Archivum Lithuanicum 23

ARCHIVUM *Lithuanicum 23*

Turinys (Contents)

Diskusijos, apžvalgos, pastabos (Discussions, Surveys, Notes)

Mortimer Drach
Johann Wolfgang Goethe University, Frankfurt

CorDon – A Deeply Annotated Digital Corpus of the Works of Kristijonas Donelaitis

1. INTRODUCTION. The project "Digital Old Lithuanian: Corpus of Kristijonas Donelaitis (1714–1780)" (CorDon) has created a dynamic web presence which serves to:

1) make the complete Lithuanian-language works of Kristijonas Donelaitis freely available in a deeply annotated[1] online corpus,
2) align them with the manuscript as well as historical editions and translations, serving as a digital edition for readers,
3) enable textological, philological and linguistic queries and analyses,
4) provide a user-friendly search interface and output visualisation suited to the requirements of historical corpora.

CorDon was headed by Jolanta Gelumbeckaitė and Armin Hoenen at the Institute for Empirical Linguistics, Goethe University Frankfurt.[2] It was funded by the Fritz Thyssen Foundation.[3] The website, available in Lithuanian, German and English, can be found at https://titus.fkidg1.uni-frankfurt.de/cordon. This article will give a user-oriented overview of CorDon's contents and functionality, and briefly discuss the considerations behind their implementation.

2. CORPUS OF TEXTS. The centrepiece of CorDon (and of Donelaitis's oeuvre), at 20,400 words of text, is the four-part cycle *Metai* (DM), written in metrotonic hex-

---

[1] That is, featuring linguistically in-depth annotations.

[2] I was primarily responsible for what may be termed "philological delivery" under the supervision of Jolanta Gelumbeckaitė. This included digitising and formatting the texts, creating and correcting their annotations, creating the alignment tables for the alignment dictionary (see below), etc. The project's programmer was Philipp Büch, responsible for designing the website, search interface and Reading View, and the internal tools for exporting the data into these representations. The Lexicon, alignment dictionary and visualisations were created by Armin Hoenen.

[3] https://www.fritz-thyssen-stiftung.de/en/fundings/altlitauisch-digital-corpus-des-kristijonas-donelaitis-1714-1780-akronym-cordon/

ameters. It consists of the poems "Pavasario linksmybės" (*DM PL*), "Vasaros darbai" (*DM WD*), "Rudenio gėrybės" (*DM RG*) and "Žiemos rūpesčiai" (*DM ZR*). It was created between 1765 and 1775.

It is joined by two fragments associated with the *Metai* – the so-called "Fortsetzung" (*DF*, "Continuation") and "Pričkaus pasaka apie lietuvišką svodbą" (*DPP*) – as well as six *pasakos*, fables (*DP*): "Lapės ir gandro česnis" (*DP LG*), "Rudikis jomarkininks" (*DP RJ*), "Šuo didgalvis" (*DP SD*), "Pasaka apie šūdvabalį" (*DP PS*), "Vilks provininks" (*DP WP*) and "Aužuols gyrpelnys" (*DP AG*). In total, these texts add up to a corpus of around 24,000 words:[4]

Text	Word count	Text	Word count
DM PL 1765–1775	4,577	*DPN LG* 1896	541
DM WD 1765–1775	4,905	*DPN RJ* 1896	354
DMN RG 1896	6,059	*DPN SD* 1896	218
DMN ZR 1896	4,642	*DPN PS* 1896	379
DM F 1765–1775	199	*DPN WP* 1896	280
DMN PP 1896	1,490	*DPN AG* 1896	335
total			23,979

Table 1. Annotated texts in CorDon

DM PL.WD.F have been annotated in CorDon based on a diplomatic transcription of the surviving manuscript[5] which was created for the SLIEKKAS corpus.[6] The autographs of the remaining poems were lost (at the latest) during the upheavals of the Napoleonic Wars between 1807 and 1812.[7] Around 1794, Friedrich Hohlfeldt (1763–1829) made a copy of the Donelaitis manuscript. However, the fate of this copy, kept until the Second World War in the library of Königsberg University, is unknown today.

The first edition of the *Metai* by Ludwig Jedemin Rhesa[8] (1776–1840) contains a German translation as well as comments by Rhesa (*DMRh* 1818, 135–162). In 1824,

---

4 Not counting the parallel texts from other editions and translations (cf. table 2), which have not been annotated.

5 LLTIBR: F 1 – 5259; earlier signature of the Preußisches Staatsarchiv in Königsberg: Msc. A 120a-f. fol.

6 *Senosios lietuvių kalbos tekstynas* (SLIEKKAS) (https://titus.fkidg1.uni-frankfurt.de/sliekkas/index.html). This transcription was originally created by Goetz Keydana on the basis of the 1977 academic print edition (Donelaitis 1977, 89–261), and has been available in the TITUS database since

1992 (https://titus.fkidg1.uni-frankfurt.de/texte/etcs/balt/lit/donelait/donmetai/donme.htm). For SLIEKKAS, it was revised and corrected based on a new digitisation and analysis of the manuscript (cf. Gelumbeckaitė 2015, 23ff).

7 Cf. Krištopaitienė, Vaicekauskas 2014, 92. Rhesa remarks as much in describing his unsuccessful search for these manuscripts on p. XVI of the preface to *DMRh* 1818.

8 Better known in Lithuania as Martynas Liudvikas Rėza.

Rhesa published the first edition of the *pasakos*, without a German translation or accompanying texts (*DPRh* 1824). The second print edition of the *Metai* and the *pasakos* was published by August Schleicher in 1865 (*DMSch, DPSch* 1865). It also includes the first edition of *Pričkaus pasaka* (*DPPSch* 1865). While Schleicher did not provide a German translation of the Lithuanian texts, his edition does include a comprehensive Lithuanian-German glossary (*DMSch* 163–331). The third edition, which includes all Lithuanian texts by Donelaitis and a new German translation in hexameters, was published in 1869 by Georg Heinrich Ferdinand Nesselmann (*DMN, DPN, DPPN* 1869).

In addition to the above historical editions and German translations, the corpus includes two modern translations into other languages: an Italian translation of the *Metai* and *pasakos* by Adriano Cerri (*DMC, DPC* 2014), and an English translation of the *Metai* by Nadas Rastenis (*DMR* 1967).

Text	Poems included	Form
DM PL.WD.F. 1765–1775	PL,WD,F	manuscript
DMRh 1818	PL,WD,RG,ZR	edition & German translation
DPRh 1824	LG,RJ,SD,PS,WP,AG	edition
DMSch/DPSch/DPPSch 1865	all	edition
*DMN/DPN/DPP*N 1869	all	edition & German translation
DMR 1967	PL,WD,RG,ZR	English translation
DMC/DPC 2014	PL,WD,RG,ZR,LG,RJ,SD, PS,WP,AG	Italian translation

Table 2. Texts available in CorDon's Reading View

While none of the three 19th-century print editions can be considered wholly accurate to the manuscript, these are the only editions based directly on Donelaitis's manuscript and Hohlfeldt's copy. Nesselmann's edition deviates the least from the preserved original text; therefore, a diplomatic transcription of the Nesselmann text has been used as the basis for the CorDon annotation of those texts not preserved in manuscript form.[9]

---

9 The digitised text of *DMN RG.ZR* 1869 was available through SLIEKKAS, again based on Keydana's 1992 transcription of the 1977 academic print edition. The *pasakos* were present, in varying stages of completion, thanks to the work of students of the Old Lithuanian course of the Baltic linguistics programme at Goethe University, Frankfurt. These were corrected and completed, the remainder (*DM F, DPPN*) transcribed and annotated from scratch.

3. PRESENTATION OF TEXTS. CorDon presents Donelaitis's texts in two main ways: through the configurable *Reading View* and a versatile search GUI, designed to make the information of all annotation layers accessible for queries without requiring the user to learn a query language or the minutiae of the annotation scheme. Additionally, a lexicon/concordance has been generated, as well as a number of experimental visualisations.

The CorDon tagset used in annotating the texts is based on the SLIEKKAS tagset for Old Lithuanian,[10] representing essentially the subset of SLIEKKAS tags relevant to the language of Donelaitis in particular.[11] The SLIEKKAS annotation scheme is in turn modelled on that developed by the DFG collaborative project *Referenzkorpus Altdeutsch*[12] and the LOEWE Center "Digital Humanities"[13] in Frankfurt, making CorDon part of a wider "universe" of historical corpora.[14] The annotation scheme should thus be familiar to anyone who has worked with texts from the SLIEKKAS or *Referenzkorpus Altdeutsch* corpora. The texts were annotated semi-automatically using *Field Linguist's Toolbox*, which builds a dictionary of known forms as a text is annotated.

The lexical annotation comprises the following layers:
- *transliteration* – Transliteration of the word form into the modern Lithuanian alphabet
- *modernised form* – Modern Lithuanian equivalent of the form
- *lemma* – Citation form
- *accented lemma*
- *language*[15]
- *gloss of lemma* (*lt.*) – Lithuanian gloss of the meaning of the word in context

The grammatical annotation layers are as follows:
- *part of speech lemma*
- *part of speech form* – Part of speech of the concrete form in the text according to its function

---

10 For more on SLIEKKAS and the digitalisation and annotation of Donelaitis's texts, cf. Gelumbeckaitė 2015. SLIEKKAS also offers comprehensive metadata in its ELAN format versions of the transcribed manuscript texts. These include textological annotations such as correction layers (cf. p. 7f of the SLIEKKAS tagset documentation). This information has not been duplicated in CorDon.

11 In the process, some informal annotation practices, such as the use of commas or slashes in annotating corrected misprints, alternate lemmata, or ambiguous forms, had to be formalised to ensure consistent search results. The same applies to details unimportant to human readers of the annotations but significant in automatic parsing, such as the precise order of tags (is it *ja_Act_Past_Masc_Sg_Nom* or *ja_Act_Past_Masc_Nom_Sg*?) or the use of spaces after commas.

12 http://www.deutschdiachrondigital.de/

13 http://www.digital-humanities-hessen.de/

14 Cf. Gelumbeckaitė 2015, 21f.

15 This layer is intended for tagging the language of a token in multilingual texts, like the postils of the SLIEKKAS corpus. As such it is redundant in CorDon at present, since the texts are entirely in Old Lithuanian.

- *morphology – lemma* – Categories inherent to the lemma, such as inflection class or grammatical gender
- *morphology – form* – Specification of the inherent categories based on the concrete form in the text
- *inflection* – Variable inflectional characteristics of the concrete form in the text. Shows the endings of the form if it displays dialectal or orthographic peculiarities.

Transliteration is purely "mechanical": non-standard characters are systematically replaced with modern Lithuanian equivalents, but the forms are not otherwise normalised.[16] This is done on the *modernised form* layer, which gives modern standard formal equivalents. Thus, *PL* verse 5 *gaiwìdămĭ* is transliterated as "gaivįdami" and modernised as "gaivindami". Note that obsolete vocabulary is not replaced here. It is instead glossed with a modern equivalent or an explanation on the *gloss* layer.

While specialised dictionaries such as Kabelka (1964) and those included in *DMSch* and *DMN* were consulted, the final arbiter for lemmatisation was the *LKŽ*. If the lemma corresponding to a word form had an entry in the *LKŽ*, the form was lemmatised accordingly. If not, it was subsumed under a lemma which did have an entry. Where there were systematic differences between forms attested in the manuscript and Nesselmann's edition, they were lemmatised based on the evidence of the manuscript.[17]

Regarding the annotation criteria and system of grammatical annotation, the kind reader is referred to the documentation on the CorDon website and the documentation and publications of the SLIEKKAS project.

3.1. THE READING VIEW. The Reading View is accessed via the Texts page (fig. 1). This page gives an overview of the texts in the corpus and provides citations in several styles (APA, MLA, Chicago, Harvard) and formats (text, BibTeX) through the ▣ button. It also offers the annotated texts for download in ELAN's .eaf format, encoded as TEI xml, and (eventually) as a *Toolbox* package. Clicking on a text's title will open that text in the Reading View. The entries in the Editions column link to a digital facsimile of the edition in question.

The Reading View itself (fig. 2, 3) is an HTML representation of the annotated text using various JavaScript and CSS-based functionalities. Its layout was modelled

16 Though doubled consonants (indicating a short preceding vowel, in the German fashion) are replaced with single ones: thus, *buddĭnŏ > budino*.

17 For example, *PL* verse 2 *Truſùs* (fig. 2) is lemmatised as "trūsas" (not "triūsas"), as are all other occurences of "tr(i)ūsas" in the corpus, because the ms. consistently shows the non-palatalised form, and the *LKŽ* contains a lemma "trūsas". It is however **glossed** "triūsas". Conversely, *PL*

verse 180 *pleſdéndămĭ* does not qualify for a lemma "plesdenti" despite being attested in the ms., since this lemma is not included in the *LKŽ*. The form was thus lemmatized as "plezdenti" (and glossed "plazdenti"). This approach opens the door to interoperability with other resources making use of the *LKŽ* and also allows CorDon to offer users a link to the *LKŽ* entry of each lemma in its lexicon (cf. section 4).

371 C o r D o n – A D e e p l y A n n o t a t e d
D i g i t a l C o r p u s o f t h e W o r k s
o f K r i s t i j o n a s D o n e l a i t i s

Subcorpus	Text	Download		Editions	Manuscript
Metai	➤ *Pavasario linksmybės*		[TEI]	Rhesa (1818)	Autograph Donelaitis
				Schleicher (1865)	
				Nesselmann (1869)	
	➤ *Vasaros darbai*		[TEI]	Rhesa (1818)	Autograph Donelaitis
				Schleicher (1865)	
				Nesselmann (1869)	
	➤ *Rudenio gėrybės*		[TEI]	Rhesa (1818)	lost
				Schleicher (1865)	
				Nesselmann (1869)	
	➤ *Žiemos rūpesčiai*		[TEI]	Rhesa (1818)	lost
				Schleicher (1865)	
				Nesselmann (1869)	
Metų fragmentai	➤ *Fortsetzung*		[TEI]	Schleicher (1865)	Autograph Donelaitis
				Nesselmann (1869)	
	➤ *Pričkaus pasaka apie lietuvišką svodbą*		[TEI]	Schleicher (1865)	lost
				Nesselmann (1869)	
	➤ *Lapės ir gandro česnis*		[TEI]	Rhesa (1824)	
				Schleicher (1865)	

Figure 1. A section of the Texts page

on that of the *Lesekorpus Altdeutsch* (LeA, https://titus.uni-frankfurt.de/lea/).[18] It shows the token-level annotations as well as the parallel texts which have been aligned on a verse level with the "base" text (i.e. the diplomatic transcription of the manuscript where it survives, else the text of *DMN/DPN/DPPN* 1869). Toggles enable the user to display any combination of annotation layers, parallel editions and translations alongside the "base" text. Clicking on any word of the annotated text opens its entry in the lexicon (cf. section 3.3). For the manuscript texts (*DM PL.WD.F* 1765–1775) every fifth verse has been linked to a digital facsimile of the manuscript (fig. 4).[19]

18 LeA is designed to be an accessible, readable repesentation of the annotated texts of the *Referenzkorpus Altdeutsch* (ReA, https://www.deutschdiachrondigital.de/) (Mittmann, Plate 2019).

19 Used with the kind permission of the Institute of Lithuanian Literature and Folklore, Vilnius.

Figure 2 (screenshot):

text selection: Pavasario linksmybės

annotation layers: ☐ transliteration ☑ modernized form ☑ lemma ☐ accented lemma ☐ language ☑ part of speech lemma ☐ part of speech form ☐ morphology – lemma ☐ morphology – form ☑ inflection ☐ gloss of lemma (lt.)

PL_3r_1(1)	Iau	Saulėlė	wėl	atkópdāmā	buddinō	Swieta	,
Nesselmann (1869)	Jaú saulėlė wėl atkópdāmā bùdīnō swėtą ,						
modernized form	jau	saulelė	vėl	atkopdama	budino	svietą	-
lemma	jau	saulelė	vėl	atkopti	budinti,budyti	svietas	-
part of speech lemma	ADV	NA	ADV	V	V	NA	-
inflection	-	Sg_Nom_(e)	-	Fem_Sg	Ind_Past_3	Sg_Acc_(a)	-
Rhesa (1818)	Wiederum wendet die Sonne den Lauf und wecket die Welt auf,						

PL_3r_2(2)	Ir	Žiemõs	Šzaltõs	Trufūs	pargráudāmā	jūkės	.
Nesselmann (1869)	`Ir žemõs szaltõs triusùs pargráudāmā jůkės .						
modernized form	ir	žiemos	šaltos	trūsus	pargriaudama	juokėsi	-
lemma	ir	žiema	šaltas	trūsas	pargriauti	juoktis	-
part of speech lemma	KO	NA	ADJ	NA	V	V	-
inflection	-	Sg_Gen	Pos_Fem_Sg_Gen	Pl_Acc	Fem_Sg	Ind_Past_3_(-)	-
Rhesa (1818)	Spottet des Winters, des kalten, zerstörend was mühsam er bau'te.						

Editions: ☑ Nesselmann (1869) ☐ Rhesa (1818) ☐ Schleicher (1865)
Translations: ☐ Cerri (2014) (it.) ☐ Nesselmann (1869) (dt.) ☐ Rastenis (1967) (en.) ☑ Rhesa (1818) (dt.)

Figure 2. The first two verses of *PL* in the Reading View, with a selection of annotation layers and two parallel texts (edition *DMN*, translation *DMRh*) toggled for display

Figure 3 (screenshot):

PL_3r_1(1)	Iau	Saulėlė	wėl	atkópdāmā	buddinō	Swieta	,
Nesselmann (1869)	Jaú saulėlė wėl atkópdāmā bùdīnō swėtą ,						
Rhesa (1818)	Iau Saulėlė wėl atkópdama buddino Swieta						
Schleicher (1865)	Jaú saulėlė vėl atkópdama bùdino svêtą						
transliteration	iau	saulele	vėl	atkopdama	budino	svieta	,
modernized form	jau	saulelė	vėl	atkopdama	budino	svietą	-
lemma	jau	saulelė	vėl	atkopti	budinti,budyti	svietas	-
accented lemma	jaū	saulėlė (2)	vėl	atkópti,-kópia,-kópė	būdinti (būdyti),-ina,-ino	svietas (1)	-
language	olt	olt	olt	olt	olt	ólt	-
part of speech lemma	ADV	NA	ADV	V	V	NA	-
part of speech form	ADV	NA	ADV	VSP	V	NA	-
morphology - lemma	-	ė8_Fem	-	-ti,-ia,-ė	-ti,-a,-o	a1_Masc	-
morphology - form	-	ė_Fem	-	o	o-Past	a_Masc	-
inflection	-	Sg_Nom_(e)	-	Fem_Sg	Ind_Past_3	Sg_Acc_(a)	-
gloss of lemma (lt.)	-	saulė	-	kilti aukštyn	žadinti	pasaulis	-
Cerri (2014)	Risalendo, il solicello risvegliava già la terra						
Nesselmann (1869)	Wiederum hob sich die Sonne empor und weckte die Welt auf ,						
Rastenis (1967)	The climbing sun again was wakening the world						
Rhesa (1818)	Wiederum wendet die Sonne den Lauf und wecket die Welt auf,						

Figure 3. The first verse of *PL* displayed in the Reading View, with all annotation layers and parallel texts enabled
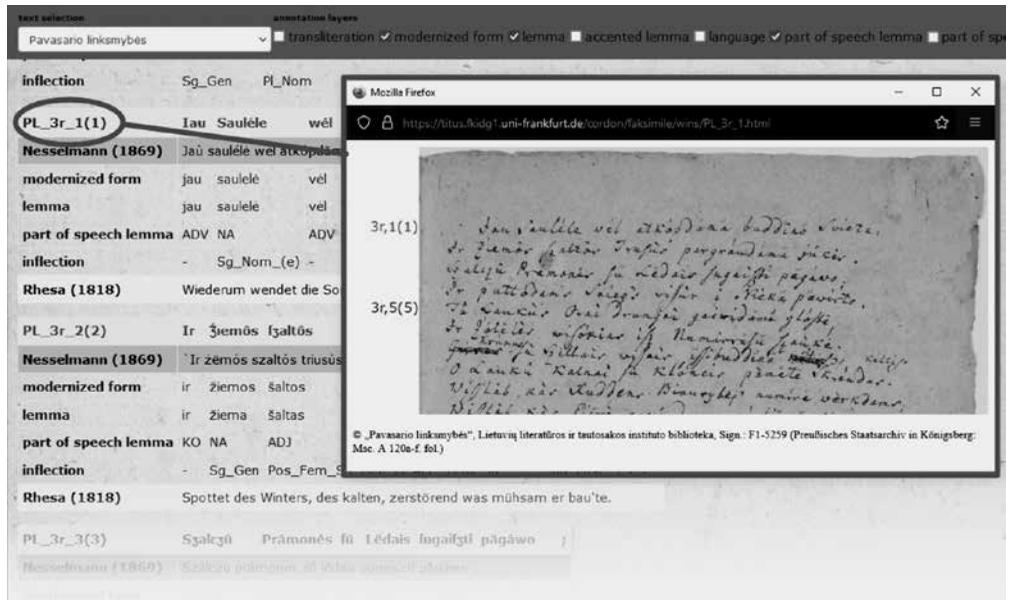
Figure 4. Every fifth verse line (highlighted in blue) in the Reading View
of *DM PL.WD.F* links to a facsimile of the manuscript, which pops up
in a separate window for side-by-side comparison

The Reading View can thus be arranged to suit a variety of needs. Linguists can view the deeply annotated base text. Scholars of Donelaitis may wish to synoptically display the text of various editions and translations. Readers and learners meanwhile can display the normalised forms and glosses to aid understanding of the original text.

The relatively straightforward approach of aligning the parallel texts with the "base" text on a verse level is reliable and would allow for the fairly painless integration of additional parallel texts (in particular, translations) in future.[20] It should however be noted that it carries some limitations when it comes to the representation of the parallel texts, which in this case concern *DMRh* 1818, and to a lesser extent *DMSch* 1865. Where a parallel text omits a verse that is present in the "base" text, this has simply been indicated by "----". However, transpositions and additions[21] relative to

---

[20] The parallel texts are formatted for import as plain .txt files, each line of verse preceded by four numbers indicating (1) the page, (2) line, and (3) verse number of the verse in the translation and (4) the number of the corresponding "base" text verse. Collaborators can supply aligned texts by adhering

to this simple format, needing no special knowledge of the workings of CorDon.

[21] Both *DMRh* and *DMSch* work the text of the "Fortsetzung" into *RG*. *DMSch* also adds two verses relative to *DMN*, namely *DMSch RG* verse 541 "Vis blogýn eídams sù manìm senésnis pastójo;" and *DMSch*
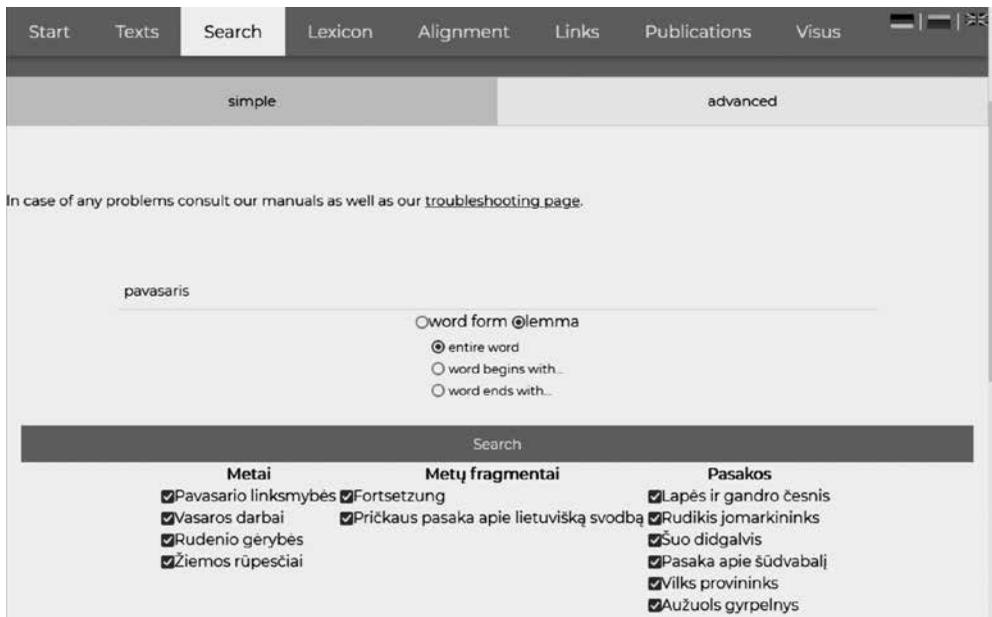
Figure 5. A *simple* search for the lemma *pavasaris* in all texts of the corpus
(selectable at the bottom)…

the "base" text cannot be adequately represented. Since each verse is aligned with its counterpart in the "base" text, transposed verses appear where they would in the "base" text, while added verses, lacking a counterpart, are omitted entirely. Users looking to read *DMRh* 1818 and *DMSch* 1865 as they originally appear will need to consult the facsimile linked on the Texts page.

3.2. THE SEARCH. The search GUI has two modes, *simple* and *advanced*. The *simple* search (fig. 5, 6) for basic queries finds lemmata or concrete word forms in the text. Its functionality is limited, but should be immediately recognisable and understandable for any user looking to get started with the corpus.

The *advanced* search meanwhile takes advantage of the systematic CorDon tagset annotation to allow users to generate complex queries using a cascade of drop-down menus. These menus only display those grammatical categories relevant to the currently selected part of speech and search type (word form or lemma). Fig. 7 shows the options offered for verbs with the search switched to "word form" and "lemma", respectively. An advanced search query may thus be narrowed step by step to find specific grammatical forms. This can be combined with a text string search.

ZR verse 642 "neí, kitĕms dárbus atlìkt, į̃ baúdžavą rèngtis." For the much more

complex situation in *DMRh* see Krištopaitienė 2007, and Krištopaitienė 2014.

database version	0.2-alpha
SQL-Query	show SQL_Query
selected texts	PL, F, LG, WD, PP, RJ, RG, SD, ZR, PS, WP, AG
Hits	21

1. PL_2r_1

	Pawasario	Linksmybės
modernized form	pavasario	linksmybes
lemma	pavasaris	linksmybe
part of speech lemma	NA	NA
inflection	Sg_Gen	Pl_Nom

2. PL_3r_31(31)

	Irgĭ		Pawáťáris	ant	wiffů		Laukû	pãťíródė	.
modernized form	ir	gi	pavasaris	ant	visų		lauku	pasirodė	-
lemma	ir	gi	pavasaris	ant	visas		laukas	pasirodyti	-
part of speech lemma	KO	PTK	NA	AP	PI		NA	V	-
inflection	-	-	Sg_Nom	-	Masc_Pl_Gen		Pl_Gen	Ind_Past_3	-

3. PL_4r_11(93)

	O	fʒtai	,	kàd	més	wèl	linkſmi	Pawáťárĭ	fʒwencʒiam'	.
modernized form	o	štai	-	kad	mes	vėl	linksmi	pavasarį	švenčiame	-
lemma	o	štai	-	kad	aš	vėl	linksmas	pavasaris	švęsti	-
part of speech lemma	KO	PTK	-	KO	PPER	ADV	ADJ	NA	V	-
inflection	-	-	- -		Pl_Nom	-	Pos_Masc_Pl_Nom	Sg_Acc_(i)	Ind_Pres_Pl_1_(-)	-

4. PL_4v_26(150) Kúda wiſſ ir álkſtantĭ Pawáťárĭ randa ;

Figure 6. …and the results. Annotation layers can be toggled as in the Reading View. Clicking on the verse ID at the start of a line will open the location in the corpus. Clicking on a word opens its lexicon entry

For each grammatical category (e.g. number), the user may choose to leave it unspecified (any number), or click the checkbox to its left in order to include it (number: singular) or exclude it (number: *not* singular) fig. 8).

Using the ⊡OR button at the bottom of the search box, multiple formal categories can be specified in the same query. The example in fig. 9 shows a search for tokens which: end in <a>, are nouns, and are in either the Nominative or the Vocative case. A maximum of seven categories can be specified in one query in this way.

The search can also find sequences of up to five tokens, where the parameters for each position in the sequence can be set individually. This is accomplished by clicking the + buttons on either side of the search box to open an additional search box there. Fig. 10 shows such a multi-token search for the lemma *būras*, preceded by any conjunction, followed by any verb. Fig. 11 shows the results.[22]

---

22 NB: The search finds *sequences* of tokens; they are not necessarily syntagms. Note also that punctuation marks are treated as tokens. For the purposes of the search, the sequence "Sveiks, dieve duok" therefore consists of **four**, not three, tokens. The search would have to be set to the sequence [*adjective*]–[*any, i.e. a "blank" search box*]–[*noun*]–[*verb*] in order to turn up this particular phrase.
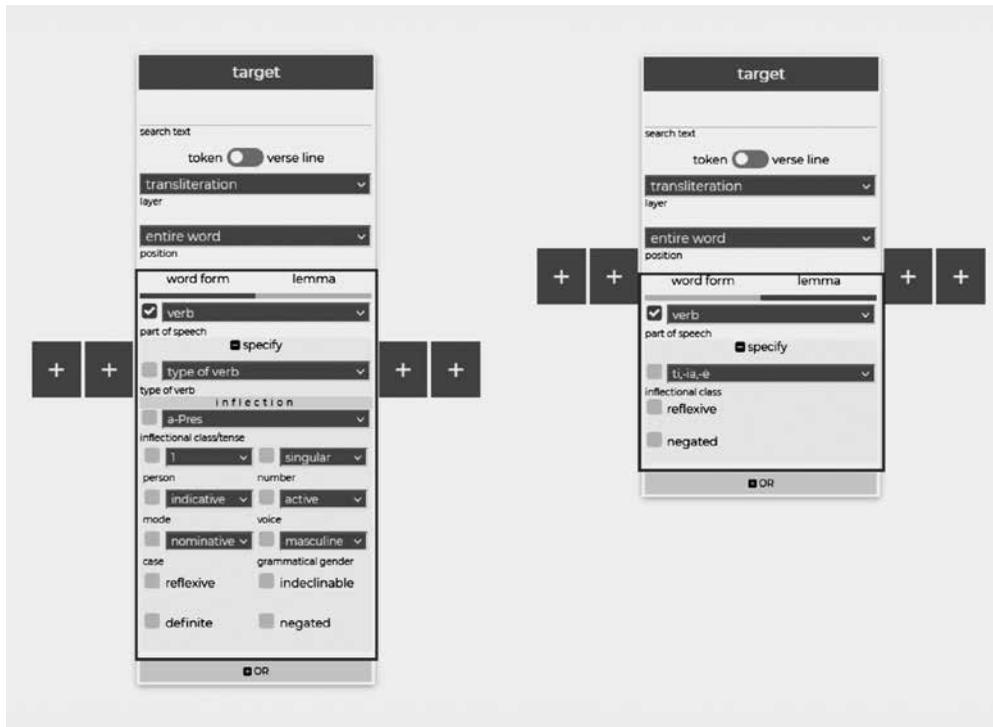
Figure 7. Context-sensitive options in the advanced search interface, showing
only those categories relevant to verb word forms (left) or lemmata (right)

The advanced search is a fairly intuitive tool, but some attention is required when first setting up a search query, to avoid confusion and misleading results. Fig. 12 shows the search box as it initially appears. In addition to entering the search text (if any), the user must initially choose:

a) Which text layer to search: The transliterated text, the original text (preserving all special characters such as <ſ> and <ʒ> as well as accents), the modernised text (which replaces Old Lithuanian word forms with their modern Standard Lithuanian equivalents), or the glosses.

b) The position of the searched-for characters (if any): Do they represent a whole word, or do they occur at the beginning, the end, or anywhere in the target word(s)?

c) Whether to search for concrete word forms in the text, or for lemmata.[23]

As an example for the importance of choice a) we may take the search term *būras*. Searching for the word form *būras* on the "original" or "transliteration" layer will

23 In the latter case, the choice of a) is ignored, and only the *lemma* layer will be searched.

Figure 8. Toggling categories.
From top to bottom: not
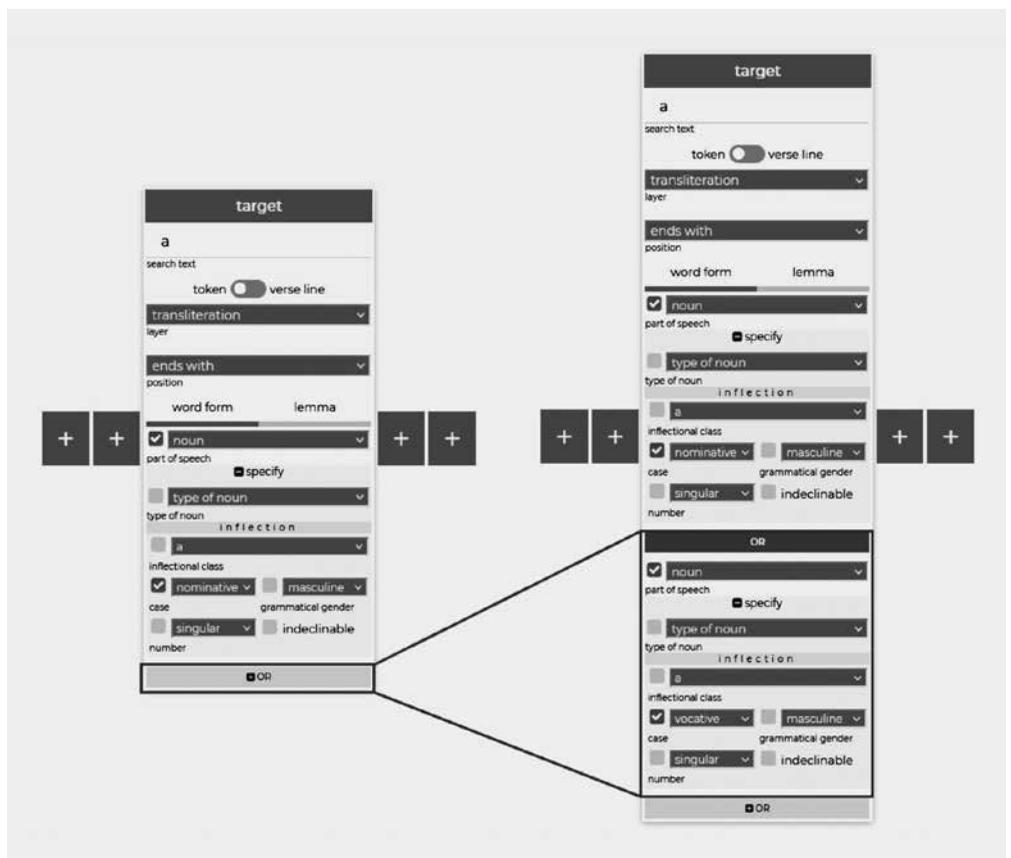singular, active, any gender

Figure 9. The OR button opens another panel in which another set of formal characteristics
can be specified for inclusion in the query. This search will find word forms which end
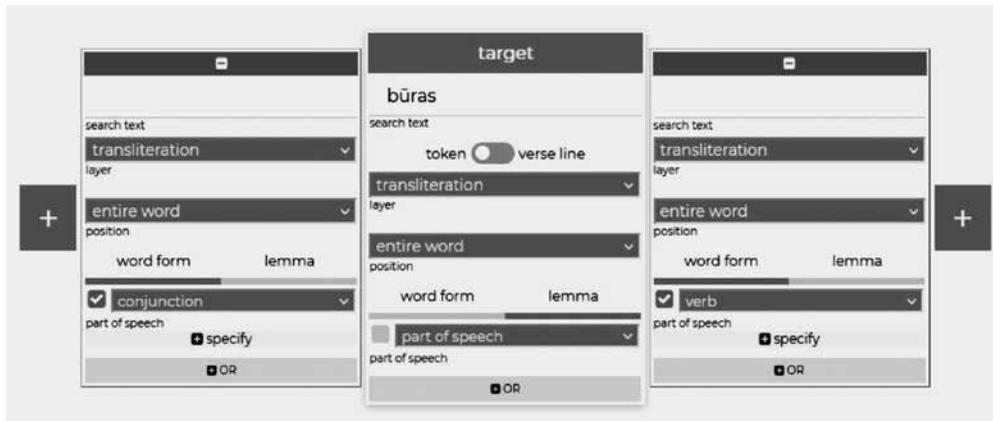in <a>, are nouns, and are in either the Nominative or the Vocative case

Figure 10. A multi-token search, specifying the sequence conjunction–Lemma: būras–verb…

Figure 11. …and the results, with the sequence highlighted

yield no results because the original orthography does not use the character <ū>, rendering the word as *bur(a)s*. Searching the "modernised form" layer, however, yields 22 results. Selecting *lemma* rather than *word* form at c), meanwhile, yields 104 results for the lemma *būras* in all its declension forms.
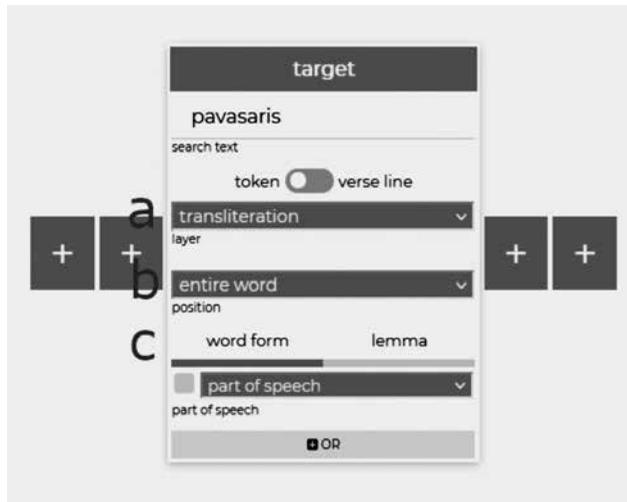
Figure 12. The advanced search box
in its initial configuration

Specifying desired forms in the search is also not entirely trivial. While the GUI prevents the user from inputting features that are impossible in principle (such as "adverb, masculine"), it cannot prevent them from inputting impossible or contradictory *combinations* of features, or otherwise misconfiguring the search (for example, by selecting the wrong layer, see above). A characteristic which may be somewhat unintuitive is that the categories offered by the search are **exclusive**, not inclusive. That is to say, a search for "noun > common noun" finds exactly those forms tagged as "common noun" (NA), but not those tagged as "common noun, adverbial" (NAA) or "common noun, predicative" (NAP). The structure of the annotations itself can also lead to some unexpected results.[24]

The site offers a trouble-shooting page and an illustrated guide to using the search which address these peculiarities, but in order to obtain reliable search results for "serious" investigations, users will still need to familiarise themselves with the annotation scheme. Ultimately, the GUI should significantly lower the barrier to entry compared to traditional solutions, and should encourage users to move beyond a simple word search. But it cannot simplify a complex subject matter below its inherent level of complexity.

---

24 Most notably when searching for lemmata. Ambiguous forms, negated reflexives, un-inflected posessive pronouns and verbs of the type *-inti/-yti* have "double" lemma entries, separated by a slash. Since the search views these entries as one "word", they will not be found when searching for (for example) "whole word: *savo*", since the lemma is "savo/savęs".

The context-sensitive search GUI distinguishes CorDon from other widely known annotated corpora[25] and corpus tools.[26] The aim was to create a search interface which makes the information of *all* annotation layers accessible without requiring the use of a query language and/or regular expressions, which can be off-putting to less technically minded scholars, or casual visitors looking to see whether the corpus suits their needs in the first place. At the same time, it does not "dumb down" the search for the sake of accessibility. By putting the options on display in an immediately readable format, which "unfolds" as the user goes through them, it avoids a situation where they are confronted with a blank search bar and may never move on beyond the familiar simple text search.[27] The options offered also indicate up-front what information can be found in the annotations. The verse structure of the texts serves as a "natural" unit for displaying the search results in context, so that they can be displayed as in the Reading View, tying together the two modes of accessing the texts.

The approach taken with CorDon's search does not depend on a specific annotation scheme or users' knowledge thereof, and can be generalised for data annotated in similar formats (e.g. ELAN).

4. THE LEXICON. The *lexicon* presents the vocabulary of Donelaitis's texts, sorted by lemma. It is structured on the concordance principle. The information in the lexicon is generated from the annotated texts, and includes the following: The lemma's part of speech, accentuation, inherent categories (like inflectional class or grammatical gender), any glosses of its meaning in its concrete occurrences in the text, a link to its entry in the *LKŽ*, and the lemma's ten nearest neighbours in lemma vector space[28] (generated using the *word2vec* tool[29]). For every lemma, the word forms

25 No comparable annotated online corpus currently exists for Old Lithuanian. The similarly annotated *Referenzkorpus Altdeutsch* (ReA, https://www.deutschdia-chrondigital.de/) relies on ANNIS for search queries. Its sibling projects *Referenzkorpus Mittelhochdeutsch* (ReM, https://www.linguistics.ruhr-uni-bochum.de/rem/) and *Referenzkorpus Mittelniederdeutsch/Niederrheinisch* (ReN, https://www.slm.uni-hamburg.de/ren) offer a "simplified ANNIS search mask" with basic functionality corresponding to CorDon's simple search. The *Lesekorpus Altdeutsch* (https://titus.uni-frankfurt.de/lea/) meanwhile, whose representation of ReA's annotated texts is the model for CorDon's Reading View, presents the texts in a very accessible way, but has no search functionality of its own.

26 Such as ANNIS (https://corpus-tools.org/annis/), with its query language AQL, whose Query Builder is hardly any less abstract than plain AQL queries, and presupposes familiarity with the annotation scheme. The situation is similar for annotated texts accessed through SketchEngine (https://www.sketchengine.eu/) and its Corpus Query Language (CQL).

27 Users are in fact presented with the advanced search GUI by default when opening the Search page, and must actively choose to switch to the simple search if basic text queries are all that interests them (though the advanced search of course supports these basic queries as well).

28 Proximity in vector space typically indicates a paradigmatic, syntagmatic and/or metrical similarity to the word in question.

29 Mikolov 2013.

Figure 13. The beginning of the lexicon entry for the lemma *būras*. Clicking on a verse ID listed under a form's occurrences will open its location in the Reading View

appearing in the corpus are given, each with its morphological properties and variant spellings. A link allows the user to directly jump to a form's location in the corpus.

5. ANALYSES AND VISUALISATIONS. The site additionally offers some experimental analyses and visualisations of the data, two of which warrant special mention.

Firstly, the texts were analysed for redundancy/text re-use using the Recurrence Analysis Function (ReAF) developed by Hoenen (2018). The main goal of this visualisation is to use textual redundancy to characterise the mode of text genesis as either *born-oral* (bardic) or *born-written*. The results for Donelaitis's texts are unsurprising, in that the ReAF analysis aligns with the expectation that they are *born-written*.[30]

[30] Cf. https://titus.fkidg1.uni-frankfurt.de/ cordon/menu/eng/visualisations.html

Figure 14. ReAF table for *DMN RG*. Each field corresponds to one verse.
The digit in the field indicates the number of words in that verse
which are repeated in another verse

Figure 15. ReAF table for *DMN RG*, highlighting the repeated line
"Àlĕ nĕdýwўkĭtės, kalbàs girdėdămĭ tókias", *RG* verses 257 and 662

_	_	PL_155_1	PL_155_2	PL_155_3	PL_155_4	PL_155_5	PL_155_6	PL_155_7	
_	_	Taip	ſûmîſʒay	bĕſîjŭkiant	,	ſʒtai	Uʒimm's	pãſîkĕlė	ExVe
LF_155_1	Als			x					
LF_155_2	mit		x						
LF_155_3	einander		x						
LF_155_4	die								
LF_155_5	Vögel								
LF_155_6	so	x		x					
LF_155_7	scherzten			x					
LF_155_8	,				x				
LF_155_9	erhob							x	
LF_155_10	ein						x		
LF_155_11	Geräusch						x		
LF_155_12	sich							x	
LF_155_13	,								
_	ExVe								

Figure 16. The alignment table for PL vrse 155

Secondly, a method was developed for the (manual) systematic token-level align-ment of the translated German text with its Lithuanian source. Each sentence (or, in our case, verse) is paired up in a table with its translation – the original text on the line and the translation in the column (fig. 16). The object of the alignment is to mark direct correspondences between the original and the translation: correspondences are marked with an "x" in the relevant field of the table (a "y" marks indivisible phrases and idioms). The alignment system covers not only cases in which one Lithuanian token corresponds to one German token (1 : 1 correspondence), but also cases in which: multiple German tokens correspond to one Lithuanian token (1 : n); multiple Lithu-anian tokens correspond to one German token (n : 1); a Lithuanian token has no equivalent in the German text (1 : 0); a German token has no equivalent in the Lith-uanian text (0 : 1); or entire phrases correspond to each other (n : m). An additional field at the end of each line and column (labelled *ExVe*, "external verse", in fig. 16) is used to enter the ID of aligned tokens that are found in another verse.

An experimental alignment of PL with its German translation from DMN 1869 was created. On the basis of this data, an *alignment dictionary* was generated,[31] which is sorted by Lithuanian lemmata and their word forms, and lists their corresponding translations (fig. 17). Two visualisations of the alignments have also been generated: a matrix with colour-coded fields (fig. 18), and a parallel vertical view of the two lan-guage versions, with aligned tokens connected by lines crossing the centre. Creating

31 The alignment tables for the remaining texts have been created, but remain to be processed.

Figure 17. The alignment dictionary entry for *juoktis*

CorDon – A Deeply Annotated
Digital Corpus of the Works
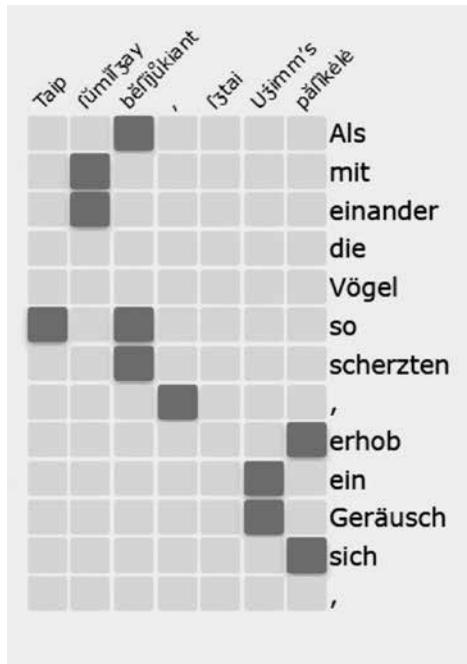of Kristijonas Donelaitis

Figure 18. The matrix visualization
of *PL* verse 155's alignment

a German equivalent would additionally require a manually proofed lemmatisation, word class assignment, and morphological analysis; this data could then serve as a basis for analyses of the translation, like a contrastive word class analysis. This would enable us to address questions such as: *Which word classes are typically translated by another? How often and in what contexts did the translator resort to a word class transition?* etc. This will need to be undertaken as part of a planned follow-up project making use of the alignment data for the remaining texts as well as the data of the lexicon.

6. A SCHOLARLY DIGITAL EDITION? A somewhat controversial question that arises when describing a digital corpus like CorDon is whether it should be considered an *edition* or "merely" an *archive* – in particular, whether the work qualifies as *critical*, representing "activities that apply scholarly knowledge and reasoning to the process of reproducing documents and transforming a document or text into an edition".[32]

It is clear in any case that CorDon is not a prototypical edition. The stemma being what it is – comparison with the extant manuscript giving no indication that variants from the other historical editions will get us closer to the text of the lost archetype

32 Sahle 2016, 24.

than *DMN/DPN/DPPN* 1869 alone does – critical editorial changes to the texts were confined to noting and correcting the occasional misprint. Where the annotated text is based on the surviving manuscript, meanwhile, it follows the SLIEKKAS text.

That said, CorDon ought to satisfy the four criteria Sahle lays out for inclusion in his catalogue of scholarly digital editions:[33]

1. Is there a full representation of the subject in question? This may be an edited text or at least a very accurate transcription. Sometimes, although this is rather an exception than the rule and depends on the specific characteristics of the material, a digital facsimile may suffice. Sometimes even a structured database can be a complete representation.

2. Is it critical? Have rules for the processing of the material been stated and substantiated? Have these rules been applied in the light of the relevant scholarly knowledge on the material, its genesis, its contexts and its reception? Does the edition add information to the representation making it more accessible, understandable and usable?

3. Is the edition of academic quality? Have the rules been applied rigorously and in a transparent manner? Are the responsibilities stated clearly? Does the edition suffice as a substitute for the previous editions or primary documents making it unnecessary to go back to them in most cases? Does it enable further scholarly research on a reliable and trustworthy basis?

4. Does the edition follow a digital paradigm? Does it make use of the possibilities of digital technology and media? Is it not printable without a major loss of content and functionality?[34]

Criteria 1 and 3 would appear to be clearly satisfied.[35] The same applies to criterion 4, in that the ability to arrange the Reading View according to the user's needs is central to the concept of CorDon, as is the in-depth search functionality. Finally, criterion 2's question, "Does the edition add information to the representation making it more accessible, understandable and usable?" can be answered in the affirmative, in spite of the absence of prototypical Lachmannian textual criticism:[36] The facsimile and parallel texts, combined with the work of "explaining what is not evident to the present-day reader", of "bridg[ing] a distance in time, a historical difference"[37] represented by the modernised word-forms, parallelised translations and

33 https://www.digitale-edition.de/exist/apps/editions-browser/index.html
34 Sahle 2016, 38.
35 Though note the limitations imposed on the presentation of transposed and added lines in the parallel texts by their verse-level alignment with the "base" text (cf. section 3.1) regarding criterum 3's question of whether "the edition suffice[s] as a substitute for the previous editions or primary documents making it unnecessary to go back to them in most cases[.]" (Sahle 2016, 38).
36 As previously argued at AIUCD 2021: https://titus.fkidg1.uni-frankfurt.de/cordon/docs/posterAIUCD2021Pisa.pdf (poster), https://titus.fkidg1.uni-frankfurt.de/cordon/docs/AIUCD2021_BOA-versione1_0.pdf (Book of Abstracts, 459f).
37 Sahle 2016, 26.

glosses of "difficult" words, fulfil the purpose of an edition generally, and Sahle's criterion in particular. CorDon thus combines the paradigms, so to speak, by serving as a corpus archive for linguists, and as an edition for readers.

7. CONCLUSION. CorDon makes accessible texts and associated data which were previously digitally available (if at all) in unwieldy text archives, using an annotation scheme and representation which have been successfully applied to prior historical corpora of Lithuanian and German, while improving accessibility with a convenient reading view and novel search GUI. As a reliably annotated corpus of a cohesive "slice" of Old Lithuanian, it represents a step in the creation of an empirical basis for a historical dictionary of Lithuanian and a historical grammar of Old Lithuanian, and may serve as a blueprint for future corpora in the same vein.

CorDon is intended as a versatile tool for a broad range of users. It can fill the role: of a digital edition for Donelaitis scholars, making the text of the manuscript and the oldest print editions available for easy reference and comparison; of a deeply linguistically annotated corpus for researchers of Old Lithuanian and comparative Indo-European linguistics; but also of an accessible way to read Donelaitis for a modern audience of native speakers, teachers and learners of Lithuanian, helping to bridge the 250-year linguistic and cultural gap. By combining these roles, it encourages users themselves to move between them, and broaden their interaction with the text and data. In this way, CorDon can serve as a common access point to the works of Kristijonas Donelaitis.

Literature and sources:

DONELAITIS, KRISTIJONAS, 1977: *Kristijonas Donelaitis*, *Raštai*, redakcinė komisija: K[ostas] Korsakas, K[ostas] Doveika, L[eonas] Gineitis, J[onas] Kabelka, K[azys] Ulvydas, Vilnius: Vaga.

DM PL.WD.F 1765–1775 – Kristijonas Donelaitis, *Metai* ("Pavasario linksmybės", "Vasaros darbai", "Fortsetzung"); LLTIBR: F 1 – 5259.

DMC 2014, *DPC* 2014 – *Kristijonas Donelaitis. Le Stagioni*, Traduzione e cura di Adriano Cerri, Novi Ligure: Joker.

DMN 1869, *DPN* 1869, *DPPN* 1869 – *Christian Donalitius*, *Littauische Dichtungen* nach den Königsberger Handschriften mit metrischer Uebersetzung, kritischen Anmerkungen und genauem Glossar herausgegeben von G[eorg] H[einrich] F[erdinand] Nesselmann, Königsberg: Hübner & Matz.

DMR 1967 – *The Seasons by Kristijonas Donelaitis*, Rendered from Lithuanian into English verse by Nadas Rastenis, Introduction and editing by Elena Tumas, Ph.D. Los Angeles: Lithuanian Day Publishers.

DMRh 1818 – *Das Jahr in vier Gesängen, ein ländliches Epos aus dem Litthauischen des Christian Donaleitis, genannt Donalitius*, In gleichem Versmaaß ins Deutsche übertragen von D. L[udwig] J[edemin] Rhesa, Prof. d. Theol., Königsberg: Königl. Hartungschen Hofbuchdrukkerei.

DMSch 1865, *DPSch* 1865, *DPPSch* 1865 – *Christian Donaleitis, Litauische Dichtungen*, Erste volständige Ausgabe mit Glossar von Aug[ust] Schleicher, St. Petersburg: Kaiserliche Akademie der Wissenschaften.

DPRh 1824 – *Aisópas arba Pásakos iß Grykonû kalbôs* perguldytos per D. L[udwig] J[edemin] Rhesa. Su Pridéjimù kellû naujû Pásakû, Karaláucʒuje: Ißspaustas ir pardůdamas pas Augustą Artungą.

Gelumbeckaitė, Jolanta, 2015: "Kristijonas Donelaitis im Referenzcorpus Altlitauisch (SLIEKKAS)", *Baltu filoloģija* 24(1), 19–50.

Hoenen, Armin, 2018: "Recurrence Analysis Function, a Dynamic Heatmap for the Visualization of Verse Text and Beyond", *Noah Bubenhofer, Marc Kupietz (eds.), Visualisierung sprachlicher Daten*, Heidelberg University Publishing.

Kabelka, Jonas, 1964: *Kristijono Donelaičio raštų leksika*, Vilnius: Mintis.

Krištopaitienė, Daiva, 2007: *Kristijono Donelaičio raštų leidimai: tekstologinės problemos*, Vilnius: Lietuvių literatūros ir tautosakos institutas.

Krištopaitienė, Daiva, 2014: "Pirmasis *Metų* leidimas: Redakcinės pataisos", *Liucija Citavičiūtė (ed.), Martynas Liudvikas Rėza, Raštai 4. Kristijono Donelaičio kūrybos publikavimas*, Vilnius: Lietuvių literatūros ir tautosakos institutas, 53–59.

Krištopaitienė, Daiva; Mikas Vaicekauskas, 2014: "Die Handschriften von Kristijonas Donelaitis: Beschreibung der Quelle", *Mikas Vaicekauskas (ed.), Kristijono Donelaičio rankraščiai*, fotografuotinis leidimas, Vilnius: Lietuvių literatūros ir tautosakos institutas.

Mikolov, Tomas et al., 2013: *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781).

Mittmann, Roland; Ralf Plate, 2019: "Das 'Referenzkorpus Altdeutsch' als Lesekorpus. Grammatisch annotierte und mit Wörterbüchern verknüpfte Texte für Lehre und Selbststudium", *Das Mittelalter* 24(1), Heidelberg University Publishing, 173–187.

Sahle, Patrick, 2016: "What Is a Scholarly Digital Edition?", *Matthew James Driscoll, Elena Pierazzo (eds.), Digital Scholarly Editing: Theories and Practices*, Cambridge: Open Book Publishers, 19–39.

Mortimer Drach
CORDON – KRISTIJONO DONELAIČIO TEKSTŲ
SKAITMENINIS GILIOSIOS ANOTACIJOS KORPUSAS

Santrauka

CorDon (https://titus.fkidg1.uni-frankfurt.de/cordon/menu/lt/startas.html) yra laisvai prieinamas leksiškai ir morfosintaksiškai anotuotas lietuvių literatūros klasiko Kristijono Donelaičio (1714–1780) lietuviškų tekstų korpusas. Anotuoti tekstai pateikiami paslankaus formato skaityklėje, kurioje galima aktyvuoti atskirus anotacijos sluoksnius, paralelius teksto leidimus ir vertimus į vokiečių, anglų ir italų kalbas. Paieškos grafinė naudotojo sąsaja įgalina paiešką visuose anotacijų sluoksniuose nenaudojant specialios užklausų kal-

bos ar reguliariųjų išraiškų. Šis paieškos metodas yra nepriklausomas nuo anotacijos sche-
mos ir gali būti apibendrintas panašiu būdu anotuotiems tekstams. Anotacijos pagrindu
sugeneruotame leksikone su konkordancija pagal antraštinį žodį (lemą) suregistruoti Do-
nelaičio tekstų žodžiai. Intuityvi CorDon paieška sukonstruota taip, kad užklausą būtų
galima palaipsniui tikslinti ir plėsti. Paieškos rezultatai pateikiami plačiam vartotojų ratui
prieinama forma. CorDon gali pasitarnauti kaip skaitmeninis leidinys Donelaičio tekstų
tyrėjoms bei tyrėjams, kaip anotuotas korpuso archyvas kalbininkėms bei kalbininkams
arba kaip skaitymo korpusas mokymo(si) procese.

Mortimer Drach
Institut für Empirische Sprachwissenschaft
Senckenberganlage 31
D-60325 Frankfurt am Main
E-Mail: drach@posteo.de