Towards the Corpus of Latvian Romani Texts: Deciphering the Manuscripts in Jānis Leimanis' Archive

Natalia Perkova[1], Kirill Kozhanov[2,3]

[1] *Uppsala University, Sweden*
[2] *Helsinki University, Finland*
[3] *Potsdam University, Germany*

Abstract

Latvian Romani is a Northeastern Romani dialect with a limited number of publicly available sources. Two large archival collections of texts in Latvian Romani, compiled primarily in the 1930s in Latvia and Estonia, have been recently digitized as images and made available online for a wider public. In our study, we focus on one of these collections, the Latvian Romani folklore texts collected by Jānis Leimanis in interwar Latvia. In this paper, we describe how initial manual transcriptions, most of which have been created with the help of a special crowdsourcing platform, were integrated in the handwritten text recognition (HTR) workflow in Transkribus. We present two HTR models trained on the basis of Leimanis' collection and discuss various issues related to the work on these texts.

Keywords

Handwritten Text Recognition, Low-resource Languages, Digitization, Latvian Romani, Folklore Texts

1. Introduction

This study focuses on the development of a corpus for Latvian Romani (Lotfitka), one of the understudied Romani dialects. It belongs to the group of Northeastern Romani dialects [1, 2] and is spoken in Latvia, Estonia, and northern Lithuania. There exist only a handful of texts published in Latvian Romani (a translation of the Gospel of John by Jānis Leimanis (1933) and several fairy tales and legends in [3, 4, 5]). The data in the Romani-Latvian-English dictionary [6] and the Romani Morpho-Syntax Database (https://romani.humanities.manchester.ac.uk/rms/, see [7]) contain only words and separate unrelated phrases. Our attempt at creating a corpus of Latvian Romani texts has been encouraged by recent digitization initiatives in several countries (Estonia, Latvia, and Finland) which resulted in providing open access to the two important archives of Latvian Romani texts, both compiled before the World War II: the collection by Jānis Leimanis, a prominent Latvian Romani personality of the interwar Latvia, at the Archive of Latvian folklore in Riga (http://garamantas.lv/en/collection/886320/Romani-folklore-collection-of-Janis-Leimanis), and the collection by Paul Ariste, a brilliant Estonian linguist, archived in the Estonian literary museum and available online at the National library of Finland (https://fennougrica.kansalliskirjasto.fi/handle/10024/87064).

Digitization of manuscripts with field notes in an endangered, understudied and obviously low-resource language variety using automatic handwritten recognition still represents a rather novel direction in applying computational methods to such data. Among numerous public models available in Transkribus, the Evenki-Russian bilingual model trained on Konstantin Rychkov's manuscripts from the 1910s seems to be the only example of similar research, see [8] for more details on this project. In our case, the texts in focus are also bilingual and handwritten, collected by the same person about a

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), March 15-18, 2022, Uppsala, Sweden
EMAIL: nofpernat@gmail.com (A. 1); kozhanov@uni-potsdam.de (A. 2)
ORCID: 0000-0003-0426-898X (A. 1); 0000-0003-3852-6617 (A. 2)

century ago and not previously published in full. Another important recent initiative focusing on digitizing old manuscripts with texts in minor languages is Manuscripta Castreniana (https://www.sgr.fi/manuscripta/), an impressive collection of texts transcribed, translated and commented by the linguist Matthias Alexander Castrén in the middle of the 19th century, see [9] for an overview.

2. Jānis Leimanis and his Latvian Romani folklore collection

Jānis Leimanis (1886-1950) was a prominent Romani figure of interwar Latvia (see [10]). Founder of the first Romani organization in Latvia, translator and activist, he was officially involved in the work of the Archives of Latvian Folklore (Latvijas Folkloras Krātuve, further LFK) in 1933–1934. In 1939, Leimanis' son Juris published a fiction book titled "Gypsies in Latvia's forests, homes and markets", primarily based on the materials collected by his father (it was recently reprinted with some commentaries by Māra Vīksna [11]). This book, however, was published in Latvian and did not truly present any text from the collection at its length.

Leimanis' archive in LFK comprises 75 copybooks (three of them currently unavailable), with about 500 folklore units of different genres (463 of them accessible) and 1254 manuscript pages in total. A diversity of genres is impressive: there are about 65 longer narrative texts (fairy tales and stories), a number of songs, short forms (often incorporated in longer texts, but still classified as separate folklore units), description of traditional practices (e.g., Latvian *pirts* 'sauna'), proverbs and parables, as well as an appendix with a list of obsolete and disappearing words. This collection is mostly bilingual, as all Romani texts have Latvian translations provided by Leimanis himself: overall, 884 pages contain such bilingual texts. This gives us a unique opportunity to compile a parallel corpus for a Romani dialect.

In 2014, LFK launched a massive crowdsourcing campaign on deciphering the handwritten manuscripts which were scanned and uploaded to a special platform (http://garamantas.lv/en/). This initiative became a huge success in terms of the citizens' voluntary engagement into the preservation of cultural heritage (see more in [12]). As part of this campaign, Leimanis' archive was digitized and uploaded to the crowdsourcing platform as a separate collection (http://garamantas.lv/lv/collection/886320/Jana-Leimana-ciganu-folkloras-vakums). Since then, some files have been deciphered by volunteers, but by October 2020 the deciphered pages comprised about 25% of all files and only about 21% of files with Romani text. Although a dozen of languages spoken by the ethnic groups of Latvia are presented in the LFK materials, most volunteers do not master these languages and pay their attention primarily to the texts written in Latvian. This is one of the reasons why the Latvian Romani collection remained mostly non-deciphered. Our original question was whether it would be possible to accelerate this process, bearing in mind that the entire collection was written by the same person, or, in other words, it has the same handwriting.

The orthography used by Leimanis for Romani texts is based on Latvian; no extra symbols or diacritics (for instance, to mark stress) are used. Leimanis' handwriting is very clear, and the text is almost always very accurately adjusted to the lines. No doubt, such handwritten texts make an ideal case for automatic recognition.

3. Transkribus and the Latvian Romani HTR pipeline

One of the well-known and freely available HTR tools is Transkribus (https://transkribus.eu/Transkribus/). It is possible to use it as a desktop application with the access to the server or work directly in the browser in Transkribus Lite. The platform can be used for transcribing texts and training various text recognition models on them. Transcription can be manual or automatic, conducted with the help of the existing models (OCR models for printed texts are also available for users), see more details in [13, 14]. By now, 119 public models for at least a dozen of languages are

available for Transkribus users; a smaller number of 87 models are presented at the website[2]. Currently, there are no models available for Latvian print or handwritten texts, not to mention Latvian Romani.

In January 2021, we intensively worked on transferring the deciphered texts to the Transkribus platform as ground truth data for the Leimanis collection. Preliminary transcription experiments were conducted already in 2020. At the initial stage, the texts were taken from finished garamantas.lv transcriptions, with some minor editing in most obvious cases. The almost perfect quality of the transcriptions available at garamantas.lv accelerated the process of ground truth preparation. The major difficulty was related to the distribution of transcription blocks between the lines, as Transkribus requires a very detailed image-based reproduction of texts in transcriptions. In contrast, the guidelines for crowdsourcers ask for normalized form of erratic words to be put in square brackets. This information is not based on the scanned images, and such additions had to be removed from the transcriptions.

Even though this work was initially conducted just by one person, a trained linguist with only theoretical knowledge of Latvian Romani, we hope that access to the resources on this and other dialects, most importantly, the dictionary [6], allowed for the correct interpretation of most cases, and that the number of errors is minimal. At the same time, this person speaks Latvian at an advanced level, which allowed for more thorough checking of the Latvian part of the texts. At the later stage, two researchers who specialize in Romani dialects, including Lotfitka, joined the process of deciphering, which accelerated proofreading of automatically recognized pages. These pages were also lately approved as ground truth by the more experienced transcriber after checking the proofread pages. In this way, we managed to proofread several more folklore units by the middle of 2021. The priority was given to bilingual texts, with monolingual ones left for the later deciphering stage.

Due to the standard format of the copybooks used by Leimanis, the standard size of the files is about 20–34 lines (usually two-page layouts with the Latvian Romani text on the left and the Latvian translation on the right side). It takes about 6–10 minutes to copy and paste a previously deciphered text and distribute it correctly across recognized lines, as well as to correct minor things (e.g., add strikethrough annotation for some words which were not relevant for the transcribers, but which are more important for the HTR training in Transkribus). Figure 2 shows what the text from Figure 1 looks like after having been transcribed.

It is claimed that training a good Transkribus model requires about 15000 words, or 75 pages, of ground truth material. In our case, an average file with a two-page layout has about 160 words. This gives the following preliminary calculation: at least 94 files (15000/163) are needed to train the HTR model. As there were more deciphered files available at garamantas.lv, it was decided to add at least 200 deciphered pages as ground truth transcriptions.

The first model, *Leimanis_test*, was trained on 212 pages and validated on 23 pages. About 30 pages in the training data are copybook covers, which have only several shortened lines and describe metadata on the copybook content. The second model, *Leimanis_updated*, has the first model as its base model. Only several pages were additionally transcribed and used as ground truth data. The proportion of train and validation data was shifted to the improvement of the latter in terms of size. Still, using the base model and a higher number of epochs (100 in the second model vs. 50 in the first model) resulted in a considerable improvement of quality, as character error rate (CER) was already below the threshold of 5%. Both models were trained with the CITlab HTR+ method. The comparison of our models is given in Table 1 below.

[2] https://readcoop.eu/transkribus/public-models/ (accessed at 15/02/2022).

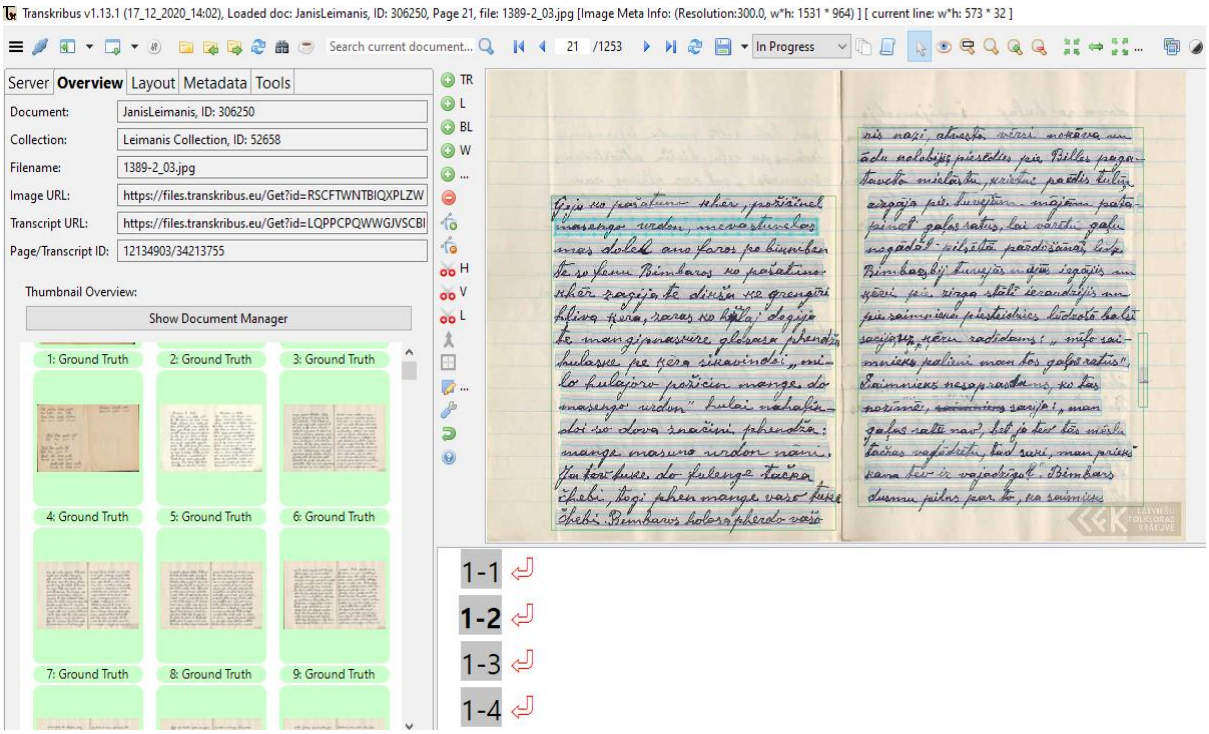

Figure 1: Page 21, unit 1389-02-03 (copybook 2, page 3) without transcription

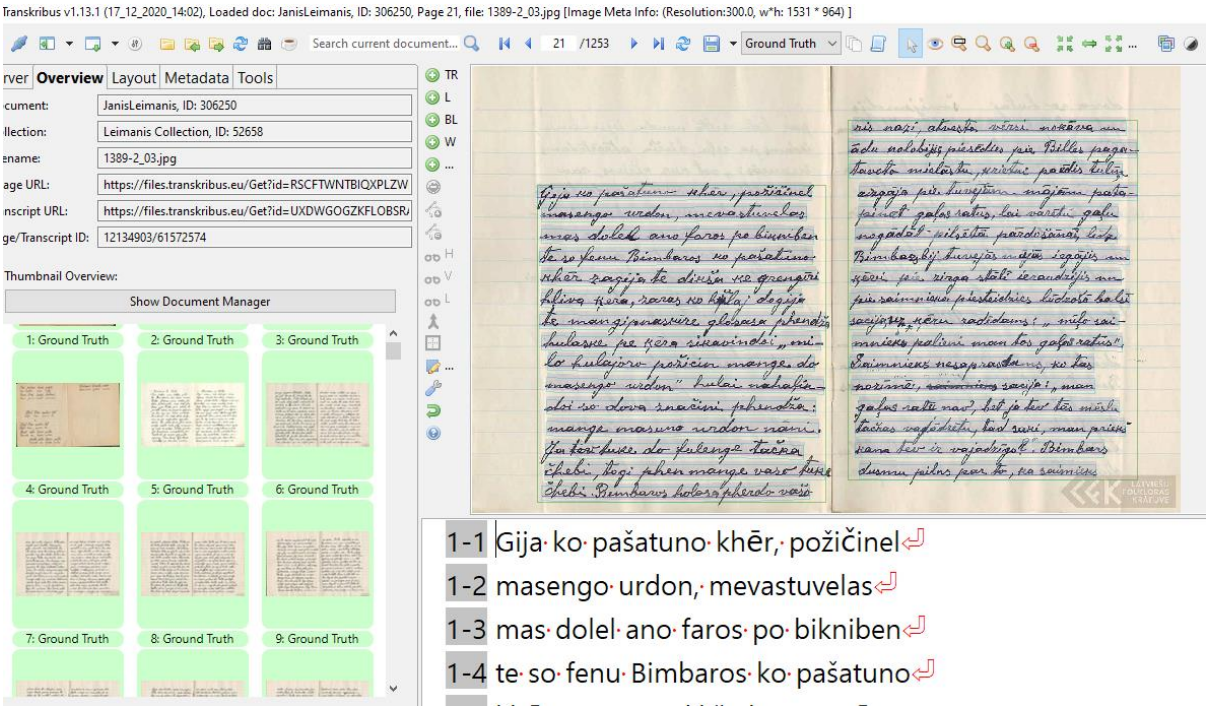

Figure 2: Page 21, unit 1389-02-03 (copybook 2, page 3) with transcription (ground truth)

Table 1
Comparison of two HTR models trained on Leimanis' archive

Model	Leimanis_test	Leimanis_updated
Base model	no	Leimanis_test
Epochs	50	100
Train set(pages)	212	144
Validation set (pages)	23	97
Train set (words)	25890	16952
Train set (lines)	4911	3136
CER on train set	3.51%	1.76%
CER on validation set	8.44%	3.95%

A preliminary layout analysis had been previously performed with the help of the CITlab Advanced tool with default settings. For ground truth pages, it was always manually corrected, when necessary, but for many pages it is still left uncorrected: this is seen as a preparatory step before each automatic page recognition procedure, so it is conducted consecutively.

After obtaining the HTR model and transcribing the rest of the files automatically, one should have a look at the quality of the obtained transcriptions. There is some chance that the initial quality of the model would not be enough, then adding more manual transcription could probably improve the training. In any case, in order to continue work with the texts, one needs to correct all possible errors in automatic transcriptions. The project page gives information on character error rate about 5% for some of the available models, which, of course, requires some postprocessing of the automatically annotated files.

For evaluation of the quality of recognition of the two available models and their comparison, we launched HTR on a page outside the ground truth data (page 1245 in our collection, originally notebook 68, page 5[3]). It is a fragment of the unit 426 (http://garamantas.lv/lv/unit/400700/LFK-1389-426), a fairytale 'How a baron married a Roma girl, and how she preferred forest and her people' ("Kā skaisto čigānieti apprecēja barons, un ka tai mežs un tautieši labāk patika..." / "Sir šukārune romane čha lija baronos, te sir lake, vešs te roma fidīr kamdžapes nasir filačin te...").

Figure 3 shows the result of automatic layout analysis (CITlab Advanced with default settings). The text in notebooks is very consistently and clearly written line-to-line, and corrections are minor, in this case only as a superscript word in line 5 on the right page. The layout recognition works very well, and in this case all the errors are minor and expected and belong to the regularly occurring types. First, the very first line is split into two parts; this happens occasionally, but still not too frequently to noticeably disturb the process of post-correction. Second, superscripts are usually assigned separate lines in automatic layout analysis. This is not incorrect from a purely visual perspective, but for structural reasons we want to incorporate such superscript words in the lines where they truly belong by merging them with the next line. As Transkribus provides additional markup for superscript and subscript (also for bold, italic, underlined, and strikethrough, among others), the post-corrected result can be easily used for training next models and for the restoration of the original.

[3]Here by pages we mean double pages as scanned and uploaded to garamantas.lv; page numbers are also reflected in the file names (1389-68-05 at the website and 1389_68_05.jpg in our collection).

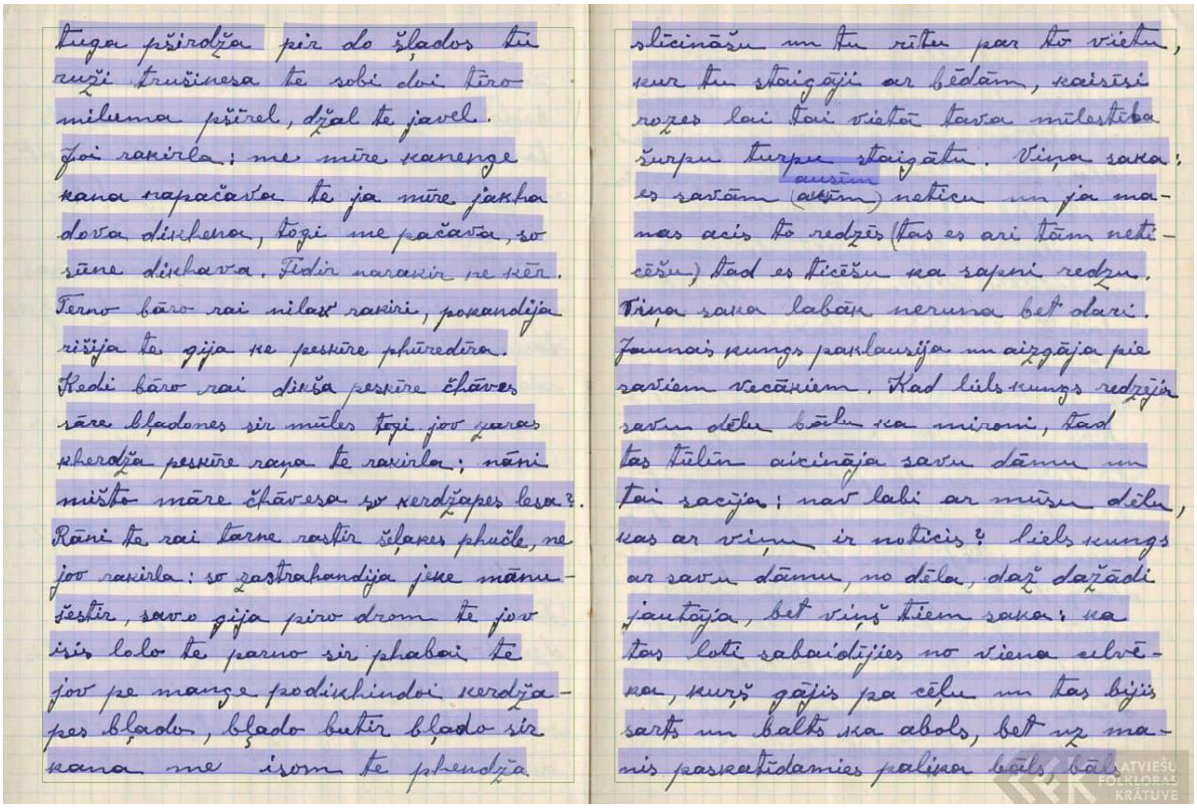

Figure 3: Notebook 68, page 5 (unit 426): the result of the automatic layout analysis

The Transkribus interface has an option for visual and metrics comparison of any compatible versions of the same pages. For instance, Table 2 shows how the processing of the same page (after the automatically proposed layout has been checked for errors and corrected) differs in terms of quality compared to the proofread GT page; values for both CER and WER (word error rate) are provided. Language models are generated on the basis of training data in the process of the HTR model training; they can be later used in recognition of particular pages. In the official tutorial, it is said that using language models does not always result in better recognition quality[4]. In our case, the better model performs with higher quality without language model added, but considering such a tiny dataset (one double page), this difference is explained by just a two-symbol improvement in the word under the superscript: *laizīm*) vs. (*aiīcm*) vs. the correct form (*acīm*). Still, the qualitative difference between the two models is more clearly reflected in the results, see Figure 4 (there are four other errors for *Leimanis_test* below the last line ending by *redzēja*; the better model has no errors in the same fragment).

Table 2

Comparison of recognition results by the HTR models trained on our corpus (with and without automatically generated language models) and ground truth data for page 1245 (1389-68-05)

Model	CER	WER
Leimanis_updated + LM	1.00	2.70
Leimanis_updated	0.86	2.36
Leimanis_test + LM	1.50	4.73
Leimanis_test	1.64	5.41

[4] See the tutorial available at https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/ (accessed on 15/02/2021).

Recognition errors are often related to diacritics either in the Latvian or the Latvian Romani text; sometimes diacritics can also be incorrectly added, as in our case with *sūne* (incorrect *šūne*). In addition, some problematic cases are really challenging, as they are related to less standard letter combinations occasionally made in handwritten texts (e.g., when one letter is distorted, or when borderlines between adjacent letters are not as clear as they normally are). Sometimes, capitalization errors are made, if Leimanis doesn't make clear size difference between the first capital letter and the following one (see the error *Viņa* / *viņa* in Figure 4). Finally, sometimes different letters are indeed similar in their graphic forms, therefore automatic recognition models make mistakes like with human recognition, as we as transcribers also sometimes struggle, trying to decipher some letters; in our case, the capitalized F in the word *Fidir* got interpreted as either L or T.

4. Conclusions

It has been shown how the opportunities provided by Transkribus can be successfully exploited with respect to unique handwritten Latvian Romani data. Initial volunteer-based manual deciphering efforts have been successfully integrated in the development of automatic handwritten recognition for Latvian Romani; we are thankful for being granted access to such preliminary ground truth data.

The character of handwriting and accurate original copybook layout made it possible to train recognition models, the best of which has very satisfactory quality, so that the further deciphering process is based on post-correction of automatically recognised pages. Errors made by the model *Leimanis_updated* are not numerous and in many cases can be compared to similar difficulties faced by ordinary transcribers.

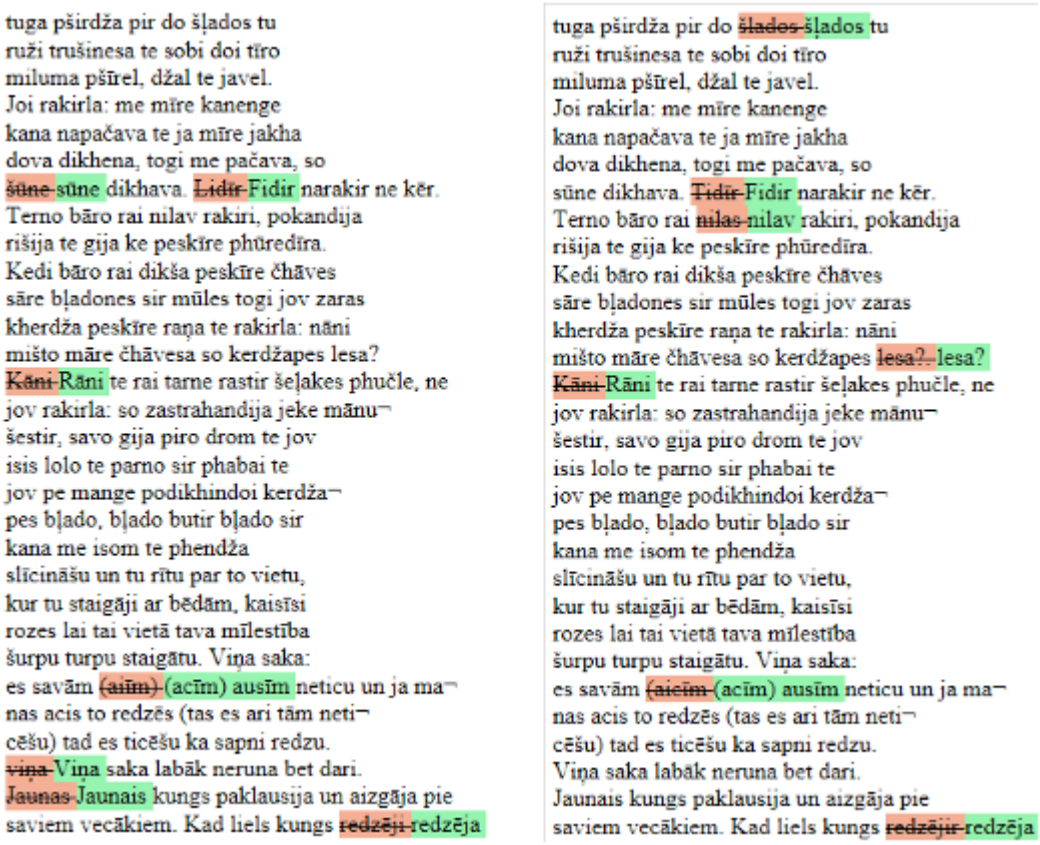

Figure 4: Visual comparison of best results for two models (Leimanis_updated to the left and Leimanis_test + LM to the right)

We hope that our initial effort at digitizing Leimanis' archive will result in the development of a bigger Latvian Romani corpus. This would, of course, imply more serious normalization of texts, additional translations (at least in English), developing morphological annotation for Latvian Romani, etc. The bilingual character of Leimanis' collection makes it possible to consider the compilation of a parallel corpus, with such options as, for instance, word alignment. It is also worth mentioning that for several texts later translations are available in [15] and in the materials from Pasakas.net[5]. Various Latvian Romani materials are already available in our repository (https://github.com/LatvianRomani/Lotfitka).

5. Acknowledgements

We are thankful to Sanita Reinsone for giving access to the collection and kind support and to Ieva Tihovska for her help with text transcriptions and sharing her knowledge about the history of Latvian Roma people and Jānis Leimanis in particular. We would also like to recognize the contribution by Anette Ross who helped us with proofreading the automatically recognized transcriptions and gave us access to various texts in Latvian Romani.

6. References

[1] Y. Matras, Romani: A linguistic introduction, Cambridge University Press, Cambridge, 2002.
[2] A.Tenser, Northeastern Group of Romani Dialects, Ph.D. thesis, The University of Manchester, 2008.
[3] P. Ariste, Romenge paramiši: Mustlaste muinasjutte, Tartu, 1938.
[4] P. Ariste, Supplementary review concerning the Baltic Gypsies and their dialect, Journal of the Gipsy Lore Society, Third Series, 43, 1/2 (1964) 35-37.
[5] P. Ariste, Einige Märchen Čuchný-Zigeuner, Tartu Riikliku Ülikooli Toimetised 309 (1973) 5-40.
[6] L. Mānušs, J. Neilands, K. Rudevičs, Čigānu-latviešu-angļu un latviešu-čigānu vārdnica, Zvaigzne ABC, Riga, 1997.
[7] Y. Matras, Ch. White, V. Elšik, The Romani morpho-syntax (RMS) database, in: M. Everaert, S. Musgrave, A. Dimitriadis (Eds.), The use of databases in crosslinguistic studies, Mouton de Gruyter, Berlin, 2009, pp. 329–362.
[8] A. Arkhipov, A. Barinskaya, R. Shtefura, Using handwritten text recognition on bilingual Evenki-Russian manuscripts of Konstantin Rychkov, Scripta & eScripta 21 (2021) 233-244.
[9] N. Partanen, J. Rueter, M. Hämäläinen, K. Alnajjar, Processing M. A. Castrén's materials: Multilingual typed and handwritten manuscripts, in: M. Hämäläinen, K. Alnajjar, N. Partanen, J. Rueter (Eds.), Proceedings of the Workshop on Natural Language Processing for Digital Humanities, The Association for Computational Linguistics, Stroudsburg, 2021, pp. 47–54.
[10] I. Tihovska, Jānis Leimanis and the Beginnings of Latvian Roma Activism, in: E. Marushiakova, V. Popov (Eds.), Roma Portraits in History: Roma Civic Emancipation Elite in Central, South-Eastern and Eastern Europe from the 19th Century until World War II, Brill | Schöningh, Leiden, 2022. doi: https://doi.org/10.30965/9783657705191_011
[11] J. Leimanis, Čigāni Latvijas mežos, mājās un tirgos, Zinātne, Rīga, 2005 [1939].
[12] S. Reinsone, Searching for deeper meanings in cultural heritage crowdsourcing, in: P. Hetland, P. Pierroux, L. Esborg (Eds.), A History of Participation in Museums and Archives: Traversing Citizen Science and Citizen Humanities, Routledge, 2020. doi:10.4324/9780429197536-14
[13] P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, Transkribus – a platform for transcription, recognition and retrieval of document images, IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2017, pp. 19–24.
[14] G. Mühlberger, L. Seaward, M. Terras, S. Ares Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G.

[5] As the original domain was taken by other people, a duplicated archival copy is currently stored at asakas.net.

Heyer, L. Hirvonen, T. Hodel, M. Jokinen, Ph. Kahle, M. Kallio, Fr. Kaplan, Fl. Kleber, R. Labahn, E. M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J.-L. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver, Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. Andreu Sánchez, Ph. Schofield, G. Sfikas, Chr. Sieber, N. Stamatopoulos, T. Strauß, T. Terbul, A. Héctor Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Weidemann, H. Wurster, K. Zagoris. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, Journal of Documentation 75, 5 (2019) 954-976. doi: 10.1108/JD-07-2018-0114

[15] S. Brice (Ed.), Laimes puķe: čigānu tautas pasakas, Sprīdītis, Rīga, 1992.