Eesthetic: A Paralex Lexicon of Estonian Paradigms

Sacha Beniamine[1], **Mari Aigro**[2], **Matthew Baerman**[1], **Jules Bouton**[3] **and Maria Copot**[4]
[1]University of Surrey, Surrey Morphology Group, Guildford, United Kingdom
[2]University of Tartu, Tartu, Estonia
[3]Université Paris Cité, Laboratoire de Linguistique Formelle, CNRS, France
[4]The Ohio State University, Columbus, USA
[1]{s.beniamine, m.baerman}@surrey.ac.uk, [2]mari.aigro@ut.ee,
[3]jules.bouton@u-paris.fr, [4]maria.copot.s@gmail.com

Abstract

We introduce Eesthetic, a comprehensive Estonian noun and verb lexicon sourced from the Ekilex database. It documents 5475 nouns inflecting for 28 paradigm cells and 5076 verbs inflecting for 51 cells, and comprises a total of 452885 inflected forms. Our openly accessible machine-readable dataset adheres to the Paralex standard. It comprises CSV tables linked by formal relationships. Metadata in JSON format, following the Frictionless standard, provides detailed descriptions of the tables and dataset. The lexicon offers extensive linguistic annotations, including orthographic forms, automatically transcribed phonemic transcriptions, non-canonical morphological phenomena such as overabundance and defectiveness, rich mapping of the paradigm cells and feature-values to other notation schemes, a decomposition of phonemes in distinctive features, and annotation of inflection classes. It is suited for both monolingual and comparative research, enabling qualitative and quantitative analysis. This paper outlines the creation process, rationale, and resulting structure, along with our set of rules for automatic grapheme to phoneme (g2p) conversion.

Keywords: Estonian, morphology, paradigms, lexicon, inflection

1. Introduction

We present Eesthetic[1], a large open inflected lexicon of Estonian[2] nouns and verbs derived from Ekilex (Kallas et al., 2022). It documents 5475 nouns inflecting for 28 paradigm cells and 5076 verbs inflecting for 51 cells, and comprises a total of 452885 inflected forms given in orthographic and phonemic (automatically transcribed) notation. The dataset is structured and formatted as a set of CSV files following the emerging Paralex standard (Beniamine et al., 2023). It is suited for both qualitative and quantitative investigations into the inflectional properties of the Estonian language.

Over the past decade, studies of inflectional morphology have increasingly turned towards computational methods in order to precisely measure typological variables such as predictability, complexity, inflection class, and analogical structure (Stump and Finkel, 2013; Ackerman and Malouf, 2013; Bonami and Beniamine, 2016; Sims and Parker, 2016; Naranjo and Bonami, 2021; Pellegrini, 2020; Copot and Bonami, in press, among others). Progress in this field is still held back by the limited availability of data documenting full inflectional systems. In particular, it remains impossible to accurately assess distributional properties of these measurements across the world's languages, because the existing datasets (e.g. for Latin Pellegrini and Passarotti, 2018; Oto-Manguean languages Feist and Palancar, 2015) are overwhelmingly restricted to a non-representative sample in which certain languages or language families are overrepresented (e.g. Russian Beniamine and Brown, 2019), and in particular Romance (e.g. for French: Bonami et al., 2014, Italian: Pellegrini, 2020; Portuguese: Perdigão et al., 2021; Spanish: Herce, 2023). This inflected lexicon of Estonian, a Uralic language, diversifies the set of available resources.

Machine-readable lexicons of inflected forms are also invaluable for data-driven studies of individual systems. The rich, sizeable inflectional system of Estonian shows complexity in many dimensions, on the one hand mixing fusional and agglutinative elements (Laakso, 2021), and on the other hand including numerous instances where the system deviates from canonical expectations of inflection. These include multiple inflection classes and extensive stem alternations (Viks, 1992; Blevins, 2007, 2008), a three grade syllable quantity (and length) system involved in inflectional exponence (Ehala, 2003; Viht and Habicht, 2019), and pervasive 'overabundance' (parallel forms in cells; Thornton, 2011; Aigro and Vihman, in press) . Such a dataset is needed in studies aiming to quantitatively describe and explain these phenomena.

Moreover, inflectional systems are known for being the site of complex analogical change, a

[1]Archived on Zenodo under the DOI https://zenodo.org/doi/10.5281/zenodo.8383522
[2]ISO 639-3 code `ekk` ; Glottocode `esto1258`;

LREC-COLING 2024, pages 5526–5537
20-25 May, 2024.

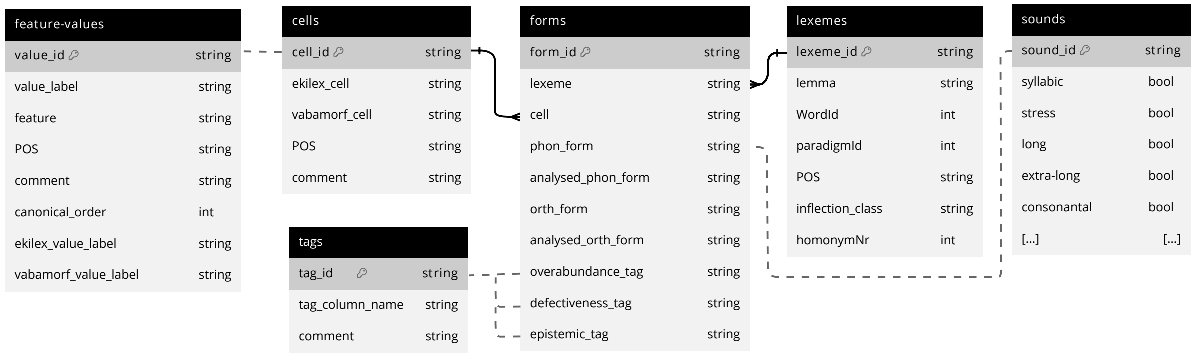

Figure 1: Relational schema of the dataset. Foreign key relations are indicated with a plain line, and vocabulary relations by a dashed line. Primary keys are indicated by a darker background and a key symbol. The list of columns of the sounds file is truncated.

phenomenon which confounds the results of the traditional comparative method of historical linguistics (Hock, 2021, chapter 6). Thus, large inflected datasets provide opportunities to investigate the analogical structure and evolution of inflectional systems. Finally, this lexicon can constitute a useful resource for any NLP applications which contend with Estonian morphology.

In summary, our contributions are:

- A large lexicon of Estonian inflectional morphology, covering both Nouns and Verbs, and following the Paralex standard.
- Rich linguistic annotation meant to maximize reusability and interoperability;
- Generated phonological forms, provided along with orthographic ones;
- A set of linguistically motivated rules for the phonemic transcription of Estonian.

2. Related work and applications

This work adds to the set of large lexicons of inflected forms available for both computational and qualitative linguistic analysis (see Introduction). To the best of our knowledge, no such lexicons previously existed for Estonian nouns and verbs, even though Estonian benefits from a very rich data ecosystem. The Ekilex dictionary system (Kallas et al., 2022) combines and integrates numerous lexical databases and dictionaries (Langemets, 2020; Eesti Keele Instituut, 2019, 2014, 2023, among others), serving them through a dedicated API and online interface. Its morphological component (Eesti Keele Instituut, 2023), accessed through the REST API, served as the basis for this work (see § 3). For frequencies, we relied on the Estonian National Corpus 2021 (Kallas and Koppel, 2022), the largest available text collection in Estonian. We used the Vabamorf tools (Filosoft, 2015) for morphological generation and analysis.

The only somewhat similar dataset for Estonian exists as part of the set of Unimorph lexicons (Kirov et al., 2016, 2018; McCarthy et al., 2020; Batsuren et al., 2022). Although these are widely used for NLP applications, they are not suitable for use in linguistic investigation without further work-intensive processing and annotation (Malouf et al., 2019). One of the issues is that these datasets are often automatically extracted from Wiktionary, which has consequences on both data quality and homogeneity. In addition, they only provide triplets of cell, lexeme and orthographic form, leaving out crucial phonological information. Finally, they do not annotate non-canonical phenomena such as overabundance and defectiveness in reliable ways. The current work therefore constitutes a much richer resource with more detailed annotation (see Figure 1), in addition to being two orders of magnitude larger (10551 vs 886 lexemes in the Unimorph Estonian lexicon).

3. Building the lexicon

We built the lexicon in four steps: (1) we selected the most frequent 5000 nouns and verbs (§3.1); (2) we extracted the corresponding lexemes and their paradigms from the Ekilex API (§3.2); (3) we enriched their orthographic representations using the Vabamorf software (§3.3); and (4) we generated phonemic transcriptions for all forms (§3.4).

3.1. Frequency measurement

We counted nominal and verbal lemma frequency from the National Estonian Corpus 2021 (Kallas and Koppel, 2022)[3]. These are used in order to select 5000 lemmas in each category, as described in the next section. and selected the most frequent 5000 entries for each part of speech.

[3]The Web sections were excluded as they are generally noisier, leading to poorer lemmatisation.

3.2. Paradigm extraction

We queried the Ekilex API endpoints `/api/word/search/` (to obtain word identifiers `.[].wordId`) and `word/details/` in order to select the most frequent lemmas in each category. We excluded any words explicitly marked as foreign, prefixes, suffixes, and very short words (often letters). We obtain a list of 5000 lemmas for each category. Due to homonyms, this is more than 5000 distinct lexemes (distinguished by separate word IDs).

We then queried `paradigm/details/` in order to obtain the inflected forms of each lexeme. We filtered out uninflected words. For each word ID, the paradigm details endpoint returns a list of distinct paradigms with unique paradigm IDs (`.[].id`). We created a lexeme for each combination of a word ID and a paradigm IDs (i.e. flexemes in the sense of Fradin and Kerleroux 2003; Thornton 2018; Pellegrini 2023). This is particularly relevant to words belonging to multiple inflection classes, where each distinct paradigm leads to a separate flexeme. Each paradigm presents a list of forms (`.[].forms[]`) which constitute the basis for our dataset (see § 4). Of particular interest are the orthographic form (`.[].forms[].value`) and the annotated orthographic form (`.[].forms[].displayForm`). The latter includes diacritic markers indicating phenomena which are not predictable from the sole orthographic form, such as the quantity system which distinguishes three syllable quantities (see § 3.4 for details). A grave accent (<`>) is placed before the vowels of syllables in quantity 3 (exs 1, 4). An acute accent (<´>) marks unpredictable stressed syllables (ex 1). Both of these diacritics are in parentheses if their realization can be omitted. An opening square bracket marks the boundary between the stem and inflectional ending (ex 2). A plus sign indicates word-internal boundaries in compounds (ex 3), while a subscript plus sign does so specifically in loan words. A straight apostrophe marks the palatalization of the previous consonant (ex 4).

(1) <´oman`ik> ['omɑnʲ'ikː] 'owner'[4]

(2) <sõna[sse> [sɤnɑsːe] 'word' ILL.SG

(3) <h`oo+`aeg> [h'oːː'ɑ͜eːk] 'season'

(4) <k`ot't> [k'otʲːː] 'bag'

3.3. Vabamorf annotations

The annotated orthographic form is very useful for phonemic transcription as it provides information missing from plain orthography. However, some of these marks are occasionally missing in Ekilex forms. In order to fill in missing information, we used the Vabamorf `etsyn` software. It takes part of speech, orthographic lemma, desired paradigm cell and plain orthographic form as input, outputting a list of annotated forms, marked in a notation which we converted to the Ekilex system described above. If one of these forms matched the plain orthographic form from Ekilex, we merged them. For most symbols, we simply added marking from both sources. For phonological quantity, we selected the form with the richest annotation. If the two sources disagreed on quantity 3, we selected the Ekilex version. A few examples are given in Table 1.

3.4. Phonological transcription

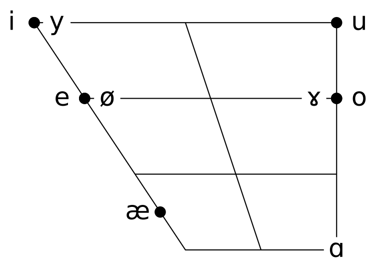

Figure 2: IPA vowels used in the lexicon.

The most work-intensive step was the elaboration and refinement of rules for (annotated) orthography to phonemic transcription.

Estonian phonology is characterized by a system involving three syllable "quantity" levels: short (quantity 1), long (quantity 2) and extra-long (quantity 3) (hereafter Q1, Q2 and Q3 Asu et al., 2016). The three levels can be expressed on most consonants (listed in Table 2) and all vowels (Figure 2). The system comprises a large number of diphthongs (Table 3), which by default count as long, and can all be Q3. In most cases, we write long characters with one length marker (eg. [ɑː], [pː]) and extra-long with two (eg. [ɑːː], [pːː]). However, diphthongs and cluster initial consonants in Q3 are marked only with a single length marker (eg. [ɑ͜uː]) as they are by default in quantity 2. The exact nature of the quantity system, and in particular, the details of its realization (involving pitch patterns and length ratios) and relevant prosodic units (feet, syllables or mora, see Prillop, 2013, 2020) are still debated (Asu et al., 2016, p 131). The quantity system combines with a stress system: most native words are stressed on the first syllable (with Q3 syllables always receiving primary stress), but borrowed words may retain their original stress patterns while inflectional morphology may alter primary stress. The distribution of secondary stress

[4]When not specified, nominal examples are given in the nominative singular, and verbal examples in the infinitive.

lemma	gloss	cell	Vabamorf	Ekilex	merged
klattima	‘resolve’	ind.prs.impers	klat'itakse	klati[takse	klat'i[takse
implanteerima	‘implant’	ind.prs.2pl	implant´eerite	implanteeri[te	implant´eeri[te
ekshumeerima	‘exhume’	ind.pst.ipfv.1pl	\`ekshum\`eerisime	eks+hum\`eeri[sime	\`ekshum\`eeris[ime

Table 1: Examples of discrepancies and merged orthographic forms between Ekilex and Vabamorf.

	bilabial	labio-dental	alveolar	post-alv.	palatalized	palatal	lab.-velar	velar	glottal
stop	p pː pːː		t tː tːː		tʲ tʲː tʲːː			k kː kːː	
nasal	m mː mːː		n nː nːː		nʲ nʲː nʲːː			ŋ ŋː	
trill			r rː						
fricative		f fː fːː v vː vːː	s sː sːː z	ʃ ʃː ʃːː	sʲ sʲː sʲːː				h hː
approximant						j	w		
lateral			l lː lːː		lʲ lʲː				

Table 2: IPA consonants used in the lexicon.

	e	i	u	o	ɑ
i	ie̯		iu̯	io̯	iɑ̯
y		yi̯			yɑ̯
u	ue̯	ui̯		uo̯	uɑ̯
e		ei̯	eu̯	eo̯	eɑ̯
ø	øe̯	øi̯			øɑ̯
ɤ	ɤe̯	ɤi̯	ɤu̯	ɤo̯	
o	oe̯	oi̯	ou̯		oɑ̯
æ	æe̯	æi̯	æu̯	æo̯	
ɑ	ɑe̯	ɑi̯	ɑu̯	ɑo̯	

Table 3: IPA diphthongs used in the lexicon.

is more varied and complex. For practical reasons, our transcription notes primary stress only (using the IPA symbol [']). Finally, Estonian presents both lexical and phonologically conditioned palatalization of some consonants: we took great care to identify and annotate these where relevant.

We rely on the Epitran (Mortensen et al., 2018) software to perform the transcription from orthography to IPA. We devised our own set of transcription rules. Epitran proceeds in three steps, which are applied independently on input forms: (1) a set of rules (§ 3.4.1); (2) a set of non-contextual mappings from orthography to IPA (§ 3.4.2) and (3) a second set of rules (§ 3.4.3). Epitran rules are ordered and written in custom syntax resembling traditional phonological rules, employing both variables and regular expressions. We grouped them by numbered blocks for readability purposes. In addition, we implemented some pre- and post-processing to handle steps which require information beyond single forms (§ 3.4.4).

3.4.1. Initial rules

The first four rules define variables for vowels, consonants, diacritic markers and voiced sounds. We then remove parentheses, and transcribe <š>, <ž>[5] and intervocalic <sh> as [ʃ], postvocalic <ck> as [kk], and <n> before <k> or <g> as [ŋ].

The next set of rules address Q3. It is often not predictible from the orthography, which only marks it on stop consonants. For example, it is not marked in the written forms in ex 5. Thankfully, the annotation Ekilex provides for orthographic forms unequivocally indicates which part of a word is lengthened under Q3, if any. However, empirically this information is less straightforward than it seems, as it it notably “difficult to define which segments in a Q3 syllable have increased duration” (Prillop, 2020, p.155).

(5) aasa ‘loop’
 a. **Q2**: GEN.SG <aasa> [ɑːsɑ]
 b. **Q3**: PART.SG <aasa> ['ɑːːsɑ]

Rule blocks 9 to 11 focus on making the notation less ambiguous. Rule block 9 adds Q3 to monosyllabic words (Asu and Teras, 2009). For example <aeg> ‘time’ is rewritten as <‘aeg>. Rule block 10 rewrites double orthographic consonants <pp>, <tt> and <kk> as Q3 (ex 6) and rule block 11 does so for <ff> and [ʃʃ] (ex 7, see Pajusalu, 2009).

(6) a. õppur → õp::ur ‘learner’
 b. k\`ot't → k\`ot'':: ‘bag’
 c. pikkus → pik::us ‘length’

(7) a. tuʃʃ → tuʃ:: ‘shower’
 b. proff → prof:: ‘pro’

Rule block 12 uses certain geminates to mark Q2. This is done for the characters <p>, <t>, <k>, <ʃ>, and <f> in voiced environements, as well as

[5]Although officially [ʒ] is not considered part of the Estonian phonemic inventory (Asu et al., 2016, p 65), the orthographic <ž> may occasionally be pronounced as a voiced [ʒ] rather than a voiceless [ʃ]. However, this is a peripheral phenomenon and was therefore left out of the rules.

the double consonants <ss>, <mm>, <nn>, <rr>, <vv> and <ll> (exs 8, see Asu and Teras, 2009).

(8) a. lepe → lepːe 'agreement'
b. mõte → mõtːe 'thought'
c. info → infːo 'info'
d. `aadr`es's → `aadr`es'ː 'address'
e. tunnus → tunːus 'attribute'

Rule 13 addresses palatalization, converting the straight apostrophy <'> in Ekilex annotation to [ʲ]. Rule block 14 palatalizes <s>, <n>, <l>, and <t> before <i> or <j> (ex 9). In clusters, only the first consonant palatalizes. There are a few exception to this, see § 3.4.4.

Rule block 15 changes the notation of long vowels from doubling <aa> to the length symbol <aː> (ex 10). Rule 16 (ex 10) diphthongizes syllable final long <ü> to [üi] when the next syllable starts with <a>, <e> or <u> (Asu and Teras, 2009).

(9) masin → masʲin 'machine'
(10) m`üüa → m`üːa → m`üia 'sell'

Finally, rule blocks 17 and 18 handle glides. Rule 17 adds a <j> after a long <i> and before another vowel (ex 11, see Asu and Teras, 2009)). Rule 18 does the same for <u>, adding <w> before another vowel (ex 12).

(11) l`aius → l`aijus 'width'
(12) l`uːa → l`uːwa 'create'

3.4.2. Mapping

This step transcribes the vowels <a>, <ä>, <õ>, <ö>, <ü> as respectively [ɑ], [æ], [ɤ], [ø], [y]; the stress mark <´> as the standard IPA [']; the letter <c> as [k] and the consonants <b>, <d>, <g> as the unvoiced [p], [t], [k]. Hereafter, the notation is phonemic, although a number of phenomena remain to be explicitly marked, and diacritic markers persist for segmentation and quantity in the input to the post-mapping rules.

3.4.3. Post-mapping rules

The first four post-mapping rules define variables for vowels, consonants, diacritic markers, and diphthongs. Rule 5 rewrites [jj] to [ij]. Rules 6 to 14 add ligatures to diphthongs, ensuring that the two vowels do not follow a vowel and do not overlap a sequence already marked as a diphthong. The set of these diphthongs is given in Table 3.

Rule blocks 15 to 22 determine which segments are lengthened in Q3 forms. They target consonants, vowels and diphthongs (in that order), being organized from the most specific to most general. Rule block 15 lengthens [s] when they occur between [l], [m], [n], or [r] and be rs] and a voiceless consonant (ex 13). Rule block 16 lengthens [p], [k], [t], [ʃ] and [f] after a vowel and [l], [m], [n], [r] and [ŋ] as in ex 14. Rule 17 lengthens the voiced consonants [l], [m], [n], [r] and [ŋ], before [p], [t], [k], [j] and [s] as in ex 15 (Saagpakk, 1982). Rule block 18 lengthens voiceless consonants at the start of a cluster (ex 16). Rule block 19 lengthens the consonants [k], [p], [t], [f], [ʃ] and [s] after a long vowel or a diphthong (ex 17). Finally, rule block 20 lengthens the first of any two consonant cluster in intervocalic position, as in ex 18 (Prillop, 2013). Rule block 21 lengthens diphthongs in a Q3 syllable if any, otherwise long vowels (ex 19). Rule 22 targets the first consonant after the vowel, when no other sound was lengthened (ex 20).

(13) n`orskan → n'orsːkan 'to snore'
(14) l`ɑmpː → l'ɑmpːː 'lamp'
(15) `ɑlpum → 'ɑlːpum 'album'
(16) a. l`ipp → l'ipːː 'banner'
b. j`uht → j'uhːt 'leader'
(17) j`oːkː → j'oːkːː 'drink'
(18) k`orv → k'orːv 'basket'
(19) a. l`ɑ͜ul → l'ɑ͜uːl 'song'
b. h`ɑːv → h`ɑːːv 'wound'
(20) a. p`anː → p'anːː 'pan'
b. s`urːa → s'urːːa 'die'
c. kol`umn → kol'umːn 'column'

The last rule blocks handle typographic corrections. Rule block 23 ensures that palatalization is written before length. Rule 24 ensures there is at most one stress mark per syllable. Rule 25 changes all segmentation symbols to "+".

3.4.4. Adjustments

Epitran rules operate solely on the lightly annotated orthographic forms from Ekilex, with no external information. However, this is not fully sufficient for determining actual phonological forms. Thus, we refined the final transcription in the ways described below.

We identified a number of exceptions to regular palatalization (pre-mapping rule 14). First, palatalization is blocked in a number of loans (m'uster, m'ustri, 'pattern') . We annotated potential cases (n = 144) of these exceptions by hand. Second, we automatically extracted lexemes where stems were combined with derivational suffixes containing <i> (<-mine>, <-vik>, <-nik>, <-mik>) which never trigger palatalization (k'uulmine, 'listening'). To block rule pre-mapping 14 from matching, we temporarily insert a dummy character (§).

In some lexemes, the stem-final consonant palatalizes throughout the paradigm, even though only some surface forms present the conditioning context in <-i> (m'ün't, mün'di 'coin'). We identify and palatalize them automatically.

We also make six ad-hoc manual changes to lexemes where palatalization is blocked in only some places (such as *filmifestival*, where the <l> does not palatalize) or where the Ekilex data presented typographic mistakes.

3.5. Evaluation

A development set of 102 forms with both annotated orthography and expected phonemic transcription was extracted from Asu and Teras (2009); Prillop (2013); Pajusalu (2009) to help refine and elaborate these rules. These examples cover basic minimal pairs, palatalization and its exceptions, quantity alternations, and regular phonological processes. They helped outline instances where rules produced an output different from the phonemic transcriptions collected from literature, enabling us to refine the rules to accommodate exceptions and various phonological nuances.

We further performed manual evaluation of targeted samples, checking a total of 1840 transcribed forms. These checks were particularly centered on the difficult handling of loan words. They included:

- 1176 forms annotated as "loan words" in Ekilex, and where the rules generated a palatalized /s/, /n/, /l/, /t/, or /d/
- 26 words which end in <-der> or <-ter>, and which were not marked as loans
- 7 lexemes with a nominative ending in <-n> and the stem vowel <i>, which were not marked as loans
- 613 forms of words which were not marked as loans, had a semantic code which let us suspect it might be a proper noun, had palatalization within the stem, and wasn't already included in the lists above
- 18 lexemes where Quantity 3 was marked as optional

When identifying any mistakes, we retained the examples and added them to our development set to prevent regressions. This added 42 development entries, bringing the total size of the development set to 144 word pairs.

4. Structure

The dataset is structured as a set of CSV files, accompanied by rich metadata and linked both by foreign key relations (in the fashion of relational databases) and vocabulary relations (where entries from one table are made up of concatenations of indexes from another table). It follows the Paralex conventions (Beniamine et al., 2023). Figure 1 presents the tables and their relations.

4.1. Forms

The most important table is the one which documents all inflected forms of the selected verbal and nominal lexemes. Inflected forms are the intersection of a specific paradigm cell and a lexeme.

A given cell and lexeme combination might lead to multiple alternate forms (overabundance). This leads to a separate row for each overabundant form. The columns for this table are:

- **form_id**: The primary key of this table, an identifier for the row. This is also a direct reference to the Ekilex form identifier, given at the paradigm details endpoint in `.forms[].id`.
- **lexeme**: A lexeme identifier. This is a reference to the identifier column of the lexemes table.
- **cell**: A cell identifier. This is a reference to the identifier column of the cells table.
- **phon_form**: The form, automatically transcribed to phonemic notation, with phonemes separated by spaces. Each phoneme is documented as a separate row of the sounds table.
- **analysed_phon_form**: The phonemic form, with additional segmentation marking (in the form of the '+' sign).
- **orth_form**: The orthographic form, as given in `.forms[].value`.
- **analysed_orth_form**: The annotated orthographic form, with both segmentation, stress, palatalization, and quantity markers. This is the combination of information from Ekilex and Vabamorf, as described in § 3.3.
- **overabundance_tag**: Marks types of overabundance. Currently only used for nouns. We distinguish series of forms using specific morphological strategies: partitives in <sid>, <id> or involving a vocalic alternation; plurals in <i>, <d/te> or involving a vocalic alternation; genitives in either <e> or <d/te>; illatives in <sse> as opposed to the variant often called short illative or aditive. We also label plural forms of long vowel monosyllables (such as maa, 'earth') and related compounds (isamaa, 'homeland'), where Q3 is optional (Viht and Habicht, 2019, 161).
- **defectiveness_tag**: Marks rows for defective forms. In these rows, the orthographic and phonemic sequences are replaced with the code "#DEF#", following a convention established by Bonami et al. (2014) and continued in Pellegrini and Passarotti (2018).

(a) The forms table

form_id	lexeme	cell	phon_form	analysed_phon_form	orth_form	analysed_orth_form	overabundance_tag	defectiveness_tag	epistemic_tag
11948044	baar_159498_1277891	abl.pl	p ˈ ɑːː r i t e l tː	p ˈ ɑːː r i + t e l tː	baaridelt	b`aari[delt	de_te_plural		
12557674	režissöör_227240_1294244	iness.pl	r ˈ e ʃː i sː øː r e s	r ˈ e ʃː i sː øː r e + s	režissööres	r´ežissööre[s	voc_rad_plural		
13489581	mõjuvõim_205225_1319180	elat.pl	#DEF#	#DEF#	#DEF#	#DEF#		defective	
14671136	olevik_210928_1350835	ad.pl	ˈ o l e v ˈ i kːː u t e l	ˈ o l e v ˈ i kːː u + t e l	olevikkudel	´olev`ikku[del	de_te_plural		
14794532	abiraha_154768_1354149	part.pl	#DEF#	#DEF#	#DEF#	#DEF#		defective	
14833016	memoriaal_201731_1355188	trans.pl	m e m o r i ˈ ɑː l e k s	m e m o r i ˈ ɑː l e + k s	memoriaaleks	memori´aale[ks	voc_rad_plural		
15532377	omaksed_211102_1373989	ill.sg	o m ɑ k s e sː e	o m ɑ k s e + sː e	omaksesse	omakse[sse	sse_illative		questionable
16057494	suruma_237732_1384516	ind.prs.2sg	s u r u t	s u r u + t	surud	suru[d			
16215704	riietuma_227516_1387666	cond.pst.impers	r iː j e tː u t u k s	r iː j e tː u + t u k s	riietutuks	riietu[tuks			
16247323	tõkestama_249008_1388323	ind.prs.impers.neg	t ɤ kː e s t ɑ tː ɑ	t ɤ kː e s t ɑ + tː ɑ	tõkestata	tõkesta[ta			
16862551	arutamine_264439_1412666	iness.pl	ɑ r u tː ɑ m i sʲ i s	ɑ r u tː ɑ m i sʲ i + s	arutamisis	arutamisi[s	voc_rad_plural		
16864723	hankimine_265204_1412723	ad.pl	h ˈ ɑ ŋ kːː i m i sʲ i l	h ˈ ɑ ŋ kːː i m i sʲ i + l	hankimisil	h`ankimisi[l	voc_rad_plural		
17804539	koroonavaktsiin_1260437_1444804	abl.pl	k o r oː n ɑ v ɑ k t s ˈ iː n e l tː	k o r oː n ɑ + v ɑ k t s ˈ iː n e + l tː	koroonavaktsiinelt	koroona+vakts´iine[lt	voc_rad_plural		
20044497	kriitika_186466_1507768	trans.pl	k r iː tʲː i k ɑ͜i k s	k r iː tʲː i k ɑ +͜ i k s	kriitikaiks	kriitika[iks	voc_i_plural		questionable

(b) The lexemes tables

lexeme_id	lemma	WordId	paradigmId	POS	inflection_class	homonymNr
diskonteerima_161559_1412380	diskonteerima	161559	1412380	verb	28	1
ebakindlus_162090_1342240	ebakindlus	162090	1342240	noun	11	1
jõulud_174997_1374721	jõulud	174997	1374721	noun	22e	1
kast_178705_1440436	kast	178705	1440436	noun	22e	2
meeleheide_201276_1430716	meeleheide	201276	1430716	noun	06	1
sünkroniseerima_240031_1483960	sünkroniseerima	240031	1483960	verb	28	1
tagasiside_241170_1313171	tagasiside	241170	1313171	noun	16	1

(c) The cells table

cell_id	ekilex_cell	vabamorf_cell	POS	comment
ind.prs.1sg	IndPrSg1	n	verb	
ind.prs.2sg	IndPrSg2	d	verb	
ind.prs.3sg	IndPrSg3	b	verb	
nom.sg	SgN	sg n	noun	
gen.sg	SgG	sg g	noun	
part.sg	SgP	sg p	noun	
ill.sg	SgIll	sg ill	noun	

Table 4: Excerpts from the forms, lexemes and cells tables

- **epistemic_tag**: Serves to mark forms which were noted as “questionable” or were marked with a star in the source.

All tag columns contain references to indexes of the tags table. In Paralex, this is considered a vocabulary relation, as it is possible to concatenate several tags on the same row, although it is not the case in this particular dataset.

4.2. Lexemes

As noted above (§ 3.2), we generated a separate lexeme for each distinct paradigm, so that words which can inflect through either of several inflection classes lead to separate lexemes. The lexemes table presents one row for each of these lexemes. Rows are uniquely identified by their primary key (**lexeme_id**), which is the concatenation of the citation form (**lemma**), the word identifier (**WordId**) and the paradigm identifier from Ekilex (**paradigmId**), separated by underscores. The identifier is not meant to be parsed, and each piece of information is provided through its own separate column. The table also presents a column indicating the part of speech (“noun” or “verb”, **POS**), using the LexInfo (McCrae et al., 2020) vocabulary, and the inflection class identifier (**inflection_class**), extracted from the paradigm details at `.[].inflectionType`. Since homonymy leads to multiple Word ID for the same lemma, Ekilex provides a homonym numbering (`.homonymNr` in the word search response) which we report in the **homonymNr** column.

4.3. Cells and Feature-values

The cells table presents one row for each nominal or verbal paradigm cell. Its primary key and identifier is given in **cell_id**. It is made of a concatenation of feature-values, each of which is documented in the feature-values table, and separated by dots, following the Leipzig glossing rules. Mappings to the Ekilex and Vabamorf conventions are provided in the columns **ekilex_cell** and **vabamorf_cell** respectively. The **POS** column reports the part of speech each cell is relevant for, and a **comment** column is also provided.

Estonian finite verbs inflect for tense (present and past) and mood (indicative, conditional, imperative and quotative). The indicative present comprises negative polarity forms. All but the imperative and negative forms inflect for subject person-number. There are also a number of non-finite forms with complete or partial case paradigms (participle, supine, gerund). All finite and non-finite paradigms except for the supine and gerund also have impersonal forms. Estonian nouns inflect for all combinations of two numbers and fourteen cases.

The feature-values table presents a single row for each value used in composition in the cells. Its identifier column is the **value_id**. Since this is an abbreviation (as per the Leipzig Glossing rules), the **value_label** column provides the full value name; while the **feature** column provides the name of the relevant feature. The table also presents **POS** and **comment** columns, as well as the value labels in the two other schemes: **ekilex_value_label** and **vabamorf_value_label**.

4.4. Sounds

The sounds table lists all phonemes used in phonological forms (**sound_id**). The other columns represent distinctive features, which are useful in order to define the properties of phonemes, as well as a similarity space between them. This file is a requirement for some linguistically motivated computational tools such as Qumin (Bonami and Beniamine, 2016; Beniamine, 2018), which calculate morphological analogies, taking into account phonological constraints and similarity. To elaborate this file, we started from a general distinctive features spreadsheet (Hayes, 2012), and adapted it to Estonian phonology. In particular, we coded more features as unvalued for segments they did not pertain to, added a palatalization feature for consonants, added a mechanism to code diphthongs,[6] added an extra-length level, and made separate high, low, front, and back features specific to vowels or consonants.

4.5. Tags

The tags table serves simply to list the codes (**tag_id**) used in tag columns (**tag_column_name**), and comment them (**comment**) in a human understandable way to clarify their meaning.

4.6. Metadata

The dataset is accompanied by a data sheet which was adapted for Paralex lexicons, and a JSON metadata file following the frictionless specification (Fowler et al., 2018). The medata is generated with the help of the dedicated Paralex package. It enables thorough verification of the data, including conformity of the CSV files and of the relational schema, enforcing vocabulary restrictions, etc.

[6] We introduce an explicit feature coding whether the sound is a diphthong, with the first phoneme coded the same way as monophthong vowels and its second phoneme with a separate set of features reserved for diphthongs

5. Comparison to Unimorph

Category		count	percentage
concordant	identical	12732	84.88%
	E. overabundant	1792	11.94%
	total	**14524**	**96.83%**
discordant	E. defective	28	0.18 %
	E. error	1	0.006%
	U. error	436	2.9%
	variants	10	0.06 %
	total	**475**	**3.16 %**

Table 5: Concordance of orthographic forms between Unimorph (U.) & Eesthetic (E.)

The unimorph dataset for Estonian[7], extracted from wiktionary, inflects 211 verbal lexemes (5076 in Eesthetic) and 675 nominal lexemes (5475 in Eesthetic). The resources have respectively 211 and 287 lexemes in common.

The two lexicons describe slightly different paradigm structures. In nouns, the Unimorph dataset includes an accusative case. Because it is entirely syncretic with the geneitive, we do not consider it a separate case (Viks, 1992; Blevins, 2008). In verbs, Unimorph includes 49 compound cells which use an auxiliary, such as the pluperfect (ex 21), whereas we only include synthetic forms. Moreover, the Unimorph dataset distinguishes two syncretic cells for singular and plural in the imperative present third person, whereas we count a single cell. Conversely, a few cells present in Eesthetic are absent from the Unimorph dataset: the infinitive, genitive and the nominal inflection of the supine; the synthetic forms of the past conditional (for which Unimorph lists only the compound forms); and the past quotative.

(21) oli lugenud
AUX read.PTCP.PST.PERS
He had read (IND.PLUPRF.3SG)

For this common subset, we compared the orthographic forms across resources automatically, then did manual error analysis. Results are given in Table 5. Of 14999 evaluated cell/lexeme combinations, 96.83% are congruent, although in 11.94% of these, the Eesthetic lexicon provides extra overabundant forms. In 3.16% of cases, the resources disagree, with most cases (2.9%) being errors in Unimorph (including both incorrect forms and wiktionary parsing errors). In one case, Eesthetic had an incorrect form (0.006%). In a minority of cases, either Unimorph inflects a form which we annotate as defective (0.18%), or both resources present a variant acceptable form (0.06%).

We conclude that Eesthetic is vastly more reliable than the wiktionary-extracted Unimorph data for Estonian, in addition to being much larger and more richly annotated.

[7] https://github.com/unimorph/est

6. Conclusion and future work

We described the design and construction of a large inflectional lexicon for Estonian nouns and Verbs, providing rich annotation of complex morphological phenomena, and listing forms in both orthographic and phonemic notation. It is released under the CC BY-SA 4.0 license and can be accessed on Zenodo (10.5281/zenodo.8383522).

The dataset follows the Paralex standard (Beniamine et al., 2023), and is accompanied by Frictionless metadata (Fowler et al., 2018). Compliance with these standards makes it adhere not only to the FAIR principles (Wilkinson et al., 2016), but also to the DeAR principles (Beniamine et al., 2023): it contributes to the **De**centralized standardization of resources in computational morphology; data quality is ensured by extensive **A**utomated validation, performed as continuous pipelines; finally it is presented with a user-friendly web interface as a static Paralex website, which is generated by a continuous pipeline, and thus **R**evised as soon as the data is updated. This is in stark contrast with current practices in linguistics, where database websites are often created by contractors at the term of the project, locking in a specific version of the data, with little possibilities of updating it without further external contracting. The move towards revisable pipelines is crucial for the long term management of linguistic data.

Given the size of the lexicon, manual transcriptions were out of the question. Our automated transcriptions were devised in a transparent and linguistically motivated way, guided by a carefully crafted development set, and went through several rounds of manual verification.

Although we took great care to identify edge cases, we expect inaccuracies to occasionally occur with borrowings which are not explicitly identified in Ekilex data, as well as with occasional exceptions to systematic palatalization.

We currently do not report the lexeme frequencies calculated in corpus (§ 3.1). Indeed, homonymy makes it difficult to assign frequencies to a specific lexeme (one lemma found in corpus might correspond to more than one lexeme in the dataset). Although it is complex to find reasonable practical solutions to this, future versions of the dataset could benefit from indicating frequencies for lexemes, cells or inflected forms.

7. Ethical Statement

To the best of our knowledge, there are no ethical concerns of this dataset.

8. Acknowledgements

This work was partially funded by a British Academy International Newton Fellowship (reference: NIF23\100218) and a Leverhulme Early Career Fellowship (ECF-2022-286).

9. Bibliographical References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429–464.

Mari Aigro and Virve-Anneli Vihman. in press. Realised overabundance in Estonian noun paradigms: A corpus study. *Word Structure*.

Eva Liina Asu, Pärtel Lippus, Karl Pajusalu, and Pire Teras. 2016. *Eesti keele hääldus [Estonian pronounciation]*. University of Tartu Press, Tartu.

Eva Liina Asu and Pire Teras. 2009. Estonian. *Journal of the International Phonetic Association*, 39(3):367–372.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Sacha Beniamine. 2018. *Classifications flexionnelles: Étude quantitative des structures de paradigmes*. Ph.D. thesis, Université Sorbonne Paris Cité - Université Paris Diderot.

Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. Paralex: a DeAR standard for rich lexicons of inflected forms. In *Presentation at International Symposium of Morphology*. https://www.paralex-standard.org.

James P. Blevins. 2007. Conjugation classes in Estonian. *Linguistica Uralica*, 43(4):250–267.

James P. Blevins. 2008. Declension classes in Estonian. *Linguistica Uralica*, 44(4):241–267.

Olivier Bonami and Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Olivier Bonami, Gauthier Caron, and Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In *Actes du quatrième Congrès Mondial de Linguistique Française*, pages 2583–2596.

Maria Copot and Olivier Bonami. in press. Behavioural evidence for implicative paradigmatic relations. *Mental Lexicon*.

Martin Ehala. 2003. Estonian Quantity: Implications for Moraic Theory. In Satu Manninen and Diane Nelson, editors, *Generative Approaches to Finnic and Saami Linguistics*, pages 51–80. Center for the Study of Language and Information (CSLI).

Dan Fowler, Jo Barratt, and Paul Walsh. 2018. Frictionless Data: Making Research Data Quality Visible. *International Journal of Digital Curation*, 12(2):274–285.

Bernard Fradin and Françoise Kerleroux. 2003. Troubles with lexemes. In Geert Booij, Janet DeCesaris, Angela Ralli, and Sergio Scalise, editors, *Selected papers from the thirdMediterranean Morphology Meeting*, page 177–196. IULA – Universitat PompeuFabra.

Borja Herce. 2023. VeLeSpa: An inflected verbal lexicon of Peninsular Spanish and a quantitative analysis of paradigmatic predictability. Pre-print.

Hans Henrich Hock. 2021. *Principles of Historical Linguistics*. De Gruyter Mouton, Berlin, Boston.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Johanna Laakso. 2021. Language contact and typological change: The case of Estonian revisited. *Word Structure*, 14(2):226–245.

Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2019. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics*.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Matías Guzmán Naranjo and Olivier Bonami. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa: a journal of general linguistics*, 6(1).

Karl Pajusalu. 2009. Dynamics of Estonian phonology. *STUF - Language Typology and Universals*, 62(1-2).

Matteo Pellegrini. 2020. Using LatInfLexi for an Entropy-Based Assessment of Predictability in Latin Inflection. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 37–46, Marseille, France. European Language Resources Association (ELRA).

Matteo Pellegrini. 2023. Flexemes in theory and in practice. *Morphology*, 33:361–395.

Matteo Pellegrini and Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, page December, Aachen.

Külli Prillop. 2013. Feet, Syllables, Moras and the Estonian Quantity System. *Linguistica Uralica*, 49(1):1.

Külli Prillop. 2020. Morae in Estonian. A reply to Natalja Kuznetsova's paper "Estonian word prosody on the Procrustean bed of morae". *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 10(2):151–183.

Paul Friidrih Saagpakk. 1982. *Eesti-inglise sõnaraamat*. Yale linguistic series. Yale Univ. Pr., New Haven, Conn. [u.a.].

Andrea Sims and Jeff Parker. 2016. How inflection classes work: On the informativity of implicative structure. *Word Structure*, 9(2):215–239.

Gregory T. Stump and Raphael Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge.

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*, pages 358–381. Oxford University Press (OUP).

Anna M. Thornton. 2018. Troubles With Flexemes. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo, and Fiammetta Namer, editors, *The lexeme in descriptive and theoretical morphology*, chapter 13, page 303–321. Language Science Press.

Annika Viht and Külli Habicht. 2019. *Eesti keele sõnamuutmine [Estonian inflection system]*. University of Tartu Press, Tartu.

Ülle Viks. 1992. *Väike vormisõnastik [A short form dictionary]*. Estonian Academy of Sciences, Tallinn.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

10. Language Resource References

Sacha Beniamine and Dunstan Brown. 2019. *Inflected lexicon of Russian Nouns in IPA notation*. Online repository.

Eesti Keele Instituut. 2014. *BED: Eesti keele põhisõnavara sõnastik 2014. [The Basic Estonian Dictionary 2014]*. Sõnaveeb.

Eesti Keele Instituut. 2019. *ECD: Eesti keele naabersõnad 2019. [The Estonian Collocations Dictionary 2019]*. Sõnaveeb.

Eesti Keele Instituut. 2023. *Eesti Keele Instituudi eesti keele morfoloogiline andmebaas 2023. [Morphological database of Estonian 2023]*. Sõnaveeb.

Timothy Feist and Enrique L. Palancar. 2015. *Oto-Manguean Inflectional Class Database*. University of Surrey.

Filosoft. 2015. *Vabamorf: a set of open-source morphological tools for Estonian*. Online repository.

Bruce Hayes. 2012. *Spreadsheet with segments and their feature values*. Distributed as part of course material for Linguistics 120A: Phonology I at UCLA.

Kallas, Jelena and Koppel, Kristina. 2022. *Eesti keele ühendkorpus 2021 - vert (Estonian National Corpus 2021)*. Center of Estonian Language Resources.

Kallas, Jelena and Langemets, Margit and Koppel, Kristina. 2022. *EKI ühendsõnastik 2022 (EKI combined dictionary 2022)*. Center of Estonian Language Resources. API documentation: `https://github.com/keeleinstituut/ekilex/wiki/Ekilex-API`; user interface: `https://sonaveeb.ee/`.

Langemets, Margit. 2020. *Eesti keele sõnaraamat 2019 (veebisõnaraamat) [The Dictionary of Estonian 2019]*. Center of Estonian Language Resources.

John Philip McCrae and Philipp Cimiano and Paul Buitelaar. 2020. *Lexinfo v3.0*. An ontology of types, values and properties to be used with the OntoLex-Lemon model.

Matteo Pellegrini. 2020. *LeFFI: Inflected lexicon of Italian verbs*. Online repository.

Perdigão, Fernando and Beniamine, Sacha and Luís, Ana R. and Bonami, Olivier. 2021. *European Portuguese Verbal Paradigms in Phonemic Notation*. Zenodo. Supplementary material for Beniamine, Bonami and Luís (2021).