Realised overabundance in Estonian noun paradigms: A corpus study
A forthcoming article by Aigro & Vihman

October 11, 2023
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.*

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.*
What is overabundance? II

Example:

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 fázok / fázom</td>
<td>fázunk</td>
</tr>
<tr>
<td>P2 fázol</td>
<td>fáztok</td>
</tr>
<tr>
<td>P3 fázik</td>
<td>fáznak</td>
</tr>
</tbody>
</table>

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


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.
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 differentiation 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\(^4\).

---


“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.”

---


**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
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.
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

---

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

<table>
<thead>
<tr>
<th>Coinitial</th>
<th>Static</th>
<th>Cofinal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>Elative</td>
<td>Inessive</td>
</tr>
<tr>
<td>External</td>
<td>Ablative</td>
<td>Adessive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Illative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allative</td>
</tr>
</tbody>
</table>

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\(^6\).

Paradigm and overabundance

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

---

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, päre ('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] ~ [jalːka]
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

<table>
<thead>
<tr>
<th>Partitive plural</th>
<th>maja (‘house’)</th>
<th>sündmus (‘event’)</th>
<th>kuu (‘month’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-id</td>
<td>mäid</td>
<td>sündmuseid</td>
<td>kuid</td>
</tr>
<tr>
<td>-sid</td>
<td>majasid</td>
<td>kuusid</td>
<td></td>
</tr>
<tr>
<td>vocalic</td>
<td>maju</td>
<td>sündmusi</td>
<td></td>
</tr>
</tbody>
</table>

**Table: Overabundance patterns for the PART.PL cell**

---

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 500,000 word corpus\textsuperscript{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:

\[
f_{L,El.Pl} \geq 2 \quad \text{and} \quad f_L \geq \frac{2f_N}{f_{El.Pl}}
\]
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 \( a_1, a_2, \ldots, a_n \) for a given cell of a given overabundant lexeme, the probability that the attestations reflect this overabundance:

\[
P(\exists (i, j) \in \{1, 2, \ldots, n\} : a_i \neq a_j)
\]

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.

---

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\textsuperscript{10} to compute these results over 5478 lexemes.

<table>
<thead>
<tr>
<th>Case</th>
<th>Realised OA</th>
<th>Possible OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>elat.pl</td>
<td>8.35 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>ill.sg</td>
<td>5.49 %</td>
<td>62.10 %</td>
</tr>
<tr>
<td>ad.pl</td>
<td>5.21 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>all.pl</td>
<td>5.09 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>iness.pl</td>
<td>4.89 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>part.pl</td>
<td>4.09 %</td>
<td>37.11 %</td>
</tr>
<tr>
<td>trans.pl</td>
<td>3.92 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>abl.pl</td>
<td>3.88 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>ill.pl</td>
<td>3.07 %</td>
<td>89.16 %</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Case</th>
<th>Realised OA</th>
<th>Possible OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>elat.pl</td>
<td>8.35 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>ill.sg</td>
<td>5.49 %</td>
<td>62.10 %</td>
</tr>
<tr>
<td>ad.pl</td>
<td>5.21 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>all.pl</td>
<td>5.09 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>iness.pl</td>
<td>4.89 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>part.pl</td>
<td>4.09 %</td>
<td>37.11 %</td>
</tr>
<tr>
<td>trans.pl</td>
<td>3.92 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>abl.pl</td>
<td>3.88 %</td>
<td>89.16 %</td>
</tr>
<tr>
<td>ill.pl</td>
<td>3.07 %</td>
<td>89.16 %</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>Long pattern</th>
<th>Short pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ill.sg</td>
<td>-sse</td>
<td>short illative</td>
</tr>
<tr>
<td>Part.pl</td>
<td>-sid</td>
<td>-u, -i, -e, etc.</td>
</tr>
<tr>
<td>All.pl</td>
<td>te/de-le</td>
<td>-i/u/a-le</td>
</tr>
<tr>
<td>Ad.pl</td>
<td>te/de-l</td>
<td>-i/u/a-l</td>
</tr>
<tr>
<td>Ela.pl</td>
<td>te/de-st</td>
<td>-i/u/a-st</td>
</tr>
<tr>
<td>In.pl</td>
<td>te/de-s</td>
<td>-i/u/a-s</td>
</tr>
</tbody>
</table>

Table: Formatives used in the different overabundant cells

Realised overabundance in Estonian noun paradigms: A corpus study
Long vs. short forms II

<table>
<thead>
<tr>
<th></th>
<th>Long pattern</th>
<th>Short pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ill.sg</td>
<td>-sse</td>
<td>short illative</td>
</tr>
<tr>
<td>Part.pl</td>
<td>-sid</td>
<td>-u, -i, -e, etc.</td>
</tr>
<tr>
<td>All.pl</td>
<td>te/de-le</td>
<td>-i/u/a-le</td>
</tr>
<tr>
<td>Ad.pl</td>
<td>te/de-l</td>
<td>-i/u/a-l</td>
</tr>
<tr>
<td>Ela.pl</td>
<td>te/de-st</td>
<td>-i/u/a-st</td>
</tr>
<tr>
<td>In.pl</td>
<td>te/de-s</td>
<td>-i/u/a-s</td>
</tr>
</tbody>
</table>

Table: Formatives used in the different overabundant cells

Realised overabundance in Estonian noun paradigms: A corpus study
Long vs. short forms III

<table>
<thead>
<tr>
<th></th>
<th>Long pattern</th>
<th>Short pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ill.sg</td>
<td>24 %</td>
<td>76 %</td>
</tr>
<tr>
<td>Part.pl</td>
<td>37 %</td>
<td>62 %</td>
</tr>
<tr>
<td>All.pl</td>
<td>94 %</td>
<td>6 %</td>
</tr>
<tr>
<td>Ad.pl</td>
<td>76 %</td>
<td>24 %</td>
</tr>
<tr>
<td>Ela.pl</td>
<td>94 %</td>
<td>6 %</td>
</tr>
<tr>
<td>In.pl</td>
<td>93 %</td>
<td>7 %</td>
</tr>
</tbody>
</table>

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

- 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).
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


