

Studie k problematice korpusů ve vztahu k romštině¹

Korpusy romštiny

Při první chvatné rešerši se zdá, že korpusů romštiny není zatím ve světě mnoho. Zde uvádím informace o třech.

1 Korpus ROMI

Z adresy <http://lidemesta.cz/archiv/cisla/13-2011-1/romi-prvni-rozsahla-databanka-romskeho-etnolektu-cestiny.html> přejímám:

„ROMI je rozsáhlá databanka romského etnolektu češtiny. Vzniká jako subkorpus rozsáhlé databanky CZESL, jejímž cílem je zachytit češtinu cizinců a Romů. Databanka má sloužit především pro pedagogické účely - v první řadě jako zdroj pro analýzu jazykových kompetencí těchto skupin uživatelů češtiny a její využití ve výuce: ke zmapování hlavních problémů, při sestavování cvičebnic, při explanaci konkrétních jazykových jevů, vyhledání relevantních příkladů a podobně. Tomuto zaměření odpovídá jednak uživatelská přístupnost databanky (snadné vyhledávání, široce uživatelsky zaměřený přepis nahrávek a textů ad.), jednak věkové vymezení mluvčích a další zohledňované parametry zaměřené na školní prostředí.

ROMI představuje naprosto ojedinělý, rozsáhlý soubor textů a nahrávek romských mluvčích češtiny od předškolního věku do cca 26 let, který přináší poprvé v dějinách české lingvistiky a romistiky takto rozsáhlý jazykový materiál romského etnolektu. ROMI bude přístupný ve formě přepisu textů a nahrávek širší odborné veřejnosti (podobně jako např. Český národní korpus), nahrávky a originály textů pak v určitých případech (např. za účelem fonetického rozboru). Jazykový materiál je unikátní nejen svým rozsahem (k 17. 3. 2011 obsahuje 2 466 písemných textů a 497 zhruba 10-20minutových

¹ Tento text představuje excerpt z dokumentu *Analýza potenciálu jazykových technologií při revitalizaci menšinových jazyků se zaměřením na romštinu*, zpracovaného pod vedením T. Svobody jako součást stejnojmenného projektu podpořeného v dotačním programu Podpora implementace Evropské charty regionálních či menšinových jazyků v r. 2015. Vzhledem ke stěžejní relevanci textu jej pro účely aktuálního projektu přejímáme, s drobnými úpravami.

nahrávek, přičemž sběr dat probíhá od října 2009 a pokračovat bude cca do října 2011), ale také celorepublikovým zaměřením (dosud jediná systematická studie romského etnolektu M. Bořkovcové [Romský etnolekt češtiny. Signeta, Praha 2006] se zaměřuje v první řadě na jednu konkrétní komunitu obývajících v době výzkumu pražský Smíchov) a pestrostí zkoumaných prostředí. Do projektu se zapojila řada základních škol všech typů (běžné základní školy, základní školy speciální a praktické), ale také řada individuálních spolupracovníků z neziskových organizací romských i neromských, romští pedagogičtí asistenti i individuální výzkumníci. Jako unikátní zdroj nejen jazykových dat se osvědčila spolupráce s jedním romským sdružením, jehož členové dosud nahráli několik desítek nahrávek přímo ve své komunitě (a v nahrávkách pokračují). Cílem projektu totiž není jen shromáždit materiál „zvenku“, ale zapojit do vybudování databanky i romské mluvčí.

2 Korpus severocentrální romštiny jako součást korpusu InterCorp

Autor tohoto textu navrhl asi před čtyřmi lety pracovníkům **Ústavu Českého národního korpusu** ([viz https://www.korpus.cz/](https://www.korpus.cz/)), který je součástí Karlovy univerzity v Praze, vytvoření korpusu vytvořeného z romských textů, nebo ještě spíše korpusu romsko-českého, který by se stal součástí vícejazyčného korpusu **InterCorp**.

Korpus InterCorp je hlavním výstupem stejnojmenného projektu, jehož cílem je vybudovat rozsáhlý paralelní synchronní korpus, pokrývající co největší počet jazyků.

Byl jsem ustanoven koordinátorem tohoto korpusu pro romštinu. Práce na romském korpusu jsou ovšem na úplném počátku. Pro korpus jsem naskenoval zatím asi 10 romských děl, většinou z krásné literatury, a připravil s pomocí Elišky Bokové, spolupracovnice ČNK dvě díla pro zařazení do databáze InterKorpu.

Pro zařazení do korpusu byla v roce 2015 naskenována a k dalšímu zpracování připravena tato díla:

Fabiánová, Tera: Čavargoš : [romaňi paramisi] = Tulák : [romská pohádka] / Tera Fabiánová, Milena Hübschmannová ; [ilustrovala Renata Fučíková] Vyd. 1. Apeiron, 1991

God'aver lava phure Romendar = Moudrá slova starých Romů / [přísloví sebrali Milena Hübschmannová ... et al. ; přispěli Marta Bandyová ... et al. ; přeložila a uspořádala Milena Hübschmannová]. 2., rozš. vyd., v nakl. Apeiron 1. vyd. Praha : Apeiron, c1991

Z druhého díla uvádím ukázkou výskytu slova „lav“ (slovo), vytvořenou zatím je ručně ne mém počítači:

God'aver lava phure Romendar	Moudrá slova starých Romů
------------------------------	---------------------------

Maribnaha na kereha čhavoreha nič, ča laveha.	Ranami dítě nevychováš - jenom slovem.
Lav šaj avel tho maro the čhuri.	Slovo může být chlebem i nožem.
Gule lavendar na čafoha.	Sladká slova tě nenasytí.
Ma dikh pro lava, dikh pro vasta.	Nevšímej si slov, ale rukou.
Te našti des maro, de choča lačo lav.	Nemůžeš-li dát chleba, dej alespoň dobré slovo.
Tiri buťi tut bararel, na tire lava.	Povýší tě činy, ne slova.
Andro muj gule lava, e čhuri andre baj.	V ústech sladká slova, v rukávu nůž.
Lačo lav sar maro.	Dobré slovo je jako chleba.

Obrázek č. 1. Ukázka výskytu slova „lav“ v díle „Godaver lava“.



Obrázek č. 2. Ukázka obrazovky nástroje Intertext, s jehož pomocí se zarovnávají uložené texty. V tomto případě jde o text „Čavargoš“.

V roce 2016 se plánuje zařadit dalších asi šestnáct děl. Budou to například:

Fabiánová, Tera: *Sar me phiravas andre škola = Jak jsem chodila do školy. 1. vyd. České Budějovice : ÚDO ve spolupráci se Společenstvím Romů na Moravě, 1992*

Giňa, Andrej: *Paťiv : ještě víme, co je úcta : vyprávění, úvahy, pohádky. Vyd. 1. Praha : Triáda, 2013*

Horvátová, Agnesa: *Pal e Bari Rama the aver paramisa = O Velké Ramě a jiné příběhy. Praha : Signeta, 2003*

Hübschmannová, Milena: *Romské hádanky : hin man ajsi čhaj, so-*. Vyd. 2., přeprac., *Ve Fortuně 1*. Praha : Fortuna, 2003

Po Židoch Cigáni : svědectví Romů ze Slovenska 1939-1945. Vyd. 1. *Triáda*, 2005-

Oláh, Vlado: *Le khameskere čhave = Děti slunce : romská próza a poezie*. Vyd. 1. Praha : Matice romská, 2003

O evangelijum le Jaňustar. Vyd. 1. Praha : Česká biblická společnost, 1997

Pal oda, so kerenas le devleskere bičhade = Skutky apoštolů. 1. romsko-české vyd. Praha : G plus G : Česká biblická společnost : Matice romská, 2000

Rád bych zařadil také texty z romských a romistických časopisů a novin.

Korpus může sloužit např. při praktickém používání jazyka (mimo jiné při překládání), lze s jeho pomocí sledovat a předpovídat vývoj romštiny, vytvářet jeho pomocí (pravděpodobně reprezentativnější, než klasickým způsobem) další jazykové pomůcky a nástroje, jako mimo jiné slovníky (včetně frekvenčního a retrogradního a frazeologického) a korektor pravopisu.

3 Korpus olašské romštiny

Jako vhodné se jeví zahájit práce také na korpusu druhého významného dialektu romštiny v České republice, totiž olašské romštiny. Olašských textů je publikováno značně méně než textů v severocentrálním dialektu, je však pravděpodobné, že jich bude v budoucnu přibývat.

4 Korpus romštiny ve sbírce Pangloss Collection

Na webové stránce http://lacito.vjf.cnrs.fr/pangloss/languages/Romani_en.htm je údaj o romském korpusu věnovaného romštině v Řecku (podle uvedené stránky se jedná o dvě varianty: olašskou romštinu a romštinu „balkánskou“). Korpus, který se nezaměřuje jen na romštinu, obsahuje 3 vyprávění zaznamenaná badatelkou Evangelia Adamou ve zvukové a grafické podobě, přičemž romský text je doprovázen souběžným anglickým překladem. Zvuková podoba je opatřena tagy. Uvedeme ukázkou prvního příběhu, nazvaného **The louse and the Rom (Veš a Rom)**.

Lacito Home > Pangloss Home

Citation

The louse and the Rom [ⓘ]

Language: Romani

Researcher(s): Adamou, Evangelia

Your browser does not support the audio tag Continuous playing:

Transcription by sentence

Phonologic Whole text transcription Glosses

Translation by sentence

en Whole text translation en

Words in italics = Words from contact languages

S1 **jek naj sas duj naj sas ek zamano sas ek patifaj**

jek naj sas duj naj sas ek zamano sas ek patifaj

one be.NEG.3SG was.3SG two be.NEG.3SG was.3SG one time was.3SG one king

Once upon a time there was a king.

Obrázek č. 3: Ukázka z korpusu řecké romštiny ve sbírce Pangloss Collection

5 Korpus Opus

Bezplatný korpus Opus na stránkách <http://opus.lingfil.uu.se/> obsahuje velké množství romských textů. Jak je možno tento korpus využívat, a o jaké romské texty jde, musím teprve zjistit.

Home / Query / WordAlign / Wiki [books] [DGT] [DOCC] [ECB] [EMEA] [EUbooks] [EU] [Europarl] [GNOME] [Hans] [JRC] [KDE4/doc] [MBS] [MultiUN] [NCv9] [OO/OO3] [subs/12/13] [ParCor] [PHP] [SETIMES] [SPC] [Tatoeba] [TEP] [TedTalks] [Tanzil] [Ubuntu] [UN] [WikiSource] [WMT]

OPUS

... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving...
Contributions are very welcome! Please contact <jorg.niedemann@lingfil.uu.se >

Search & download resources: rom (Romany) any (-- any language --) all

Language resources: click on [tmx] [moses] [xces] [lang-id] to download the data (raw = untokenized, true = truecaser model, TM = phrase-based translation model)

corpus	doc's	sent's	src tokens	trg tokens	XCES/XML	raw	TMX	Moses	mono	raw	true	TM	dic	freq	Browse Files
Ubuntu	3	32	78	86	[xces ace rom]	[ace rom]	[tmx] [moses]	ace rom	ace rom						[sample] [xml/ace][xml/rom]
Ubuntu	1	0	0	0	[xces ady rom]	[ady rom]	[tmx] [moses]	ady rom	ady rom						[sample] [xml/ady][xml/rom]
Ubuntu	3	32	77	86	[xces af rom]	[af rom]	[tmx] [moses]	af rom	af rom						[sample] [xml/af][xml/rom]
Ubuntu	2	6	9	13	[xces ak rom]	[ak rom]	[tmx] [moses]	ak rom	ak rom						[sample] [xml/ak][xml/rom]
Ubuntu	3	32	76	86	[xces am rom]	[am rom]	[tmx] [moses]	am rom	am rom						[sample] [xml/am][xml/rom]
Ubuntu	2	32	92	86	[xces an rom]	[an rom]	[tmx] [moses]	an rom	an rom						[sample] [xml/an][xml/rom]
Ubuntu	3	32	73	86	[xces ar rom]	[ar rom]	[tmx] [moses]	ar rom	ar rom						[sample] [xml/ar][xml/rom]
Ubuntu	1	0	0	0	[xces ar rom]	[ar rom]	[tmx] [moses]	ar rom	ar rom						[sample] [xml/ar][xml/rom]
Ubuntu	1	0	0	0	[xces ar rom]	[ar rom]	[tmx] [moses]	ar rom	ar rom						[sample] [xml/ar][xml/rom]
Ubuntu	3	32	80	86	[xces as rom]	[as rom]	[tmx] [moses]	as rom	as rom						[sample] [xml/as][xml/rom]
Ubuntu	3	32	86	86	[xces ast rom]	[ast rom]	[tmx] [moses]	ast rom	ast rom						[sample] [xml/ast][xml/rom]
Ubuntu	3	31	71	83	[xces az rom]	[az rom]	[tmx] [moses]	az rom	az rom						[sample] [xml/az][xml/rom]
Ubuntu	1	0	0	0	[xces ba rom]	[ba rom]	[tmx] [moses]	ba rom	ba rom						[sample] [xml/ba][xml/rom]
Ubuntu	3	32	88	86	[xces be rom]	[be rom]	[tmx] [moses]	be rom	be rom						[sample] [xml/be][xml/rom]
Ubuntu	1	0	0	0	[xces be rom]	[be rom]	[tmx] [moses]	be rom	be rom						[sample] [xml/be][xml/rom]
Ubuntu	1	0	0	0	[xces be rom]	[be rom]	[tmx] [moses]	be rom	be rom						[sample] [xml/be][xml/rom]
Ubuntu	3	32	85	86	[xces bg rom]	[bg rom]	[tmx] [moses]	bg rom	bg rom						[sample] [xml/bg][xml/rom]
Ubuntu	3	32	76	86	[xces bn rom]	[bn rom]	[tmx] [moses]	bn rom	bn rom						[sample] [xml/bn][xml/rom]
Ubuntu	3	32	33	86	[xces bo rom]	[bo rom]	[tmx] [moses]	bo rom	bo rom						[sample] [xml/bo][xml/rom]
Ubuntu	3	32	89	86	[xces br rom]	[br rom]	[tmx] [moses]	br rom	br rom						[sample] [xml/br][xml/rom]
Ubuntu	1	0	0	0	[xces br rom]	[br rom]	[tmx] [moses]	br rom	br rom						[sample] [xml/br][xml/rom]
Ubuntu	3	32	78	86	[xces bx rom]	[bx rom]	[tmx] [moses]	bx rom	bx rom						[sample] [xml/bx][xml/rom]
Ubuntu	1	0	0	0	[xces byn rom]	[byn rom]	[tmx] [moses]	byn rom	byn rom						[sample] [xml/byn][xml/rom]
Ubuntu	3	32	89	86	[xces ca rom]	[ca rom]	[tmx] [moses]	ca rom	ca rom						[sample] [xml/ca][xml/rom]

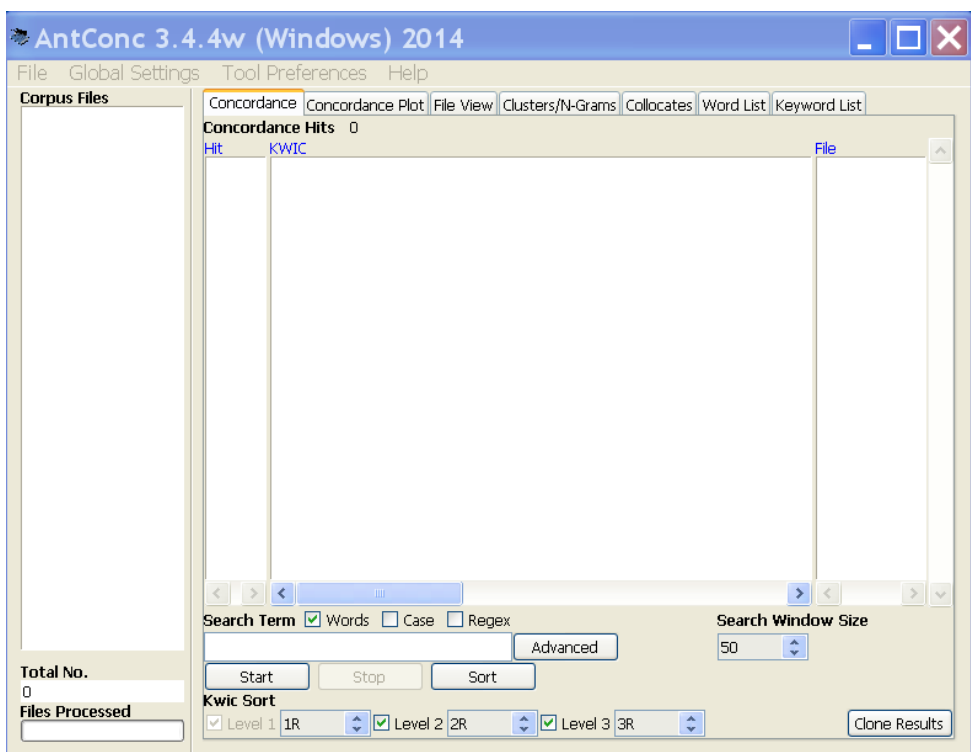
Latest News

- 2015-10-15: New versions of TED2013, NCv9
- 2014-10-24: New JRC-Acquis
- 2014-10-20: NCv9, TED talks, DGT, WMT
- 2014-08-21: New Ubuntu, GNOME
- 2014-07-30: New Translated Books
- 2014-07-27: New DOGC, Tanzil
- 2014-05-07: Parallel coref corpus ParCor

Obrázek č. 4. Začátek soupisu romských textů v korpusu Opus

6 Korpus AntConc

Uživatel romštiny, který chce romštinu, respektive texty v jiných jazycích podrobit zkoumání, může využít bezplatný korpus AntConc autora Laurence Anthonyho, působícího na Faculty of Science and Engineering na Waseda University v Japonsku. Korpus lze stáhnout z adresy <http://www.laurenceanthony.net/software.html> spolu s dalšími užitečnými bezplatnými nástroji vytvořenými autorem.



Obrázek č. 5. Výchozí obrazovka korpusu AntConc 3.4w