AIUCD 2021

DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale

10° congresso annuale PISA 19-22 gennaio

DIGITAL PUBLIC HUMANITIES

OPEN CULTURE

RETI SOCIALI

TECH ECONOMY

E-PARTICIPATION

TECNOLOGIE ASSISTIVE

Versione PROVVISORIA del contributo presentato al Convegno Annuale

DISCLAIMER

Questa versione dell'abstract non è da considerarsi definitiva e viene pubblicata esclusivamente per facilitare la partecipazione del pubblico al convegno AIUCD 2021

Il Book of Abstract contenente le versioni definitive e dotato di ISBN sarà disponibile liberamente a partire dal 19 gennaio sul sito del convegno sotto licenza creative commons.

Both Digital Edition and Corpus Archive: The works of Kristijonas Donelaitis

Philipp Büch[1], Mortimer Drach[1], Jolanta Gelumbeckaitė[1], Armin Hoenen[1], Adriano Cerri[2], Vanessa Wadlinger[1]

[1] Emp. Ling. Inst., GU Frankfurt, Germany {buech, drach, gelumbeckaite, hoenen}(«»)em.uni-frankfurt.de
[2] Università di Pisa, Dipartimento di Filologia, Letteratura e Linguistica, Italy

ABSTRACT

In this paper and on the poster we present the CorDon project and its web-presence,[1] see http://titus.fkidg1.uni-frankfurt.de/cordon/docs/posterAIUCD2021Pisa.pdf .

KEY WORDs

Donelaitis, Online Edition, Corpus, Old Lithuanian

1. INTRODUCTION

Subject of the project are the texts in metrotonic hexameters of the Lithuanian poet Kristijonas Donelaitis (*1714 - †1780), whose works have been linguistically deeply annotated in this and previous projects. The works consist of 4 texts from the cycle *Metai* (The Seasons), 2 associated fragments, and 6 fables (*Pasakos*). These had to be checked, analyzed, and made available on the web.

2. ANNOTATIONS, EDITIONS AND TRANSLATIONS

Donelaitis' texts, comprising around 30,000 tokens with around 9,500 types, are the most extensive autochtonous Old Lithuanian texts. They have been annotated with approx. 330,000 annotations on 11 annotation layers*: transliteration, modernized form, lemma, accented lemma, language, part of speech lemma, part of speech form, morphology lemma, morphology form, inflection, gloss*. In addition to the representation of these base texts and their annotations, the project also provides two other early parallel Lithuanian text editions, two early translations into German, and two modern translations into Italian and English.

3. FUNCTONALITIES, EXPECTATIONS, USERS

The website is primarily designed to serve the needs of researchers and teachers of (Old) Lithuanian, linguists and Indo-Europeanists as well as interested laypersons. The texts were digitized according to the principles of diplomatic literal transcription. The annotation was carried out using the program *Toolbox*[2], and further revised in the annotation software ELAN[3] [1] which then exports the data to a table-based HTML structure, easily customizable through some JavaScript functionality so as to dynamically toggle any annotation layer. This potentially generic way to create a web corpus had been provided to us by P. Hinkelmanns (Salzburg) who developed the concept for the *Lesekorpus Altdeutsch* (Old High German Reading Corpus) [2].[4] Within CorDon, we included a facsimile (as [3, p.27] discusses, this is expected for digital editions) and the Lithuanian editions and/or all translations (parallel text edition). Using a database connected to the web interface, we provide a comprehensive search function with a PDF export so as to search for specific occurrences or patterns within the texts. The search mask is GUI-based and has 2 modes: basic and extended, the latter allowing the specification of any annotation configuration for up to 5 tokens in a row. The same data ordered in a different way is presented as a lexicon/concordance and augmented by additional information such as each lemma's neighbors in lemma vector space. A subset of this data is online as RDF using OntoLex-Lemon. Finally, we provide some visualizations such as the ubiquitous word cloud as well as some more analytical types.

4. ACKNOWLEDGEMENTS

We would like to thank the Fritz Thyssen Stiftung for their financial support.

[1] http://titus.fkidg1.uni-frankfurt.de/cordon/start.html

[2] https://software.sil.org/toolbox

[3] https://archive.mpi.nl/tla/elan

[4] http://titus.uni-frankfurt.de/lea

BIBLIOGRAPHY

[1] Brugman, H., and A. Russel (2004). Annotating Multimedia/ Multi-modal resources with ELAN. *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.*

[2] Mittmann, R., and R. Plate (2014). "Das ‚Referenzkorpus Altdeutsch‘ als Lesekorpus. Grammatisch annotierte und mit Wörterbüchern verknüpfte Texte für Lehre und Selbststudium". Das Mittelalter 24(1) 2019: 173–187.

[3] Sahle, P. (2016). "What is a Scholarly Digital Edition?" In: Driscoll, M. J., and E. Pierazzo. *Digital Scholarly Editing*. OpenBookPublishers. pp. 19-41.