Silga, Sviķe, *Dr. philol.*
Ventspils University of Applied Sciences

Report on ELEXIS Transnational Research Visit Grant at the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH-OeAW). (Vienna, Austria March 16-20 2020)

Travel Grant: Call 3

Project title: "German-Latvian LSP Glossary of Kawall's "Dieva radījumi pasaulē" and its Original Work"

Introduction

My visiting grants project proposal is part of a larger project aiming to research Latvian botanical terminology used in H. Kawall's work "God's Creatures in the World" ("Dieva radījumi pasaulē"). This work (a textbook) is one of the first translations from German into Latvian, in which the author mentions Latvian botanical terms for the first time ever, in addition to the terms of zoology and mineralogy, the book has a separate chapter – Plant Kingdom (Augu valsts). A detailed research of botanical lexis requires a digital corpus of language material on which to compare and study special lexis used in the original language and translation. Therefore, the **goal** of the research stay is to create a bilingual digital LSP corpus based on the original book in German and its translation into Latvian, along with the aim to compile a bilingual LSP glossary that includes a collection of special botanical vocabulary used in H. Kawall's translation and original work.

In this report, I will present the workflow of my research visit, i.e. preparatory work, used books, support and main work at the ACDH-OeAW, materials studied and tools used to the achieve the goal – create two corpora and a glossary. I will summarize the conclusions and single out some possible solutions for the further research process.

Workflow and description of steps for performing specific tasks

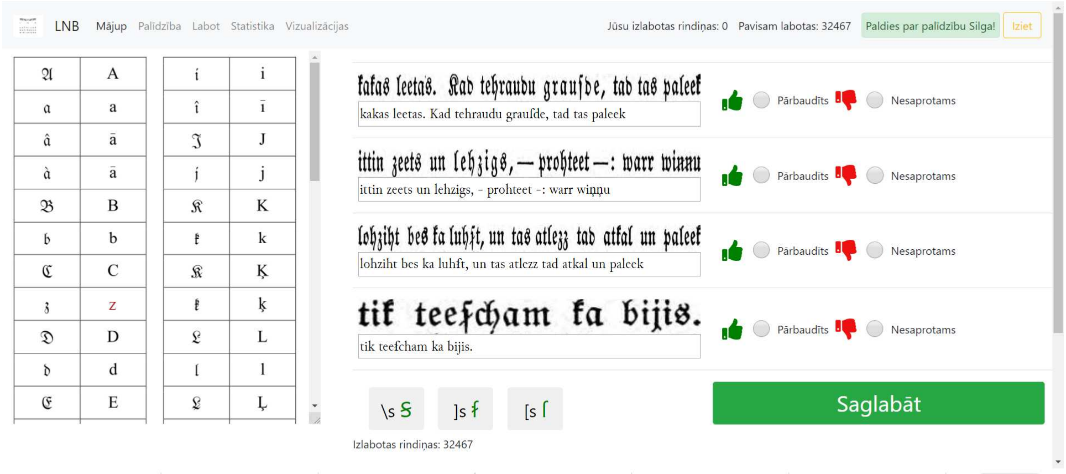
Conducting preparatory work before the research visit.

Preparation of two printed books: H. Kawall “Deewa raddijumi pasaulē” („Dieva radījumi pasaulē” (DRP)) (1860) and “Die Naturgeschichte für Kinder und Elementarschüler, oder erster Unterricht über das Mineralreich, Pflanzenreich und Tierreich, mit über 300 kolorierten Abbildungen” (1855). Scanning and saving both books in PDF format.

Digitising paper books (OCR scans).

For solving theoretical issues, finding tools and methods of digitisation of both scanned books the support of the National Library of Latvia (NLL) was sought. The scanned documents had to be prepared in OCR (optical character recognition) format, i.e. in such a format to make it possible to find, edit and process specific fragments of the text using the search function. The documents could not be pictures, such as jpg format. Doc-Works programme environment was used to process texts by working with the scanned material. The texts of both books are written in the old script, and it was the greatest challenge of the project. The new supervised machine learning approach offered by NLL was used to digitise the texts, and these two books were the first ones to be processed like this. Initially the text was recognised by an untrained algorithm. Next, the text was edited.

For the computer to recognise text accurately, a sample of the correct representation of the text must be made first – at least 10 000 perfect, human-edited lines. Each line must be checked by two writers editing individual lines. You can do it and view the process on the NLL website: https://frakturs.lnb.lv/ (see Fig. 1)



Jūsu izlabotas rindiņas: 0 Pavisam labotas: 32467

𝔄	A	i	i
𝔞	a	î	ī
â	ā	𝔍	J
à	ā	j	j
𝔅	B	𝔎	K
𝔟	b	𝔨	k
ℭ	C	𝔎	Ķ
𝔷	z	𝔨	ķ
𝔇	D	𝔏	L
𝔡	d	𝔩	l
𝔈	E	𝔏	Ļ

kakas leetas. Kad tehraudu grauſde, tad tas paleek
kakas leetas. Kad tehraudu grauſde, tad tas paleek

ittin zeets un lehzigs, — prohteet —: warr winnu
ittin zeets un lehzigs, - prohteet -: warr wiņņu

lohziht bes ka luhſt, un tas atlezz tad atkal un paleek
lohziht bes ka luhft, un tas atlezz tad atkal un paleek

tik teeſcham ka bijis.
tik teefcham ka bijis.

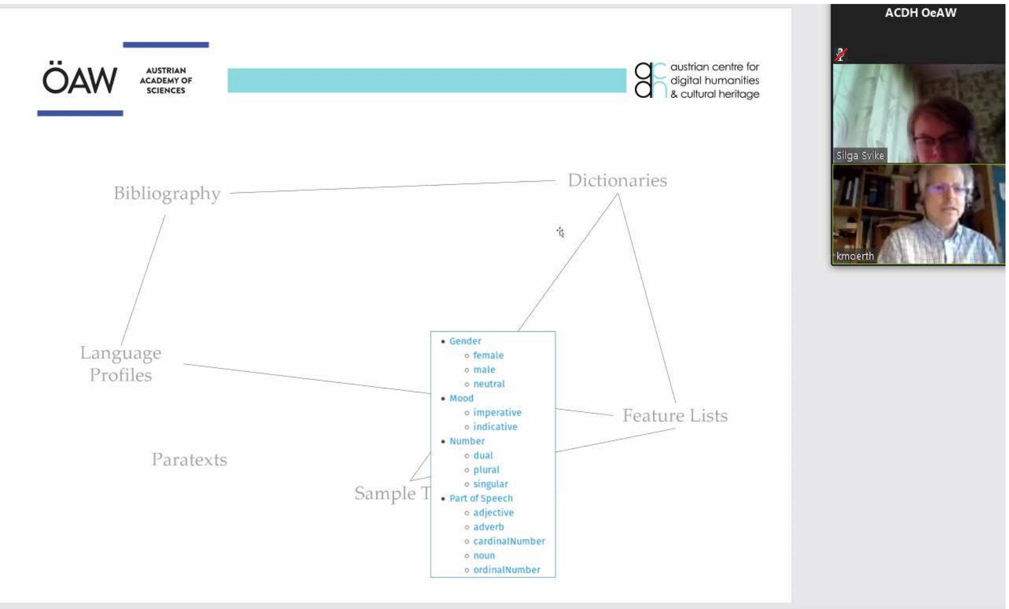
Izlabotas rindiņas: 32467

Figure 1. Editing Gothic texts manually

To enable automatic recognition, the Tesseract text recognition software was used. It works using the LSTM (long short-term memory) neural network model. LSTM operates more accurately than the early neural network models, and it is well-suited for script and speech recognition. LSTM belongs to the deep learning algorithms. The Tesseract software is a completely new solution of the year 2019, and it was used in this project as an experiment. This program used the German text model as the basic model, where the letters have diacritical marks.

During research visit.

Participation in the virtual meeting with deputy head of the ACDH-OeAW, Dr. Karlheinz Mörth, who shared expertise regarding best practice and standards in lexicography (see Figure 2). During the virtual meeting there was a possibility follow insights into different projects carried out in the ACDH-OeAW e.g. Vienna Corpus of Arabic varieties (VICAV). It was a good opportunity to read the study “Best practices for lexicography – intermediate report” by C. Tiberius, R. Costa, T. Erjavec, S. Krek, J. McCrae, C. Roche, T. Tasovac, 31 January 2020 (available at: http://ejuz.lv/estpractices).



ÖAW
AUSTRIAN ACADEMY OF SCIENCES
austrian centre for digital humanities & cultural heritage
ACDH OeAW
Silga Svike
kmoerth

Bibliography

Dictionaries

Language Profiles

Paratexts

Sample T

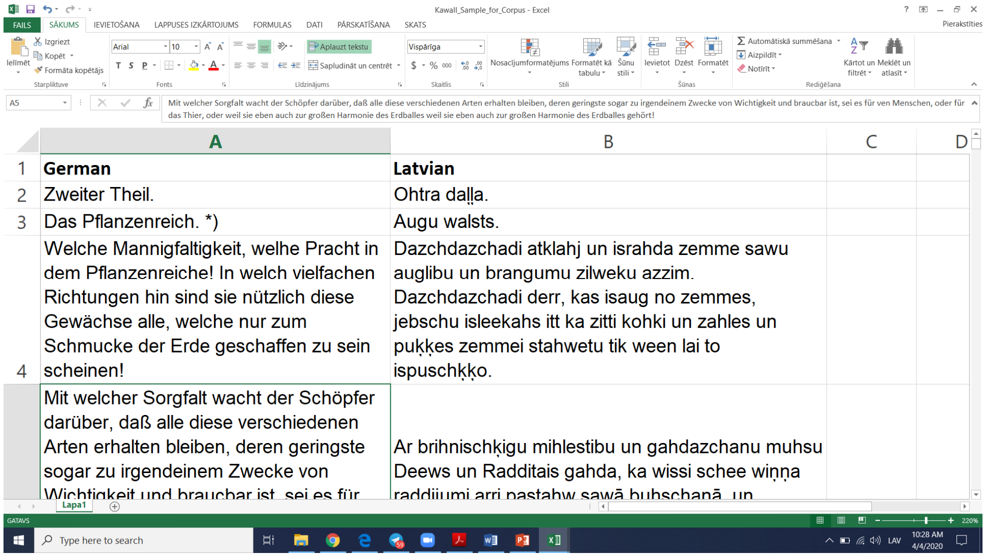
Feature Lists

- Gender
  - female
  - male
  - neutral
- Mood
  - imperative
  - indicative
- Number
  - dual
  - plural
  - singular
- Part of Speech
  - adjective
  - adverb
  - cardinalNumber
  - noun
  - ordinalNumber

Figure 2. Virtual Meeting with Dr. Karlheinz Mörth

Aligning the original work and its translation.

Aligning was performed by taking a sample from Chapter 2 of both books and manually copying the original work and translation segments into Excel Sheets (see Figure 3), as Excel is one of the formats required for future work with Sketch Engine. As the German text was relatively erroneous and required a lot of manual editing work, only a small part of the text could be processed in this way during the research visit.



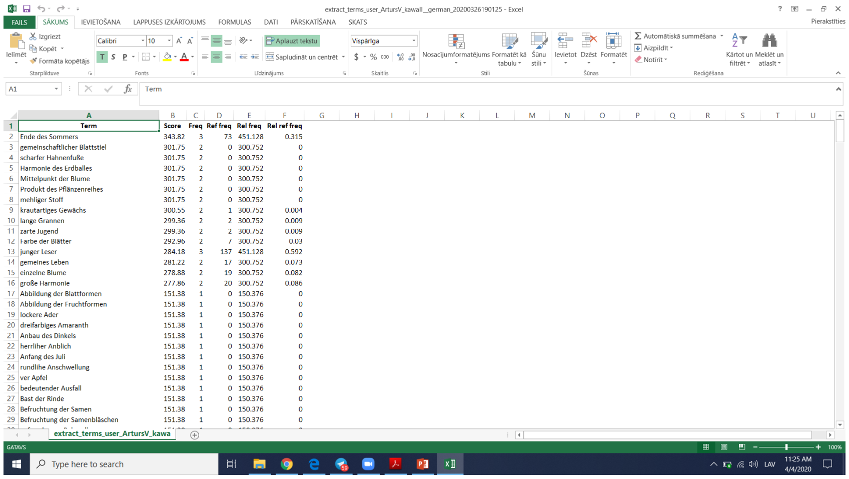
Mit welcher Sorgfalt wacht der Schöpfer darüber, daß alle diese verschiedenen Arten erhalten bleiben, deren geringste sogar zu irgendeinem Zwecke von Wichtigkeit und braucbar ist, sei es für ven Menschen, oder für das Thier, oder weil sie eben auch zur großen Harmonie des Erdballes weil sie eben auch zur großen Harmonie des Erdballes gehört!

	A	B	C	D
1	**German**	**Latvian**		
2	Zweiter Theil.	Ohtra daļļa.		
3	Das Pflanzenreich. *)	Augu walsts.		
4	Welche Mannigfaltigkeit, welhe Pracht in dem Pflanzenreiche! In welch vielfachen Richtungen hin sind sie nützlich diese Gewächse alle, welche nur zum Schmucke der Erde geschaffen zu sein scheinen!	Dazchdazchadi atklahj un israhda zemme sawu auglibu un brangumu zilweku azzim. Dazchdazchadi derr, kas isaug no zemmes, jebschu isleekahs itt ka zitti kohki un zahles un puķķes zemmei stahwetu tik ween lai to ispuschķķo.		
	Mit welcher Sorgfalt wacht der Schöpfer darüber, daß alle diese verschiedenen Arten erhalten bleiben, deren geringste sogar zu irgendeinem Zwecke von Wichtigkeit und brauchbar ist, sei es für	Ar brihnischķigu mihlestibu un gahdazchanu muhsu Deews un Radditais gahda, ka wissi schee wiņņa raddijumi arri pastahw sawā buhschanā, un		

Figure 3. Aligned texts in Excel Sheets

Extracting data for a bilingual glossary.

To obtain data for the glossary, the Sketch Engine software was used. Texts in Excel tables were first uploaded there, and two text corpora were created from those – German and Latvian language corpus. Using the functionalities of Sketch Engine extractions, keyword lists and term lists with detailed information were created, e.g. score, frequency (see Figure 4 for German). They are intended for future use in creating glossaries and in terminology research.



A1 | Term

	A	B	C	D	E	F
1	Term	Score	Freq	Ref freq	Rel freq	Rel ref freq
2	Ende des Sommers	343.82	3	73	451.128	0.315
3	gemeinschaftlicher Blattstiel	301.75	2	0	300.752	0
4	scharfer Hahnenfuße	301.75	2	0	300.752	0
5	Harmonie des Erdballes	301.75	2	0	300.752	0
6	Mittelpunkt der Blume	301.75	2	0	300.752	0
7	Produkt des Pflänzenreihes	301.75	2	0	300.752	0
8	mehliger Stoff	301.75	2	0	300.752	0
9	krautartiges Gewächs	300.55	2	1	300.752	0.004
10	lange Grannen	299.36	2	2	300.752	0.009
11	zarte Jugend	299.36	2	2	300.752	0.009
12	Farbe der Blätter	292.96	2	7	300.752	0.03
13	junger Leser	284.18	3	137	451.128	0.592
14	gemeines Leben	281.22	2	17	300.752	0.073
15	einzelne Blume	278.88	2	19	300.752	0.082
16	große Harmonie	277.86	2	20	300.752	0.086
17	Abbildung der Blattformen	151.38	1	0	150.376	0
18	Abbildung der Fruchtformen	151.38	1	0	150.376	0
19	lockere Ader	151.38	1	0	150.376	0
20	dreifarbiges Amaranth	151.38	1	0	150.376	0
21	Anbau des Dinkels	151.38	1	0	150.376	0
22	herrliher Anblich	151.38	1	0	150.376	0
23	Anfang des Juli	151.38	1	0	150.376	0
24	rundlihe Anschwellung	151.38	1	0	150.376	0
25	ver Apfel	151.38	1	0	150.376	0
26	bedeutender Ausfall	151.38	1	0	150.376	0
27	Bast der Rinde	151.38	1	0	150.376	0
28	Befruchtung der Samen	151.38	1	0	150.376	0
29	Befruchtung der Samenbläschen	151.38	1	0	150.376	0

Figure 4. Extracted Terms from German corpus

Short conclusions and future prospects after research visit.

To sum up the experience, it must be concluded that the prepared text material plays a significant role. In this case the text was in old script in two languages (German and Latvian). The greatest challenge of this research was to successfully prepare the old script material, and it needs to be further developed. Another objective would be to research the available software for editing German Gothic script, as manual editing work is time-consuming. The result of the research visit are digitised versions of two sections of printed books mentioned above and extracted keyword and term lists for the text samples. The extracted material will be further analysed by frequency data, as well as clarifying the most popular word collocations. In this way it will be possible to identify specific translation models and strategies in terminology translation developed by H. Kawall. The extracted Latvian botanical terms used in H. Kawall's work will also be analysed in relation to the contemporary botanical terminology, and the results of the research in the form of an article will be included in the multilingual “New Botanical Dictionary” (a mobile app prototype), developed at Ventspils University of Applied Sciences.

Special thanks to: Dr. Tanja Wissik, Dr. Karlheinz Mörth and Mag. Anna Woldrich, as well as for the possibility of this remote stay during the ELEXIS visiting grant, taking into consideration the restrictions made due to the corona crisis.