UNIVERSITY OF SURREY

UNIVERSITY OF TARTU
Institute of Estonian and General Linguistics

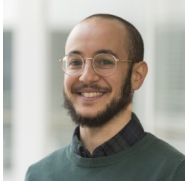
Université Paris Cité

THE OHIO STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

Eesthetic: A Paralex Lexicon of Estonian Paradigms

LREC-COLING 2024

Sacha Beniamine



Mari Aigro

Matthew Baerman

Jules Bouton

Maria Copot

Eesthetic

- a **large lexicon** of Estonian inflection

	lexemes	cells	forms
Nouns	5475	28	193 931
Verbs	5076	51	258 954
total	10551	79	452 885

- **standardized** in the Paralex format (csv)
- with rich **linguistic information**
- for **quantitative & qualitative** morphology

Estonian paradigms

case / number	SG	PL
NOM	sõna	sõnad
PART	sõna	sõnu / sõnasid
GEN	sõna	sõnade
ILL	sõnna / sõnasse	sõnusse / sõnadesse
INESS	sõnas	sõnus / sõnades
ELAT	sõnast	sõnust / sõnadest
ALL	sõnale	sõnule / sõnadele
AD	sõnal	sõnul / sõnadel
ABL	sõnalt	sõnult / sõnadelt
TRANS	sõnaks	sõnuks / sõnadeks
TERM	sõnani	sõnadeni
ESS	sõnana	sõnadena
ABESS	sõnata	sõnadeta
COM	sõnaga	sõnadega

sõna, N., ‘word’, class 17u

Estonian paradigms

case / number	SG	PL
NOM	sõna	sõnad
PART	sõna	sõnu / sõnasid
GEN	sõna	sõnade
ILL	sõnna / sõnasse	sõnusse / sõnadesse
INESS	sõnas	sõnus / sõnades
ELAT	sõnast	sõnust / sõnadest
ALL	sõnale	sõnule / sõnadele
AD	sõnal	sõnul / sõnadel
ABL	sõnalt	sõnult / sõnadelt
TRANS	sõnaks	sõnuks / sõnadeks
TERM	sõnani	sõnadeni
ESS	sõnana	sõnadena
ABESS	sõnata	sõnadeta
COM	sõnaga	sõnadega

sõna, N., ‘word’, **class 17u**

case / number	SG	PL
NOM	aas	aasad
PART	aasa	aasu / aasasid
GEN	aasa	aasade
ILL	aasa / aasasse	aasusse / aasadesse
INESS	aasas	aasus / aasades
ELAT	aasast	aasust / aasadest
ALL	aasale	aasule / aasadele
AD	aasal	aasul / aasadel
ABL	aasalt	aasult / aasadelt
TRANS	aasaks	aasuks / aasadeks
TERM	aasani	aasadeni
ESS	aasana	aasadena
ABESS	aasata	aasadeta
COM	aasaga	aasadega

aas, N., ‘loop’, **class 22u**

Estonian paradigms

Quantity 1 | Quantity 2 | Quantity 3

case / number	SG	PL
NOM	aas /ˈɑːs/	aasad
PART	aasa /ˈɑːsɑ/	aasu / aasasid
GEN	aasa /ɑːsɑ/	aasade
ILL	aasa /ˈɑːsɑ/	aasusse / aasadesse

aas, N., ‘loop’, class 22u

case / number	SG	PL
NOM	sõna /sɤnɑ/	sõnad
PART	sõna /sɤnɑ/	sõnu / sõnasid
GEN	sõna /sɤnɑ/	sõnade
ILL	sõnna /sˈɤnːɑ/	sõnusse / sõnadesse

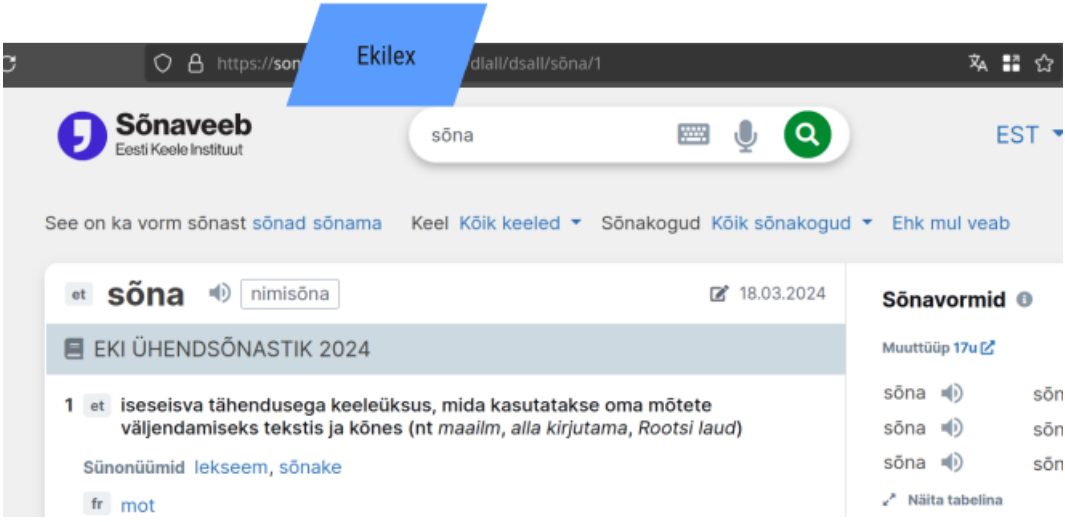
sõna, N., ‘word’, class 17u

Estonian paradigms

case / number	SG	PL
NOM	sõna	sõnad
PART	sõna	sõnu / sõnasid
GEN	sõna	sõnade
ILL	sõnna / sõnasse	sõnusse / sõnadesse
INESS	sõnas	sõnus / sõnades
ELAT	sõnast	sõnust / sõnadest
ALL	sõnale	sõnule / sõnadele
AD	sõnal	sõnul / sõnadel
ABL	sõnalt	sõnult / sõnadelt
TRANS	sõnaks	sõnuks / sõnadeks
TERM	sõnani	sõnadeni
ESS	sõnana	sõnadena
ABESS	sõnata	sõnadeta
COM	sõnaga	sõnadega

sõna, N., ‘word’, class 17u

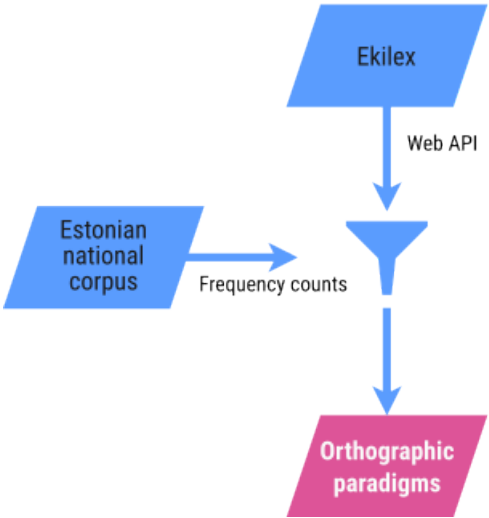
Lexicon creation (I)



Lexicon creation (I)

Estonian national corpus

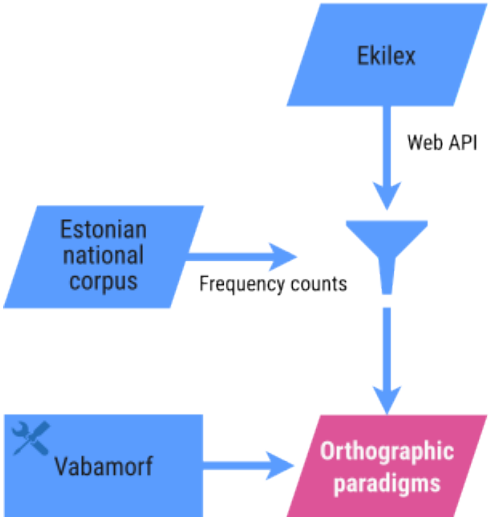
Lexicon creation (I)



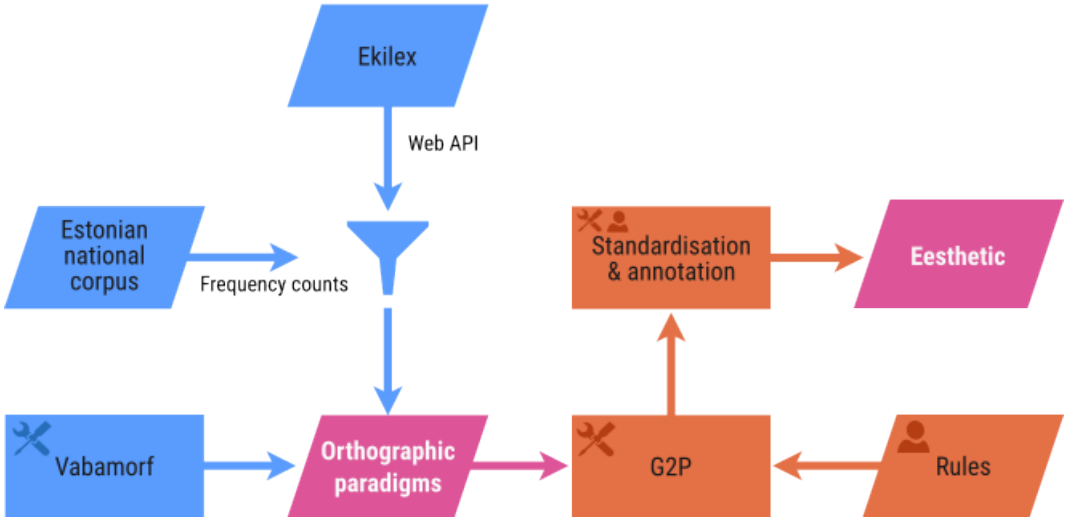
Annotated orthography

(1)	<´oman`ik> ['omanʲ'ikː]	‘owner’ NOM.SG
(2)	<sõna[sse> [sɤnasːe]	‘word’ ILL.SG
(3)	<h`oo+`aeg> [h'oː:'a̯eːk]	‘season’
(4)	<k`ot't> [k'otʲːt]	‘bag’
(5)	aasa	‘loop’
	a. <aasa> [aːsa]	GEN.SG
	b. <`aasa> [aː:sa]	PART.SG

Lexicon creation (II)



Lexicon creation (II)



Grapheme to phoneme

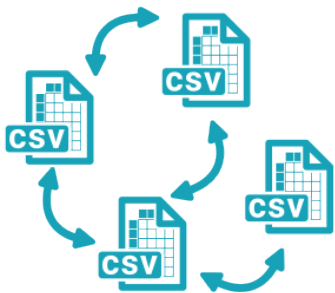
Custom rules for Epitran (Mortensen et al., 2018)

```
n -> ŋ / _ [kg]
```

- Pre-processing rules:
  - Transcribe orthographic conventions (quantity, length, palatalization)
  - Apply regular phonological rules
- Non-contextual mapping
- Post-processing rules:
  - Add ligatures to diphthongs
  - Identify which sounds to lengthen under Q3
  - Cleanup typography
- Adjustments

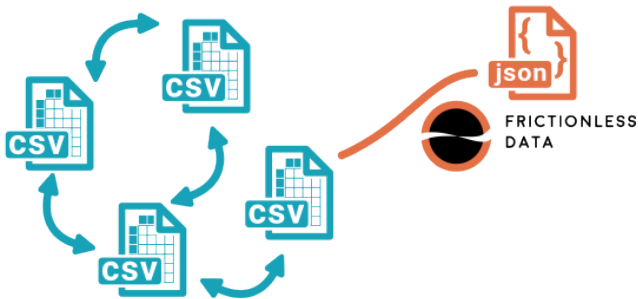
Paralex standard

DATA



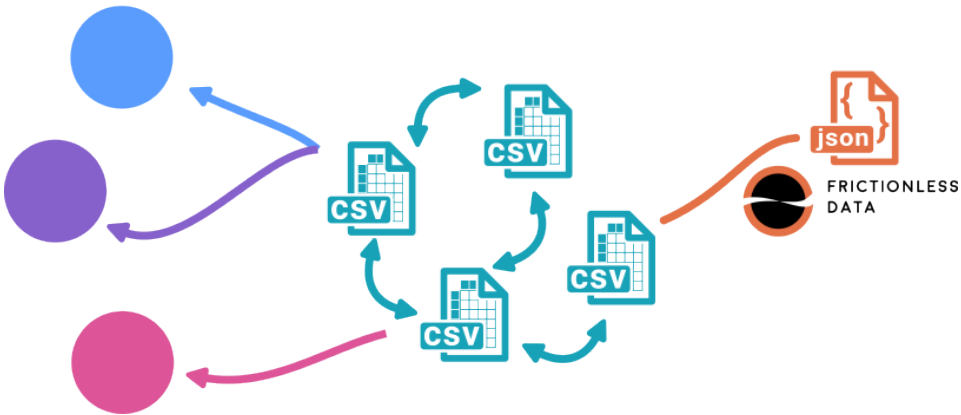
Paralex standard

DATA METADATA



Paralex standard

LINKED DATA METADATA



Paralex standard

Paralex
Home
Standard
Specs
Ontology
Background
Principles
What are metadata ?
Why the long form ?
How to
Link sources
Structure paradigms
Use comments and tags
Express sounds
Tutorial
FAQ
Datasets

Paralex: lexicons of morphological paradigms

Paralex is a standard for morphological lexicons which document inflectional paradigms.

It strives to provide data which is FAIR, so it can be used automatically, CARE, so it respects and empowers language communities, and DeAR (our own set of principles), so we can create a virtuous data ecosystem. It was inspired by the Cross-Linguistic Data Formats (CLDF) standard, and adheres to a similar philosophy and the same the design principles. We aim to keep the two standards compatible in order to facilitate inter-operability.

A **paralex** lexicon is a set of tables written as comma separated value (csv) files. It follows a relational model, tables are written in long form, metadata is written using the frictionless standard, and the tables respect pre-defined conventions. An ontology is also provided to allow converting paralex lexicons into RDF lemon/ontolex lexicons.

The standard is meant for sharing and interfacing, but not necessarily for data input. The expectation is for data creators to first input data through any convenient means, then convert the result into the standardized structure for publishing and sharing.

Table of contents
1 Contributors
2 Version

http://www.paralex-standard.org

DeAR Principles

To publish high quality, easily citable, scientifically impactful data, useful for the long term:

- **Decentralized**
  - Centralized data is neither robust nor long-lasting
  - International collaboration must be incentivised
- **Automatically validated**
  - Manual curation of large datasets is necessary but error-prone
  - Automatic quality control of structure and content is crucial
- **Revisable**
  - Data evolves, its presentation must be updateable
  - Seamless updates of showcases can be generated from a single data source

Relational Structure

form_id	lexeme	cell	phon_form	analysed_phon_form	orth_form	analysed_orth_form	overabundance_tag	defectiveness_tag	epistemic_tag
14618895	sõna_238850_1349429	nom.sg	s ɤ n ɑ	s ɤ n ɑ	sõna	sõna			
14618896	sõna_238850_1349429	gen.sg	s ɤ n ɑ	s ɤ n ɑ	sõna	sõna			
14618897	sõna_238850_1349429	part.sg	s ɤ n ɑ	s ɤ n ɑ	sõna	sõna			
14618898	sõna_238850_1349429	ill.sg	s ˈ ɤ n:: ɑ	s ˈ ɤ n:: ɑ	sõnna	s'õnna	aditive		
14618899	sõna_238850_1349429	ill.sg	s ɤ n ɑ s: e	s ɤ n ɑ + s: e	sõnasse	sõna[sse	sse_illative		
14618900	sõna_238850_1349429	iness.sg	s ɤ n ɑ s	s ɤ n ɑ + s	sõnas	sõna[s			
14618901	sõna_238850_1349429	elat.sg	s ɤ n ɑ s t	s ɤ n ɑ + s t	sõnast	sõna[st			
14618902	sõna_238850_1349429	all.sg	s ɤ n ɑ l e	s ɤ n ɑ + l e	sõnale	sõna[le			
14618903	sõna_238850_1349429	ad.sg	s ɤ n ɑ l	s ɤ n ɑ + l	sõnal	sõna[l			
14618904	sõna_238850_1349429	abl.sg	s ɤ n ɑ l t:	s ɤ n ɑ + l t:	sõnalt	sõna[lt			
14618905	sõna_238850_1349429	trans.sg	s ɤ n ɑ k s	s ɤ n ɑ + k s	sõnaks	sõna[ks			
14618906	sõna_238850_1349429	term.sg	s ɤ n ɑ nʲ i	s ɤ n ɑ + nʲ i	sõnani	sõna[ni			
14618907	sõna_238850_1349429	ess.sg	s ɤ n ɑ n ɑ	s ɤ n ɑ + n ɑ	sõnana	sõna[na			
14618908	sõna_238850_1349429	abess.sg	s ɤ n ɑ t: ɑ	s ɤ n ɑ + t: ɑ	sõnata	sõna[ta			
14618909	sõna_238850_1349429	com.sg	s ɤ n ɑ k ɑ	s ɤ n ɑ + k ɑ	sõnaga	sõna[ga			
14618910	sõna_238850_1349429	nom.pl	s ɤ n ɑ t	s ɤ n ɑ + t	sõnad	sõna[d			
14618911	sõna_238850_1349429	gen.pl	s ɤ n ɑ t e	s ɤ n ɑ + t e	sõnade	sõna[de	de_te_genitive		
14618912	sõna_238850_1349429	part.pl	s ɤ n u	s ɤ n u	sõnu	sõnu	voc_partitive		
14618913	sõna_238850_1349429	part.pl	s ɤ n ɑ sʲ i t	s ɤ n ɑ + sʲ i t	sõnasid	sõna[sid	sid_partitive		
14618914	sõna_238850_1349429	ill.pl	s ɤ n ɑ t e s: e	s ɤ n ɑ + t e s: e	sõnadesse	sõna[desse	de_te_plural		
14618915	sõna_238850_1349429	ill.pl	s ɤ n u s: e	s ɤ n u + s: e	sõnusse	sõnu[sse	voc_rad_plural		
14618916	sõna_238850_1349429	iness.pl	s ɤ n ɑ t e s	s ɤ n ɑ + t e s	sõnades	sõna[des	de_te_plural		
14618917	sõna_238850_1349429	iness.pl	s ɤ n u s	s ɤ n u + s	sõnus	sõnu[s	voc_rad_plural		
14618918	sõna_238850_1349429	elat.pl	s ɤ n ɑ t e s t	s ɤ n ɑ + t e s t	sõnadest	sõna[dest	de_te_plural		
14618919	sõna_238850_1349429	elat.pl	s ɤ n u s t	s ɤ n u + s t	sõnust	sõnu[st	voc_rad_plural		
14618920	sõna_238850_1349429	all.pl	s ɤ n ɑ t e l e	s ɤ n ɑ + t e l e	sõnadele	sõna[dele	de_te_plural		
14618921	sõna_238850_1349429	all.pl	s ɤ n u l e	s ɤ n u + l e	sõnule	sõnu[le	voc_rad_plural		
14618922	sõna_238850_1349429	ad.pl	s ɤ n ɑ t e l	s ɤ n ɑ + t e l	sõnadel	sõna[del	de_te_plural		
14618923	sõna_238850_1349429	ad.pl	s ɤ n u l	s ɤ n u + l	sõnul	sõnu[l	voc_rad_plural		
14618924	sõna_238850_1349429	abl.pl	s ɤ n ɑ t e l t:	s ɤ n ɑ + t e l t:	sõnadelt	sõna[delt	de_te_plural		
14618925	sõna_238850_1349429	abl.pl	s ɤ n u l t:	s ɤ n u + l t:	sõnult	sõnu[lt	voc_rad_plural		
14618926	sõna_238850_1349429	trans.pl	s ɤ n ɑ t e k s	s ɤ n ɑ + t e k s	sõnadeks	sõna[deks	de_te_plural		
14618927	sõna_238850_1349429	trans.pl	s ɤ n u k s	s ɤ n u + k s	sõnuks	sõnu[ks	voc_rad_plural		
14618928	sõna_238850_1349429	term.pl	s ɤ n ɑ t e nʲ i	s ɤ n ɑ + t e nʲ i	sõnadeni	sõna[deni	de_te_plural		
14618929	sõna_238850_1349429	ess.pl	s ɤ n ɑ t e n ɑ	s ɤ n ɑ + t e n ɑ	sõnadena	sõna[dena	de_te_plural		
14618930	sõna_238850_1349429	abess.pl	s ɤ n ɑ t e t: ɑ	s ɤ n ɑ + t e t: ɑ	sõnadeta	sõna[deta	de_te_plural		
14618931	sõna_238850_1349429	com.pl	s ɤ n ɑ t e k ɑ	s ɤ n ɑ + t e k ɑ	sõnadega	sõna[dega	de_te_plural		

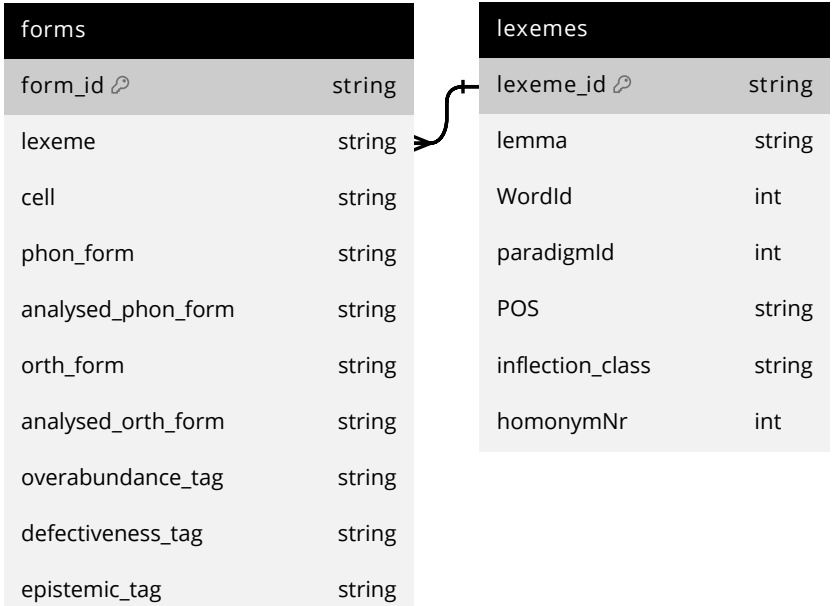
Relational Structure

forms	
form_id	string
lexeme	string
cell	string
phon_form	string
analysed_phon_form	string
orth_form	string
analysed_orth_form	string
overabundance_tag	string
defectiveness_tag	string
epistemic_tag	string

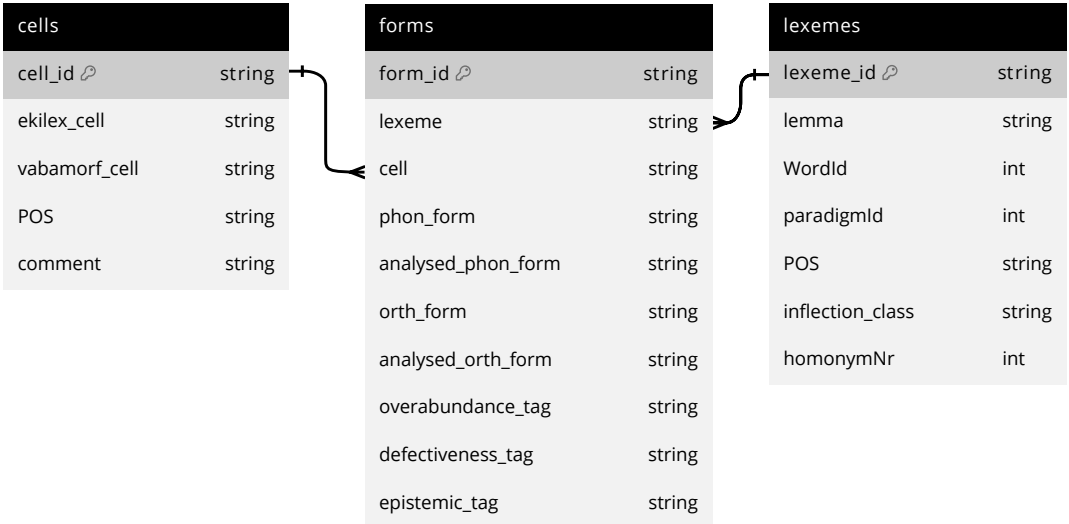
Relational Structure

lexeme_id	lemma	wordId	paradigmId	POS	inflection_class	homonymNr
sõna_238850_1349429	sõna	238850	1349429	noun	17u	1
sahkerdama_230449_1387659	sahkerdama	230449	1387659	verb	27	1
gastroleerima_166864_1388631	gastroleerima	166864	1388631	verb	28	1
vastukaja_254709_1479772	vastukaja	254709	1479772	noun	17	1
sinetama_233880_1386716	sinetama	233880	1386716	verb	27	1
servima_232818_1387793	servima	232818	1387793	verb	28	1
pärija_223664_1348672	pärija	223664	1348672	noun	1	1
tipphetk_244967_1477638	tipphetk	244967	1477638	noun	22i	1
ammendama_157058_1385304	ammendama	157058	1385304	verb	27	1
töö_250035_1373819	töö	250035	1373819	noun	26i	1
transleerima_246364_1389340	transleerima	246364	1389340	verb	28	1
ägisema_261520_1386191	ägisema	261520	1386191	verb	27	1
kristalliseerima_186562_1412429	kristalliseerima	186562	1412429	verb	28	1
vaade_252049_1371583	vaade	252049	1371583	noun	6	1
kõpsutama_189238_1384885	kõpsutama	189238	1384885	verb	27	1
jäätis_175576_1288507	jäätis	175576	1288507	noun	11	1
rutiin_1111368_1444852	rutiin	1111368	1444852	noun	22e	2
pesu_215900_1352908	pesu	215900	1352908	noun	17	2
põlistuma_222793_1506098	põlistuma	222793	1506098	verb	27	1
veerg_255264_1297645	veerg	255264	1297645	noun	22e	1
sonkima_235110_1387555	sonkima	235110	1387555	verb	28	1

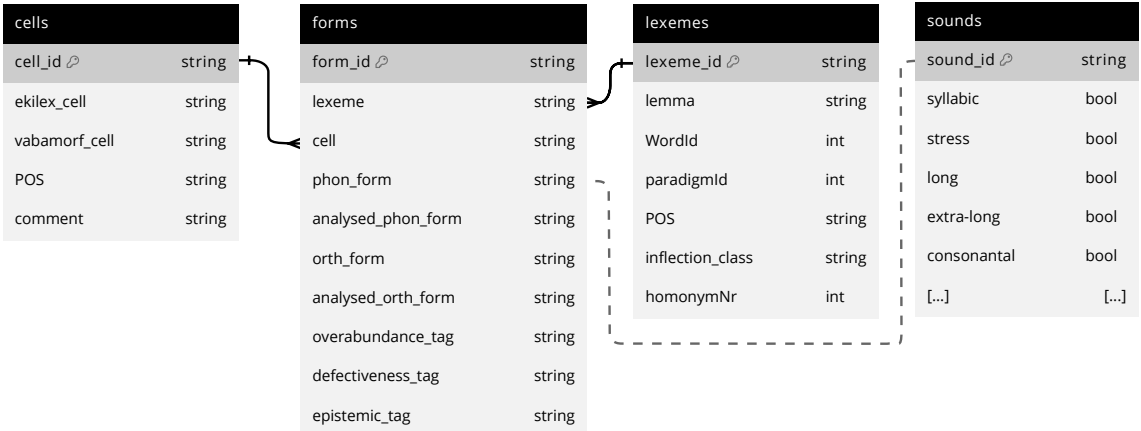
Relational Structure



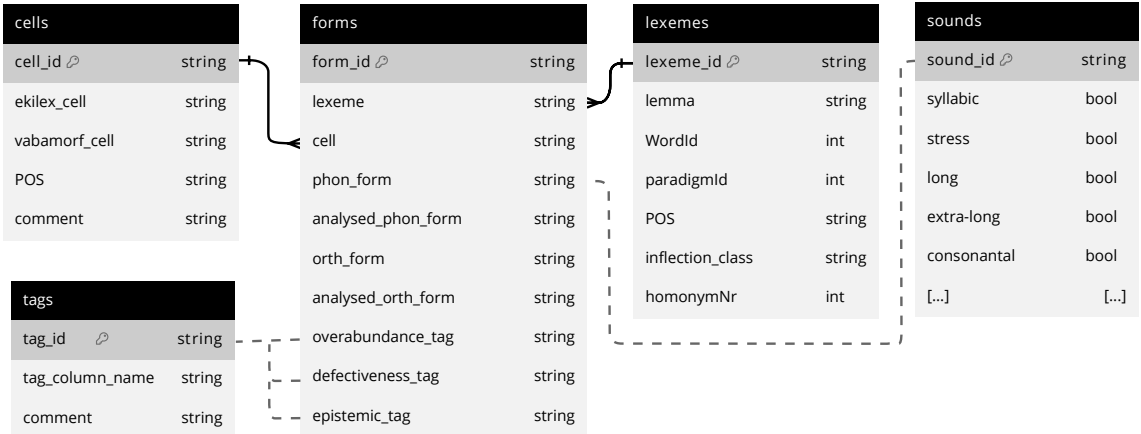
Relational Structure



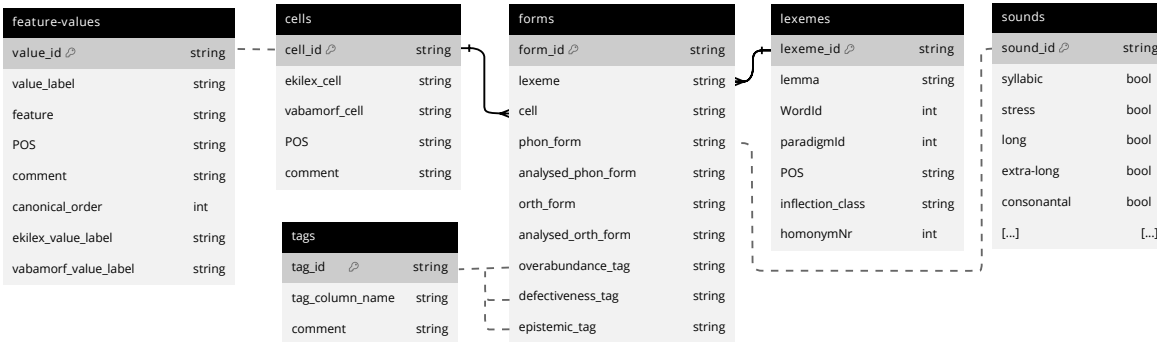
Relational Structure



Relational Structure



Relational Structure



Evaluation

- **Validation**: frictionless & paralex
- **G2P**: development set; targeted manual verifications
- **Paradigms**: Comparison to Unimorph

Conclusion: Eesthetic lexicon

DOI 10.5281/zenodo.8383522

- DeAR Paralex lexicon for Estonian N & V
- Large (5k lexemes per POS)
- Rich linguistic annotation
- Generated phonological forms
- Linguistically motivated G2P rules

Conclusion: Eesthetic lexicon

DOI 10.5281/zenodo.8383522

- DeAR Paralex lexicon for Estonian N & V
- Large (5k lexemes per POS)
- Rich linguistic annotation
- Generated phonological forms
- Linguistically motivated G2P rules

Thank you !