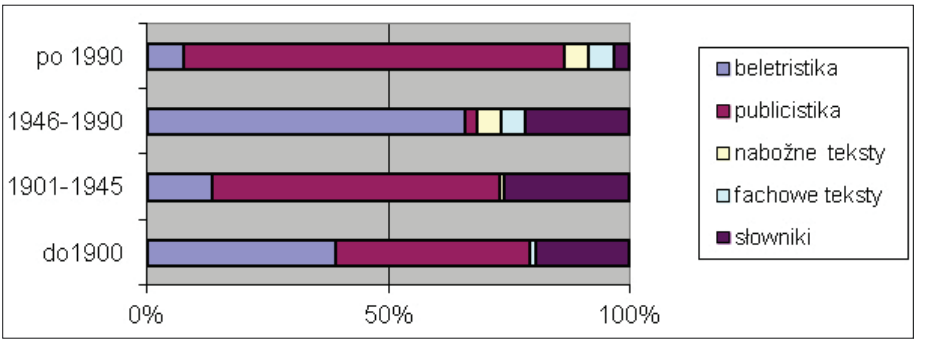
[1] Zestajane časowe formy kaž na př. perfekt *sym spěwał* liča jako dwě słowje.
[2] Procentualne podźěle su so wuličili na zakładźe ličby tokenow (hlej horjeka).

a časopis „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, tola wón dowola we wěstej měrje tež diachroniski přistup.



Wobr. 4. Podźěl tekstow ze wšelakich dobow

Hdyž wobhladamy sej rozdźělenje tekstowych družin na jednotliwe časowe wotrězki, pytnjemy, zo njeje jich podźěl přeco jenaki. Při dalšim rozšěrjenju korpusa budźe so wosebje na wurunanje disproporcijow mjez wšelakimi tekstowymi typami a časowymi dobami dźiwać dyrbjeć, najtrěbniše je wudospołnjenje korpusa za druhu połojcu 20. lětstotka ze zběrku publicistiki.



Wobr. 5. Podźěl tekstowych družin po časowych dobach

Dokelž so hornjoserbski tekstowy korpus běžnje wudospołnja, dyrbi so rjec, zo njejedna so wo tak mjenowany referenčny korpus, to rěka wo statiski korpus,

z kotrehož móža so konstantne statistiske daty dobywać. Wuhódnoćenja wotbłyšćuja jenož aktualny staw w tekstach korpusa.

Přistup k hornjoserbskemu tekstowemu korpusej

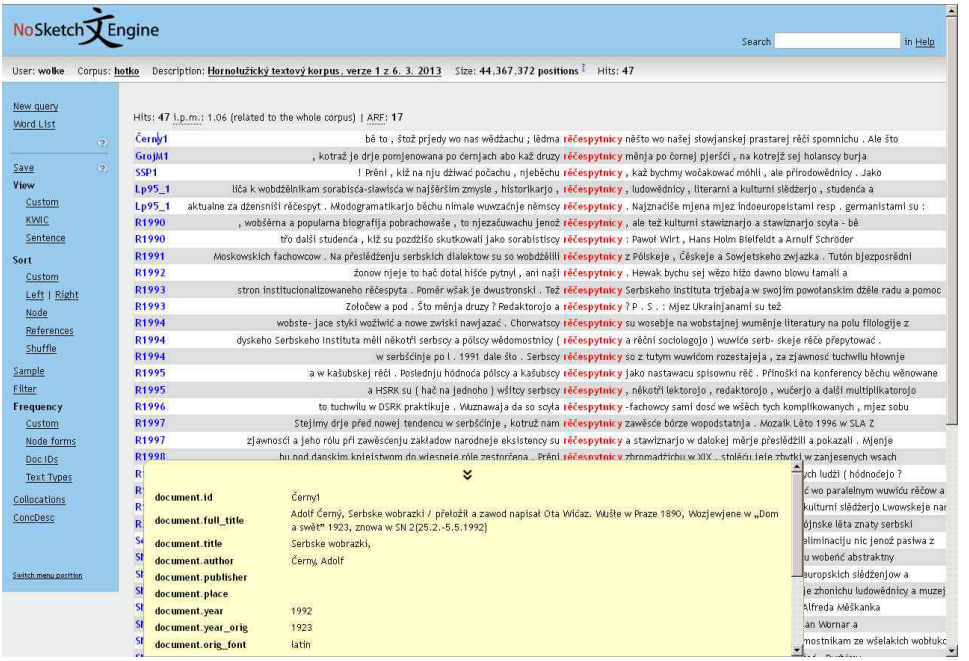
HOrnjoserbski Tekstowy KOrpus, skrótka HOTKO, je zaměstnjeny na stronje Serbskeho instituta pod menijowym dypkom *online-publikacije*. Za serbskich wužiwarjow smy přihotowali nadrobnu dokumentaciju a pomocne informacije w hornjoserbskej rěči, kotrež namakaće, hdyž nakliknjeće jednotliwe dypki pod krótkim zawodnym tekstom.

Wobr. 6. Přistup k HoTKo na stronje http://www.serbski-institut.de/cms/os/48/hornjoserbski

Dokelž zmóžnja so přepytowanje hornjoserbskeho tekstoweho korpusa přez kooperaciju z Institutom za Češki narodny korpus na Filozofiskej fakulće Karloweje uniwersity w Praze (www.korpus.cz), dyrbja sej wužiwarjo na jeho stronach wosobinske wužiwarske konto założić. Za tutu proceduru namakaja so na mjenowanej internetowej stronje Serbskeho instituta wotpowědne serbskorěčne wujasnjenja (menijowy dypk: přistup ke korpusej).

Štóž je sej na te wašnje přistup zarjadował, może z pomocu wotprašowanskeho programa Češkeho narodneho korpusa NoSketchEngine za wustupowanjom wěstych słowow w zapřijatych tekstach pytać – při tym namakaja so awtomatisce konkordancy za słowa, jich formy abo kombinacije. Koděrowane su tež metainformacije wo žórłach: Podaća wo awtorje, titlu, městnje a času wuchadźenja móža so jako skrótšenka na lěwym boku abo tež dokładnišo

we wosebitym woknješku pokazować. Też šěrši kontekst hač do 100 słowow hodźi so pokazać.



Wobr. 7. Konkordanca zdobyta w HoTKo z rozšěrjenym podaćom žórła

W tutych konkordancach płaća pytanske a statistiske możnosće programa NoSketchEngine: Hodźi so na př. přepytować, kak husto wěste słowa abo słowne formy wustupuja a w kajkim susodstwje so wosebje jewja, może so też zwěsćić, hač su za wěstych awtorow typiske, na wěste družiny tekstow abo snano na wěstu časowu dobu wobmjezowane. Program zmóžnja też pytanje słowow w definowanych kontekstach a dowoli indiwiduelny wuběr, w kotrych tekstach ma so pytać, z pomocu definowanja wosobinskich subkorpusow.

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 możnosće wosebje za rěčespytne prašenja z wobłuka syntaksy a morfologije. Za namakanje po możnosći wšěch formow někajkeho pytaneho słowa móža sej wužiwarjo pomhać z tak mjenowanymi regularnymi wurazami (jendź. regular expressions), w kotrychž wužiwaja so znamješka, kotrež funguja jako naměstniki za wšelake warianty – za

to wobsteja wšelake w nadrobnosćach so wotchilace standardy. Za HoTKo płaćace zasady, kotrež bazuja wězo na zasadach Češkeho narodneho korpusa, su w lisćinje na stronje Serbskeho instituta pod dypkom „regularne wurazy" wopisane.

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 zależi na tym, zo móža teksty korpusa nimo prostych OCR-zmylkow tež rozšěrjene rěčne zmylki dokumentować – hranica mjez zmylkom a nowej, so šěrjacej formu tež za rěčespytnika druhdy cyle wótra njeje.

Dalše informaciske srědki wo hornjoserbskej rěči w interneće

Přestajeny hornjoserbski tekstowy korpus dyrbi so widźeć w konteksće dalšich projektow Budyskeho rěčespytneho wotrjada Serbskeho instituta. Ćežišćo tworja tu prašenja słowoskłada a leksikografije. Centralny zaměr předstaja projekt wobšěrneho noweho němsko-hornjoserbskeho słownika, kotryž tuchwilu w našim wotrjedźe koncipujemy – z tym reagujemy na potrjeby serbskeje rěčneje praksy, kotraž žada sej wjac hač dwaj lětdźesatkaj po wudaću poslednjeho tajkeho kompendija (Jenč, Michałk & Šěrakowa, 1989–1991) nowy a metodologisce dalewuwity pomocny grat. Planowany słownik je primarnje za digitalne wužiwanje w interneće mysleny. Njeměrka pak to, zo njebudźe so z njeho sekundarnje tež ćišćany słownik generěrować dać. Technologisce budźe so wón zepěrać na nazhonjenja wotrjada za delnjoserbske slědźenje, a tež koncepcionelnje matej wobaj słownikaj mjezsobu kompatibelnej być. Tole wšak je – hladajo na hornjoserbšćinu – tuchwilu hišće hudźba přichoda. Zrealizowane su porno tomu hižo wěste předdźěła.

Projekt *Němsko-hornjoserbskeho słownika noweje leksiki* wobdźěłany wot Helmuta Jenča, Anje Pohončoweje a Jany Šołćineje (Jenč, Pohončowa & Šołćina, 2006), wobsahowacy někak 12000 hesłow, drje njeje ani za internet koncipowany ani w nim přistupny, jako zběrka, registrowaca nowu leksiku po wudaću němsko-hornjoserbskeho słownika 1989–1991, pak ma so wón jako krok po puću k planowanemu internetowemu słownikej widźeć.

Prěni z digitalnych, online přistupnych poskitkow SI za hornjoserbšćinu, kiž chcu tu mjenować, je *Hornjoserbski frazeologiski słownik* w interneće. Wón je nastał na zakładźe hornjoserbsko-němsko-ruskeho frazeologiskeho słownika,

kotryž smój hromadźe z Anatolijom Iwčenku w lěće 2004 wudałoj (Ivčenko & Wölke, 2004). Internetna wersija, kotraž je wot 2005 přistupna (http://www.serbski-institut.de/cms/de/50/Obersorbisches-phraseologisches-Woerterbuch), wobmjezuje so na hornjoserbšćinu a němčinu, dowoli pak za to wobšěrniše pytanske možnosće wuchadźejo z wobeju rěčow. Nimo toho so tuta digitalna wersija běžnje z přidatnymi frazeologizmami wudospołnja.

Specialnu leksikografisku zběrku předstaja datowa banka geografiskich mjenow w hornjoserbšćinje a němčinje (Geografiske mjena hornjoserbsce, n.d.; http://www.serbski-institut.de/cms/os/396/Geografiske-mjena-hornjoserbsce). Wona je nastała w zhromadnym dźěle z Rěčnym centrumom WITAJ, hdźež nastawaja serbske wučbnicy, mjez nimi tež tajke za geografiju. Tuchwilu su w datowej bance předewšěm zezběrane tak mjenowane eksonymy (něhdźe 2000), potajkim geografiske mjena za objekty zwonka Serbow, tola planujemy ju rozšěrić tež wo hornjoserbske geografiske mjena z Łužicy. Wotprašowanje datow je možne wuchadźejo z němčiny, ale tež wuchadźejo z hornjoserbšćiny. Systematiske wobdźěłanje tutoho wosebiteho dźěła słowoskłada w cyłku je zakład za jeho po možnosći dospołne a zdobom konsekwentne a we sebi konsistentne předstajenje w planowanym němsko-hornjoserbskim słowniku a služi na druhim boku stabilizaciji normy spisowneje rěče.

Zwonka rěčespytneho wotrjada je so w nadawku Serbskeho instituta zdźěłała zběrka w interneće přistupnych hornjoserbskich dwurěčnych słownikow SERDIS – Serbske digitalne słowniki. SERDIS nasta na iniciatiwu a z koordinaciju přez Franca Šěna w zhromadnym dźěle Serbskeho instituta w Heidelbergu skutkowacym neurobiologom Jurjom Brankačkom a ukrainskim programowarjom Vladimirom Kukušku. Online přistupne, programowane w Java®, namakaja so tam digitalne wersije hornjoserbsko-němskeju słownikow Křesćana Bohuwěra Pfula (Pful, 1866) a Jurja Krala (Kral, 1927), hornjoserbsko-ruskeho słownika Konstantina Trofimowiča (Trofimowič, 1974) a dwě dalšej mjeńšej leksikografiskej zběrce. W SERDIS móže so principielnje w dwěmaj směromaj pytać – tola pytanje wuchadźejo z ciloweje rěče je skerje wobmjezowane dla přesnadneje hłubokosće a nadrobnosće analyzy słownikowych artiklow. Za dalewuwiće tutoho online-poskitka by derje było, so orientować na metodach, kiž su so nałožowali za wotpowědny delnjoserbski poskitk.

Hornjoserbski tekstowy korpus twori ważny materialowy zakład za wšě leksikografiske dźěła rěčespytneho wotrjada Serbskeho instituta. Intensiwnje je so wón hižo wužiwał za wšitke tři runje mjenowane hižo zrealizowane pro-

jekty: słownik noweje leksiki (Jenč et al., 2006), frazeologiski słownik (Ivčenko & Wölke, 2004) a datowu banku geografiskich mjenow, nimo toho też zwonka SI za jendźelsko-hornjoserbski słownik Měrćina Straucha a Edwarda Wornarja (Wornar & Strauch, 2007). Porno tradicionalnym metodam leksikografije před dobu kompjutera – jako dźěłaše so ze statiskimi papjerjanymi kartotekami – mamy z tym wulke lěpšiny. Hladajo na wobjim (tuchwilu 36 mio słowow porno 1 mio kartkow za dwuzwjazkowy němsko-hornjoserbski słownik Jenča et al. 1989–1991) je digitalny korpus wo wjele spušćomniše wuchadźišćo. Za digitalny korpus steja k dispoziciji wjele diferencowaniše móžnosće analyzy, na př. po frekwency a konteksće. Nimo toho njeje digitalny korpus lokalnje wjazany na městno kartoteki, ale je přistupny zjawnje po cyłym swěće, a to nic jenož za leksikografiske, ale tohorunja za druhe rěčespytne předewzaća, a z přiběracym mnóstwom zapřijatych tekstow też přeco lěpje za rešerše literaturowědneho abo stawizniskeho razu.

Bibliography

Dolnoserbski. (n.d.). Retrieved December 19, 2013, from www.dolnoserbski.de

Geografiske mjena hornjoserbsce. (n.d.). Retrieved from http://www.serbski-institut.de/cms/os/396/Geografiske-mjena-hornjoserbsce

Ivčenko, A., & Wölke, S. (2004). *Hornjoserbski frazeologiski słownik* [Obersorbisches phraseologisches Wörterbuch; *Верхнелужицкий фразеологический словарь*]. Budyšin: Ludowe nakładnistwo Domowina. Retrieved from http://www.serbski-institut.de/cms/de/50/Obersorbisches-phraseologisches-Woerterbuch

Jenč, H., Michałk, F., & Šěrakowa, I. (1989–1991). *Němsko-hornjoserbski słownik* (Vol. 1–2). Budyšin: Ludowe nakładnistwo Domowina.

Jenč, H., Pohončowa, A., & Šołćina, J. (2006). *Němsko-hornjoserbski słownik noweje leksiki*. Budyšin: Ludowe nakładnistwo Domowina.

Kral, J. (1927). *Serbsko-němski Słownik hornjołužiskeje rěče* [Wendisch-deutsches Wörterbuch der oberlausitzer Sprache]. Budyšin: Maćica Serbska.

Pful, K. B. (1866). *Łužiski serbski słownik: Lausitzisch Wendisches Wörterbuch*. Budyšin: Maćica Serbska.

Šěn, F. (2013). Sorabistika a nowe medije. In M. Milewska-Stawiany, & S. Wölkowa (Eds.), *Leksikologiske přinoški II: IV. seminar serbskeje słowotwórby a leksiki* [IV Seminarium Słowotwórstwa i Słownictwa Łużyckiego] : *Uniwersytet Gdański. Serbski institut 31.5. – 1.6.2012* (pp. 131–141). Budyšin: Serbski institut.

Šołćina, J., & Wornar, E. (1996). Čehodla trjebamy jednoreč̌ny hornjoserbski słownik? In E. E. Siatkowska, & J. Molas (Eds.), *Sprawy łużyckie w ich słowiańskich kontekstach* (pp. 15–22). Warszawa: Instytut Filologii Słowiańskiej UW.

Trofimowič, K. (1974). *Hornjoserbsko-ruski słownik* [*Верхнелужицко-русский словарь*]. Budyšin: Ludowe nakładnistwo Domowina.

Wölkowa, S. (2013). Hornjoserbski tekstowy korpus w nowej formje. *Serbska Šula*, *66*(2), 44–47.

Wornar, E., & Strauch, M. (2007). *Jendźelsko-hornjoserbski šulski słownik* [English-Upper Sorbian Learner's Dictionary]. Budyšin: Ludowe nakładnistwo Domowina.

Bibliography (transliteration)

Dolnoserbski. (n.d.). Retrieved December 19, 2013, from www.dolnoserbski.de

Geografiske mjena hornjoserbsce. (n.d.). Retrieved from http://www.serbski-institut.de/cms/os/396/Geografiske-mjena-hornjoserbsce

Ivčenko, A., & Wölke, S. (2004). *Hornjoserbski frazeologiski słownik* [Obersorbisches phraseologisches Wörterbuch; Verkhneluzhitskiĭ frazeologicheskiĭ slovar']. Budyšin: Ludowe nakładnistwo Domowina. Retrieved from http://www.serbski-institut.de/cms/de/50/Obersorbisches-phraseologisches-Woerterbuch.

Jenč, H., Michałk, F., & Šěrakowa, I. (1989–1991). *Němsko-hornjoserbski słownik* (Vol. 1–2). Budyšin: Ludowe nakładnistwo Domowina.

Jenč, H., Pohončowa, A., & Šołćina, J. (2006). *Němsko-hornjoserbski słownik noweje leksiki*. Budyšin: Ludowe nakładnistwo Domowina.

Kral, J. (1927). *Serbsko-němski Słownik hornjołužiskeje rěče* [Wendisch-deutsches Wörterbuch der oberlausitzer Sprache]. Budyšin: Maćica Serbska.

Pful, K. B. (1866). *Łužiski serbski słownik: Lausitzisch Wendisches Wörterbuch*. Budyšin: Maćica Serbska.

Šěn, F. (2013). Sorabistika a nowe medije. In M. Milewska-Stawiany & S. Wölkowa (Eds.), *Leksikologiske přinoški II: IV. seminar serbskeje słowotwórby a leksiki* [IV Seminarium Słowotwórstwa i Słownictwa Łużyckiego]*: Uniwersytet Gdański. Serbski institut 31.5. – 1.6.2012* (pp. 131–141). Budyšin: Serbski institut.

Šołćina, J., & Wornar, E. (1996). Čehodla trjebamy jednoreč̌ny hornjoserbski słownik? In E. E. Siatkowska & J. Molas (Eds.), *Sprawy łużyckie w ich słowiańskich kontekstach* (pp. 15–22). Warszawa: Instytut Filologii Słowiańskiej UW.

Trofimowič, K. (1974). *Hornjoserbsko-ruski słownik* [Verkhneluzhitsko-russkiĭ slovar']. Budyšin: Ludowe nakładnistwo Domowina.

Wölkowa, S. (2013). Hornjoserbski tekstowy korpus w nowej formje. *Serbska Šula*, *66*(2), 44–47.

Wornar, E., & Strauch, M. (2007). *Jendźelsko-hornjoserbski šulski słownik* [English-Upper Sorbian Learner's Dictionary]. Budyšin: Ludowe nakładnistwo Domowina.

The Upper Sorbian text corpus and further sources of information with regard to Upper Sorbian in the Internet

Summary

In the present era of globalisation and the omnipresence of the Internet, Sorbian linguistics faces new challenges along the lines "What is not in the Internet, does not exist" The demand for digital sources of information with regard to Upper and Lower Sorbian and those accessible online as working tools and reference points for language practice and as a source for academic research increases. As a result of this ongoing development, the *Foundation for the Sorbian People* established a workgroup called "Sorbian in the new media" at the end of 2012, which has pointed out the creation of an online German-Upper Sorbian dictionary as the major task in this field of activities. The focus of this article, however, is the Upper-Sorbian text corpus HoTKo, which has been created by the *Sorbian Institute* and which has been made available in co-operation with the Institute of the Czech National Corpus at the Charles University in Prague. The article presents the history and development of the corpus, its extent and shape as well as its link to or incorporation into further planned digital projects of the *Sorbian Institute* with regard to the Upper Sorbian language.

Keywords: computational lexicography; corpus linguistics; digitalization; text corpus; Upper Sorbian language

Słowa kluczowe: digitalizacja; język górnołużycki; korpus językowy; leksykografia komputerowa; lingwistyka korpusowa