Realised overabundance in Estonian noun paradigms: A corpus study

A forthcoming article by Aigro & Vihman

October 11, 2023

Starting point

Mari Aigro and Virve Vihman (forthcoming). “Realised overabundance in Estonian noun paradigms: A corpus study.” In: *Word Structure* 16.2. DOI: 10.3366/word.2023.0227

Plan

Overabundance

Estonian Morphology

The Corpus study

What is overabundance? I

Thornton 2011:

> *When a cell in a paradigm is filled by two or more synonymous forms which realize the same set of morpho-syntactic properties, as in the case of the two English past tense forms burnt and burned.*[1]

A now widely used definition:

> *The situation in which two (or more) inflectional forms are available to realize the same cell in an inflectional paradigm.*[2]

What is overabundance? II

Example:

	Singular	Plural
P1	fázok / fázom	fázunk
P2	fázol	fáztok
P3	fázik	fáznak

Table: Overabundance in the paradigm of Hungarian *fázik* ’to feel cold’

[1] Anna M. Thornton (2011). “Overabundance (Multiple Forms Realizing the Same Cell): A Non-canonical Phenomenon in Italian Verb Morphology.” In: *Morphological autonomy: perspectives from Romance inflectional morphology*. Ed. by Martin Maiden et al. Oxford ; New York: Oxford University Press, pp. 358–381, p. 223.

[2] Anna M. Thornton (2019). “Overabundance: A Canonical Typology.” In: *Competition in Inflection and Word-Formation*. Ed. by Franz Rainer et al. Studies in Morphology 5. Cham: Springer, pp. 223–258. DOI: 10.1007/978-3-030-02550-2_9, p. 223.

Framework I

- Scale ?
  One may want to reduce OA to a social phenomenon. Each individual would produce only one form for a given cell, but different individuals might produce different forms for the same cell, according to:
  - Diaphasic variation
  - Distratic variation
  - Diatopic variation
  - Diachronic variation
  - Diamesic variation
- Reality ?
  One may consider that OA in speech production only reflects noise, i.e. errors from the speaker.

Framework II

In fact, there is some OA at the individual scale: not a defeat of descriptive linguistics, but an interesting challenge. How to account for it?

- Biological insight: “ecological niche differentiation”. OA is a transitory phenomenon and differenciation should always occur at some point[3].
- Behavioural insight: an individual isn’t a machine. The speaker is exposed to a non-uniform environment and may itself produce different forms he has been exposed to.
- Morphological insight: OA is true morphology. It can be partially motivated and then contribute to the definition of inflectionnal classes[4].

[3] Mark Aronoff (2019). “Competitors and Alternants in Linguistic Morphology.” In: *Competition in Inflection and Word-Formation.* Ed. by Franz Rainer et al. Studies in Morphology 5. Cham: Springer International Publishing, pp. 39–66. DOI: 10.1007/978-3-030-02550-2_2.

[4] Matías Guzmán Naranjo and Olivier Bonami (2021). “Overabundance and inflectional classification: Quantitative evidence from Czech.” In: *Glossa: a journal of general linguistics* 6.1 (1). DOI: 10.5334/gjgl.1626.

Canonicity

“The canonical approach requires clear definitions. We take these to their logical end points, in order to construct a theoretical space. The convergence of criteria fixes a canonical point from which the phenomena actually found can be calibrated.”[a]

[a] Greville G. Corbett (2008). “Determining morphosyntactic feature values: The case of case.” In: *Case and Grammatical Relations: Studies in honor of Bernard Comrie*. Ed. by Greville G. Corbett and Michael Noonan. Typological Studies in Language 81. Amsterdam: John Benjamins, pp. 1–34. DOI: `10.1075/tsl.81`, p. 4.

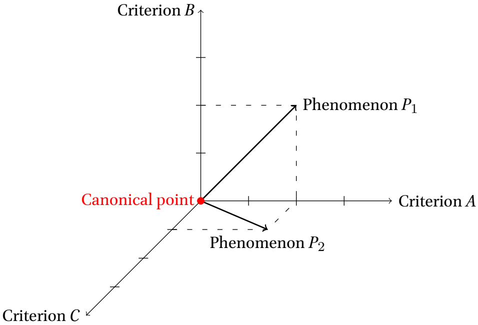


Figure: Representation of the “Theoretical space”

Canonicity and overabundance I

OA is an irregularity. Therefore a “canonical” kind of OA is...

- Exceptionnal and unpredictable (in the paradigm)
- Exceptionnal and unpredictable (in the lexicon)
- Perfectly balanced between the cell-mates
- Perfectly unmotivated

Canonicity and overabundance I

OA is an irregularity. Therefore a “canonical” kind of OA is...

- Exceptionnal and unpredictable (in the paradigm)
- Exceptionnal and unpredictable (in the lexicon)
- Perfectly balanced between the cell-mates
- Perfectly unmotivated

This defines the canonical point. It is probably not filled.

Canonicity and overabundance II

A more formal description by Thornton (2019) gives:

1. Uniqueness of cell (canon: 1 cell)
2. Uniqueness of lexeme (canon: 1 lexeme)
3. Frequency Ratio Between the Cell Mates (canon: 1:1 ratio)
4. Conditions of use (canon: no conditions)
   - Geo-socio-stylistic conditions
   - Grammatical conditions

Plan

Overabundance

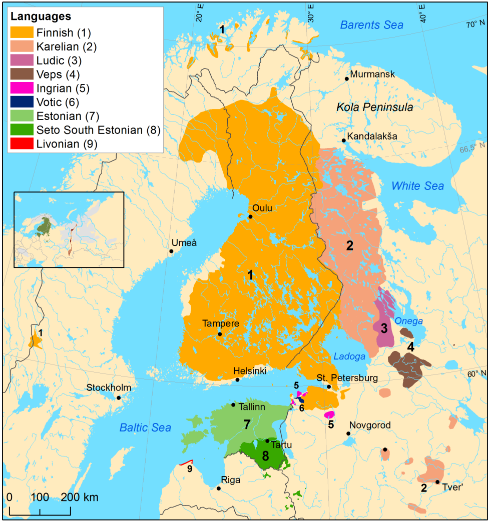
Estonian Morphology

The Corpus study

Questions

- Where does overabundance appear in Estonian?
- Why does Estonian overabundance appear to be interesting?

Language area[5]



[5] Timo Rantanen et al. (2021). *Geographical database of the Uralic languages 1.0*. Version v1.0. Zenodo. DOI: 10.5281/ZENODO.4784188.

Case system

- 14 cases:
  - 3 grammatical cases : nominative, genitive, partitive.
  - 6 local cases (see below)
  - 5 semantic cases : terminative, comitative, abessive, translative, essive.
- 2 numbers: singular and plural

	Coinitial	Static	Cofinal
Internal	Elative	Inessive	Illative
External	Ablative	Adessive	Allative

Table: Estonian locative case system

Main features

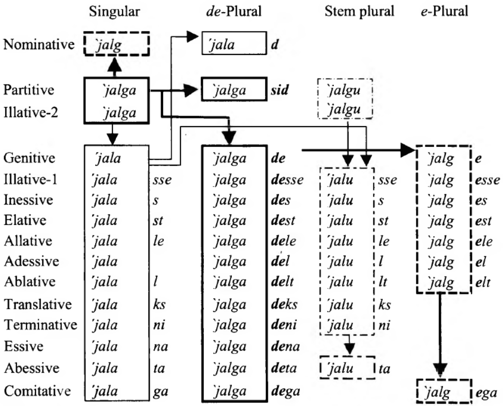
Commonly assessed features of Estonian morphology:

- “Agglutinative” features:
  - Semantic and local cases
  - -de/te plural
  - -sid partitive plural
  - + the verbal morphology
- “Fusional” features:
  - Grammatical cases
  - Short illative
  - -i plural
  - -i partitive plural

Endless debates about the typology of the language[6].

[6] Johanna Laakso (2021). “Language contact and typological change: The case of Estonian revisited.” In: *Word Structure* 14.2, pp. 226–245. DOI: 10.3366/word.2021.0188.

Paradigm and overabundance



	Singular		*de*-Plural		Stem plural		*e*-Plural	
Nominative	`jalg		´jala	d				
Partitive	`jalga		`jalga	sid	`jalgu			
Illative-2	`jalga				`jalgu			
Genitive	´jala		`jalga	de			`jalg	e
Illative-1	´jala	sse	`jalga	desse	´jalu	sse	`jalg	esse
Inessive	´jala	s	`jalga	des	´jalu	s	`jalg	es
Elative	´jala	st	`jalga	dest	´jalu	st	`jalg	est
Allative	´jala	le	`jalga	dele	´jalu	le	`jalg	ele
Adessive	´jala		`jalga	del	´jalu	l	`jalg	el
Ablative	´jala	l	`jalga	delt	´jalu	lt	`jalg	elt
Translative	´jala	ks	`jalga	deks	´jalu	ks		
Terminative	´jala	ni	`jalga	deni	´jalu	ni		
Essive	´jala	na	`jalga	dena				
Abessive	´jala	ta	`jalga	deta	´jalu	ta		
Comitative	´jala	ga	`jalga	dega			`jalg	ega

Figure: A (very) theoretical paradigm of *jalg* 'foot'[7]

[7] Mati Erelt (2003). *Estonian language*. Linguistica Uralica 1. Tallinn: Estonian Academy Publishers, p. 38.

Illative singular

The *short form*: the ancient one, resulting from phonetic erosion of **sen* (Proto-Finnic). Different barely predictable patterns :

- puu, maa, pää ~ puhu, maha, pähe (’tree’, ’earth’, ’head’)
- keel, meel ~ keelde, meelde (’language’, ’spirit, mind’)
- tuba, jalg, taevas ~ ‘tuppa, ‘jalga, ‘taeva (’room’, ’foot’, ’sky’)

The *new form* in *-sse*: results from analogy and reanalysis of older short illatives of sigmatic stems (Nom. hambas~ Ill. hambasse). Easy to segment : jala-sse, jalga-de-sse (foot.SG-ILL, foot-PL-ILL)

NB: ‘ marks a suprasegmental distinction [jalk] ~ [ˈjɑl:kɑ]

Partitive plural I

The *vocalic partitive* (also -i or stem partitive): reflects the old *-j/i* pluralizer, which is fused to the stem vowel (Gen. päeva ~ Part. päevi). The partitive ending **tA* was lost (phonetic erosion).

The *-id* partitive: reflects the old **tA* suffix under certain phonotactic conditions : hamba-id (tooth-PART.PL).

The *sigmatic partitive* in *-sid*: results from analogy and reanalysis of sigmatic stems (Gen. punase ~ Ill. *punasid). Easy to build on vocalic stems : maja-sid (house-PART.PL).

Partitive plural II

Partitive plural	maja ‘house’	sündmus ‘event’	kuu ‘month’
-id		sündmuseid	kuid
-sid	majasid		kuusid
vocalic	maju	sündmusi	

Table: Overabundance patterns for the PART.PL cell[8]

[8] Mari Aigro and Virve Vihman (forthcoming). “Realised overabundance in Estonian noun paradigms: A corpus study.” In: *Word Structure* 16.2. DOI: 10.3366/word.2023.0227, p. 160.

Semantic and local cases in plural

Built upon :

- The *genitive plural* form: *Gen.* jalga-de → *Ine.* jalga-de-s
- The *partitive plural* form: *Part.* ‘jalgu → *Ine.* jalu-s [with degree alternation]

A kind of “row-effect” : the whole plural paradigm may be affected by the choice of the baseform.
Question: consistency across semantic and local cases?

Other cases

- Genitive plural in -e : gen.pl kirjanike *vs.* kirjanikkude ‘writer’
- Plural cells for lexemes which can belong to two different inflectional classes : in.pl äärmusis *vs.* äärmuseis.
- Lexeme-specific overabundance phenomena : el.pl. kodunt *vs.* kodust ‘house’
- Partitive singular for some lexemes, when the declension pattern is not clear enough : part.sg aastasada *vs.* aastasadat ‘century’
- Stem-based variants: all.pl. päikestele *vs.* päiksetele ‘sun’

Why is this interesting ?

- Motivation of OA:
  - Internal: OA derives from other cells in the paradigm (local cases).
  - External: OA formatives derive from diachronic and diatopic variation (ill.sg, part.pl).
- Morphological status:
  - Class-specific: holds for some lexemes, can be used to define a class.
  - Cross-class: holds for all lexemes.

Plan

Overabundance

Estonian Morphology

The Corpus study

- Extent and distribution
- Patterns and frequency ratios
- Closing remarks

Research questions

Focus on the canonicity of Estonian OA :

- Extent and distribution of OA?
  - How many lexemes display OA in a given cell ?
  - In how many cells of a lexeme is there overabundance ?
- Patterns of usage and frequency ratios?
- Does syntactic function condition OA?

Methodological concerns I

- Previous study relied on a 500000 word corpus[9]. Not enough.
- Thus : *Balanced corpus of Estonian*, 15-million word corpus.
- Reason of this : attestation rate of some cells are really low, so you need a big corpus to be able to study the whole paradigm.

Methodological concerns II

Let *L* be a lexeme. How many occurrences of *L* do we need in the corpus, to have at least 2 occurrences for the *El.Pl* cell ?
Let's assume that the distribution of the cells is the same across the Lexicon:

$$\frac{f_{L.El.Pl}}{f_L} = \frac{f_{El.Pl}}{f_N}$$

$$f_L = \frac{f_{L.El.Pl} \cdot f_N}{f_{El.Pl}}$$

The Authors consider that at least 2 values are needed for a cell to decide whether it is overabundant or not.
Therefore :

$$f_{L.El.Pl} \geqslant 2 \quad \text{and} \quad f_L \geqslant \frac{2f_N}{f_{El.Pl}}$$

Methodological concerns III

In a given cell of a given lexeme, how many occurrences do we need to be *sure* that this paradigm cell is overabundant or not ?
If we have n occurrences $a_1, a_2, ..., a_n$ for a given cell of a given **overabundant** lexeme, the probability **that the attestations reflect this overabundance :**

$$P(\exists(i,j) \in \{1, 2, ..., n\} : a_i \neq a_j)$$

Let's note it $P(OA)$. We assume that there are only two competing forms A and B, with frequencies f_A and f_B, then according to Bernouilli & Poisson :

$$P(OA) = 1 - {f_A}^n - (1 - f_A)^n$$

Methodological concerns IV

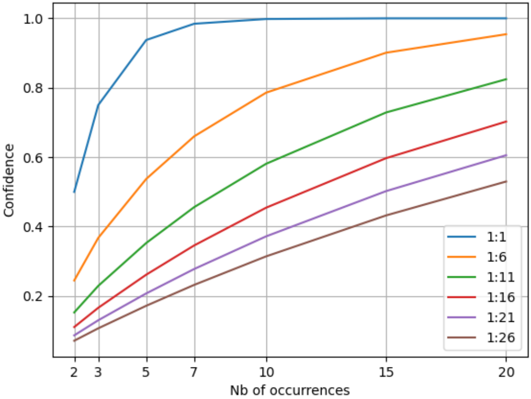


Figure: Confidence in the identification of overabundant lexemes for a given cell, as a function of the expected ratio and of the number of attestations

Methodological concerns V

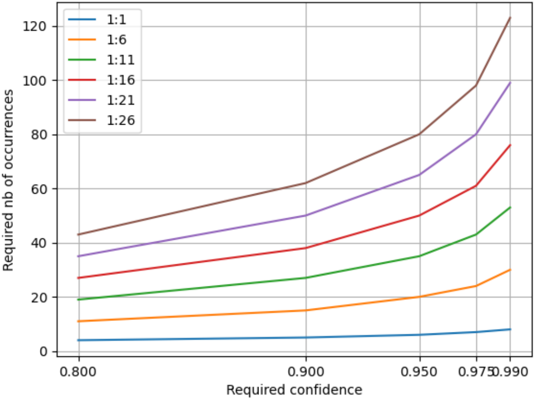


Figure: Required number of occurrences to be able to assess with a given confidence whether a given cell of a given lexeme has overabundance or not.

[9] Heiki-Jaan Kaalep (2010). “Mitmuse osastav eesti keele käändesüsteemis.” In: *Keel ja kirjandus* 2, pp. 94–111.

Preliminary conclusion I

Is this a problem ? Not really, as long as you are cautious with the results.

- This method leads to an underestimate of the overabundance rate at the scale of the lexemes.
- Frequency ratios will also probably be a little different (closer to 1:1, because less extreme cases are represented)

The only rule is that you should be very prudent when comparing these results.

Preliminary conclusion II

How to improve the statistical value of these results ?

- Take statistics into account when building the sub-corpus. But you will have less forms, especially with a little corpus.
- Use a really big corpus... (500000, 15000000, what's next ?) Consider the TERM.PL. Approximately 1700 attestations in the corpus, for all the lexemes... Very few lexemes will have more than two forms for this cell, whereas you need at least 6 forms (for a rate of 1:1 and a confidence of 0.95). It's hard to prove that a lexeme is overabundant for such a cell.
- Study languages with less paradigm cells (e.g. Czech).
- Consider OA per-inflection class rather than per-lexeme.

Preliminary conclusion III

What is “realised overabundance”?

- OA as immediately available in the corpus? (the approach of the Authors in this article)
- OA as it can be extrapolated from the corpus?

A corpus shows only a small part of the combinatorial potential of the language. Thus, probabilistic approaches are required to account for the flaws of the corpus.

The Corpus study

Extent and distribution

Results

NB: The possible OA is the percentage of lexemes that *could* display overabundant forms for a given cell. I used the Estonian paralex database[10] to compute these results over 5478 lexemes.

Case	Realised OA	Possible OA
elat.pl	8.35 %	89.16 %
ill.sg	5.49 %	62.10 %
ad.pl	5.21 %	89.16 %
all.pl	5.09 %	89.16 %
iness.pl	4.89 %	89.16 %
part.pl	4.09 %	37.11 %
trans.pl	3.92 %	89.16 %
abl.pl	3.88 %	89.16 %
ill.pl	3.07 %	89.16 %

Table: Realised OA as computed by Aigro & Vihman *versus* expected OA in Estonian paradigms.

[10] Sacha Beniamine et al. (Sept. 27, 2023). *Estonian Paradigms in Phonemic Notation.* Version v1.0.1. Zenodo. DOI: 10.5281/zenodo.8392744.

Remarks I

The Authors conclude what follows:

> *In terms of Uniqueness of Lexeme, the attested overabundance in the Estonian corpus is much more "canonically" deviant than potential overabundance: out of the tens of thousands of noun lemmas described as having parallel forms available, the corpus turned up just over a thousand lexemes with attested parallel forms in one cell or another.* (*Aigro and Vihman forthcoming, p. 172*)

Remarks II

This is a clear case where the results are underestimated.

- If we assume a Zipfian distribution :
  - A lot of lexemes with only 2 forms attested in the given cell.
  - Few lexemes with a lot of attestations
- In the first case, the decision to classify a lexeme as non-overabundant is highly questionable.
- Unfortunately, the Authors do not give enough information to quantify the uncertainty on these results.

Conclusion : a consolidated Realised OA-rate would probably be somewhere between the observed Realised OA-rate and the Expected OA.

Types of overabundance

Case	Realised OA	Possible OA
elat.pl	8.35 %	89.16 %
ill.sg	5.49 %	62.10 %
ad.pl	5.21 %	89.16 %
all.pl	5.09 %	89.16 %
iness.pl	4.89 %	89.16 %
part.pl	4.09 %	37.11 %
trans.pl	3.92 %	89.16 %
abl.pl	3.88 %	89.16 %
ill.pl	3.07 %	89.16 %

Table: Realised OA as computed by Aigro & Vihman *versus* expected OA in Estonian paradigms.

Nb of overabundant cells

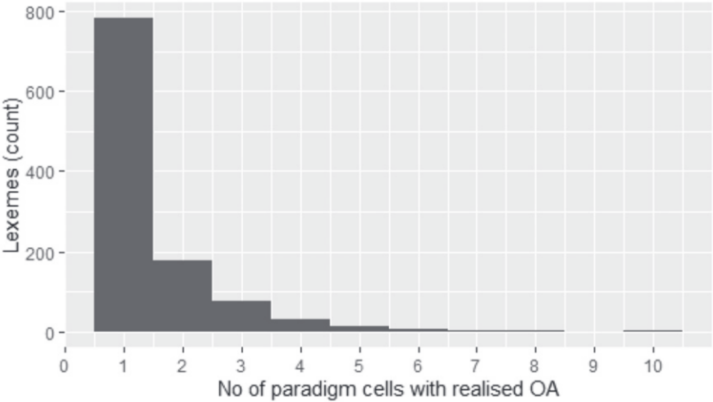


Figure: Table given by Aigro & Vihman

Again, probably an underestimate.

The Corpus study

Patterns and frequency ratios

Long *vs.* short forms I

- Productive forms (long forms) should be overwhelming for unfrequent lexemes.
- Unproductive forms (short forms) should be available for frequent lexemes only.

Same behaviour expected for all cases.

	Long pattern	Short pattern
Ill.sg	-sse	short illative
Part.pl	-sid	-u, -i, -e, etc.
All.pl	te/de-le	-i/u/a-le
Ad.pl	te/de-l	-i/u/a-l
Ela.pl	te/de-st	-i/u/a-st
In.pl	te/de-s	-i/u/a-s

Table: Formatives used in the different overabundant cells

Long *vs.* short forms II

	Long pattern	Short pattern
Ill.sg	-sse	short illative
Part.pl	-sid	-u, -i, -e, etc.
All.pl	te/de-le	-i/u/a-le
Ad.pl	te/de-l	-i/u/a-l
Ela.pl	te/de-st	-i/u/a-st
In.pl	te/de-s	-i/u/a-s

Table: Formatives used in the different overabundant cells

Long *vs.* short forms III

	Long pattern	Short pattern
Ill.sg	24 %	76 %
Part.pl	37 %	62 %
All.pl	94 %	6 %
Ad.pl	76 %	24 %
Ela.pl	94 %	6 %
In.pl	93 %	7 %

Table: Formatives used in the different overabundant cells

Attestation rates I

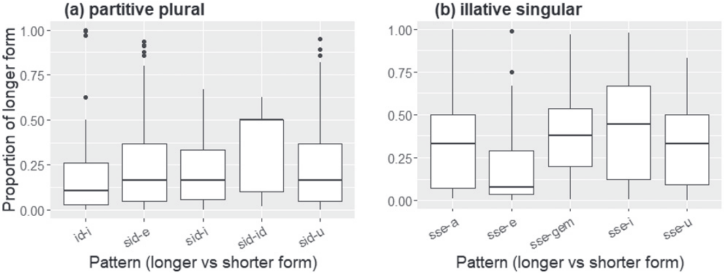


Figure: OA-ratio as computed by Aigro & Vihman for the most frequent patterns of each cell

Attestation rates II

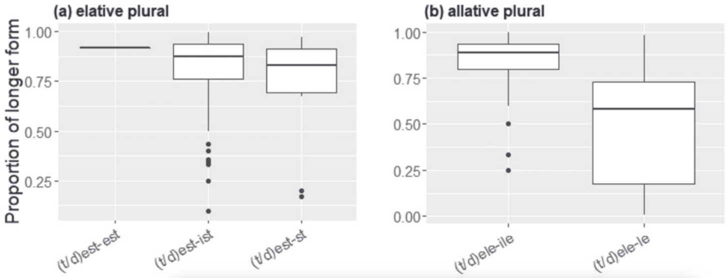


Figure: OA-ratio as computed by Aigro & Vihman for the most frequent patterns of each cell

The Corpus study

Closing remarks

Closing remarks I

- Methodological concerns: bear in mind that the quality of the results clearly depends on the probabilistic approach.
- Results are interesting, providing :
  - An overview of the patterns of overabundance and of their distribution.
  - A clear-cut distinction in the behaviour of local cases *vs.* part.pl and ill.sg.
  - A temporary proof that there is no good syntactic conditioning (yet).

Closing remarks II

Further ideas of the Authors:

- Semantic coherence of local cases ?
- Elative, Allative and Ablative display a high rate of OA *and* are highly used as verbal arguments.
- Further studies needed to understand Conditions of Use.
- + Focusing on the user, study the correlation between the Part.pl and the forms of the local cases ? (implicative patterns in OA).

Literature I

Aigro, Mari and Virve Vihman (forthcoming). “Realised overabundance in Estonian noun paradigms: A corpus study.” In: *Word Structure* 16.2. DOI: 10.3366/word.2023.0227.

Aronoff, Mark (2019). “Competitors and Alternants in Linguistic Morphology.” In: *Competition in Inflection and Word-Formation*. Ed. by Franz Rainer et al. Studies in Morphology 5. Cham: Springer International Publishing, pp. 39–66. DOI: 10.1007/978-3-030-02550-2_2.

Beniamine, Sacha et al. (Sept. 27, 2023). *Estonian Paradigms in Phonemic Notation*. Version v1.0.1. Zenodo. DOI: 10.5281/zenodo.8392744.

Corbett, Greville G. (2008). “Determining morphosyntactic feature values: The case of case.” In: *Case and Grammatical Relations: Studies in honor of Bernard Comrie*. Ed. by Greville G. Corbett and Michael Noonan. Typological Studies in Language 81. Amsterdam: John Benjamins, pp. 1–34. DOI: 10.1075/tsl.81.

Erelt, Mati (2003). *Estonian language*. Linguistica Uralica 1. Tallinn: Estonian Academy Publishers.

Kaalep, Heiki-Jaan (2010). “Mitmuse osastav eesti keele käändesüsteemis.” In: *Keel ja kirjandus* 2, pp. 94–111.

Literature II

Laakso, Johanna (2021). “Language contact and typological change: The case of Estonian revisited.” In: *Word Structure* 14.2, pp. 226–245. DOI: 10.3366/word.2021.0188.

Naranjo, Matías Guzmán and Olivier Bonami (2021). “Overabundance and inflectional classification: Quantitative evidence from Czech.” In: *Glossa: a journal of general linguistics* 6.1 (1). DOI: 10.5334/gjgl.1626.

Rantanen, Timo et al. (2021). *Geographical database of the Uralic languages 1.0*. Version v1.0. Zenodo. DOI: 10.5281/ZENODO.4784188.

Thornton, Anna M. (2011). “Overabundance (Multiple Forms Realizing the Same Cell): A Non-canonical Phenomenon in Italian Verb Morphology.” In: *Morphological autonomy: perspectives from Romance inflectional morphology*. Ed. by Martin Maiden et al. Oxford ; New York: Oxford University Press, pp. 358–381.

— (2019). “Overabundance: A Canonical Typology.” In: *Competition in Inflection and Word-Formation*. Ed. by Franz Rainer et al. Studies in Morphology 5. Cham: Springer, pp. 223–258. DOI: 10.1007/978-3-030-02550-2_9.