

Building a database of academic text: challenges and solutions

Djuddah Leijen, Helen Hint, Helena Lemendik, Marleen Kirsipuu



UNIVERSITY
OF TARTU

BWRITE

Academic Writing in the Baltic States:

Rhetorical Structures through
culture(s) and languages



UNIVERSITY
OF TARTU



TEXTA

Background

Writing in the Baltic countries

- Determine the writing tradition(s) in the three Baltic countries:
 - Estonian
 - Latvian
 - Lithuanian
- Measure how local writing traditions differ from English-centered Anglo-American tradition
 - the assumption that the way academic texts are written is culture-dependent

The Bwrite project

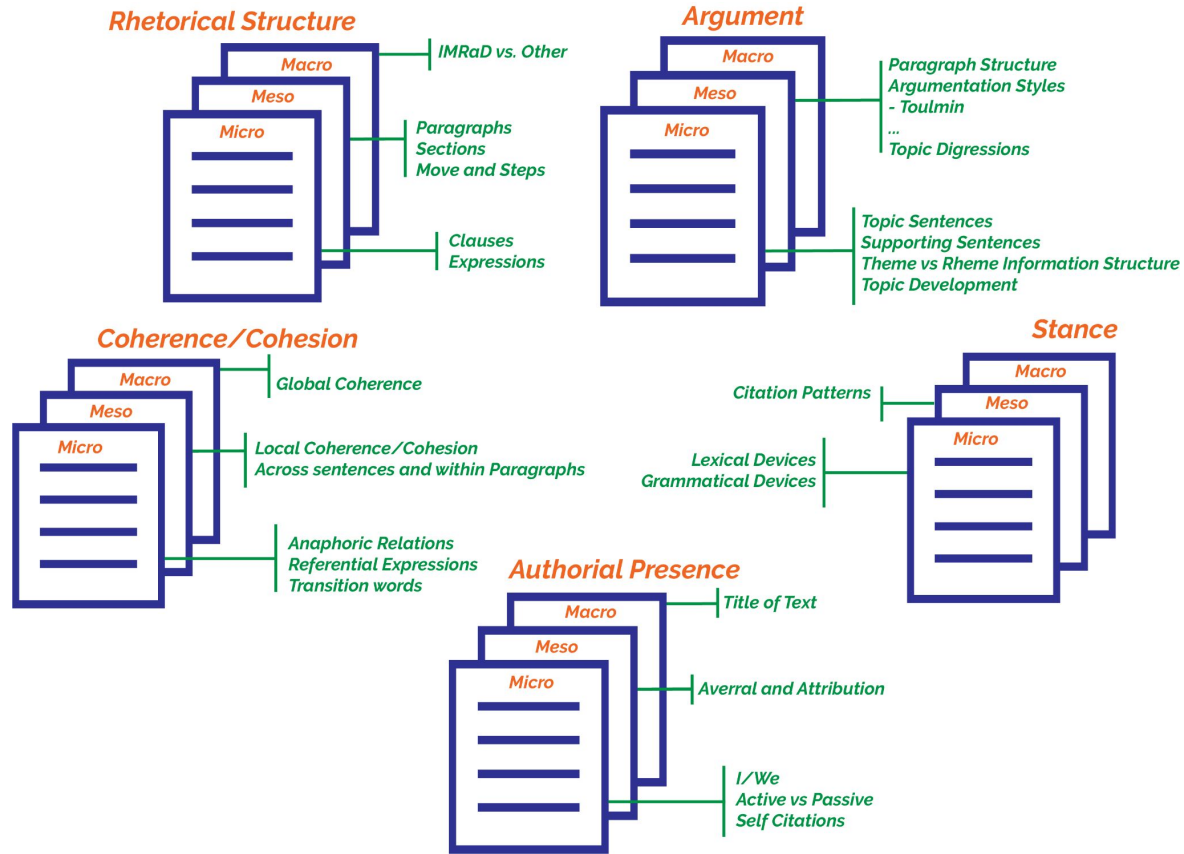
Our aim is to:

- **develop a research method** that allows one to determine which **features of a text are** indicators related to
- **genre, discipline, culture and experience.**
- Furthermore, we aim to provide **strong empirical results**
- that will allow **writers and instructors** of writing to **better apply** those specific text features for teaching and writing.

Current aims

- Create a representative and meaningful structured **corpus of academic texts**.
- **Develop a model** which can be used to measure specific features of texts which could help us describe how these occur or vary across different types of texts and/or languages.
- Apply machine learning methods on the created dataset to train and **test each feature** from the model **individually**, and the features in **combination**.

Five-feature model (~writing tradition)


















The corpus: what it contains

Texts

- All academic texts freely accessible in the internet
- Web scraping
 - From various universities
 - From various disciplines:
 - humanities
 - social sciences
 - other (less represented - lack of journals / opportunities to publish in local languages)

Result: tons of PDFs in Nextcloud

<input type="checkbox"/>	Name ▲			Size
<input type="checkbox"/>	 DSpace files			22.7 GB
<input type="checkbox"/>	 Journal articles			9.2 GB
<input type="checkbox"/>	 Proceedings			269.4 MB
<input type="checkbox"/>	 Student work			77.5 GB
<input type="checkbox"/>	 Yearbooks			Pending

Data overview: document counts

Category	Estonian	Latvian	Lithuanian*	Total sum
Number of all documents in corpus	34,928.00	6,409.00	33,747.00	75,084.00
Number of all student work documents	27,833.00	3,147.00	0.00	30,980.00
Number of BA theses	16,295.00	153.00	0.00	16,448.00
Number of MA theses	9,420.00	2.00	0.00	9,422.00
Number of PhD theses	1,246.00	2,992.00	0.00	4,238.00
Number of other/general student work documents	872.00	0.00	0.00	872.00
Number of journal articles	6,227.00	2,711.00	33,747.00	42,685.00
Number of yearbooks	734.00	0.00	0.00	734.00
Number of proceedings	134.00	551.00	0.00	685.00

** Lithuanian: we are currently readjusting the data, all theses got cleaned away accidentally.*

Data overview: document lengths

	Length of documents			Amount of sentences in document			Amount of words		
All docs	<i>Estonian</i>	<i>Latvian</i>	<i>Lithuanian</i>	<i>Estonian</i>	<i>Latvian</i>	<i>Lithuanian</i>	<i>Estonian</i>	<i>Latvian</i>	<i>Lithuanian</i>
<i>min</i>	2,002	2,047	7	6	7	1	168	320	2
<i>max</i>	2,127,643	2,220,928	3,555,962	17,053	21,750	25,391	323,226	410,569	641,415
<i>avg</i>	95,809	190,159	37,392	804	1,300	271	14,714	32,821	6,548
<i>sum</i>	3,346,417,026	1,218,727,953	1,261,851,932	28,089,940	8,331,330	9,155,446	513,939,884	210,348,172	220,969,256

Technical process



Download PDF documents



Upload JSON documents

BWrite Data Processing Pipeline

Step 1.



Raw Data

Download PDF
documents



Parses PDFs into JSON
documents & adds file system
information.

Texta DocParser



Stanza

Adds linguistic information:
sentence boundaries, lemmas
& POS-tags.

Texta Multilingual
Processor



Splits documents into
sentences and adds metadata
from file system information.

Post-processing Script

Upload JSON
documents

TEXTA TOOLKIT

Analysis Environment

BWrite Data Processing Pipeline

Nextcloud: the beginning

1. Using Python to scrape available data from the digital repositories of universities
 - a. BeautifulSoup library
2. Uploading the downloaded files to cloud
 - a. Creating a folder structure of downloaded files

Lessons learned:

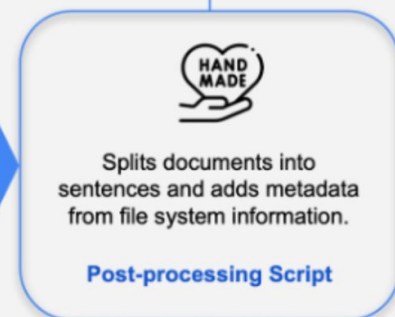
- Gather easily accessible metadata
 - a. Year, university, institution, tags



Download PDF documents



Upload JSON documents



Step 2.

BWrite Data Processing Pipeline

Making PDF-s machine readable (1)

1. Download PDF-s from the cloud;
2. Parse documents from PDF to text
 - a. PyMuPDF fitz module
3. Process texts with Texta Toolkit's MLP (MultiLingual Preprocessor) module:
 - a. Lemmatize texts;
 - b. Detect language;
 - c. Add POS tags;
 - d. Add representation of scripted languages.
4. Submit computations to the UT High Performance Computing Center (HPC Center) cluster.

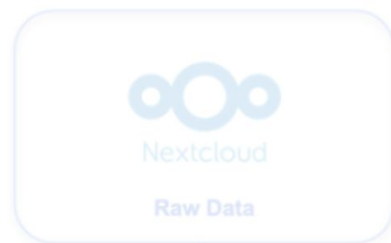
Making PDF-s machine readable (2)

Input:

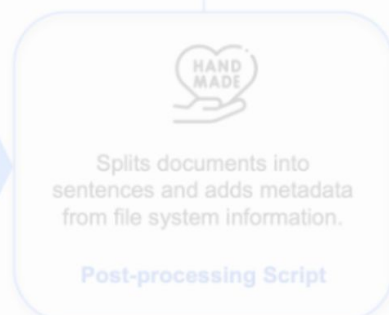
```
{  
  "texts": ["Mis su nimi on?"]  
}
```

Output:

```
[  
  {  
    "text": {"text": "Mis su nimi on ?", "lang": "et", "lemmas": "mis sina nimi olema ?", "pos_tags": "P P S V Z"},  
    "texta_facts": []  
  }  
]
```



Download PDF documents



Step 3.



Upload JSON documents

BWrite Data Processing Pipeline

Uploading texts to Texta Toolkit

Texta Toolkit - No-Code Machine Learning & NLP Platform,
Research & Data Automation

What it does?

- Data preparation
- Training Text Classifiers
- Training Entity Extractors
- Live data analysing

Benefits:

- Proven track record
- All NLP tools in one software
- Graphical user interface
- Open source (for now...)

Cleaning the data

Projects

Search

Lexicons (0)

Models (0)

Tools (5)

Indices
texta-1-import-project-3

Projects
3. Bwrite Raw Dat...

marleen

Search
Constraints

Simple search

Advanced search

Select Fields

text;

Operator Match Stop

and phrase 0

Type here

Searcher options

Highlight facts: ☒

Only highlight matching facts: ☐

Highlight searcher matches: ☒

Show short version: ☐

Search

Saved Queries

☐ Search description Edit

☐ sent is bigger than 5

☐ slides_summaries

☐ not proceedings

☐ too_long_docs_et_long_docs

Toggle Columns
properties.extension, st...

Toggle drawer

Export results

Items per page: 10 1 - 10 of 648/38978

properties.extension	stats.len_chars	stats.len_sents	stats.len_words	text ↑	text_mlp.language.analysis	text_mlp.lemmas	texta_facts
.html	307	7	58	Nextcloud This application requires JavaScript for correct operation. Please enable JavaScript and reload the page. Nextcloud File not found The document could not be found on the server. Maybe the share was deleted or has expired? Back to Nextcloud Nextcloud – a safe home for all your data	en	Nextcloud this application require javascript for correct operation . please enable javascript and reload the page . nextcloud file not find the document could not be find on the server . maybe the share be delete or have expire ? back to Nextcloud Nextcloud - a safe home for all you datum	Genre BA Thesis University Eesti Lennuakadeemia
.html	307	7	58	Nextcloud This application requires JavaScript for correct operation. Please enable JavaScript and reload the page. Nextcloud File not found The document could not be found on the server. Maybe the share was deleted or has expired? Back to Nextcloud Nextcloud – a safe home for all your data	en	Nextcloud this application require javascript for correct operation . please enable javascript and reload the page . nextcloud file not find the document could not be find on the server . maybe the share be delete or have expire ? back to Nextcloud Nextcloud - a safe home for all you datum	Genre BA Thesis University Eesti Lennuakadeemia
.html	307	7	58	Nextcloud This application requires JavaScript for correct operation. Please enable JavaScript and reload the page. Nextcloud File not found The document could not be found on the server. Maybe the share was deleted or has expired? Back to Nextcloud Nextcloud – a safe home for all your data	en	Nextcloud this application require javascript for correct operation . please enable javascript and reload the page . nextcloud file not find the document could not be find on the server . maybe the share be delete or have expire ? back to Nextcloud Nextcloud - a safe home for all you datum	Genre BA Thesis University Eesti Lennuakadeemia

Biznesa augstskolas Turība konferenču rakstu krājums

*Radīt nākotni:
komunikācija,
izglītība,
business*

XIV starptautiskā
zinātniskā
konference

 **Turība**
Biznesa augstskola

2011
ISSN

Table of Contents. Saturs

Māris Andžāns. PROSPECTS OF REGIONALIZATION OF SECURITY IN THE CYBERSPACE: CASE OF THE BALTIC STATES.....	14
Introduction.....	14
1. Theoretical Aspects of Regionalization of Security and the Concept of Security	15
2. Dependence on Use of Information Technologies of the Baltic States	16
3. Cyber Threats to the Baltic States	18
Conclusions.....	20
Bibliography	21
Mytko Antonina. PROSPECTS AND CHALLENGES OF INFORMATION DEMOCRACY ..	25
Introduction.....	25
1. Information Globalization in ICT-space	26
2. Information Democracy in Political Processes	27
3. Impact of Information Democracy on the Moral and Ethical Principles of Humanity	28
4. Rapid Growth of Media Products in the Globalized World	29
5. Information Democracy and the Global Economy	30
Conclusions.....	30
Bibliography	31
Dalia Perkūnienė, Danutė Kleiniene. THE IMPACT OF CULTURE CENTER ACTIVITIES ON THE COMMUNITY LEISURE ORGANIZATION	33
Introduction.....	33
1. Strategic Vision of the Cultural Policy in Lithuania	34
2. Results of the Research	35
Conclusions.....	37
Bibliography	38
Sigita Šimbelytė, Giedrius Nemeikis. THE USAGE PROBLEMS OF ELECTRONIC DOCUMENT AS EVIDENCE IN THE CIVIL PROCESS IN LITHUANIA	40
Introduction.....	40
1. Electronic Document as Evidence in the Civil Process in Lithuania	41
2. The Problem of Primary and Secondary Documents	43
3. The Problem of Primary and Secondary Documents: Particular Cases	44
4. Electronic Signature as a Guarantor of Document's Authenticity	45
5. Admissibility and Evaluation of Electronic Information as Evidence	46
Conclusions.....	48
Bibliography	48
Anda Komarovska. THE ACCESSIBILITY OF TOURIST INFORMATION IN FOREIGN LANGUAGES IN RIGA AND LATVIA	51

- 8 -

Adding metadata

1. Adding metadata as texta facts
 - a. Genre (BA thesis, MA thesis, Journal Article...)
 - b. University
 - c. Publication
2. Splitting text to find additional metadata
3. Using Regex to find desired patterns
 - a. Discipline
 - b. Faculty
 - c. Year
 - d. ...

Search Constraints

☐ Simple search ☒ Advanced search

Select Fields

text;

Operator	Match	Slop
and	phrase prefix	0

Type here

Tartu Ülikool



text

TARTU ÜLIKOOL | VIILJANDI KULTUURIAKADEEMIA
Kultuurhariduse osakond
kultuurikorralduse õppekava

TARTU ÜLIKOOL | KARJÄÄRIPÄEV
Loov-praktiline lõputöö
Juhendaja:
Kaitsmisele lubatud:
Viljandi 2017

```
modified_titles = list()

for sentence in sentences:
    for string in sentence:
        match = re.match("\s?\d{1,3}?\d{1,2}?\.\d{1,2}?\.\s?([A-ZÕÄÖÜ][a-zõäöü\s\d]{0,80})", string)
        if match:
            print("match")
            if len(match.group()) > 7:
                modified_titles.append(match.group(1))
```

Challenges

You can't make assumptions

Search Constraints

☐ Simple search ☒ Advanced search

Select Fields

text;

Operator Match Slop
and phrase prefix 0

Type here
Tartu Ülikool




text	
3 UURIMISTÖÖ TULEMUSED	27
3.1 Dokumendialüüsi tulemused	27
3.1.1 Tartu Ülikool	
27	
3.1.2 Tallinna Tehnikaülikool	
30	
3.1.3 Tallinna Ülikool	
33	
3.1.4 Eesti Maaülikool	
36	
3.2 Intervjuude tulemused	39
3.2.1 Tartu Ülikool	
39	
3.2.2 Tallinna Tehnikaülikool	
43	
3.2.3 Tallinna Ülikool	
46	
3.2.4 Eesti Maaülikool	
49	

Challenges

You must know your data

VILNIAUS UNIVERSITETAS
MEDICINOS FAKULTETAS
REABILITACIJOS, FIZINĖS IR SPORTO MEDICINOS KATEDRA

Tvirtinu:
Vilniaus universiteto Medicinos fakulteto
Reabilitacijos, fizinės ir sporto medicinos katedros
studijų programų komiteto pirmininkas
prof. 
Data:



**Mokyklinio amžiaus mergaičių laikysenos ir liemens raumenų funkcijų
sąsajos esant idiopatinei skoliozei**

KINEZITERAPIJOS BAKALAURO BAIGIAMASIS DARBAS

Darbo vadovė: 

Darbo priėmimo data:

Parašas


VILNIUS, 2017



Challenges

You must know your data


VILNIAUS UNIVERSITETAS
MEDICINOS FAKULTETAS
REABILITACIJOS, FIZINĖS IR SPORTO MEDICINOS KATEDRA

Tvirtinu:
Vilniaus universiteto Medicinos fakulteto
Reabilitacijos, fizinės ir sporto medicinos katedros
studijų programų komiteto pirmininkas
prof. 
Data:



**Mokyklinio amžiaus mergaičių laikysenos ir liemens raumenų funkcijų
sąsajos esant idiopatinei skoliozei**

KINEZITERAPIJOS BAKALAURO BAIGIAMASIS DARBAS

Darbo vadovė: 
Darbo priėmimo data:
Parašas

VILNIUS, 2017 

TARTU ÜLIKOOL
Sotsiaalteaduste valdkond
Ühiskonnateaduste instituut
Ühiskonna- ja infoprotsesside analüüsi õppekava

Marleen Kirsipuu

**ENIMKASUTATUD UURIMISEETODID TARTU ÜLIKOOLI SOTSIAAL- JA
HUMANITAARTEADUSLIKES LÕPUTÕODES: LOOMULIKU KEELE
TÕÕTLUSEL PÕHINEV UURIMUS**

Magistritöö


Juhendajad:
Kairi Kasearu, PhD
Raul Sirel, MA


Tartu
2023 

Challenges


You must know your data

VILNIAUS UNIVERSITETAS
MEDICINOS FAKULTETAS
REABILITACIJOS, FIZINĖS IR SPORTO MEDICINOS KATEDRA

Tvirtinu:
Vilniaus universiteto Medicinos fakulteto
Reabilitacijos, fizinės ir sporto medicinos katedros
studijų programų komiteto pirmininkas
prof. 
Data:


Mokyklinio amžiaus mergaičių laikysenos ir liemens raumenų funkcijų
sąsajos esant idiopatinei skoliozei

KINEZITERAPIJOS BAKALAURO BAIGIAMASIS DARBAS

Darbo vadovė: 
Darbo priėmimo data:
Parašas

VILNIUS, 2017 

TARTU ÜLIKOOL
Sotsiaalteaduste valdkond
Ühiskonnateaduste instituut
Ühiskonna- ja infoprotsesside analüüsi õppekava

Marleen Kirsipuu

ENIMKASUTATUD UURIMISMEETODID TARTU ÜLIKOOLI SOTSIAAL- JA
HUMANITAARTEADUSLIKES LÕPÜTÕODES: LOOMULIKU KEELE
TÕÕTLUSEL PÕHINEV UURIMUS

Magistritöö

Juhendajad:
Kairi Kasearu, PhD
Raul Sirel, MA

Tartu
2023 

128

K. UIBU^{a1}, M. MÄNNAMAA

Teaching practices and text comprehension in students during the transition from the first to second stage of school

Krista Uibu^{a1}, Mairi Männamaa^{bc}
^a University of Tartu, Institute of Educational Science
^b Tallinn University, Institute of Psychology
^c Children's Clinic of Tartu University Hospital

Summary

Introduction

Children acquire their elementary reading and text comprehension skills at primary school. Good reading skills is not only essential in the context of language sub-skills, but it is the basis of academic success in all subjects (Cain & Oakhill, 2007). According to the OECD results in PISA 2009 (Programme for International Student Assessment), Estonian students came 10th among OECD countries in reading comprehension skills and even higher, fifth, among European countries (Tire et al., 2010).

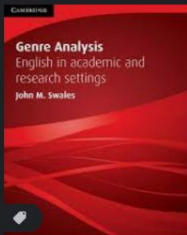
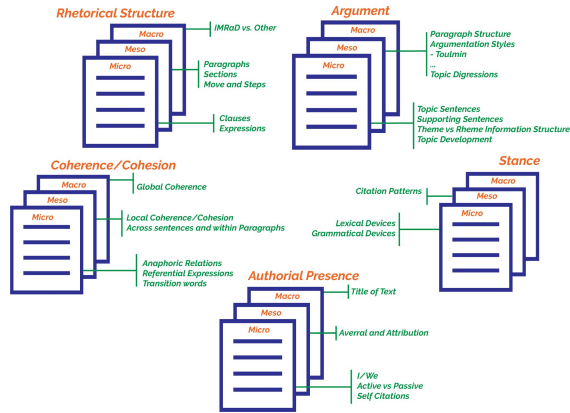
Despite these high results in international studies, there are still students who have difficulties with age-appropriate text comprehension (Henno et al., 2007; Soodla & Kikas, 2010; Tire et al., 2010). Some children who struggled with text comprehension had satisfactory or even good reading skills (Lervåg & Aukrust, 2010). Among poor performers there are children with very different abilities (Cain & Oakhill, 2006), and their prevalence is higher in the case of low abilities (Reynolds & Turek, 2012). In addition to varying abilities, differences in text comprehension have also been found between boys and girls, mostly in favour of girls (Logan & Johnston, 2010; Tire et al., 2010).

There is no common agreement on which abilities and skills are most essential in text comprehension. Verbal skill has been considered a good indicator of text comprehension (Berninger et al., 2006; Echols et al., 1996; Pečjak et al., 2011). Vocabulary and previous knowledge play an important role in text comprehension (Broek & Espin, 2012; McKeown & Beck, 2004). What makes a text easier to comprehend is some knowledge of semantics, syntactic and grammatical constructions (Cain & Oakhill,

¹ Institute of Education, Faculty of Social Sciences and Education, University of Tartu, Salmela 1a, 50103 Tartu, Estonia; krista.uibu@ut.ee



What follows?



Amazon.com: Genre A...
amazon.com · In stock



John SWALES | Professor e...
researchgate.net



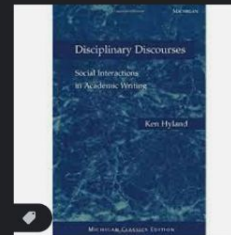
Social Interactions in Academic Writ...
amazon.com · In stock



routeedge.com · In stock · press.uitm.edu



Intro...



Social Interactions in Acade...
amazon.com · In stock



essential bookshelf: Aca...
cambridge.org



Research Genres: Expl...
amazon.com · In stock



<https://www.bwrite.ut.ee>