**Ruben Pellar**

**Studie k problematice korpusů ve vztahu k romštině**[1]

# 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 obývající 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/),** 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, spolupracovnici 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 ukázku výskytu slova „lav" (slovo), vytvořenou zatím je ručně ne mém počítači:

| | |
|---|---|
| Goďaver lava phure Romendar | Moudrá slova starých Romů |

| | |
|---|---|
| Maribnaha na kereha čhavoreha ňič, ča laveha. | Ranami dítě nevychováš - jenom slovem. |
| Lav šaj avel tho maro the čhuri. | Slovo může být chlebem i nožem. |
| Gule lavendar na čaľoha. | Sladká slova tě nenasytí. |
| Ma dikh pro lava, dikh pro vasta. | Nevšímej si slov, ale rukou. |
| Te našťi des maro, de choča lačho 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. |
| Lačho lav sar maro. | Dobré slovo je jako chleba. |

**Obrázek č. 1. Ukázka výskytu slova „lav“ v díle „Goďaver 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.

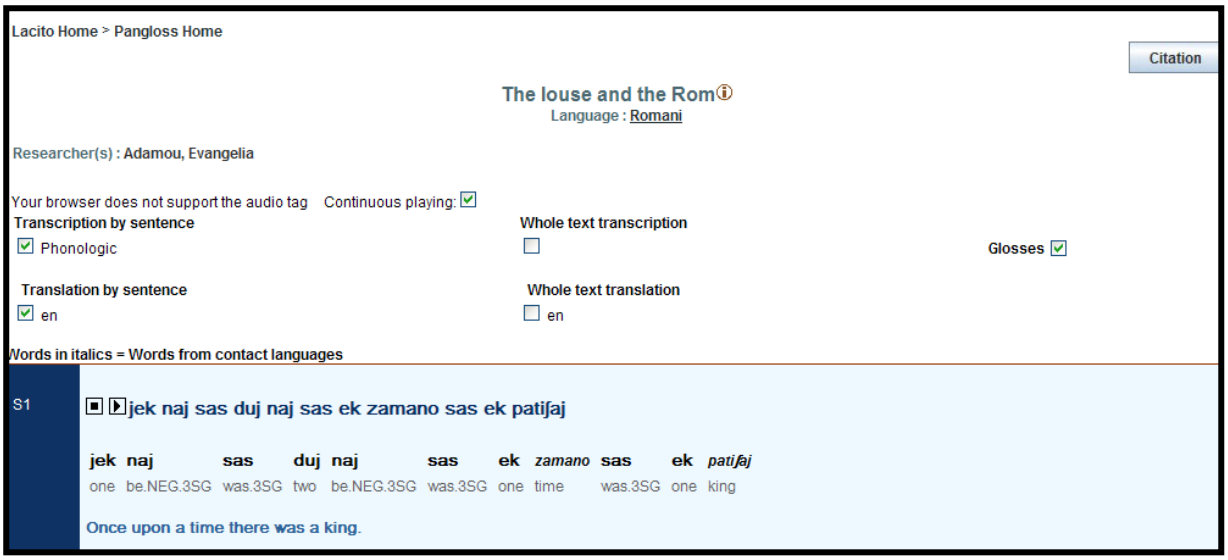
Korpus může sloužit např. při praktickém používání jazyka (mimo jiné při překládání), lze s jeho pomocí sledovat a předpovídat vývoj romštiny, vytvářet jeho pomocí (pravděpodobně reprezentativnější, než klasickým způsobem) další jazykové pomůcky a nástroje, jako mimo jiné slovníky (včetně frekvenčního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

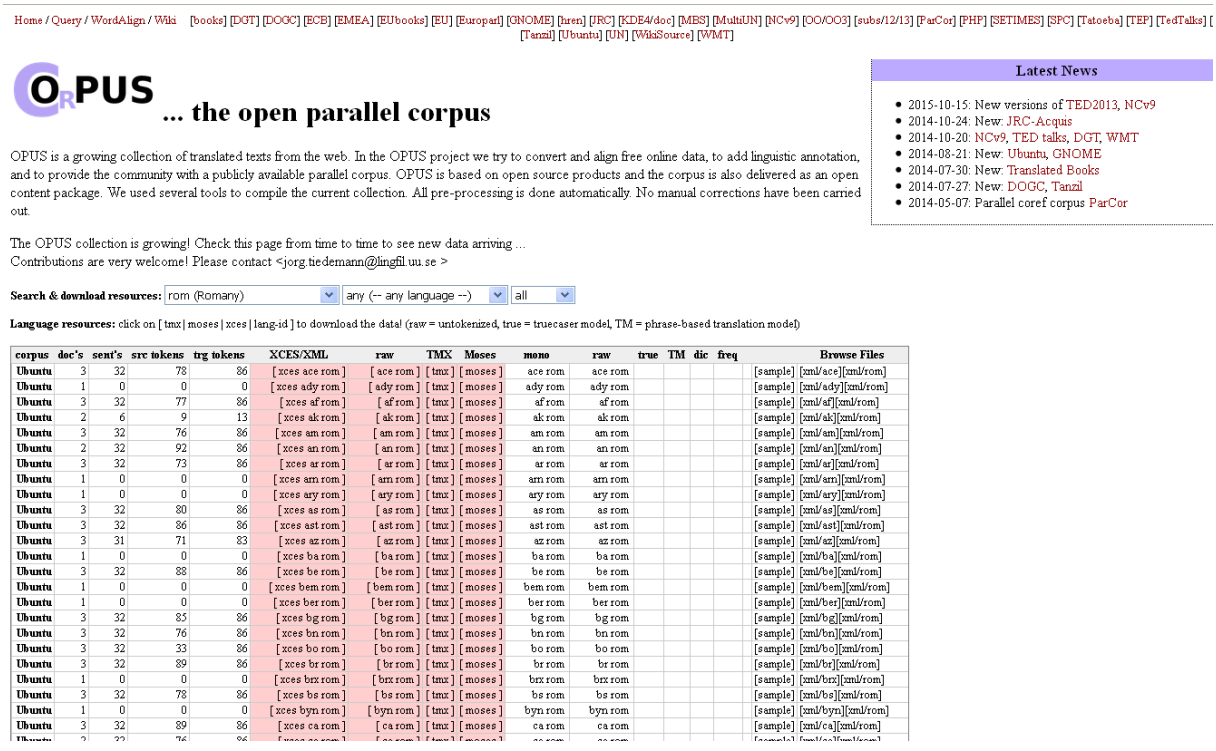
Na webové stránce http://lacito.vjf.cnrs.fr/pangloss/languages/Romani_en.htm je údaj o romském korpusu věnovaného romštině v Řecku (podle uvedené stránky se jedná o dvě varianty: olašskou romštinu a romštinu „balkánskou“. Korpus, který se nezaměřuje jen na romštinu, obsahuje 3 vyprávění zaznamenaná badatelkou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




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

## 5 Korpus Opus

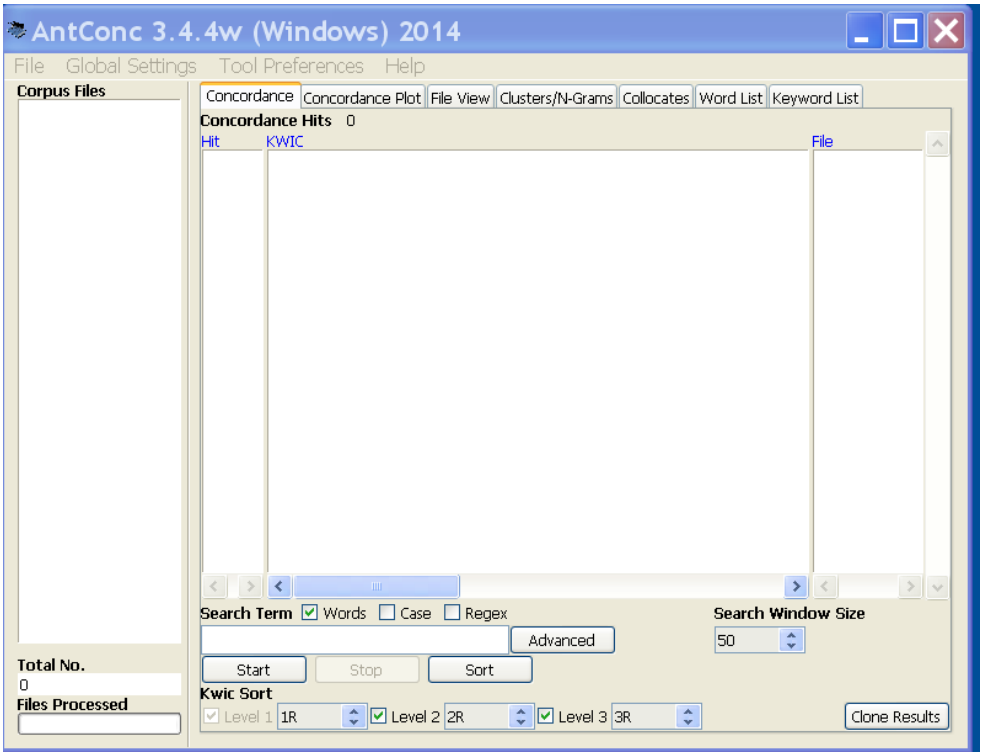
Bezplatný korpus Opus na stránkách http://opus.lingfil.uu.se/ obsahuje velké množství romských textů. Jak je možno tento korpus využívat, a o jaké romské texty jde, musím teprve zjistit.




*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 Univesity 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*