A Computational Model of Latvian Morphology

Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma
University of Latvia Institute of Mathematics and Computer Science
peteris@ailab.lv, lauma@ailab.lv, laura@ailab.lv

Abstract

In this paper we describe a computational model of Latvian morphology that provides a formal structure for Latvian word form inflection and has been implemented in software for generation, analysis and lemmatization of Latvian word forms. The work was motivated by the need for a NLP inflection model that can cover all the complexity of the Latvian language and explicitly enumerate and handle the many exceptions to the general Latvian inflection principles. This is an evolution of earlier work, extending the initial proof of concept model to properly cover Latvian language. We provide a set of morphological paradigms that differ from current linguistic tradition, a set of systematic stem changes and combine it with an extensive lexicon that includes paradigm information and structured morphological attributes for 118 000 lexemes. This model has been applied on both dictionary and corpora data, demonstrating that it provides a good coverage for modern Latvian literary language. We also consider that there is a good potential to extend this also to the related Latgalian language.

Keywords: morphology, Latvian, computational model

1. Introduction

Computational linguistics work on synthetic flexive languages like Latvian benefits from an explicit treatment of morphology, being able to link an arbitrary written word (the *surface form*) to the appropriate lexeme and the morphological attributes encoded by the inflectional word form.

Natural language processing often considers morphology in the context of three separate problems handled by separate specialized tools:

- morphological analysis – providing information about a given word surface form;
- morphological synthesis – building all the relevant inflectional forms of a single lemma;
- spell-checking and word lists – a list of valid word forms or validation of whether a particular word form is valid in the given language.

Although these problems sometimes have contradictory requirements (for example, certain word forms should not be considered valid in spell-checking but are used in practice and should be recognized in text analysis), this can be handled within a single model by appropriate configuration.

The work was motivated by the need for a NLP inflection model that can cover all the complexity of Latvian language and handle the many exceptions to the general Latvian inflection principles, preferably by separating them into a limited number of explicitly listed systematic groups.

Since earlier work on Latvian morphology described in section 2, there has been considerable progress in advancing a computational model of Latvian morphology that we now consider reasonably complete in covering the Latvian standard language according to validation on both dictionary and corpora data.

We propose a formal structure for word form inflection and treatment of morphological attributes and a software implementation of this model for generation, analysis, and lemmatization of Latvian word forms, including a set of morphological paradigms that differ from current Latvian linguistic tradition, a set of systematic stem changes and combine it with an extensive lexicon that includes paradigm information and structured morphological attributes for 118 000 lexemes.

Latvian is a synthetic, inflective Indo-European language with a relatively rich morphology that has around 1.5 million native speakers. Many of Latvian endings overlap, creating homoforms, so Latvian morphological analysis is inherently ambiguous. Corpus analysis shows that approximately 50% of the words in Latvian running text have multiple possible morphological interpretations.

It should be noted that the theoretical basis in the morphology of the Latvian language is very well developed (Ceplīte and Ceplītis, 1991; Auziņa et al., 2013; Kalnača and Lokmane, 2021), however, the division into categories and paradigms varies for different researchers, especially for less homogeneous groups, for example, the division of irregular verbs into subgroups (Fennel, 1980; Andronovs, 1997; Nau, 1998), and also in our model the paradigms split partially differs from these sources, as it is more driven by matching implementation of inflection rules than the derivation or etymology properties of words.

The rest of this paper is organized as follows. Chapters 2 and 3 describe the related work on computational morphology for Latvian and other languages. Chapter 4 describes the proposed computational model structure, and Chapter 5 de-

LREC-COLING 2024, pages 221–232
20-25 May, 2024.

scribes the implementation of this model for Latvian language-specific elements. Chapter 6 describes the validation of this model and the final Chapter 7 describes ongoing and future work.

2. Related work on Latvian

The earliest experiments with automated Latvian morphological analysis started in the 1970s, implementing the basic noun and adjective paradigms (Drīzule, 1978).

Morphemic analysis, proposed by Krūze-Krauze (1996, 1998); Sarkans (1996), was an attempt to implement and extend "A Derivational dictionary of Latvian" (Fennel, 1985), allowing to generate and analyze Latvian words by defining rules that define the ways how different morphemes may combine together to make a single word. However, the system quickly reached an unmaintainable size of rules and special cases while still having a low coverage, as it fails to recognize a large portion of words – especially loanwords adopted from other languages such as Greek, Latin, or English.

With the advent of personal computers, since the 1990s there have been several attempts to create morphological systems for Latvian based on the main linguistic rules for word endings and morphemes (Greitāne, 1994; Kreslins, 1996; Vasiļjevs et al., 2004; Eger and Sējāne, 2010). A common problem for such systems is that different parts of speech have forms with overlapping morphological suffixes, and treating all stems as equally possible causes significant ambiguity, adding many analysis candidates that are not valid words.

However, the experience with those systems indicated that proper handling of Latvian inflection does require certain lexeme-specific information, such as whether a particular verb stem exists, or whether some noun is masculine or feminine, so they were followed by multiple lexicon-based analysis systems (Skadiņa, 2004; Paikens, 2007; Deksne, 2013). The obvious limitation is that such a structured lexicon has to be created and maintained, and it will never cover 100% of possible words.

There have also been multiple morphosyntactic tagging solutions for Latvian, often based on the disambiguation of options provided by some morphological analysis model – Pinnis and Goba (2011); Paikens et al. (2013); Ņikiforovs (2015); Paikens (2016); Treimanis (2023).

To summarize, before this work, there were multiple approaches that provided some solutions for Latvian morphology which all handle the some general principles of Latvian inflection, but they vary greatly in their capability to model and accurately address less common linguistic phenomena. In this work, we extend the approach of Paikens (2007) to improve the coverage of the Latvian language.

3. Related work on other flexive languages

Looking at the treatment of morphology for other flexive languages, we can observe a similar evolution of approaches as we did in the previous section for Latvian.

There have been many implementations of rule-based morphological models – Hajic et al. (2001) for Czech, Zinkevičius (2000) for Lithuanian, Kaalep (1997) for Estonian, Yuret and Türe (2006) for Turkish, etc. This is especially popular in earlier years, likely due to limitations of computer capacity.

Those are then often followed by introduction of lexicon data to provide a more accurate representation and limit ungrammatical interpretations (Kaalep and Vaino, 2001; Woliński, 2006, 2014; Straková et al., 2014; Korobov, 2015), while still usually keeping the morphological rules to handle any out of vocabulary words.

A popular approach for developing morphological analysis systems has been finite state transducers (Linden et al., 2011; Yona and Wintner, 2008; Kaalep et al., 2018). Our lexicon-based proposal is relatively similar, as a prepared FST would also include a lexicon with a stem list categorized in relatively fine-grained paradigm equivalents. One of the existing systems for Latvian morphology (Deksne, 2013) and Lithuanian (Mackevičiūtė, 2004) was developed using this approach.

Afterward the research focus of morphological analysis for those languages moves on from morphological analysis towards disambiguation of the options (Pajarskaitė et al., 2004; Daudaravičius et al., 2007; Straková et al., 2014), however, the underlying morphological models (or, at the very least, their lexicon data) still keep being advanced and fine-tuned to improve coverage and accuracy, as indicated by ongoing publications on their performance for, for example, Lithuanian (Kapočiūtė-Dzikienė et al., 2017; Boizou et al., 2018; Bielinskiene et al., 2016).

4. Structure of the computational model

The key components of this model that we will define in this chapter are paradigms, lexemes, stems, endings, stem change rules, lemmas, and attributes.

Wordforms are modeled as consisting of a fixed starting part – which we define as 'stem' – followed by an 'ending' specifying the inflectional form. In certain cases, inflection may require conditional modifications to the stem, so endings can also define an optional 'stem change rule'. Any indeclin-

able words are considered to have their ending as an empty string and the whole word as its 'stem'.

We define a 'paradigm' as a set of 'inflections' defining all the wordforms that can be generated from each lexeme. In our model an 'inflection' specifies how the wordform is built, combining a stem, the ending of that inflection, and an optional stem change rule. The inflection also lists the morphological attributes specified by that inflectional form, and the paradigm lists certain attributes that apply to all inflectional forms in that paradigm. A 'lexeme' in this context is a stem that belongs to a certain inflectional paradigm, and optionally certain lexical attributes that may also influence its inflection. Usually there a single stem is sufficient to fully define the lexeme, but in some paradigms (e.g. Latvian first conjugation verbs) the lexemes include multiple separate stems as required for different inflectional forms.

The terms used in our model are drawn from Latvian linguistic tradition, but they are not exact equivalents. What we call 'stems' in our model also includes prefixes and infixes which do change during inflection. Stem change rules include not only alterations to the end of the stem but also adding prefixes such as negation or superlative. Wordforms can be not only inflectional forms but also highly regular derivatives.

In this approach, most of the morphological complexity is not encoded in software but rather in the data tables which specify paradigms and inflections. Stem change rules are currently implemented as code for transformations in both directions - from the original stem to the wordform for synthesis, and in the opposite direction for analysis, validating any ambiguous transformations versus the lexicon.

It is possible to define that some inflections will be recognized in analysis, but not generated for synthesis, as there are some regular wordforms that are sometimes used in practice but are considered ungrammatical.

To define an instance of our model for a new language, one must implement the following things – the set of relevant morphological attributes, the paradigm definitions including the ending data, any applicable stem change rules, and a lexicon: an appropriate amount of lemmas with their appropriate paradigms, as well as the closed word classes and any irregular wordforms, if needed.

4.1. Morphological analysis

The algorithm for wordform analysis consists of the following steps:

- Select the set of possible inflections (and associated paradigms) based on suffixes of the wordform;
- For each ending, obtain the candidate stem by removing the suffix;
- Apply the stem change rule (if any) defined by the inflection to get the lemma stem from the wordform stem;
- In the lexicon, look up lexemes matching the paradigm of that inflection and that lemma stem;
- For each lexeme found, assemble the wordform data by combining the morphological attributes provided by the paradigm, inflection and lexeme.

This approach can be considered equivalent to a special case of a particular finite state transducer (FST) implemented in software, as opposed to a generic transducer platform that can execute different transducers, but making it easier to implement conditional transformations for the phonological changes of Latvian stems, which were an issue for FST systems according to (Deksne, 2013).

Some systematic phenomena of written language that do not rely on words – numbers, dates, URLs, email addresses, initials, etc. – are identified using regular expressions instead of this paradigm system.

If no valid interpretation of the wordform is found in the lexicon, for Latvian we also attempt to apply certain systematic derivations which traditionally are excluded from dictionaries – prefixed verbs and noun diminutives. If no matches are still found, we attempt a 'brute force' guessing by creating lexeme candidates for every possible matching ending.

The guessing of words outside the defined lexicon is an important part of obtaining good coverage for language analysis, but to reduce analysis ambiguity we make a simplifying assumption that if any analysis options are found in the lexicon, then these results form an exhaustive analysis. It is indeed possible that the intended wordform really is out of lexicon but happens to be a homoform of some lexicon entry, but this is relatively rare, causing issues mostly when common nouns are used as surnames (potentially with the opposite grammatical gender) or company names.

No matter if matches were found in the lexicon or generated purely by morphological rules, ambiguity in this part of the analysis is unavoidable. More than half of words in Latvian running text have multiple valid options. For example, word *roku* can be interpreted as a word form of *roks* ('rock music'; masculine noun), two word forms (singular accusative and plural genitive) of *roka* ('hand'; feminine noun), or a form of *rakt* ('to dig'; verb). This ambiguity can be resolved only by morphosyntac-

tic tagging, taking into account the surrounding sentence context.

In addition to the attributes, full analysis results also require identifying the lemma of this word form. In the proposed model each inflection points towards a 'lemma-inflection' (possibly itself) which should be used for lemmas of that wordform. In this model, different forms of the same lexeme may have different lemmas. For example, any adjective lexemes will have both masculine and feminine forms, but Latvian tradition is to lemmatize adjectives to the same gender as the original word form, resulting in two possible lemmas for the same lexeme.

4.2. Morphological synthesis

The same model is used in the opposite direction for the synthesis of inflectional wordforms.

Assuming a known lexeme from the lexicon (which includes the lemma, the inflectional paradigm, and structured attributes), wordform synthesis consists of generating a wordform for every ending in the inflectional paradigm, combining the lexeme stem (modified by an ending-determined stem change rule, if any) with the ending, and merging all the attributes from the paradigm, ending and lexeme. This is followed by filtering the wordforms based on lexeme-defined restrictions such as certain nouns that are plural-only or singular-only.

If only the lemma is available, then the relevant paradigm and lexeme are identified by performing morphological analysis as described in the previous section; however, there is unsolvable ambiguity in the case of homoforms.

The lexicon, inflectional paradigm, and ending data is the same as for morphological analysis, but the code containing stem change rules is different as it needs to perform the opposite transformation. Also, there's a conceptual difference in treating optional variations – for analysis, our model tries to accept as much as possible in order to obtain good coverage for text containing unusual forms, but for synthesis, we choose to err on the side of caution and generate only recommended forms. For example, the future 3rd person archaic plural form *darīsit* ('[you'll] do') of the verb *darīt* ('to do') is recognized in analysis but we generate only the contemporary version *darīsiet*.

5. Latvian morphology in this model

The modeling of Latvian in this structure was started in Paikens (2007), however, over the years since that initial work, continuous application of this model for corpora and digital dictionaries indicated many less frequent phenomena that required further additions and adjustments to the set of paradigms, endings, and other attributes.

The major differences from the initial implementation are the following:

- A systematic split of Latvian 5th and 6th declension nouns and 3rd conjugation verbs in separate paradigms according to whether a particular stem change should apply for each word;
- Implementation of restricted paradigms for adjectives and verb participles which have nominalized and thus have a limited set of possible inflections and a non-verbal lemma;
- Implementation of rarely used reflexive noun forms;
- Review of all the edge cases in semi-irregular verb inflections and stem changes, covering all Latvian literary language verbs except the unique *būt* ('to be');
- Analysis and generation of certain less common alternate forms for the same inflection;
- Exhaustive enumeration of the closed word classes based on the existing dictionaries for Latvian.

5.1. Latvian paradigms and stem change rules

The paradigm set was initially based on the standard Latvian split of noun declinations and verb conjugations, and one paradigm per each other part of speech, however a more fine-grained split is required to have sufficient information on how to inflect certain words. The full list of implemented paradigms is given in Table 1. We preferentially attempted to implement any systematic inflection variations as stem change rules, and resort to separate paradigms only if inflection requires external information – in which case we sorted the relevant lexemes into separate paradigms.

For nouns we had to subdivide the traditional 6 Latvian noun declensions based on gender (3rd, 4th and 5th declension), nominative ending (1st and 2nd declension), genitive ending (2nd declension) and the presence or absence of stem changes (2nd, 5th and 6th declensions).

For adjectives we use two paradigms respectively for adjectives ending on *-s* or *-š* in singular nominative, as well as several paradigms for adjectives that lack the positive degree forms either because of nominalization (*rožveidīgie* 'rosidae', lit. 'rose-likes') or semantic reasons (*galvenais* 'the main'). While for most adjectives both genders (masculine and feminine) are regular, for paradigms lacking the positive degree sometimes the feminine or masculine forms are not possible, thus, in this case we needed a separate paradigm for masculine and feminine nominalized adjectives.

For numerals we use one paradigm for ordinal numbers and three paradigms for cardinal numerals, as the set of wordforms depends on whether the numeral is grammatically singular, plural, or indeclinable. For declinable numerals, the same paradigm contains both grammatical genders. If a word denoting a number can be inflected for case, but not for gender, e.g., *nulle* 'zero', the inflection table matches the equivalent noun table, so we model them as a special case of noun paradigms.

For verbs, we started with three paradigms each for direct and reflexive verbs – one per each traditional verb conjugation. However, it turned out to be useful to separate the 3rd conjugation paradigms depending on whether the present simple forms contain a stem change (*sacīt, saku* 'to say, [I] say') or no (*lasīt, lasu* 'to read, [I] read'). Furthermore, to accurately model the highly irregular 1st conjugation we opted for storing three stems for each lexeme: the infinitive stem, present stem, and past stem, as all the other wordforms can be derived from one of those three stems. For example, lexeme *aust* with meaning 'weave' is defined with stems `aus`, `auž`, `aud`, but partial homonym *aust* with meaning 'rise' – with stems `aus`, `aust`, `aus`. We also have a separate paradigm for the highly irregular verb *būt* 'to be', grouping it with its many prefix derivatives *pabūt* ('to be [for a little bit]') , *sabūt* 'to be [for some time]'), and others.

For adverbs, we currently have two paradigms – a paradigm for adverbs that have positive, comparative, and superlative degree forms just as Latvian adjectives, and a paradigm for adverbs that have only a single form.

The other paradigms are for indeclinable words, including the genitive-only invariant nouns (*augstpapēžu* 'high-heeled'), indeclinable loanword nouns (*kanoe* 'canoe'), indeclinable loanword adjectives (*rozā* 'pink'), prepositions, conjunctions, interjections, particles, abbreviations and Roman numbers.

We also have a separate paradigm to contain a limited number of 'hard coded' irregular forms. Those are most notably used to model some pronouns (most Latvian pronouns are can be inflected according to noun paradigms), the highly irregular numeral *trīs* 'three' and some regional dialect forms like *eima* ('let's go').

Some systematic derivations, like adverbs derived from related adjectives (*ātrs* →*ātri*; 'quick' →'quickly') or the *-šana* derivation of action nouns (*rakt* →*rakšana*; 'to dig' →'digging') are implemented along with ordinary inflectional forms, adding an extra set of endings to the paradigm with a specific part of speech attribute which will override the paradigm default part of speech. This approach matches the Latvian dictionary tradition of not listing these derived words as separate entries unless they develop a separate meaning.

To complement the set of given paradigms our Latvian model also features a large set of rules for stem changes. These rules are used to model multiple separate phenomena of Latvian and can be roughly grouped as follows:

- Consonant changes before the ending (*brālis* ('brother' – *brāļa* 'brother's', *mape* 'folder' – *mapju* 'folders"; *kož* 'bites' – *kod* 'bit', but also *beidzu* 'end' – *beigušais* '[the person that has] ended'; *pūst* 'blow' – *pūzdams* '[while] blowing'; *dzimušais* 'born' – *dzimusī*) '[the person that was] born';
- Removing the final stem letter (*lasīt* 'to read' – *lasu* '[I] read'; *nest* 'to carry' – *nešana* 'carrying' (not *nessšana* as would be regular));
- Adding additional infix vowel depending on the last consonant of the stem (*lekt* 'to jump' – *leķšos*, but *lauzt* 'to break' – *lauzīšos*);
- Modeling long or short vowel in verb endings (*darīt* 'to do' – *darāms*'[can be] done', but *sēdēt* 'to sit' – *sēžams* '[can be] sat');
- Optional vocative ending when addressing people (*Kristīne* – *Kristīn* or *Kristīne*);
- Debitive prefix *jā-* and superlative prefix *vis-* (*dara* 'does' – *jādara* 'has to do'; *labākais* 'best' – *vislabākais* 'very best');
- Comparative infix *-āk-* (*labs* 'good' – *labāks*) 'better';
- Shortening of repetitive affixes (*pēdējais* 'last' – *pēdējam* '[to the] last one' instead of *pēdējajam*, *zaļais* 'green' – *zaļajam* '[to the] green one');
- Special one-of-a-kind change for one form of verb *iet* 'to go' and its prefix derivations (3rd person *iet* 'goes', not *ej*);
- Stem change rules for combinations of above.

5.2. Morphological attributes and tagset

Each wordform in our model can be characterised by a lemma and a set of attributes. Table 2 provides an overview of the attributes used for Latvian. These attributes can include not only purely morphological features related to change of form, such as number and case, but also various lexical features such as verb type or pronoun type, or other metadata provided by the source lexicon.

We also define an encoding of these wordform attributes into a standardized morphological tag for usage in corpus tagging and NLP tools. For Latvian corpora we have adopted a positional

Table 1: A complete list of Latvian inflectional paradigms

ID	Example	Description
noun-1a	tēvs	1st declension nouns *-s*
noun-1b	ceļš	1st declension nouns *-š*
noun-2a	dzenis	2nd declension nouns *-is*
noun-2b	tētis	2nd declension nouns *-is* without stem change
noun-2c	ūdens	2nd declension nouns *-s*, genitive matches nominative
noun-2d	suns	2nd declension nouns *-s*, genitive does not match nominative
noun-3m	medus	3rd declension nouns *-us*, masculine
noun-3f	Markus	3rd declension nouns *-us*, feminine
noun-4m	puika	4th declension nouns *-a*, masculine
noun-4f	liepa	4th declension nouns *-a*, feminine
noun-5ma	Egle	5th declension nouns *-a*, masculine
noun-5mb	balamute	5th declension nouns *-a*, masculine with stem change exception
noun-5fa	egle	5th declension nouns *-a*, feminine
noun-5fb	pase	5th declension nouns *-a*, feminine with stem change exception
noun-6a	sirds	6th declension nouns *-s*
noun-6b	auss	6th declension nouns *-s* with stem change exception
noun-r1	klausītājies	Reflexive nouns *-ājies*, *-ējies*
noun-r2	vēlējumies	Reflexive nouns *-umies*
noun-r3	acīsskatīšanās	Reflexive nouns *-šanās*
noun-0	kino	Indeclinable nouns
noun-g	augstpapēžu	Invariant genitive noun
adj-1	balts	Adjectives *-s*
adj-2	zaļš	Adjectives *-š*
adj-infl	lillā	Indeclinable adjectives
adjdef-m	jaundzimušais	Nominalized definite adjectives/participles, masculine
adjdef-f1	mēnessērdzīgā	Nominalized definite adjectives, feminine
adjdef-f2	cietusī	Nominalized participles *-usī*, feminine
part-1	pusjucis	Nominalized participles *-is/-usi*
part-2	pusapģērbies	Nominalized participles *-ies/-usies*
part-3	pusjokodams	Nominalized participles *-dams/-dama*
part-4	pusjokodamies	Nominalized participles *-damies/-damās*
verb-1	rakt	1st conjugation verbs
verb-1i	būt	Irregular *to be* and its derivatives
verb-2	spēlēt	2nd conjugation verbs
verb-3a	lasīt	3rd conjugation verbs
verb-3b	sacīt	3rd conjugation verbs with stem change exception
verb-1r	rakties	1st conjugation reflexive verbs
verb-2r	spēlēties	2nd conjugation reflexive verbs
verb-3ra	lasīties	3rd conjugation reflexive verbs
verb-3rb	lasīties	3rd conjugation reflexive verbs with stem change exception
adverb	vienmēr	Adverbs without comparative forms
adverb-2	ātri	Adverbs with comparative forms
ord	trešais	Ordinal numerals
card-1	viens	Cardinal numerals, singular
card-2	divi	Cardinal numerals, plural
card-infl	divarpus	Indeclinable cardinal numerals
number	123	Numeric sequences
pron	es	Pronouns
prep	uz	Prepositions
conj	un	Conjunctions
particle	jā	Particles
abbr	u.c.	Abbreviations
punct	??!	Punctuation
excl	ak	Interjections
foreign	adaggio	Foreign insertions
hardcoded	nav, trīs, + ∂, ©	Exceptions and forms not fitting in any other group

tagset (initially described by Levāne and Spektors (2000)) with attributes specific to the part of speech, i.e. the first letter of the tag always specifies the part of speech, but the following positions encode only the attributes corresponding to the part of speech, so the length of the tag varies from 1, e.g. for particle, to 11 for verb. The approach is based on the MULTEXT-East standard (Erjavec, 2012), adapted to Latvian. The tagset contains 13 part of speech classes: 10 PoS classes correspond to the word classes defined in Latvian Grammar (Kalnača and Lokmane, 2021) – 5 for declinable word classes (nouns, adjectives, verbs, pronouns, numerals) and 5 for indeclinable word

Table 2: Attributes used in Latvian morphological analyses

Attribute	Relevant POS	Applies to	Values and usage notes
Part of speech		Paradigm	Values: noun, verb, adjective, pronoun, adverb, numeral, proposition, conjunction, interjection, abbreviation, particle, residual, punctuation
Gender	Noun, adjective, verb (participle), numeral, pronoun	Form	Feminine, masculine or not applicable
Number	Noun, adjective, verb, numeral, pronoun	Form	Singular, plural or not applicable
Case	Noun, adjective, numeral, pronoun, verb (participle)	Form	Nominative, genitive, dative, accusative, locative, vocative or not applicable
Definiteness	Adjective, numeral, verb (participle)	Form / Paradigm	Definite or indefinite ending
Degree	Adjective, adverb, verb (participle)	Form	Positive, comparative or superlative degree, or not applicable for some adverbs
Noun type	Noun	Lexeme	Proper or common noun
Declension	Noun	Paradigm	6 traditional declensions, reflexive nouns, fixed genitive form, indeclinable loanwords
Deminutive	Noun	Lexeme	Lexical feature to distinguish deminutives
Pronoun type	Pronoun	Lexeme	Personal, possessive, demonstrative, indefinite, interrogative, relative, definite
Pronoun negation	Pronoun	Lexeme	Indicates lexically negative pronouns
Numeral type	Numeral	Lexeme / Paradigm	Cardinal, ordinal or fraction
Structure of numeral	Numeral	Lexeme	Simple or compound word
Person	Verb, pronoun	Paradigm	1st, 2nd, 3rd or not applicable
Tense	Verb	Form	Present, past, future or not applicable
Mood	Verb	Form	Indicative, relative, conditional, debitive, imperative, infinitive or participle
Voice	Verb	Form	Active, passive or not applicable
Transitivity	Verb	Lexeme	Transitive or intransitive
Reflexive	Verb	Form	Reflexive or direct (reflexive noun is as a separate value in Declension attribute)
Negation	Verb	Form	Indicates which verb forms use negative prefix
Participle declinability	Verb (participle)	Form	Inflected, partially inflected or indeclinable
Verb type	Verb	Lexeme	Main verbs, auxiliaries, copula, semantic modifiers, and verbs that can function as copula
Conjugation	Verb	Paradigm	3 traditional conjugations or irregular
Prepositional adverb	Adverb	Lexeme	Yes or no, can adverb function as preposition
Prepositional position	Preposition	Lexeme	Preposition or postposition (lexico-syntactic property)
Syntactic function of a conjunction	Conjunction	Lexeme	Coordinating or subordinating
Residual type	Residual	Lexeme	Foreign material, number written in digits, URl or other (symbols etc.)
Abbreviation type	Abbreviation	Lexeme	Common noun, proper noun, adjective, verbal, adverbial, discourse
Punctuation type	Punctuation	Lexeme	Comma, quote, stop, bracket, dash, colon, other
Alternative form	All flexible POS	Form	Does a given irregular wordform overrides the one given in paradigm or provides an alternative
Systematic derivation	All	Form	Used to distinguish inflectional forms from regular derivatives included in a paradigm
Lemma properties	Noun, adjective, adverb	Lexeme	Indicates that the lemma differs from the default paradigm-provided lemma, values: plural, singular, feminine, comparative

classes (adverbs, prepositions, particles, conjunctions, interjections), and separate classes for abbreviations, punctuation and residuals.

5.3. Lexicon

We use the Tēzaurs.lv lexicographic database system (Grasmanis et al., 2023) as the lexicon for this model. Tēzaurs.lv is a monolingual, multifunctional dictionary in ongoing development, currently containing almost 400 thousand entries. Out of those, 118 000 lexemes are annotated with morphological data and are used in this model.

Tēzaurs.lv is primarily designed as an explanatory dictionary for public consumption, but it also integrates this model in order to display inflection tables in the dictionary entries, and also provides a platform for maintaining the lexemes of this model. For this purpose lexicographers specify the morphological paradigm of each lexeme, and certain attributes that influence the inflection – singulare tantum / plurale tantum, nonstandard lemmas, etc. Also, any extra attributes annotated in the lexicon will be added to the wordform analysis, with some of them (such as flags denoting proper nouns or modal verbs) appearing in the morphological tag.

If certain irregular or regional dialect words have some nonstandard inflectional forms (either replacing the regular form, or as an optional alternative), those can also be defined in this lexicon platform.

6. Validation and known limitations

This implementation of morphological tools for Latvian was used for several applications which provided useful feedback about the quality and completeness of the morphological model.

The morphological analysis functionality was validated by its usage in the development of the Latvian Treebank (Rituma et al., 2023; Zeman et al., 2021) and integration in the morphosyntactic tagger (Paikens et al., 2013) that was used for automated tagging of the balanced corpora of Latvian (Levāne-Petrova and Darģis, 2018; Levāne-Petrova et al., 2023) and Latvian Corpora Collection (Saulīte et al., 2022). It was also applied in various practical NLP pipelines (Paikens, 2014; Znotiņš and Cīrule, 2018; Barzdins et al., 2020) for semi-automatic named entity inflection and extending homonym disambiguation towards complete word sense disambiguation; Large Language Models and D-Wave Quantum Hybrid Solvers promise reduced need for human verification in these pipelines.

Annotation of the Latvian Treebank involved manual review and correction of automatically tagged morphosyntactic attributes and lemmas, thus providing gold standard data for morphological analysis. Currently, the analysis candidates provided by the model include the human-chosen tag in 99.3% cases and human-chosen lemma in 99.6% cases, with the differences mostly caused by the interpretation of proper names, brand names, and other foreign inclusions. 97.9% of the words in Latvian Treebank had a matching entry in the Tēzaurs.lv lexicon, the lemmas, and paradigms for the rest were 'guessed' as described in section 4.1.

The main current usage of the synthesis model is in providing inflection tables for lexemes of the Tēzaurs.lv online dictionary (Spektors et al., 2023). It has also been used for generating inflections of named entities in a knowledge base project (Paikens, 2014) and generating wordlists for language games and spell checking, and is available as an API for various niche use cases.

The structure of the morphological model was also validated by including data from the Dictionary of Latvian Literary Language (Tors Laimdots Ceplītis (ed), 1972–1996) and Dictionary of Contemporary Latvian Language (Jērāne et al., 2023).

As both Tēzaurs.lv and the corpora collection are widely used public resources, they receive a substantial amount of user feedback so any mistakes and inaccuracies are rapidly detected and reported, thus providing a practical validation of both the model principles and lexical data.

In this process, we have identified certain phenomena in Latvian which don't fit the model well and can be implemented only as explicitly listed exception forms.

A few compound nouns (such as *vecāmāte* (literally 'old-mother', meaning 'grandmother'), *šīsaule* 'this-world') can be inflected as if they were two separate words, inflecting also the ending of the first part of the compound, unlike the general Latvian principle that compound nouns are inflected as if they were a single noun.

There are also certain regional archaic words that don't match the inflection principles of Latvian literary language and in some cases, we don't have evidence to assume how they are inflected. In this model, we treat them as indeclinable, even though the relevant dialects likely did inflect these words in some manner.

7. Ongoing and future work

Currently, there is an ongoing effort to develop a morphological model for Latgalian (a closely related language to Latvian) in the framework described in this paper. This task includes developing the set of inflectional paradigms and stem change rules as well as creating an initial Latgalian lexicon with specified paradigms.

In future, we also intend to replicate our model and knowledge gathered in the *GrammaticalFramework* framework to improve its Latvian resource grammar (Paikens and Gruzitis, 2012). Other

plans include augmenting our model with pronunciation generation for each wordform, as well as extending the model to include inflection/analysis of multi-word expressions similar to what has been done for Hebrew by Al-Haj et al. (2014).

8. Acknowledgements

This work has been supported by the EU Recovery and Resilience Facility projects Language Technology Initiative (No 2.3.1.1.i.0/1/22/I/CFLA/002) and Latvian Quantum Initiative (No. 2.3.1.1.i.0/1/22/I/CFLA/001).

9. Bibliographical References

Hassan Al-Haj, Alon Itai, and Shuly Wintner. 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Aleksejs Andronovs. 1997. Pārdomas par verba locīšanu latviešu valodā. *Latvijas Zinātņu Vēstis*, 51:30–35.

Ilze Auziņa, Ieva Breņķe, Juris Grigorjevs, Inese Indričāne, Baiba Ivulāne, Andra Kalnača, Ilze Lokmane, Dace Markus, Daina Nītiņa, Gunta Smiltniece, Baiba Valkovska, and Anna Vulāne. 2013. *Latviešu valodas gramatika*. Latvian Language Institute, University of Latvia, Riga.

Guntis Barzdins, Didzis Gosko, Karlis Cerans, Oskars F. Barzdins, Arturs Znotins, Paulis F. Barzdins, Normunds Gruzitis, Mikus Grasmanis, janis Barzdins, Uldis Lavrinovics, Sinty K. Mayer, Intars Students, Edgars Celms, Arturs Sprogis, Gunta Nespore-Berzkalne, and Paikens Paikens. 2020. Pini language and PiniTree ontology editor: Annotation and verbalisation for atomised journalism. In *Proceedings of the 17th Extended Semantic Web Conference (ESWC): Posters and Demos*.

Agnė Bielinskiene, Loïc Boizou, Jolanta Kovalskaitė, and Erika Rimkutė. 2016. Lithuanian dependency treebank ALKSNIS. In *Proceedings of the Seventh International Conference Baltic HLT 2016*, pages 107–114, Amsterdam. IOS Press.

Loïc Boizou, Jurgita Kapociute-Dzikiene, and Erika Rimkute. 2018. Deeper error analysis of Lithuanian morphological analyzers. In *Human Language Technologies - The Baltic Perspective - Proceedings of the Eighth International Conference Baltic HLT 2018, Tartu, Estonia, 27-29 September 2018*, volume 307 of *Frontiers in Artificial Intelligence and Applications*, pages 18–25. IOS Press.

Brigita Ceplīte and Laimdots Ceplītis. 1991. *Latviešu valodas praktiskā gramatika*. Zvaigzne.

Vidas Daudaravičius, Erika Rimkutė, and Andrius Utka. 2007. Morphological annotation of the Lithuanian corpus. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 94–99. Association for Computational Linguistics.

Daiga Deksne. 2013. *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, chapter Finite State Morphology Tool for Latvian. Association for Computational Linguistics.

Viktorija Drīzule. 1978. Об автоматическом распознавании омонимии флексий латышского языка [On automated recognition of flexive homonymy in Latvian language]. *LZA Vēstis 1978*, pages 79–87.

Steffen Eger and Ineta Sējāne. 2010. An ensemble of classifiers methodology for stemming in inflectional languages: Using the example of Latvian. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, page 217–224, NLD. IOS Press.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, 46:131–142.

Trevor Garth Fennel. 1980. *A Grammar of Modern Latvian*. The Hauge, Mouton.

Trevor Garth Fennel. 1985. *A Derivational Dictionary of Latvian*. Helmut Buske.

Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalnina, Laura Rituma, Laune Strankale, Artūrs Znotiņš, and Normunds Grūzītis. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pages 400–418. Lexical Computing CZ s.r.o.

Inguna Greitāne. 1994. Latviešu valodas lokāmo vārdšķiru locīšanas algoritmi. (Algorithms for Latvian form generation). *LZA Vēstis 1994*, pages 32–39.

Jan Hajic, Pavel Krbec, Pavel Kveton, Karel Oliva, and Vladimir Petkevic. 2001. Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Toulouse, France. Association for Computational Linguistics.

Heiki-Jaan Kaalep. 1997. An Estonian morphological analyser and the impact of a corpus on its development. *Computers and the Humanities*, 31(2):115–133.

Heiki-Jaan Kaalep, Sjur Nørstebø Moshagen, and Trond Trosterud. 2018. Estonian morphology in

the Giella infrastructure. In *Baltic HLT*, pages 47–54.

Heiki-Jaan Kaalep and Tarmo Vaino. 2001. Complete morphological analysis in the linguist's toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pages 9–16.

Andra Kalnača and Ilze Lokmane. 2021. *Latvian Grammar*. University of Latvia Press, Riga.

Jurgita Kapočiūtė-Dzikienė, Erika Rimkutė, and Loic Boizou. 2017. A comparison of Lithuanian morphological analyzers. In *Text, Speech, and Dialogue*, pages 47–56, Cham. Springer International Publishing.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Karlis Kreslins. 1996. *A stemming algorithm for Latvian*. Ph.D. thesis, Loughborough University.

Baiba Krūze-Krauze. 1996. Latviešu valodas atvasināto vārdu morfēmiskā segmentācija (datorlingvistiska realizācija) : bakalaura darbs. Bachelor thesis, University of Latvia, Rīga.

Baiba Krūze-Krauze. 1998. Datorizēta latviešu valodas morfēmiski morfoloģiskā analīze : maģistra darbs. Master's thesis, University of Latvia, Rīga.

Kristīne Levāne and Andrejs Spektors. 2000. Morphemic analysis and morphological tagging of Latvian corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation, vol. 2*, pages 1095–1098.

Krister Linden, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen. 2011. Hfst—framework for compiling and applying morphologies. Vol. 100:67–85.

Jūrate Mackevičiūtė. 2004. Lithuanian morphological analysis system and grammar checker: Tilde's technologies in practice. In *Proceedings of Baltic HLT*.

Nicole Nau. 1998. *Latvian*, volume 217 of *Languages of the world*. Lincom Europa.

Paikens Paikens. 2007. Lexicon-based morphological analysis of Latvian language. In *Proceedings of 3rd Baltic Conference on Human Language Technologies (HLT 2007)*.

Pēteris Paikens. 2014. Latvian newswire information extraction system and entity knowledge base. In *Human Language Technologies - The Baltic Perspective*, volume 268. IOS Press.

Pēteris Paikens. 2016. Deep neural learning approaches for Latvian morphological tagging. In *Human Language Technologies – The Baltic Perspective*, volume 289. IOS Press.

Pēteris Paikens and Normunds Gruzitis. 2012. An implementation of a Latvian resource grammar in grammatical framework. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Pēteris Paikens, Laura Rituma, and Lauma Pretkalniņa. 2013. Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) NEALT Proceedings Series 16*, pages 267–278, Oslo.

Giedrė Pajarskaitė, Vilma Griciūtė, Gailius Raškinis, and Jan Kuper. 2004. Designing hmm-based part-of-speech tagger for lithuanian language. *Informatica*, 15(2):231–242.

Mārcis Pinnis and Kārlis Goba. 2011. Maximum entropy model for disambiguation of rich morphological tags. In *Systems and Frameworks for Computational Morphology, Communications in Computer and Information Science, 1, Volume 100, The 2nd Workshop on Systems and Frameworks for Computational Morphology (SFCM2011)*, pages 14–22.

Uģis Sarkans. 1996. Morphemic and morphological analysis of the Latvian language. In *Proceedings of the Forth conference on Computational Lexicography and Text Research*, pages 219–225.

Baiba Saulīte, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levane-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikans, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, and Jānis Ziediņš. 2022. Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

Inguna Skadiņa. 2004. Latviešu valodas morfoloģiskās analīzes sistēma – tās nozīme teikuma pareizrakstības pārbaudē. In *Vārds un tā pētīšanas aspekti 8*, pages 282–290.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Artūrs Treimanis. 2023. Latviešu valodas morfoloģiskā marķēšana, izmantojot dziļās mašīnmācīšanās metodes. Bachelor thesis, University of Latvia.

Andrejs Vasiļjevs, Jana Ķikāne, and Raivis Skadiņš. 2004. Development of HLT for Baltic languages in widely used applications. In *Proceedings of First Baltic Conference "Human Language Technologies – the Baltic Perspective"*, pages 198–202.

Marcin Woliński. 2006. Morfeusz - a practical tool for the morphological analysis of polish. In *Intelligent Information Systems*.

Marcin Woliński. 2014. Morfeusz reloaded. In *International Conference on Language Resources and Evaluation*.

Shlomo Yona and Shuly Wintner. 2008. A finite-state morphological grammar of hebrew. *Natural Language Engineering*, 14(2):173–190.

Deniz Yuret and Ferhan Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, pages 328–334.

Vytautas Zinkevičius. 2000. "Lemuoklis" – morfologinei analizei. *Darbai ir dienos*, 24:245–274.

Artūrs Znotiņš and Elita Cīrule. 2018. NLP-PIPE: Latvian NLP tool pipeline. In *Human Language Technologies - The Baltic Perspective*, volume 307, pages 183–189. IOS Press.

Pēteris Ņikiforovs. 2015. Latviešu valodas morfosintaktiskais marķētājs : bakalaura darbs. Bachelor thesis, University of Latvia.

10. Language Resource References

Jērāne, Santa and Kuplā, Ieva and Lejniece, Gunta and Migla, Ilga and Oldere, Laimdota and Ozola, Ārija and Požarnova, Vija and Roze, Anitra and Šmidebergs, Imants and Šnē, Dorisa and Šnē, Māra and Zuicena, Ieva and Pretkalniņa, Lauma and Auziņa, Ieva and Briede, Santa and Šmidebergs, Imants and Timuška, Agris. 2023. *Dictionary of Contemporary Latvian Language (MLVV) (2023-07-07)*. [link].

Levāne-Petrova, Kristīne and Darģis, Roberts. 2018. *Balanced Corpus of Modern Latvian (LVK2018)*. CLARIN-LV digital library at IMCS, University of Latvia, ISLRN http://hdl.handle.net/20.500.12574/11.

Levāne-Petrova, Kristīne and Darģis, Roberts and Pokratniece, Kristīne and Lasmanis, Viesturs Jūlijs. 2023. *Balanced Corpus of Modern Latvian (LVK2022)*. [link].

Rituma, Laura and Pretkalniņa, Lauma and Saulīte, Baiba and Nešpore-Bērzkalne, Gunta and Grūzītis, Normunds. 2023. *LVTB - Latvian Treebank v2.12 (2023-06-05)*. [link].

Spektors, Andrejs and Pretkalniņa, Lauma and Grūzītis, Normunds and Paikens, Pēteris and Rituma, Laura and Saulīte, Baiba and Nešpore-Bērzkalne, Gunta and Lokmane, Ilze and Klints, Agute and Stāde, Madara and Grasmanis, Mikus and Strankale, Laine and Auziņa, Ilze and Znotiņš, Artūrs and Darģis, Roberts and Bārzdiņš, Guntis. 2023. *Tēzaurs.lv 2023 (Summer Edition)*. [link].

Tors Laimdots Ceplītis (ed). 1972–1996. *Latviešu literārās valodas vārdnīca*. Zinātne. [link].

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia. 2021. *Universal Dependencies 2.9*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, ISLRN http://hdl.handle.net/11234/1-4611.