Towards the creation of a Latvian Romani corpus: on Jānis Leimanis' folklore collection and its context

Workshop "Romani in North-East Europe", March 5, 2021

Natalia Perkova (Uppsala University)

Latvian Romani

- also known as Lotfitko (lotfitka) or Čuxny
- one of the Northeastern Romani dialects (Matras 2002; Tenser 2008)
- spoken in Latvia, Estonia, northern Lithuania
- hard to estimate the number of speakers, ~2-3 dialects in Latvia, about 7800 Roms according to official data, but these numbers can be imprecise
- still an understudied dialect, very few published texts + a short grammar (in Latvian; Mānušs 1997) included in the dictionary (Mānušs, Neilands & Rudevičs 1997); 4628 lexical units in RomLex (http://romani.uni-graz.at/romlex/)
- see also the recent report (Ruka 2019)

Romu skaits Latvijā /2015. gads

* Avots: PMLP, 2015

Existing texts (or phrases) in Latvian Romanni

- Jānis Leimanis' Latvian Romani archive (digitalised by Latvian Folklore Archive) from the 1930s
- Paul Ariste's manuscripts (the 1930s-1941) and several publications: https://fennougrica.kansalliskirjasto.fi/handle/10024/87064 and https://fennougrica.kansalliskirjasto.fi/handle/10024/87067 (the Zingarica collection of the National Library of Finland), (Ariste 1938, 1964, 1973)
- Romani Morphosyntactic Database (RMS)
- modern translations of Romani fairytales from the website pasakas.net after (Brice 1992)
- also some illustrations in (Mānušs et al. 1997)
- Gospel translations (not easily available)

NB: still almost no texts in modern Romani

Paul Ariste

- Latvian Romani speakers in Estonia, mainly in Tartu area
- the shorter manuscript (61 page) with translations into Estonian and a short overview, the longer manuscript (123 pages) with Romani texts only (NB: a couple of Romani texts recorded in Finland)
- phonetic transcription, not clearly corresponding to the Latvian-based orthography used in (Mānušs et al. 1997); NB: accentuated!
- the texts are mostly still unpublished
- the larger manuscript has at least 5 of the 6 texts published in (Ariste 1938) and (Ariste 1964)

3. Sasiko rom.

Isis tšhōrôro sasiko rom. romni gija pono khēra, rom gija po matšhe. romni jawdža khêre, dikhela: rom nāne khêre. jandža masa to mā-re, tšhudža to kerjol. porazos rom zapišija pe pāšil reka. porazos uštš opre, dikhela: latšha pe moros. porazos dikhela: jawdž awri jek kâlo mānuš, si sastrono. porazos do sasiko rom kanel to nâšel. porazos joi phenela: na naš mandar! me som jakepa ovē to si tu. — rom phenela: so tu kames mandar? — joi phenla jakes: muk to rakirāw tusa. — rakir! — joi phenla jakes: me som tale phū ande bukhuta sâre fôrosa. ki tu javesa gudžvâro, dosta dolesa mo fōros tuke. talo moros isi mo fōros. — porazos joi phenla pe leste: ki tu naphen te romnāke, dosta me tusa rakirava. — jow phendža:

Romani Morphosyntax Database

- at least 5 Latvian Romani samples publicly available: 3 with Estonian codes and 2 with Latvian codes; come from Pärnu (2 samples), Paide (1 sample)m Riga (2 samples)
- also very much phonetically based, but also seems that vowel length is somewhat ignored

RMS, LV-006

404 - tašarlatyr pjasam kafija, beljveljenca pjasam čai

404 - In the morning we drink coffee and in the evening we drink tea.

405 - me somas terno, te sajek džaas po targos

405 - When I was young I used to go to the market very often.

406 - sy kerdjom piro udar, te raz bryšynt sacyndža-pe te džal

406 - Just as he opened the door, it started to rain.

407 - me štar dyis na gijam auri kxerestyr, (…?) bryšynt gija

407 - For four days I didn't go out because it was raining.

408 - javen auri, kame bryšynt na lyja te del

408 - Let's go out before it starts to rain again.

409 - peršu te jes daj, me dur džyivinau

409 - Before I came to live here I lived far away from here.

Jānis Leimanis (1886-1950)

https://www.lsm.lv/raksts/dzive--stils/vesture/romu-biedriba-koris-folkloras-vaksana-evangelizacija-un-nodarbinatiba--jana-leimana-aktivisms.a384449/ (Ieva Tihovska, 26/12/2020)

- an outstanding Latvian Romani activist
- founded the Romani society “Čigānu draugs” (A Friend of Roma), engaged in the Romani choir activities
- a Christian missionary, also translated St. John’s Gospel into Latvian Romani (1933)
- collected Romani folklore for the Latvian Folklore Archive in the 1930s

Jānis Leimanis' collection in Latvian Folklore Archive

An impressive collection of 75 copybooks (3 copybooks are not currently in the Archive's possession), 500 units (463 units accessible); 1254 files

http://garamantas.lv/lv/collection/886320/Jana-Leimana-ciganu-folkloras-vakums

- 884 files with some text in Romani

  NB: often ascribed to other persons (~metadata), but most likely just written by Leimanis himself based on his memory about those texts; he was a native speaker, though...

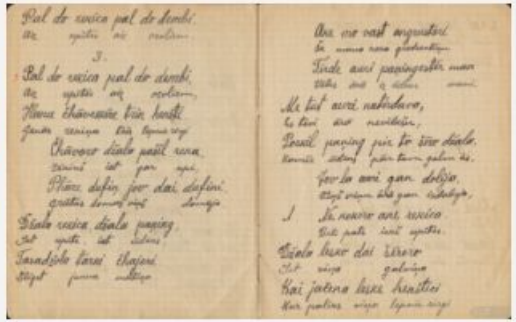
Later used for the book by his son Juris Leimanis, but not published in full length, and the book itself is in Latvian:

Leimanis, Juris. 2005 [1939]. *Čigāni Latvijas mežos, mājās un tirgos*. Rīga: Zinātne. ("Gypsies in Latvia's forests, homes and markets")

Romani folklore collection of Jānis Leimanis

Collection | Items(463) | Files(1077) | Persons(33) | Locations(3)

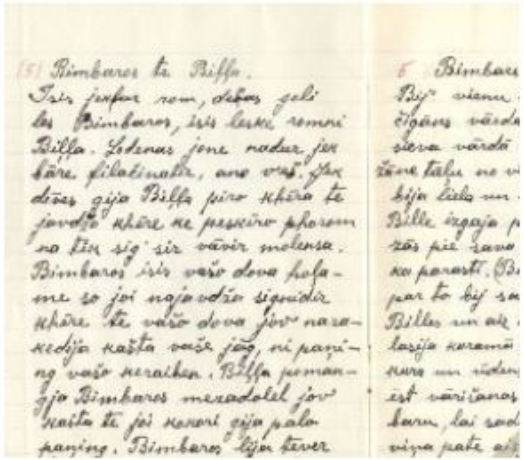
Repository	Archives of Latvian Folklore
No.	1389
Organisation	
Submitted by	Jānis Leimanis
Title	Romani folklore collection of Jānis Leimanis
Description	The most outstanding collection of Latvian Romani material - 500 folklore units in 75 volumes - was collected by Jānis (Bernis) Leimanis (1886-1950) in 1920-30's. Leimanis' manuscript is particular for two reasons - it widely represents less known Romani folklore and all of the materials are supplied with Latvian translation.

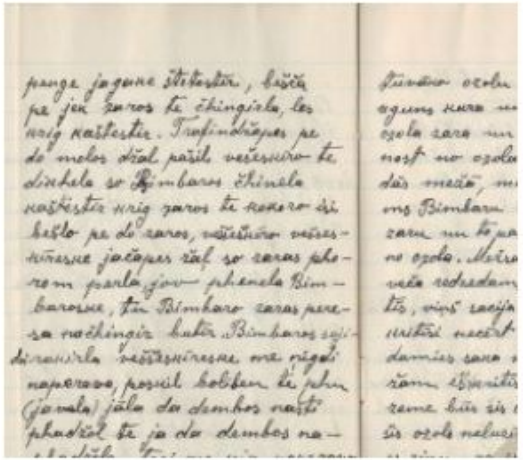


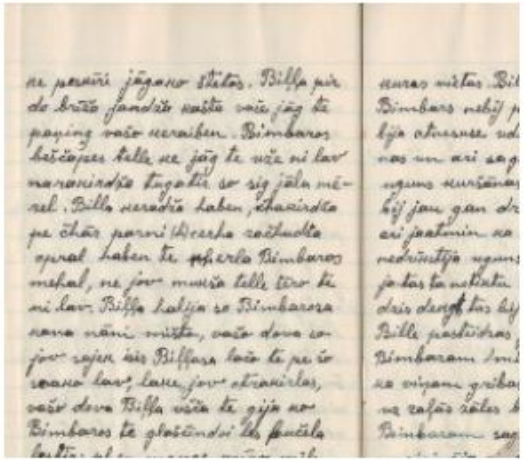
The manuscript starts with a long fairy-tale *Bimbars and Bille*, as told by Ansis Zavickis in 1902 in the Kuldīga district. Leimanis was obviously well educated because the language of the translation is brilliant and rich. The material ends in "mortal" words which are now fading out of the Romani language.

Archives of Latvian Folklore holds also Romani folklore, collected from Latvian children.

Romani folklore collection of Jānis Leimanis



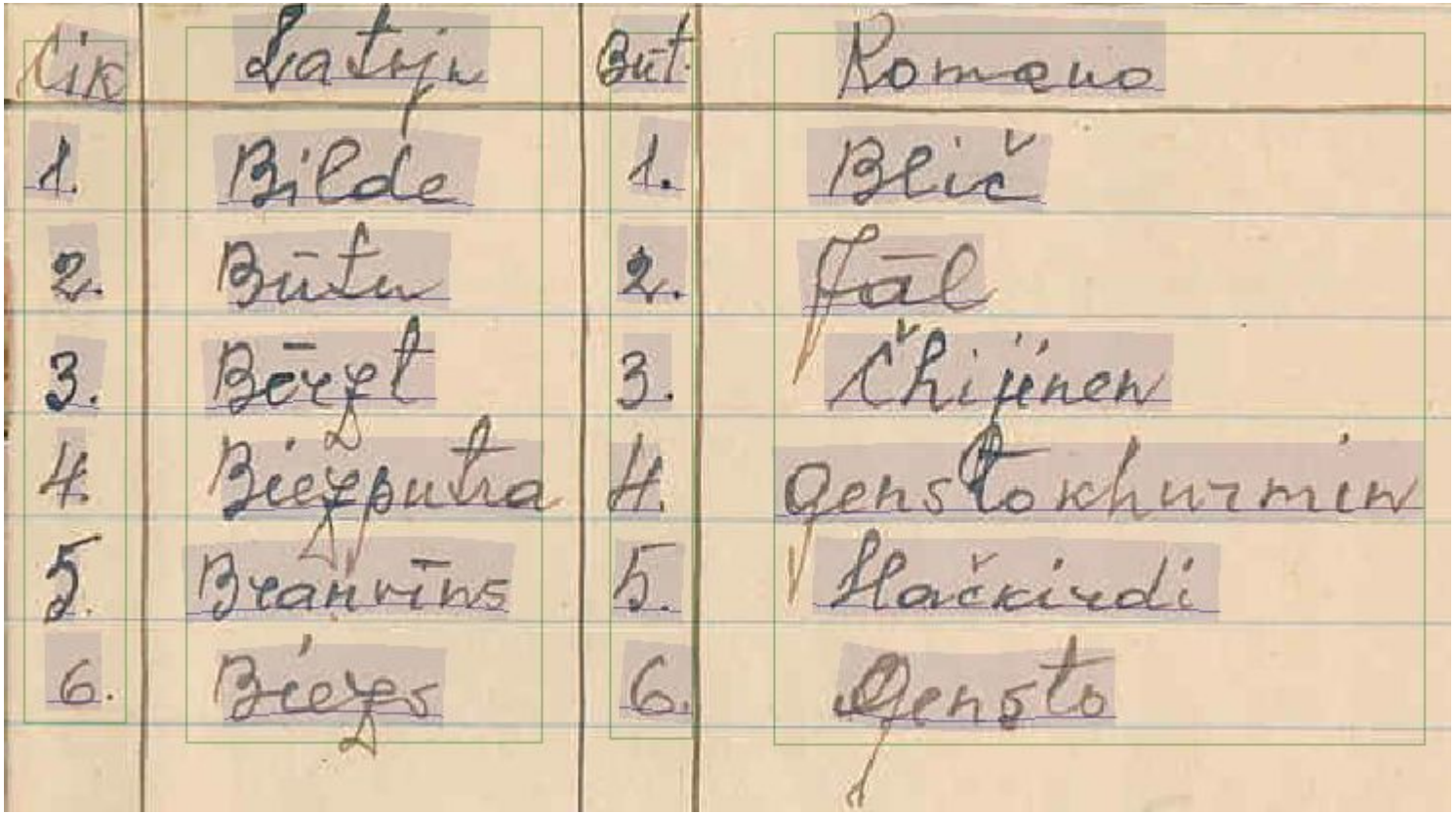




Collection | Items (463) | Files (1254) | Persons (33) | Places (172) | Editorial files

Jānis Leimanis’ collection: composition

- about 65 longer texts (fairytales and stories)
- songs
- short forms (some are parts of longer texts, but are classified as separate folklore units)
- description of some traditional practices, e.g., pirts (~sauna)
- proverbs, parables
- lists of obsolete and disappearing words



Lik	Latvijā	But.	Romeno
1.	Bilde	1.	BliČ
2.	Būtne	2.	fal
3.	Bērzs	3.	Čhijinen
4.	Bieżputra	4.	Qenstokhurmin
5.	Brāmiņš	5.	Ļačkirdi
6.	Bērns	6.	Qensto

Deciphering manuscripts: a crowdsourcing initiative

- the Leimanis collection, as well as other digitalised collections, is available at the crowdsourcing platform garamantas.lv launched in 2014 (see Reinsone 2020)
- partly deciphered (mostly thanks to Ieva Tihovska, myself and some more people), but still not sufficiently!
- October 2020: 312 of all files deciphered (~25%), 185 of files with a text in Romani (~21 %)
- now: somewhat more already available in the deciphered part + a significant shift towards more efficient deciphering!

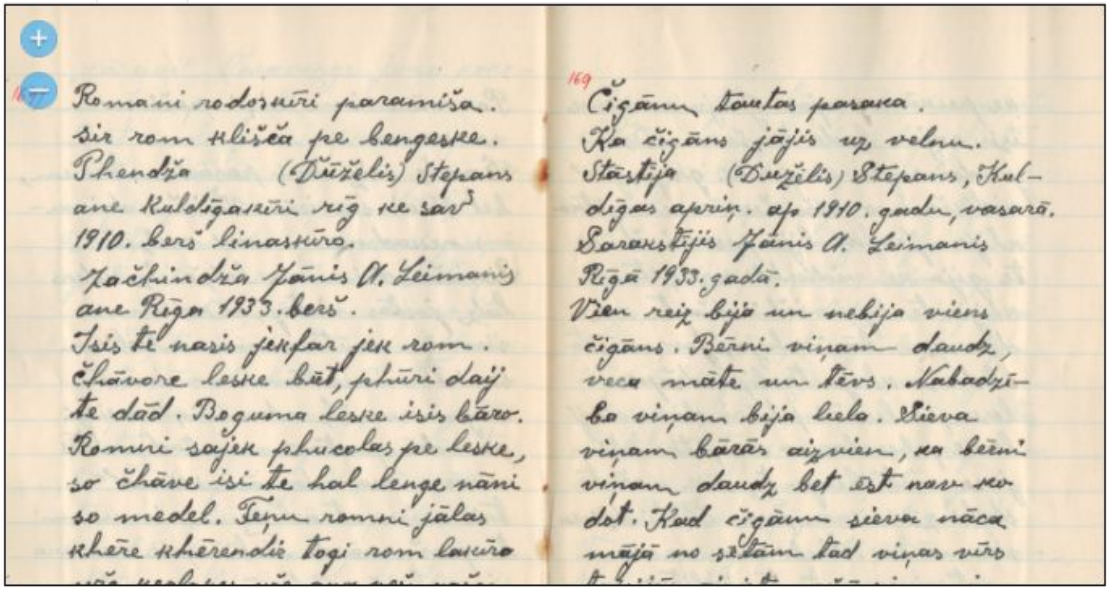
#LFK-1389-169

Previous unit | Next unit

Transcribe

4 / 8

Item | Files | Places



Romani rodoznūri paramiša.
Sir rom hliščā pre bengeske.
Phendža (Dūželis) Stepans
ane Kuldīgasnūri rig ke sar
1910. berš šinastūro.
Kačhindža Jānis A. Leimanis
ane Rīga 1933. berš.
Isis te naris jekhar jek rom.
Čhāvore leske bāt, ghtūri daj
te dād. Boguna leske isis bāro.
Romni sajek phucolas pre leske,
so čhāve isi te hal lenge nāni
so medel. Tenu romni jālas
khere khērendik togi rom lascūro

169
Čigānu šaušas parauca.
Ka čigāns jājis uz velnu.
Stāstija (Dūželis) Stepans, Kuldīgas apriņ. aja 1910. gadu, vasarā.
Sarakstījis Jānis A. Leimanis
Rīgā 1933. gadā.
Vien reiz bijā un nebijā viens
čigāns. Bērni viņam daudz,
veca māte un tēvs. Nabadzība viņam bija liela. Dieva
viņam bārās aizvien, ka bērni
viņam daudz bet ēst. nav ko
dot. Kad čigānu sieva nāca
mājā no sētām tad viņas vīrs

Transcription (text)

Transcriber	Natalia Perkova
Completed	Jā
Accepted	Nē

Transcription (text)

Transcriber	Natalia Perkova
Completed	Jā
Accepted	Nē
Updated	2021-03-03 20:06:12
Transcription (text)	**169.** Romani rodoskīri paramiša. Sir rom klišča pe bengeske. Phendža (Dūželis) Stepans ane Kuldīgakīri rig ke sav 1910. berš linaskīro. Začindža Jānis A. Leimanis ane Rīga 1933. berš. Isis te nasis jekfar jek rom. Čhāvore leske būt, phūri daij te dād. Boguma leske isis bāro. Romni sajek phucolas pe leske, so čhāve isi te hal lenge nāni so medel. Teņu romni jālas khēre khērendir togi rom lakīro uže kerlapes uže ano vešs vašo dova so romni togi butir phučolas, rom an do brīza dža * Čigānu tautas pasaka. Ka čigāns jājis uz velnu. Stāstīja (Dužēlis) Stepans, Kuldīgas apriņ. ap 1910. gadu, vasarā. Sarakstījis Janis A. Leimanis Rīgā 1933. gadā. Vien reiz bija un nebija viens čigāns. Bērni viņam daudz, veca māte un tēvs. Nabadzība viņam bija liela. Sieva viņam bārās aizvien, ka bērni viņam daudz bet ēst nav ko dot. Kad čigānu sieva nāca mājā no sētām tad viņas vīrs taisijās [taisījās] aiziet mežā pie saviem zirgiem tamdēļ ka sieva alaž [allaž] bargi rajās [rājās]
	Transcribe text

Deciphering + HTR in Transkribus

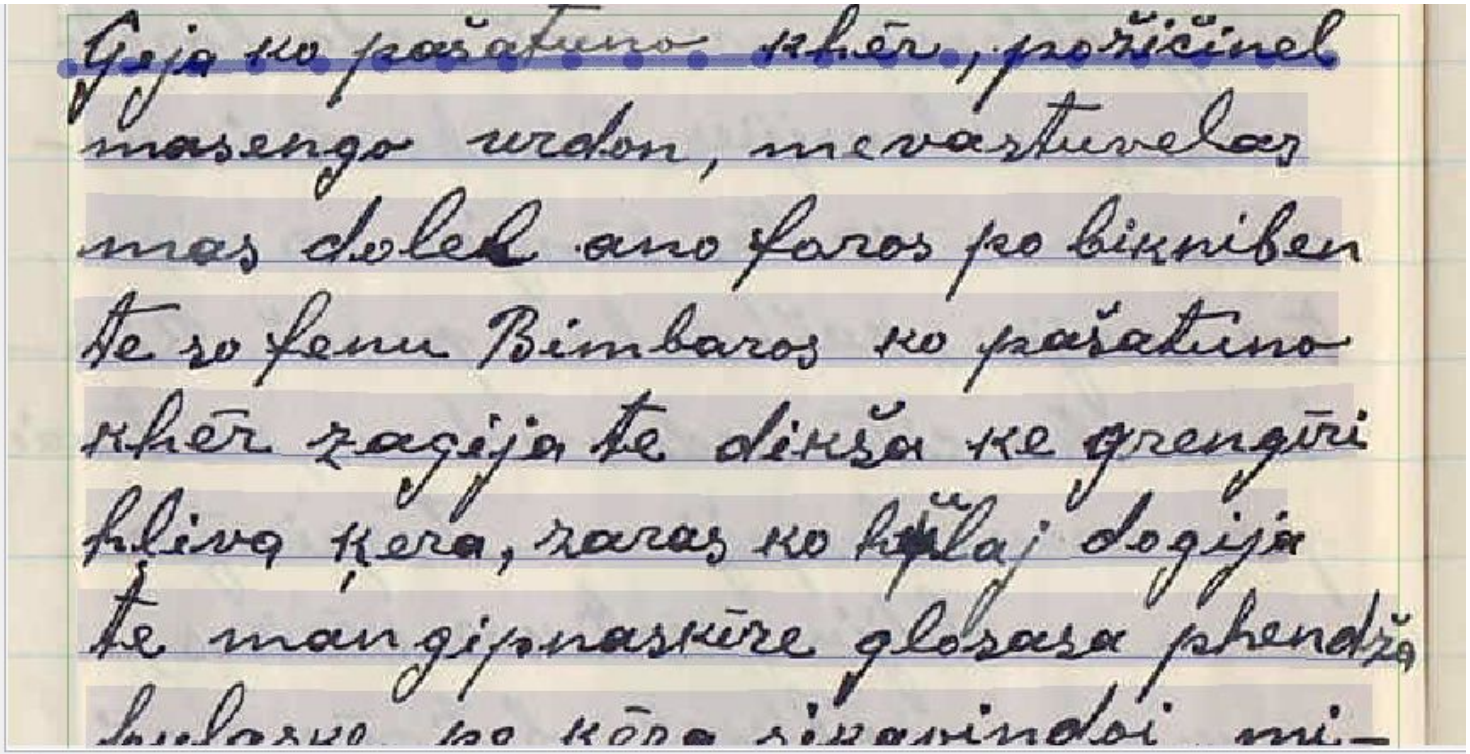
Transkribus (https://transkribus.eu/Transkribus/): software for handwritten text recognition (HTR) - train your own models or use the available ones; a highly valuable solution for manuscripts (Kahle et al. 2017)

The Leimanis collection has a strong advantage of being written by the same person, which means using the same handwriting (holds true for the majority of the texts in the collection).

- a very good case for training an HTR model

HTR in Transkribus: a workflow

- upload your files into the folder onto the server
- launch a layout analysis (already available by default)
- add your transcriptions line by line
- get a certain number of training data
- now you can train your HTR model!
- test it on some additional data (not yet deciphered pages), correct the automatic transcription manually > more training data for the next model!



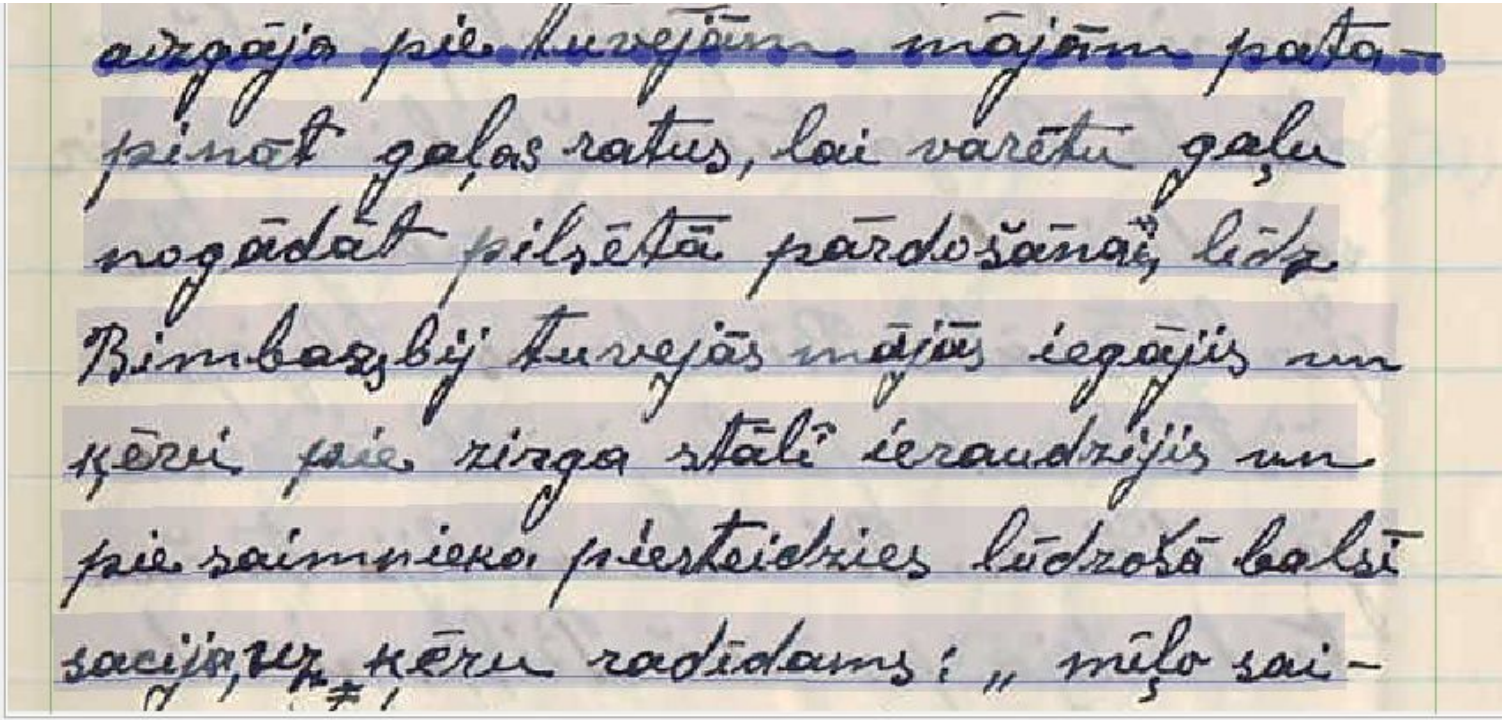
1-1 Gija· ko· pašatuno· khēr,· požičinel↵

1-2 ·masengo· urdon,· mevastuvelas↵

1-3 mas· dolel· ano· faros· po· bikniben↵

1-4 te· so· fenu· Bimbaros· ko· pašatuno↵

Unit 5, p. 21



2-4 aizgāja· pie· tuvējām· mājām· pata¬

2-5 pināt· gaļas· ratus,· lai· varētu· gaļu

2-6 nogādāt· pilsētā· pārdošānai,· līdz

2-7 Bimbars· bij· tuvejās· mājās· iegājis· un

Unit 5, p. 21

Search for...

Documents | **Fulltext (Solr)** | Tags | KWS

Search transcribed text for words or phrases

Search for: Bimbars

☑ Show word preview ☐ Current document ○ Word-based text ◉ Line-based text

☐ Case sensitive ☐ Fuzzy search

Search! | Previous page | Next page

Search results

Showing Pagehits 0 to 10 of 33

All collections | All documents | All authors | All uploaders

Context	Document	Page
rakļa, joi noko lipindža po skamin ne būt ko ēst, bet tik līdz Bimb	JanisLeimanis	15
izdzirdis ka viens stingriem soliem tuvojās, vecais Bimbars paspē	JanisLeimanis	15
atradās tuvu pie lapenes. Bimbars redzēja, ka lapenē iegāja liel ku	JanisLeimanis	15
truju mas, me han sāre džuka¬ cās prātā, gaļu varētu pārdot visu uz reizi, tas tik būtu labi. Bimbars	JanisLeimanis	25
no Bim¬ bara nūjas un no viņa saucie¬ niem, ierejās, tad Bimbars atkal jautāja: "vai rītu būs nauda	JanisLeimanis	25
?" tad suns atkal ierējās. Tad Bim¬ Bimbars gaļu no segas nogāzis, aiz¬ laidās mājā. No rīta it agri	JanisLeimanis	25
Bimbars bij klāt pēc naudas, nu vairs suņus neatradis Bim-	JanisLeimanis	25
so Biļļa jandža bakres, zaras Bimbars nodomajis ka taisnība ir ko viņam mežsārgs paregojis. Bimbars	JanisLeimanis	13
vecais Bimbars bija ieraudzijis ka atnesa priekš viņa zārku, viņš iesāka kliegt dams lūgties Billes	JanisLeimanis	13
dufindža so čačo isi (lizdami) viņai nāsīs. Bimbars izdzirdis ka ķēve izsprauslajusi viņš tulīn izstiepas	JanisLeimanis	12
otro reizi Bimbars to redzedams ka sieva aizbēgusi mežā un drīz nāks nāve, viņam palika ta ne(ob) labi	JanisLeimanis	12
ap sirds, viņš sāka saukt pēc Billes, viņa tulīn atgriezās pie Bimbara. Bimbars saņemies ma ka ari	JanisLeimanis	12
sacija: "ej un ņem, ja ir vajadzīgs, bet pēc pāru dienām gādā man viņas šurpu". Bimbars stingri	JanisLeimanis	22
apsolījās tīrākus atvest, kā tagad viņi izskatās. Bimbars paņēmis tačku un aizvilcis uz mežu. Bille par to	JanisLeimanis	22
laiku ar vērsi bija pārejos darbus izda¬ rijusi un sagatāvojusi priekš pārdošanas. Bimbars tačku atvēdis	JanisLeimanis	22
neapgrūtinātu vel jo vairak veco Bimbaru, jo viņa nezināja kas ar viņu noticis. Tad Bimbars viņai sacija: "eij	TRAINING_VALIDATION_SET_Latvi...	1
uzved mūsu lielo balto ķēvi, un piesien viņas pie mūšu ratiem un labi pabaro" Bille) Bimbars viņai vel	TRAINING_VALIDATION_SET_Latvi...	1

Training a Latvian Romani HTR model in Transkribus

It is claimed that training a good Transkribus model requires about 15000 words, or 75 pages, of ground truth material.

- the Leimanis collection: an average double page spread = about 160 words
- at least 94 GT files needed for training

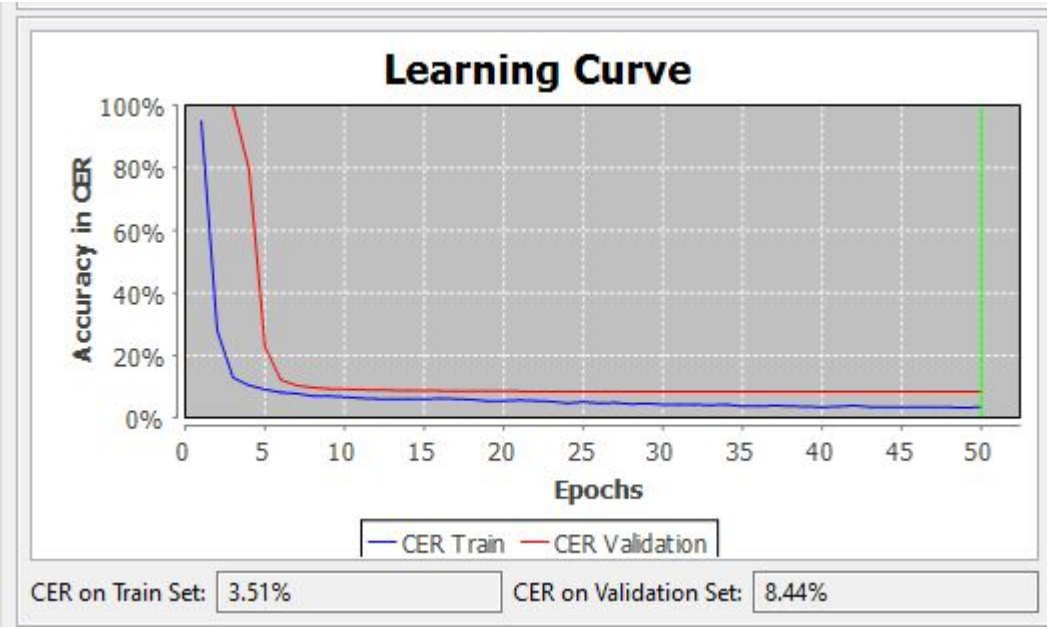
January 2021: about 200 file transcriptions put in my Transkribus project from the Garamantas deciphering

- two models trained

Training a Latvian Romani HTR model in Transkribus

Model 1: 235 pages, a training set of 25890 words and 4911 lines, 50 epochs

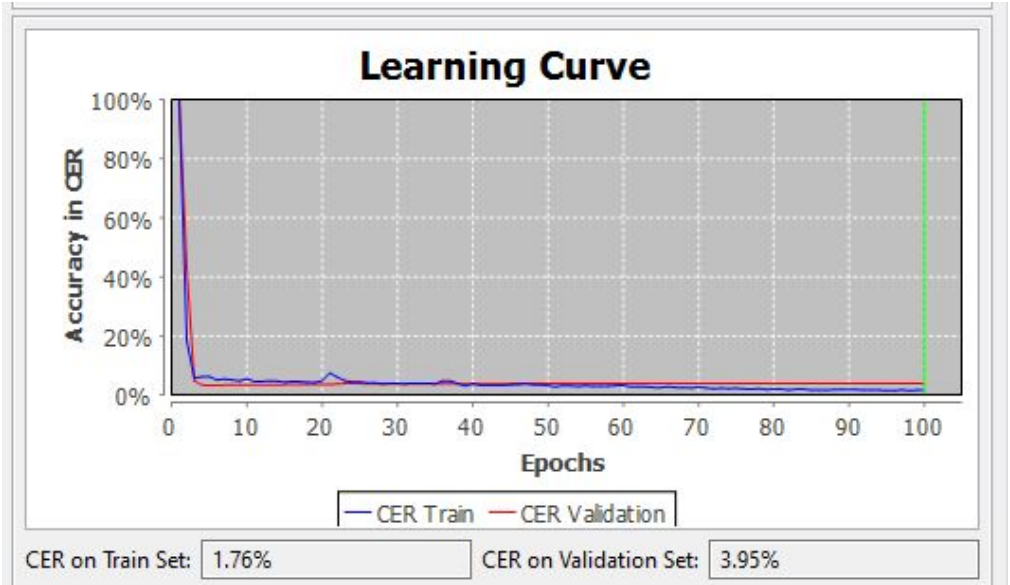
- CER 8.44% (validation)



Training a Latvian Romani HTR model in Transkribus

Model 2: 241 page, a training set of 16952 words and 3136 lines; uses Model 1 as its base model, 100 epochs

- CER 3.95% (validation)



1-12 # domārel les, ne svako molos kālo
1-13 # ~~šap~~ sap jake sig ponāšelas so feņu kuti
1-14 # golelas dikhel les. Jek šukār dīves
1-15 # tašarlatir ~~podžangadža~~ podžangadža - medžav
1-16 # lesa. Me sigo (uštom) uščom opre
1-17 # te ~~gizom~~ gijom dādesa domaras sapes.
1-18 # Dogijam žin ke bār, ne kedi jav¬
1-19 # džam ke purani bār kai pašolas
1-20 # kālo sap, an do kokori brīža kālo
2-1 # Vienu rītu kad saule bija uzkāpusi
2-2 # labu gabalu un zemi jau bija
2-3 # sasilusi no karstiem saules stariem,
2-4 # tad mans tēvs gāja zirgus apskatī¬
2-5 # ties, bet piegājis netāļu no veca ~~pīta~~ pītā
2-6 # žoga tēvs ielaudzija melnu, melnu
2-7 # čūsku ka darvu. Tēvs gāja steidzīgi
2-8 # bet uzmanīgi turēdam rokā koku
2-9 # lai ~~čusku~~ čūsku nosistu, lai ~~čusku dabo)~~ čūsku dabot
2-10 # ~~kosist~~ nosist viņš gāja vairak rītus kad čūska
2-11 # gulēja viena un tai paša vieta, bet ~~čilska~~ čūska
2-12 # tik atri varēja aizbēgt ka viņas maz
2-13 # varēja dabot redzēt. Vienā jauka rītā
2-14 # tēvs mani pamodināja, lai eimu ar
2-15 # viņu līdz. Es žilgli piecēlos un aiz¬
2-16 # gāju līdz sist čūsku. Nogājam līdz
2-17 # setai, kura bija jau sabrukusi un tai
2-18 # vietā kur ~~cuska~~ čūska guļ. Es gāju uzma¬
2-19 # nīgi, bet tiko biju tuvaku čūskai
2-20 # taj paša brīdī melnā cuska ka ~~samt.~~ samts

page 218, Unit 129
Teika par melno čūsku
(A fairytale about a black snake)

- a better model

Differences in format

Transkribus: max reproduction of original writing, all erroneous forms as they are

Garamantas: all strikethrough forms and non-essential parts in brackets unnecessary, also corrected forms are added for the Latvian translation

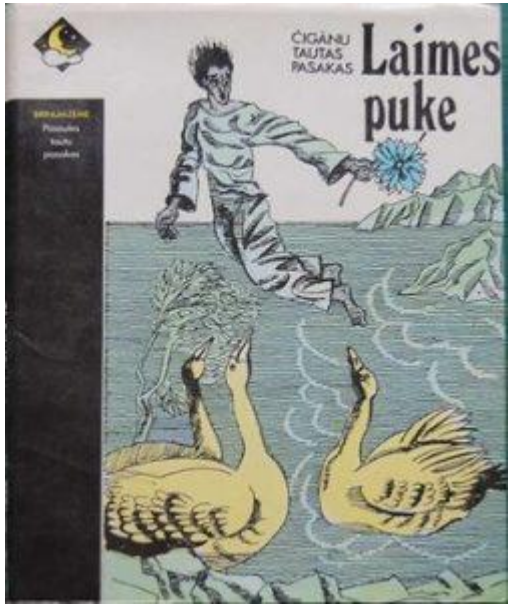
NB: various miswritten forms in the Romani part (diacritics, letters omitted, etc.)

The final corpus: ???

From the 1930s to the 2000s...

1992, “Laimes puķe” (The Happiness Flower) by Silvija Brice: translations of Romani fairytales from several languages (German, Russian, Latvian)

- Latvian texts (4 fairytales, proverbs, beliefs) come from Leimanis collection, just Latvian translations somewhat edited
- Zigeunermärchen. Eugen Diederichs Verlag. Jena 1926; Unter einem Regenbogen bin ich heut gegangen. Kassel: Erich Röth Verlag, 1976; Сказки и песни, рожденные в дороге. Москва: ГРВЛ, 1985



From the 1930s to the 2000s...

2010, “Ideju Forums”, the project ”Mazākumtautību audio pasakas internetā" - audio recordings of fairytales in various minority languages (Russian, Polish, Belarusian, Ukranian, Lithuanian, Estonian, Jewish, and Romani)

- still available at pasakas.net with translations (Anatolijs Berezovskis) http://www.pasakas.net/pasakas/citu_tautu_pasakas/ciganu-pasakas/

- 17 fairytales

From the 1930s to the 2000s...

Two fairytales at *pasakas.net* come from Leimanis' collection!

- **Kā čigāns jājis uz velna = Sir rom trādija pu bengesku dumu** ('How a Rom rode the Devil') < #LFK-1389-169, **Sir rom klišča pe bengeske**
- **Teika par čigānu baznīcu = Teika romengi khangirjatir** ('A story about a Romani church') < #LFK-1389-412 and #LFK-1389-414, **Kā čigāns gāja zivis zvejot** ('How a Rom went to do some fishing')

So, we can basically compare original Leimanis' texts with translations into modern Romani!

0	Isis te nasis jekfar jek rom.	0,204	Vien reiz bija un nebija viens čigāns.	0,204	Vienreiz bija un nebija viens čigāns.	1,97829	Sis peske te nasis peske jek rom.
1	Čhāvore leske būt, phūri daij te dād.	0,3	Bērni viņam daudz, veca māte un tēvs.	0,3	Bērnu viņam daudz, veca māte un tēvs.	1,41532	Leske sis būt čāve, phūri daj te dād.
2	Boguma leske isis bāro.	0,242857	Nabadzība viņam bija liela.	0,242857	Nabadzība viņam bija liela.	0,257143	Dživdža jov an bāri boguma.
3	Romni sajek phucolas pe leske, so čhāve isi te hal lenge nāni so medel.	0,283929	Sieva viņam bārās aizvien, ka bērni viņam daudz bet ēst nav ko dot.	0,283929	Sieva aizvien bārās uz vīru, ka bērnu gan daudz, bet ēst nav ko dot.	0,983051	Leski romni sajek pu leste khošelaspes, važdova, ki čāvore būt, a te den te hal čāvorenge nāni so.
4	Teņu romni jālas khēre khērendir togi rom lakīro uže kerlapes uže ano vešs vašo dova so romni togi butir phučolas, rom an do brīza dža ke peskīre graja.	0,297619	Kad čigānu sieva nāca mājā no sētām tad viņas vīrs taisijās [taisījās] aiziet mežā pie saviem zirgiem tamdēļ ka sieva alaž [allaž] bargi rajās [rājās]	0,202105	Kad čigāna sieva nāca mājā no sētām, tad vīrs taisījās prom mežā, pie saviem zirgiem, jo sieva allaž bargi rājās.	0,872161	Kedi javelas romeski romni khēre gavestir špīribnastir, laku rom džalas an veš, ku peski graja, važdova ki leski romni pu leste svaku molos delas goli.
5	Graja romeske isis sajek bede te šuke.	0,290909	Čigānam zirgi alaž bija slikti un vāji.	0,27	Zirgi čigānam bija slikti un vāji.	2,13362	Graja da romeske sis bede te šuke.

1930s **Brice 1992** **~2010**

Further steps towards the corpus

The corpus: a vision

- all texts have Latvian translations > a parallel corpus as a natural option
- all metadata preserved (including links to the original archive units)
- morphological annotation for both parts

The corpus: a vision

- a parallel corpus: the TsaKorpus platform (REF), used for several parallel corpora within the Russian National Corpus (including the Latvian-Russian corpus)
- all metadata preserved (including links to the original archive units) - also a standard practice for the RNC subcorpora
- morphological annotation for both parts:

  Latvian: a standard Latvian tagger from RNC, works neatly for aligned files in the standard RNC .xml format

  Romani: UniParser (Arkhangelsky et al. 2012), already applied to Russian Roma texts (http://web-corpora.net/RomaniCorpus/search/)

MAIN

Found: **683** matches, **96** documents

Джя лэ о **ловэ** .

6. **Германо А. (1935). Ганка Чямба и ваврэ роспхэныбэна. М.: Художественная литература. 102 с.** Германо А. 1935 Expand

Поракирдя ваврэ ранца, савэ пхэндлэ, со крепостно чяй Ганка ачела барэ багибныткonаса и лэла тэ плэскирэл ранякэ баро оброко, а коли Ганка закамэл тэ откинэл лэс, то банги явэла тэ отдэл пал пэстэ бут **ловэ** .

7. **Нэво дром. 1932-7** без автора 1932 Expand

Адалэ **ловэ** джяна ваш адава, собы тэ кинэс бутяритко и продовольственно ското, гавитко-хулаибныткo инвентарё и адякэ дурэдыр.

8. **Геньдеш И., Балог И. Венгерска рома. М., 1931.** Геньдеш И., Балог И. 1931 Expand

Тэло лав «свадьба » треби тэ полэс революцыя,«калаче » — значит право, **ловэ** , пхув, отлымэ австрийсконэ тагаритконэ властяса Венгриятыр.

9. **Мохначёво И. Пэрва рэнды дрэ колхозно буты (Гавитка-хулаибныткa артели). М., 1932.** Мохначёво И. 1932 Expand

Тимин бутякэ **ловэнца** дрэ колхозо запхэнэлапэ.

10. **Кинаса трактoро ваш романо колхозо // Нэво дром 1930-3** 1930 Expand

Page: First **1** 2 3 4 5 6 7 8 9 10 ... Last

Print version | Save to file

Wordform Lexeme Translation

деньги 1

Gram & Lexical Attributes

Advanced

Distance to the next token:

From 1 to 1 words

Wordform Lexeme Translation

2

Gram & Lexical Attributes

Advanced

Advanced Distance

Search Clear

Specify Subcorpus

Display Options

Search in New Window

Error Report

The corpus or more?

- all texts have Latvian translations > a parallel corpus as a natural option
- all metadata preserved (including links to the original archive units)
- morphological annotation for both parts

But what about the folklore dimension? What about adding more translations (at least in English, probably also in Russian)

- annotation for genres
- additional annotation for some relevant folklore categories
- still much to think about

(INSERT CORPUS TITLE IN messages.po) EN | RU |

LV-RU

Back to search

Search result: 1497 occurrences, 1445 sentence(s) found in approximately 83 document(s).

Мастер и Маргарита М. А. Булгаков 1929–1940

Оставалось предположить, что сонная и странная личность улетела из **дому**, как птица, не оставив по себе никакого следа.

Acīmredzot miegainā un dīvainā persona ir aizspurgusi no mājas kā putniņš —bez pēdām.

Vīrietis labākajos gados Zigmunds Skujiņš 1975

Пробегаю развалины **дома** Черноголовых, и развалин тех давно уже нет.

Es skrienu gar Melngalvju nama drupām, kuru sen jau vairs nav.

В усадьбе А. П. Чехов 1884–1885

Мейер — единственный молодой человек, который бывал в их **доме**, бывал — они это знали — ради их милого женского общества, но неугомонный старик завладел им и не отпускал его от себя ни на шаг.

Meijers bija vienīgais jaunais cilvēks, kas apmeklēja viņu māju, apmeklēja —to viņas zināja —patīkamās sieviešu sabiedrības dēļ, bet nesavaldīgais večuks saņēmis Meijeru savā varā un nelaiž ne soli projām no sevis.

Dievs. Daba. Darbs Anna Brigadere 1926–1933

И тут всплыл застарелый страх: как войдет она в чужой **дом**?

Uznāca gan arī vecā nedrošība: kā nu tik ieies svešās mājās?

Limuzīns Jāņu nakts krāsā Māra Svīre 1979

Летом это будет вообще невыносимо, надо сказать Мартыню, пусть убедит тетю, что от газа не так уж много **домов** взлетает на воздух, как ей кажется.

Vasarā tas nebūs izturams, jāsaka Mārtiņam, lai pārliecina tanti, ka gāze nemaz neuzsper gaisā tik daudz māju, kā vecajai šķiet.

ROMLEX

- University of Graz, a lexical database
- integrates various dictionaries for a number of Romani dialects
- uses a somewhat uniform tag system
- modified transcription, 3sg as a base form for verbs (ā > aa, mīļinav > miiľinel)

Latvian Romani: based on (Mānušs et al. 1997)

- some collocations also included
- no forms included (e.g., even basic wordforms, such as plural for nouns)

ROMLEX > UniParser

- the original ROMLEX .xml format (with Latvian translations only)
- a script for automatic conversion into the UniParser format for lexemes

```
-lexeme:
 lex: akhor
 stem:
 gramm: n m
 paradigm:
 transl_lv: valrieksts
 transl_en:
 transl_ru:
```

```
<entry id="e113784" dia="roml">
  <o>akhor</o>
  <pos>n m</pos>
  <g>
    <s>
      <t>
        <e>valrieksts</e>
      </t>
    </s>
  </g>
<biblio>LMA</biblio></entry>
```

UniParser

- a list of lexemes with all necessary information: stems, paradigms, translations
- a list of paradigms; productive derivations can be included
- also an option of incorrect analyses to be filtered out
- NB: loanwords can be annotated!

So far: preliminary paradigms mostly for nouns and pronouns, adjectives in progress, though only at the level of classification only (all inflectional morphology to be added); preliminary templates for all lexemes, only several are already annotated

WORK IN PROGRESS!

- what labels to use? how to normalise?

Summary

- first HTR models trained, a significant progress in deciphering; still takes some time to put the texts together and especially to put the deciphered texts back to Garamantas;
- several longer texts (fairytales) prepared for alignment; some more are already corrected automatic decipherings;
- work on the parser started;
- prepared texts are put in the folder for further alignment (alignment can be done continuously as a parallel process)
- plans for the future: experiments with word alignment (for automatic paradigm induction and possible dictionary extension)

References

Ariste, Paul. 1938. *Romenge paramiši: Mustlaste muinasjutte*. Tartu.
Ariste, Paul. 1964. Supplementary review concerning the Baltic Gypsies and their dialect. *Journal of the Gipsy Lore Society*, Third Series, 43 (1/2): 35-37.
Ariste, Paul. 1973. Einige Märchen Čuchný-Zigeune. *Tartu Riikliku Ülikooli Toimetised* 309: 5-40. Tartu.
Arkhangeskiy, Timofey, Belyaev, Oleg & Vydrin, Arseniy. 2012. The Creation of Large-Scale Annotated Corpora of Minority Languages Using UniParser and the EANC Platform. Proceedings of the 24th International Conference on Computational Linguistics (December 2012, Mumbai, India): 83-92.
Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. 2017. Transkribus – a Platform for Transcription, Recognition and Retrieval of Document Images. *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 19–24.
Leimanis, Juris. 2005 [1939]. *Čigāni Latvijas mežos, mājās un tirgos*. Rīga: Zinātne.

References

Mānušs, Leksa, Neilands, Janis & Rudevics, Karlis. 1997. *Čigānu-Latviešu-Angļu un Latviešu-Čigānu Vārdnica*. Riga: Zvaigzne ABC.
Matras, Yaron. 2005. The classification of Romani dialects: A geographic-historical perspective. In: Halwachs, D. & Schrammel, Barbara, (eds). *General and applied Romani linguistics*. Munich: Lincom Europa. 7-26.
Reinsone, Sanita. 2020. Searching for deeper meanings in cultural heritage crowdsourcing. In: Hetland, P., Pierroux, P., & Esborg, L. (eds.).. *A History of Participation in Museums and Archives: Traversing Citizen Science and Citizen Humanities* (1st ed.). Routledge.
Ruka, Elvita. 2019. *A short introduction to Latvian Romani culture*. Sava grāmata. https://www.km.gov.lv/uploads/ckeditor/files/Romi_ENG_Final.pdf
Tenser, Anton. 2008. *The Northeastern Group of Romani Dialects*. Ph.D. dissertation, University of Manchester.