Iceland Liechtenstein Norway grants

BWRITE

Determining Stance in Estonian, Latvian, and Lithuanian Academic Writing. Implications and Directions

Helen Hint, Helena Lemendik, Anna Ruskan, Baiba Egle, Christer Johansson, Nicholas Groom. Chaired by Djuddah Leijen

1632 UNIVERSITY OF TARTU

LIEPĀJAS UNIVERSITĀTE

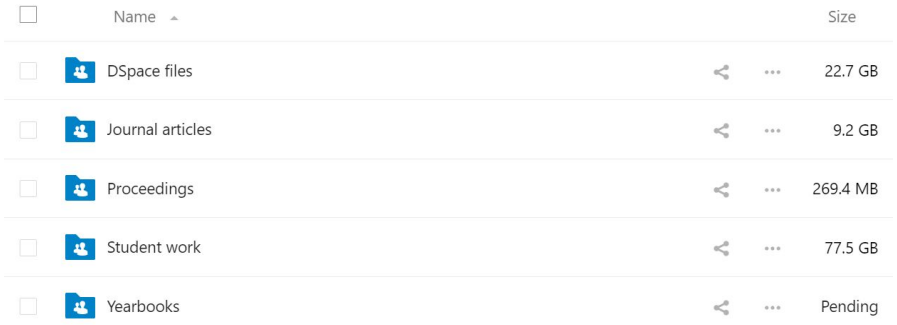
UNIVERSITAS BERGENSIS

VILNIAUS UNIVERSITETAS ·1579· UNIVERSITAS VILNENSIS

Introduction

Example of Estonian academic text

* How is writing modeled?
* What is a writing tradition? How can the features selected into the model help to capture the tradition?
* Focus on writing in other languages than in English



HELEN HINT, DJUDDAH A. J. LEIJEN, ANNI JÜRINE

EESTIKEELSE AKADEEMILISE TEKSTI TUNNUSTEST

https://doi.org/10.54013/kk772a3

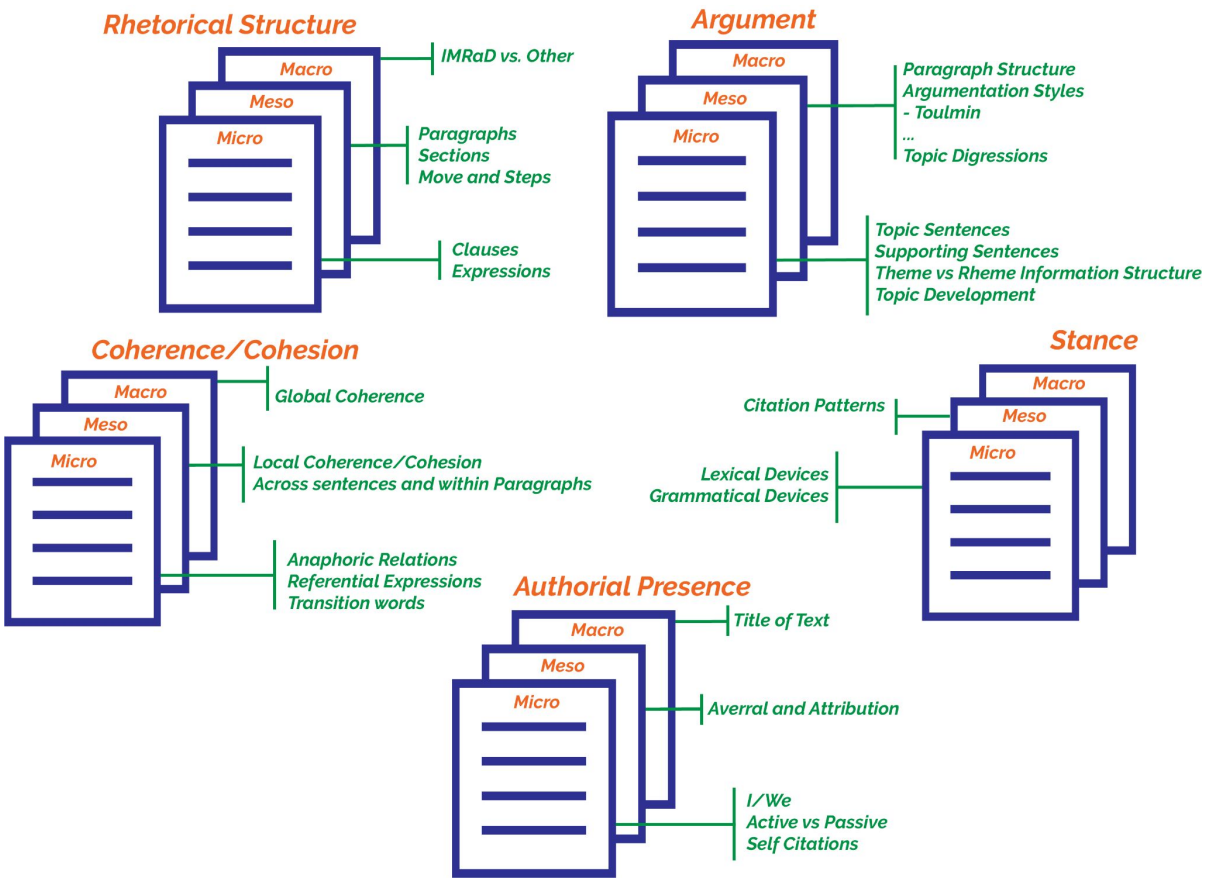
Artikkel on pühendatud Annile, säravale teadlasele ja unustamatule kolleegile

Akadeemilised tekstid on enamasti kõrgkoolides ja teadusasutustes kirjutatavad lühemad ja pikemad argumenteerivad kirjutised, mille eesmärk on (uute) teadustulemuste dokumenteerimine, levitamine ja nende üle arutlemine. Akadeemiliste tekstide hulka kuuluvad nii teadlaste teadustekstid kui ka üliõpilaste õppetöö eesmärgil kirjutatud tekstid, olenemata sellest, kas need on mõeldud trükis avaldamiseks või jäävad käsikirjadeks (Jürine jt 2014). Akadeemilised tekstid moodustavad omaette tekstiliigi ehk žanri, mida defineerime kui kindla eesmärgiga sotsiaalsete suhtlussündmuste kogumit, millele on omane kindlaks kujunenud dünaamiline suhe autori ning vastuvõtja vahel (Swales 1990: 40; Donahue 2008: 333; vt ka Kasik 2005: 8). Nagu igal tekstiliigil, on ka akadeemilisel tekstil kommunikatiivsest funktsioonist ja kultuuriruumist, aga ka keelest ja selle struktuurist sõltuvad keelelised tunnused, näiteks teatud süntaktilised mustrid, struktuur või registri markerid (Hyland 2004; Connor jt 2008; vt ka Rheindorf, Wodak 2019).

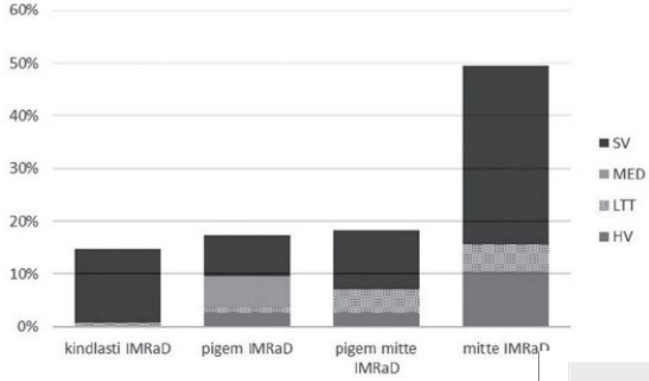
- Database of Baltic academic texts
- Compiled by web scraping

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

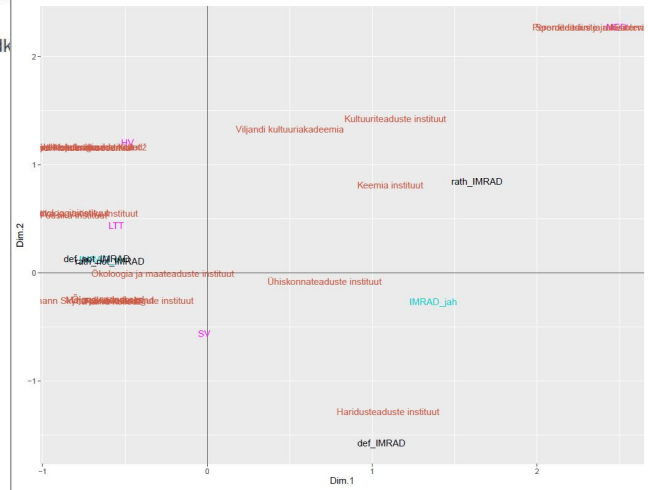
BWRITE



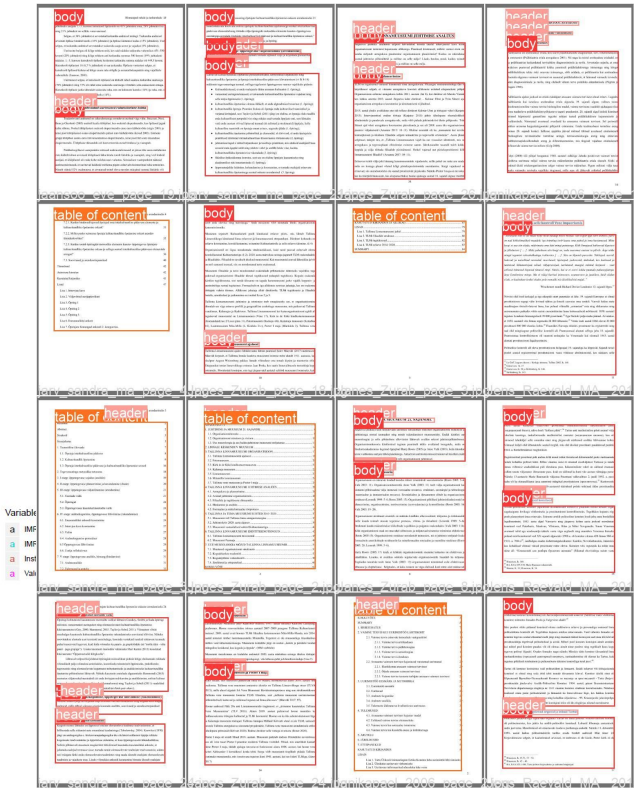
Feature 1. Rhetorical Structure: IMRaD



Joonis. Magistritööde struktuuritüübid (protsentides) teadusvaldk



YOLO (v3)[1]



Feature 2. Stance & metadiscourse

Table 3.1 *An Interpersonal model of metadiscourse*

Category	Function	Examples
Interactive	**Help to guide the reader through the text**	**Resources**
Transitions	express relations between main clauses	in addition; but; thus; and
Frame markers	refer to discourse acts, sequences or stages	finally; to conclude; my purpose is
Endophoric markers	refer to information in other parts of the text	noted above; see Fig; in section 2
Evidentials	refer to information from other texts	according to X; Z states
Code glosses	elaborate propositional meanings	namely; e.g.; such as; in other words
Interactional	**Involve the reader in the text**	**Resources**
Hedges	withhold commitment and open dialogue	might; perhaps; pos
Boosters	emphasize certainty or close dialogue	in fact; definitely; it
Attitude markers	express writer's attitude to proposition	unfortunately; I agre
Self mentions	explicit reference to author(s)	I; we; my; me; our
Engagement markers	explicitly build relationship with reader	consider; note; you

Hyland 2005

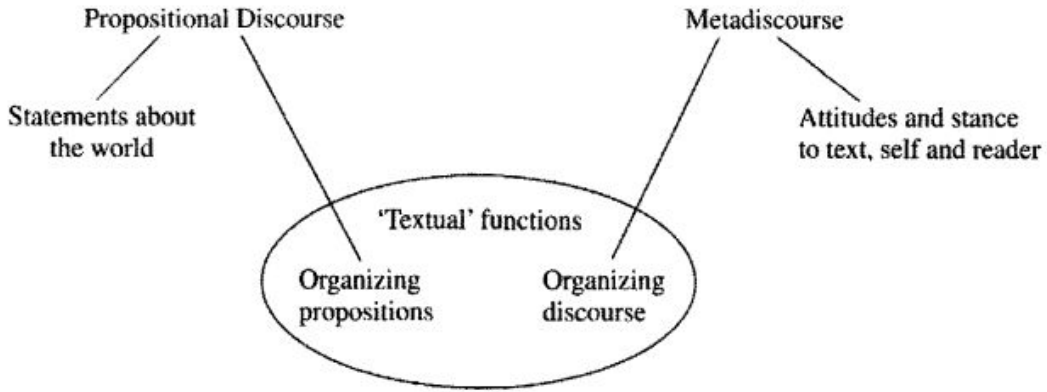


Figure 3.2 *The role of 'textual' devices in texts*

Estonian team

Metadiscourse study (incl. stance)

Aim 1: to **describe the whole paradigm** and full inventory of MD markers in academic discourse of Estonian

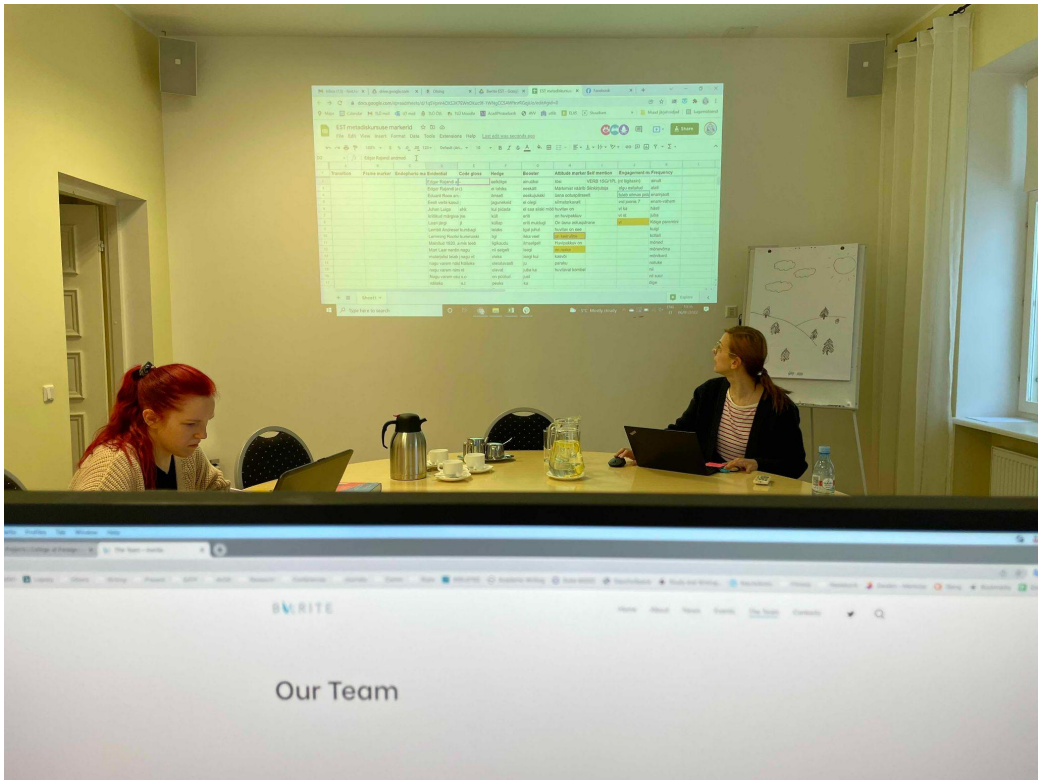
Aim 2: to **offer a transparent procedure for MD analysis**, to facilitate replication in future studies

(Hyland & interpersonal model of metadiscourse)

Method. Step 1. Corpus

3 x 7 = 21 RAs

Method. Step 2. Preliminary annotation for devising a coding scheme



Method. Step 3. Preparing the data for the analysis

No. of RAs	21
No. of journals	3
No. of RAs from each journal	7
Length of RAs in sentences	
- range	159–382
- average per article	241
- corpus size in sentences	5,058
Length of RAs in words	
- range	2,721–5,708
- average per article	4,270
- corpus size in words	89,660

Method. Step 4. Automatic data extraction and annotation

* **one-word_string** - otsib täpselt sellisel kujul (v.a see, et suurtel ja väikestel tähtedel ei tee vahet) **tähemärkide järjendit**, mis on kirjas veerus "Auto_search_marker" (nt *lisaks: lisaks, Lisaks*)

* **one-word_lemma** - otsib välja kõik sõnavormid, mille **algvormiks on veerus "Auto_search_marker" olev sõna** (nt *huvitav : huvitav, huvitava, huvitavale, Huvitavaks* jne)

* **beginning_with_string** - otsib sõnu, mis **algavad veerus "Auto_search_marker" oleva tähemärkide järjendiga** (nt *ootuspära*: ootuspärane, ootuspärasena, Ootuspäraselt* jne)

* **morphological_form** - otsib välja etteantud **sõnaliigile ja vormiinfole vastavad sõnad** (nt V_n|V_me|V_sin|V_sime: *analüüsin, vaatame, kirjeldasin, Esitasime* jne)

Method. Step 5. Manual data annotation

	Article_ID	Section	Sentence_ID	Sentence	MD_Marker	MD_Marker_lühem	Marker_category	Type_within_cat	Ling_level
210	ERY2013	res/disc	88	Kui vaadelda nende kahe näite kõrval teisi terminivalikuid ja nende põhjendusi (lisaks käsitletud töödele ka Afanasjev 2011), on võimalik täheldada teatavaid üldsuundi või tingimusi, millal kaldutakse eelistama kujundlikku varianti.	(*)	(*)	code gloss	elaboration	punct
211	ERY2013	res/disc	88	Kui vaadelda nende kahe näite kõrval teisi terminivalikuid ja nende põhjendusi (lisaks käsitletud töödele ka Afanasjev 2011), on võimalik täheldada teatavaid üldsuundi või tingimusi, millal kaldutakse eelistama kujundlikku varianti.	on võimalik täheldada	Vinf + võimalik + Vda	modal	modal	construction
212	ERY2013	res/disc	88	Kui vaadelda nende kahe näite kõrval teisi terminivalikuid ja nende põhjendusi (lisaks käsitletud töödele ka Afanasjev 2011), on võimalik täheldada teatavaid üldsuundi või tingimusi, millal kaldutakse eelistama kujundlikku varianti.	teatav*	teatav*	hedge	adv	lexical
213	ERY2013	res/disc	88	Kui vaadelda nende kahe näite kõrval teisi terminivalikuid ja nende põhjendusi (lisaks käsitletud töödele ka Afanasjev 2011), on võimalik täheldada teatavaid üldsuundi või tingimusi, millal kaldutakse eelistama kujundlikku varianti.	ka	ka	transition	addition	lexical
214	ERY2013	res/disc	89	Ohvitseride eelistus kaldub langema kujundlikule variandile siis, kui 1) on vaja eristada lähimõisteid (mõistekeskne16 argument) või kui 2) termin on juba lähtekeeles kujundlik (esmapilgul sõnakeskne argument).	1)2)3)	1)2)3)	frame marker	sequence	punct
215	ERY2013	res/disc	89	Ohvitseride eelistus kaldub langema kujundlikule variandile siis, kui 1) on vaja eristada lähimõisteid (mõistekeskne16 argument)	(*)	(*)	code gloss	elaboration	punct

Method. Step 6. Inter-rater reliability

Coder 1:

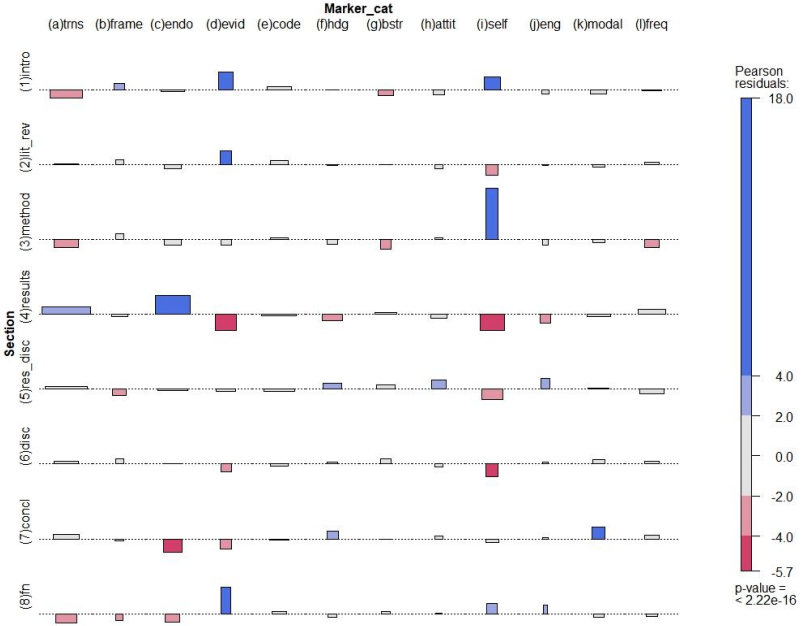
```
> kappa2(MDcoding[redacted])
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 240
   Raters = 2
    Kappa = 0.919

        z = 39.4
  p-value = 0
```

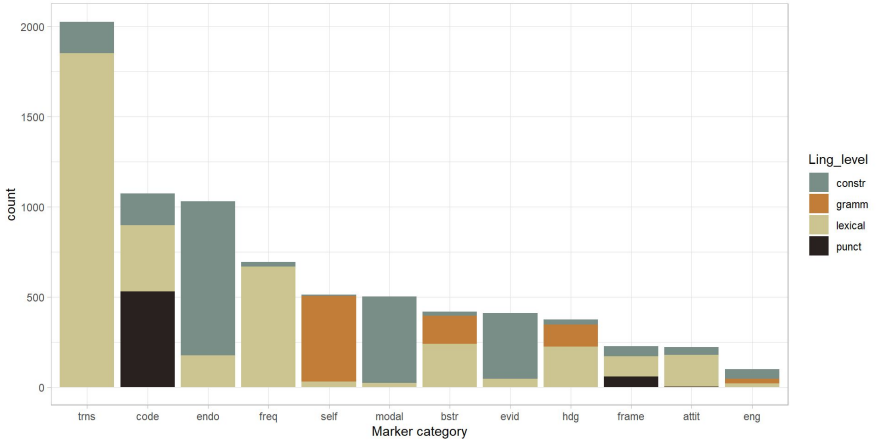
Coder 2:

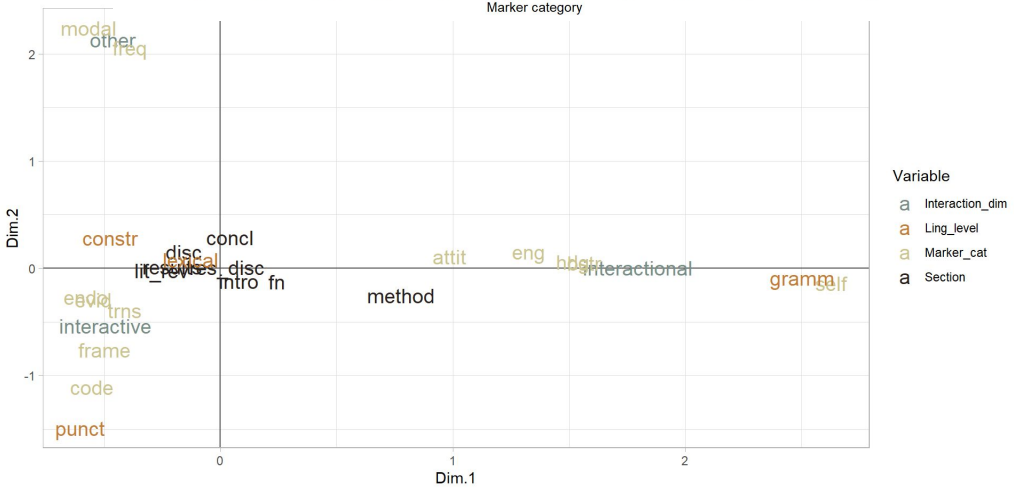
```
> kappa2(MDcoding)
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 235
   Raters = 2
    Kappa = 0.876

        z = 34.9
  p-value = 0
```

	Rater 1	Rater 2
No. of MD markers	315	284
Simple agreement (Q1)	93%	93%
Cohen's kappa (Q2)	0.919	0.876

Method. Step 7. Data analysis







BWRITE

Marker
MD
Category
Use
Text
Annotate
Interactive
Estonian
Analysis
Language
Hyland
Study
Data
Author
Differ
Model
Function
Associate
Article
Result
Section
Include
Express
Reader
Hedge
Verb
Certain
Level
Code
Academic
Self
Two
One
More
See
First
Table
Second
Item
Between
Booster
Transition
Dimension
Writer
Metadiscourse
Example
Step
Refer
Means
Research
Mention
Each
Word
Writing
Well
Form
Construction
Linguistic
Device
Pattern
Thus
Corpus
Sentence
Paper
Gloss
Structure
Present
Particular
Possible
Engage
Proposition
Frame
Aspect
Support
Base
Content
Origin
Variable
Method
Follow
Mark
Choice
Aim
Made
Select
Such
English
List
Per
Bax
Attitude
Evidential
Both
Discussed
Modal
Result
Frequency
Automatic
Person
Lexical
Figure
Process
Conclusion
Toward

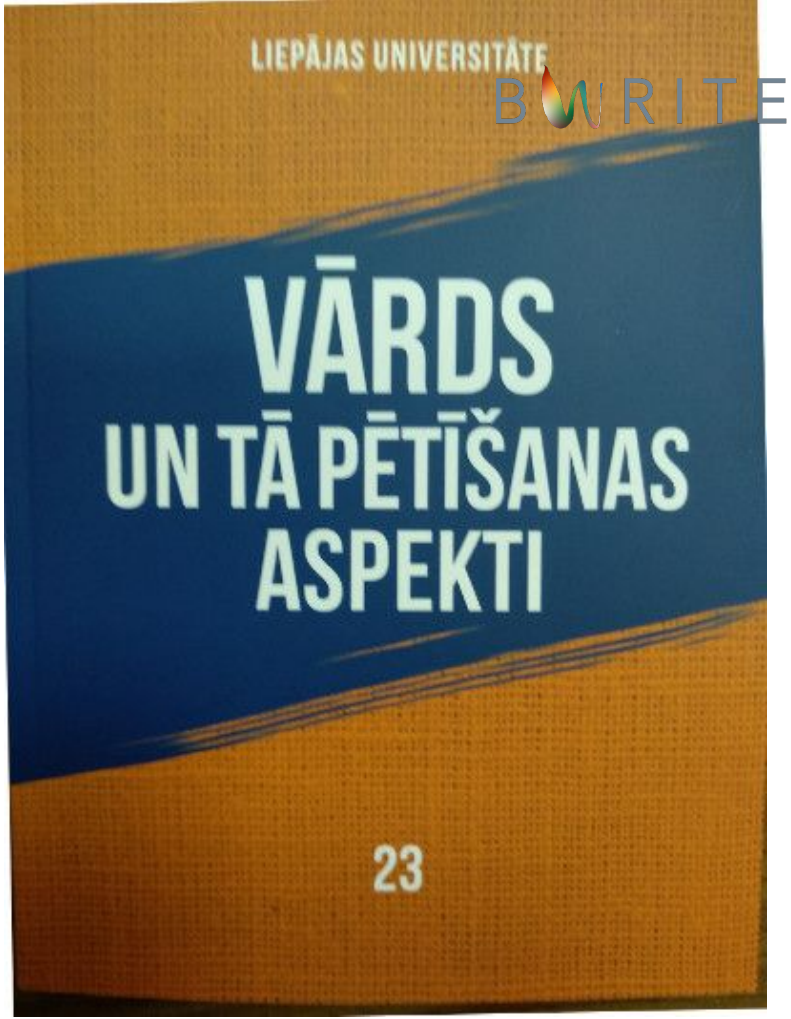
Latvian team

Latvian Metadiscourse data

Corpus: 30 articles in Linguistics

Selection criteria:

- the paper needs to have one author
- the author of the paper is a native Latvian speaker



What is the typical Latvian writing tradition?

- The major ‘’unwritten rule’’ in Latvian academic writing is that the presence of the author in their writing is neutral.
- The use of first person pronouns like *I, me* is frowned upon
- The only exception to this rule are papers written in some STEM fields, published by larger groups of researchers – in this case, the use of *we* is tolerated but not fully accepted

What do Latvian Linguists do?

1. Write about themselves in the third person, referring to themselves as ‘’the author’’. This can cause some confusion within texts, especially when citations appear. Who is supposed to be ‘’the author’’?
2. Usage of verb forms/grammatical constructions that can show the author’s presence and attitude *([I] will show, [I] found)*

Does the metadiscourse back up this tradition?

Out of 7017 total metadiscourse markers used within the 30 texts, the most common markers used are:

- Code Glosses - 2855
- Transition - 1438
- Boosters - 730
- Hedges - 428

What about self mentions? Only 78 markers can be attributed to self mention

Lithuanian team

Metadiscourse study

1) to explore interactive and interactional metadiscourse devices in Linguistics in a larger size self-compiled corpus by operationalizing Hyland's metadiscourse model (2005a);

2) to apply unified methodological procedures in data collection and annotation.

Corpus

Journal	Word count	Sentence count
Kalbotyra	46,466	1,976
Lietuvių kalba	33,896	1,684
Taikomoji kalbotyra	53,943	2,320
Total	**134,305**	**5,980**

Data annotation: dimensions and marker categories

Interactive: *transitions, evidentials, code glosses, frame markers*

Interactional: *boosters, hedges, engagement markers, self-mentions, attitude markers*

Data annotation: subcategories (1)

Transitions
additive
constrastive
consequential

Frame markers
sequencers
labellers
goal announcers
topic shifters

Endophoric markers
previewing
reviewing
visuals
examples in a text

Code glosses
reformulation
exemplification
elaboration

Evidentials
Integral

Data annotation: subcategories (2)

Hedges
modal verbs
CTPs
epistemic/ep-evid adverbials
diminishers
approximators

Boosters
lexical verbs
adverbials
emphasisers

Engagement markers
reader pronouns
directives
questions
shared knowledge

Attitude markers
adjectives
adverbials

Self-mentions
1-st person singular/plural verb forms
pronouns
the author

Data annotation

- Pilot study (10% of the sample coded)
- Three rounds of double coding (2 coders)

  97.46% inter-rater agreement

  unweighted kappa measure estimated at 0.970
- Manual annotation

Results

Interactive			Interactional		
	raw fr	%		raw fr	%
Transitions	1,867	33	**Boosters**	846	32
Evidentials	1,815	32	**Engagement markers**	793	30
Code glosses	996	18	**Hedges**	781	29
Endophorics	726	13	**Attitude markers**	173	7
Frame markers	207	4	**Self-mentions**	56	2
Total	**5,611**	**100%**	**Total**	**2,649**	**100%**

Interactive features (1)

Category	Subcategory	%
Transitions	*constrastive*	**56**
	consequence	23
	addition	21
		100
Evidentials	*non-integral*	**76**
	integral	24
		100
Code glosses	*exemplification*	**51**
	reformulation	30
	elaboration	19
		100

Interactive features (2)

Endophoric markers	*examples*	**32**
	visuals	29
	reviewing	25
	previewing	14
		100
Frame markers	*goal announcers*	**46**
	sequencers	30
	labellers	24
	topic shifters	4
		100

Interactional features (1)

Category	Subcategory	%
Boosters	*lexical verbs of showing, seeing, confirming*	**60**
	emphasisers	19
	epistemic adjectives and adverbials	11
	particles and their combinations, modal verbs, nouns and other	10
		100
Engagement markers	*directives*	**78**
	shared knowledge	15
	reader mentions	4
	questions and other markers	3
		100

Interactional features (2)

Hedges	*modal verb galėti* 'can/may'	**39**
	adverbials	22
	CTPs	18
	diminishers	10
	approximators	6
	the subjunctive mood, participles and other	5
		100
Attitude markers	*adjectives*	**81**
	adverbs/ adverbials	13
	verbs and others	6
		100
Self-mentions	*first-person singular verb inflection*	**59**
	the author	18
	pronoun	12
	first-person plural verb inflection	11
		100

Implication and larger discussion

Availability, over-representation and the conflict between traditions and explicit formats

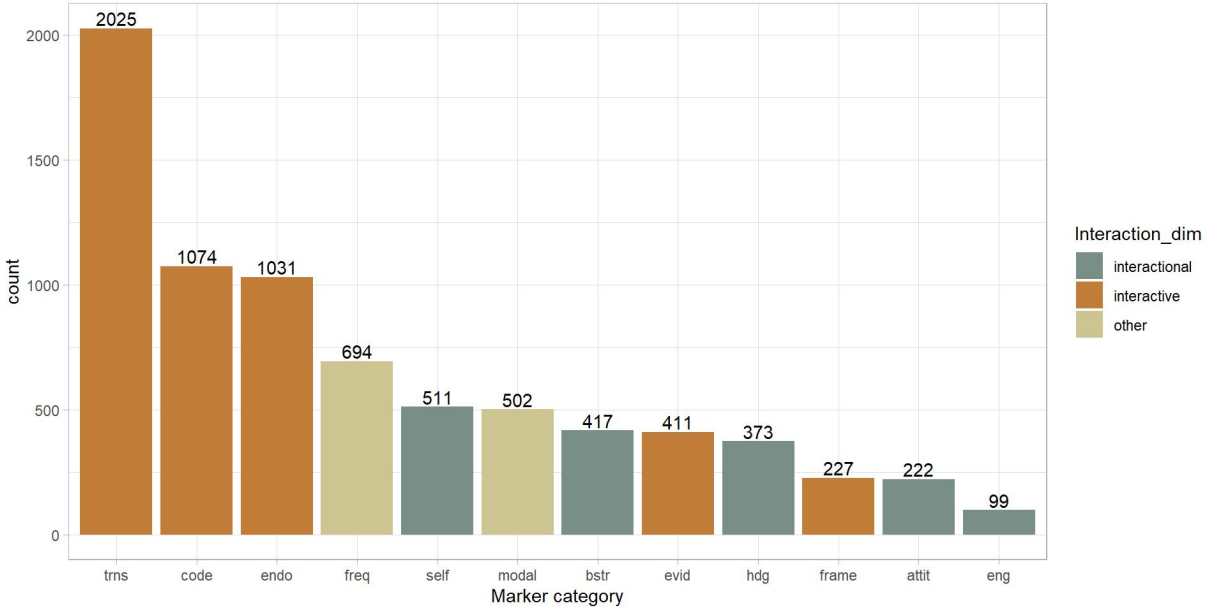
- Motivation and background for the research from the data, the interpretation of the data, and finally the argumentation and conclusions
- Within a *scientific tradition*, we signal science implicitly through similarity to other works.
- We have found out what works within our fields, without any explicit formal rules.
- The nature of traditions is that they are implicit and ubiquitous, so we are less aware of them. They are what we do.
- Preliminary findings show specific trends*

How do we know what people do when writing a scientific paper?

- Survival bias --- when we look at the published papers we will find the papers that survived the selection criteria.
- Part of the selection criteria is to follow a recognizable format.
- Detecting a writing tradition or a remnant of a writing tradition maybe still be detectable in how an article is argued.

Interactive more common than Interactional.

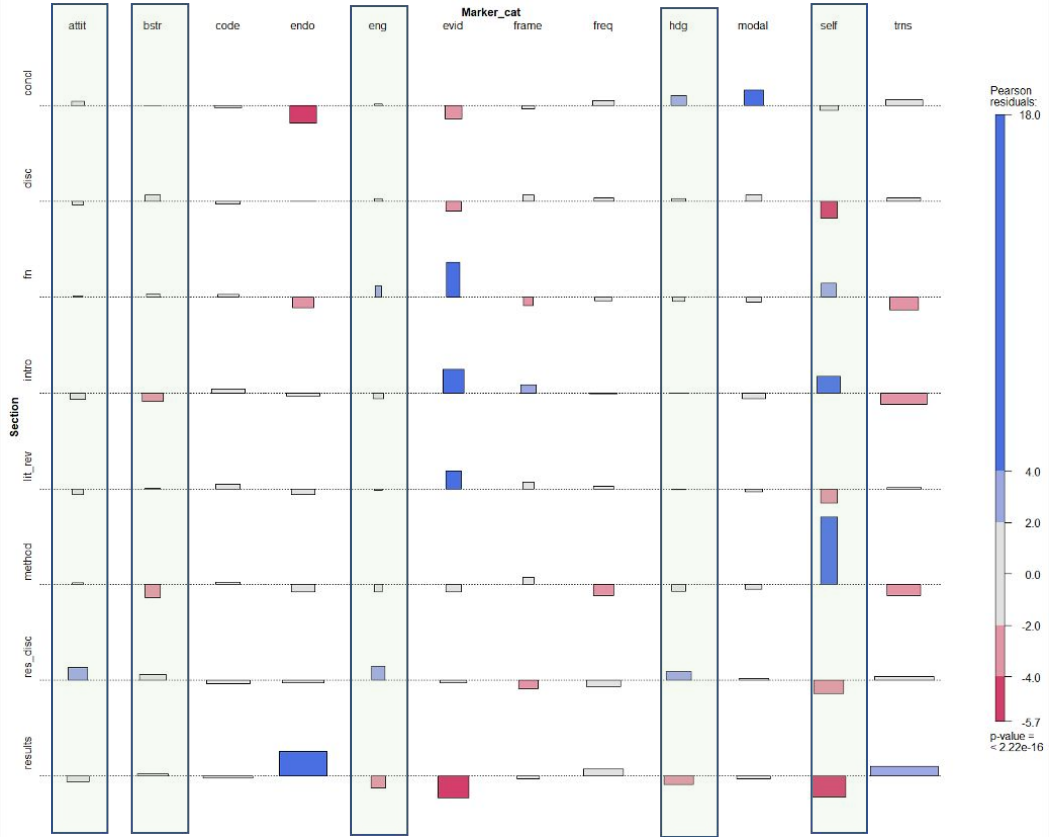
Is guiding the reader through the text more important than involving the reader, or is it possibly more difficult to involve the reader?



The next slide will show positive associations between (meta)discourse markers and sections of a text as BLUE bars.

Negative associations will be shown as RED bars.

Involvement or Guidance depends on where.



Engaging Reader marked by boxes.

Here: *Self mention* most prominent for involving the reader -- In *Method* and *Introduction*.

Guiding Reader
Results related to information *inside* text.

Intro, Lit.Rev., footnote related to information *outside* text.

Does it have to be this way

The association graph only shows how stance & metadiscourse markers show up in (**Estonian) linguistics** texts.

Will other **genres** have different preferences?

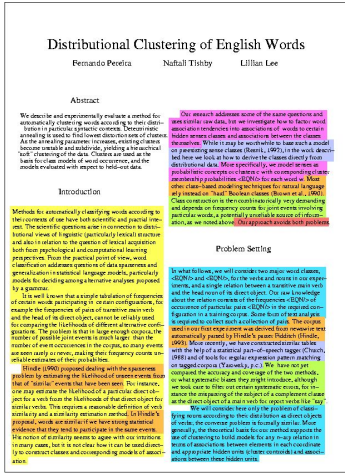
Will other **languages** result in different preferences?

Will other **writing traditions** have a different balance between *guiding* and *involving* the reader?

Is an engaging text perceived as less factual?

Is there an optimal balance?

Feature 3. tba



Distributional Clustering of English Words

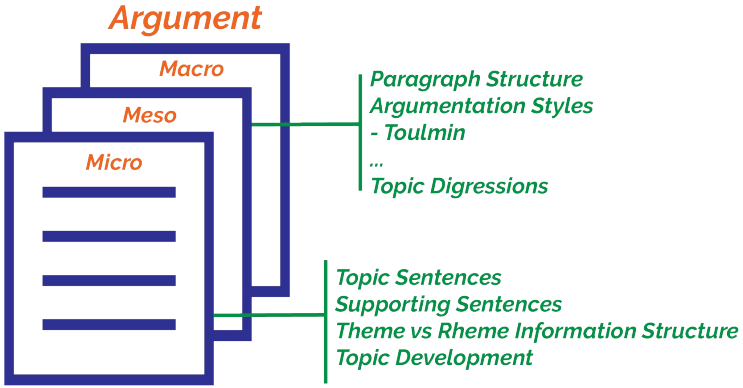
Fernando Pereira Naftali Tishby Lillian Lee

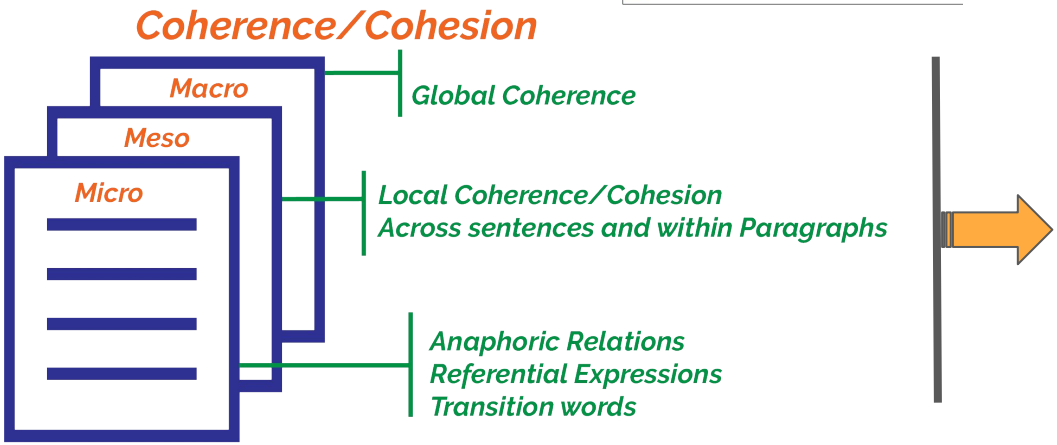
Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become natural and hierarchical, yielding a tree-structured hierarchy of the data. Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out data.

Introduction

Problem Setting

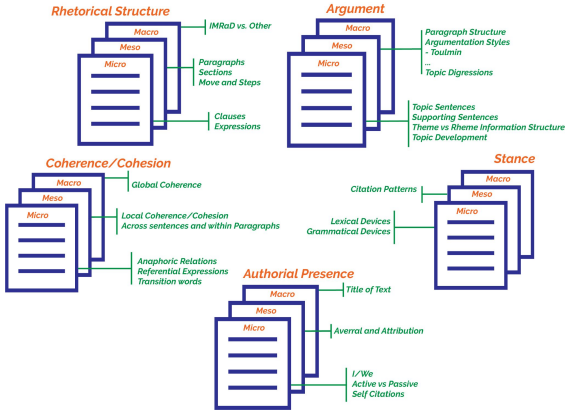




COHERENCE through

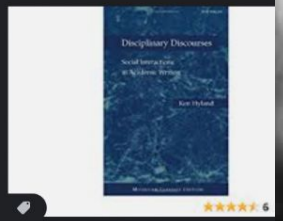
- Intertextuality
- How true are you to your community (?)
- Deictic expressions

What follows?

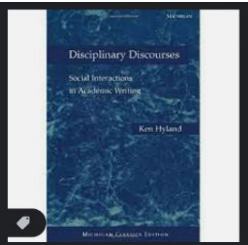


Amazon.com: Genre A...
amazon.com · In stock

John SWALES | Professor e...
researchgate.net

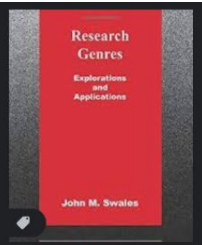


Social Interactions in Academic Writi
amazon.com · In stock



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

essential bookshelf: Aca...
cambridge.org



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

Discussion

Questions and points of discussion

Questions about …

1. … the theoretical framework
2. … methodological approach
3. … results
4. … discussion, implication, application

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

Interactive vs Interactional

Any writer needs to consider their readers.

Signs to guide the reader through the text (Interactive)

- Transitions -- express relations between clauses (this and that but also that)
- Frame markers – discourse acts. (Finally, to conclude, my aim is to …)
- Endophoric markers – refer to information in the text. (see below in Figure 4)
- Evidentials – refer to information outside of the text. (According to X, …)
- Code gloss – in other words, such as for example X, namely …

Signs to involve the reader with the text (Interactional)

- Hedges – withhold commitment. Might, perhaps, be possible, about
- Boosters – Emphasize certainty. Definitely, it is clear that, in fact …
- Attitude – Unfortunately, surprisingly, I agree …
- Self mention – reference to the author(s). I, we, our, my …
- Engagement markers – build with the reader. You can see that, we should consider