Database of Latvian Morphemes and Derivational Models: ideas and expected results

Andra Kalnača
University of Latvia
Visvalža 4a, Rīga, LV-1050
Latvia
andra.kalnaca@lu.lv

Tatjana Pakalne
University of Latvia
Visvalža 4a, Rīga, LV-1050
Latvia
tatjana.pakalne@lu.lv

Kristīne Levāne-Petrova
University of Latvia
Visvalža 4a, Rīga, LV-1050
Latvia
kristine.levane-petrova@lu.lv

Abstract

In this paper, we describe "The Database of Latvian Morphemes and Derivational Models" – a large-scale manually validated database of Latvian derivational morphology currently in development at the Department of Latvian and Baltic Studies, Faculty of Humanities, University of Latvia (project funded by Latvian Council of Science, No. lzp-2022/1-0013). The database is based on lemmas extracted from the Balanced Corpus of Modern Latvian (LVK2018) and consists of two basic interlinked parts: an annotated list of morphemes and an annotated list of lemmas containing those morphemes. Morpheme-level data include morphemes with morpheme variants (allomorphs) and manually resolved morpheme homonymy/ homography, as well as information on morpheme types and hierarchical (diachronic) relations between root morphemes. Lemma-level data for each lemma include a unique lemma ID (coinciding with the original string extracted from the corpus), a manually validated base form, as well as information on morphemic segmentation, POS, grammatical features, derivational motivation (incl. compounding) and word-family membership. The focus of the database is on providing linguistically accurate comprehensive data as a reliable basis for future work in different fields, incl. computational linguistics.

1 Introduction

Latvian (Baltic group, Indo-European language family) is a language with rich inflectional and derivational morphology. Latvian inflectional morphology is extensively documented in linguistic literature, e.g., in academic grammars (Endzelīns, 1951; Kalnača and Lokmane, 2021; Nītiņa and Grigorjevs, 2013), and, by virtue of being paradigmatic (and, as far as NLP is concerned, also synchronic), relatively readily submits to formalization, at least at the conceptual, if not at the practical, level. Over the last three decades, a number of approaches have been developed for Latvian inflectional morphology processing, resulting in solutions for wordform analysis, generation, lemmatization, POS-tagging, etc., many of them using some version of a lexicon for greater precision; for a recent proposal and an overview of previous work, see Paikens et al. (2024). Data on Latvian inflection are also available in UniMorph, which contains 136998 Latvian inflected forms corresponding to 7548 paradigms[1] (Kirov et al., 2018).

The derivational structure of words is inherently less straightforward and involves several levels of complexity (see Section 4), which need to be taken into account when developing derivational morphology processing technologies. Early computational linguistic experiments on Latvian derivational morphology have included attempts at describing possible approaches to automated morphemic segmentation of derived Latvian words and morphemic and morphological analysis, e.g., (Sarkans, 1996), but, to the best of our knowledge, no comprehensive working computational linguistic models of Latvian derivational morphology have been developed so far. It should be pointed out that up to now there has also been a lack of scientifically accurate large-scale resources (e.g., manually validated databases, lexicons) dedicated to Latvian derivational morphology that could serve as a basis for developing and testing computational linguistic, e.g., rule-based, models. The

[1] https://github.com/unimorph/lav

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 279–286
March 3-4, 2025

most complete inventory of morphemically segmented Latvian words (base forms) to date, organized into word families based on a common root or, in some cases, on a non-segmentable stem, is Baiba Metuzāle-Kangere's "Derivational Dictionary of Latvian" (a printed dictionary) (Metuzāle-Kangere, 1985).

Decisions about correct morphemic segmentation of complex words, derivational motivation or, e.g., allomorphism are not always straightforward for human linguists, and even less so for automated solutions unless the latter are trained or based on a large reliable body of data. In this paper, we describe a new digital resource (a database) dedicated to Latvian derivational morphology, currently in development and to be made freely available to the public in 2026. The "Database of Latvian Morphemes and Derivational Models" (DLMDM) is a corpus-based manually validated database in text format (`.tsv` files) with comprehensive data on the basic regularities of Latvian derivational morphology. DLMDM is designed as a general reference resource, its focus is on producing a large structured manually validated set of data accurate and consistent from the point of view of linguistic theory for the general public for all kinds of future uses, incl. as a source for NLP research.

2 Related work

Printed dictionaries of morphemes and derivational dictionaries have been around for quite some time. Particularly well represented are Slavic languages, e.g. Slovak (Sokolová et al., 1999), Czech (Slavíčková, 1975; Šiška, 1998). There are also word-family dictionaries for other languages, e.g., German (Splett, 2009; Augst, 2009). Two notable dictionaries reflecting different aspects of Latvian morphemics and derivational morphology are "A Derivational Dictionary of Latvian" (Metuzāle-Kangere, 1985) and "Latīņu un grieķu cilmes vārddaļu vārdnīca" (A dictionary of Latin and Greek word parts) (Skujiņa, 1999). Metuzāle-Kangere's dictionary is built around the concept of derivational families and is based on words extracted from two bilingual dictionaries.

The last 20 years have seen an increase in digital resources containing some sort of morphemic and/or derivational information. Such resources are often corpus-based in an effort to reflect actual contemporary language use, but differ by focus, scope and methodology (e.g. autoconstructed vs., less frequently, manually annotated). Some of the recent examples include the Database of Lithuanian Morphemics Data (Rimkutė et al., 2013), Morpholex, a lexical database for English words with morphological variables (Sánchez Gutiérrez et al., 2018), DeriNet (Vidra et al., 2019), a lexical network of word-formation relations in Czech, with autogenerated morphological segmentations of lemmas and identification of root morphs. Universal Derivations (UDer) is a collection of harmonized lexical networks of various languages capturing word formation, especially derivation, in a cross-linguistically consistent annotation scheme based on a rooted tree data structure as used in the DeriNet 2.0 database. MorphyNet is a large-scale, multilingual database that includes derivational and inflectional morphology data (over 13 million inflections and over 700 thousand derivations) for 15 languages extracted from Wiktionary and 90 thousand derivations in 271 languages inferred automatically from the combination of MorphyNet and the Universal Knowledge Core (Batsuren et al., 2021). UniMorph 2.0 contains some data on Latvian derivational morphology as supplementary structured data extracted from Wiktionary – 4235 complex words with a possible source word, a formally defined (POS:POS) word-formation model and means of derivation specified for each word[2]. The quality of these data depends on the accuracy of Wiktionary and the level of detail is limited to what is available from that resource; e.g., derivation is not distinguished from compounding and formal means of derivation are not specified as morphemes of a certain type, but rather as word-initial or word-final strings of one or more morphemes merged together. Morphemic, incl. derivational, information is also included in a number of broader scope lexical resources, e.g., the lexical database of English WordNet encodes some derivational relations, the CELEX lexical databases of English, Dutch and German contain data on the derivational and compositional structure of words. Several approaches for induction of derivational families from words extracted from large corpora have been developed, e.g. DErivBase, DErivCELEX for German, DerivBase.Hr for Croatian, etc.

[2] `https://github.com/unimorph/lav/blob/master/lav.derivations`

3 Stages of development

DLMDM is based on a case-sensitive list of 165 090 lemmas downloaded in `.xml` format from The Balanced Corpus of Modern Latvian (LVK2018) via Nosketchengine (Rychly, 2007) with zero lower frequency threshold. LVK2018 contains approximately 10 million words occurring in texts of various genres (Levāne-Petrova and Darģis, 2018) and has been chosen as the primary initial source of lemmas for the database, because it provides a snapshot of real, unidealized contemporary language use and, apart from established words, also contains novel formations (hence, the zero lower frequency threshold). Adding lemmas from other sources, e.g., other corpora, dictionaries, etc., is possible by assigning a unique lemma ID in the LEMID column and providing a source ID in the SOURCE column.

Automated pre-processing:

- Data extraction.
- Consecutive automated and semi-automated removal of invalid lemmas – removing lemmas containing characters that are not part of the Latvian alphabet and then double-matching the remaining lemmas against tēzaurs.lv (2020 spring version[3]) and an open source spelling checking dictionary[4], resulting in a list of unrecognized lemmas, which were then reviewed manually.
- Approximately 75 000 lemmas left as likely valid for further processing. The lemmas that have been filtered out include non-words, words in foreign languages, words containing spelling mistakes, erroneously generated lemmas, as well as a lot of proper names, some of which (rare or untypical for Latvian) have been left out from the final list;
- Morphological tagging[5], using a freely available tagger for Latvian.
- Rule-based automated morphemic segmentation using custom developed scripts.
- Grouping of lemmas into potential word families based on a shared root (or a non-segmentable stem) and a list of possible root allomorphs.

Further manual processing:

- Reviewing and correcting automatically generated lemma-level and word family data (see Section 5).
- Root homonymy/ homography resolution.
- Defining hierarchical relations between roots and non-segmentable stems.

The final stage of development will consist in defining and validating derivational relations between lemmas within word families.

In terms of workload, the most labour-intensive tasks have been morpheme homonymy/ homography resolution, as homographic morphemes have turned out to be pervasive in Latvian lexis, identifying synchronically non-evident allomorphs and also identifying hierarchical relations between roots and word-family membership of lemmas in non-straightforward cases.

4 Sources of complexity in data

As a manually validated database, DLMDM's primary focus is on providing comprehensive linguistically accurate data. This means accounting for all kinds of phenomena in derivational morphology, not just productive regular derivation. In this section, we outline some of the major sources of difficulty in derivational morphological analysis of existing words.

4.1 Morpheme homonymy and homography

Homonymy or homography is encountered much more often among Latvian roots and non-segmentable stems than among words. Derivational analysis without homonymy/ homography resolution may lead to incorrectly inferring derivational relations between words and, hence, to incorrect semantic interpretation (roots shown in round brackets):

(1) *(bur)-t* 'to do magic' (inherited Latvian word) – *(bur)-a* 'sail' (borrowing)
(2) *(las)-ī-t* 'to read' – *(las)-is* 'salmon' (both – inherited Latvian words)
(3) *(mat)-s* 'hair' (inherited Latvian

[3] https://github.com/LUMII-AILab/Tezaurs.git/
[4] http://dict.dv.lv/download.php?prj=lv/
[5] https://github.com/PeterisP/LVTagger.git/

word) – *(mat)-s* ‘checkmate’ – *(fiz)-(mat)-s* ‘physico-mathematical (of students)’ (borrowing)
(4) *(log)-s* ‘window’ (inherited Latvian word) – *(virus)-o-(log)-s* ‘virologist’ – *ielogoties* ‘to log in’ (both – borrowings)

E.g., the string ‘lok’ or ‘loc’ corresponds to at least 7 different roots, in Latvian, occurring in hundreds of lemmas, as in (5)–(11):

(5) lok [luok], loc [luoc] – *loks* ‘circle’, *locīt* ‘to bend’, *lokāms* ‘bendable, declinable’ (inherited Latvian words)
(6) lok [lok], loc [loc] – *lokācija* ‘location’, *lokalizācija* ‘localization’, *lokatīvs* ‘locative’, *lokomotīve* ‘locomotive’, *translocēt* ‘translocate’ (all – borrowings)
(7) lok [luok], loc [luoc] – *ķiploks* ‘garlic’ (borrowing), *ķiplokains* ‘garlicky’, *ķiplociņš* ‘garlic deminutive’, *ķiploksāls* ‘garlic salt’
(8) lok [luok], loc [luoc] – *loki* ‘green onions’, *maurloki* ‘chives’, *sīpolloki* ‘spring onions’ (all – borrowings)
(9) lok [lok] – *loka* ‘hair curl’, *lokains* ‘curly’, *lokoties* ‘to curl’, *lokšķēres* ‘curling iron’ (all – borrowings)
(10) loc [luoc] – *locis* ‘ship pilot’ (borrowing)
(11) lok [lok] – *lokauts* ‘lockout’ (borrowing)

In DLMDM, homonymous/ homographic roots are listed as separate non-related morphemes each linked to their respective word family (or subfamily).

Another problem are quasi-morphemes – sequences of characters in borrowed words graphically coinciding with existing morphemes, most notably, suffixes. Quasi-morphemes may potentially lead to incorrect segmentation, e.g. in automated morphemic segmentation approaches:

(12) *(bārd)-**ain**-is* ‘a bearded man’ (inherited Latvian word) – *(sul**ain**)-is* ‘butler’ (borrowed from Estonian *sulane*[6])
(13) *(rūp)-**est**-s* ‘concern’ (inherited Latvian word) – *(dien**est**)-s* ‘service’ (borrowed from Middle Low German *dēnest*[7])
(14) *(vair)-**og**-s* ‘shield’ (inherited Latvian word) – *(kar**og**)-s* ‘flag’ (borrowed from Old Russian[8])

Other examples of quasi-morphemes include the nouns *ceri**ņ**i* ‘lilacs’, *treni**ņ**š* ‘training’, *zāb**ak**s* ‘a boot’, etc., where as a result of phonetic adaptation the segments *-iņ-* and *-ak-* have come to resemble the Latvian suffixes *-iņ-*, *-ak-*. Quasi-morphemes are less widespread than homonymous / homographic roots.

4.2 Allomorphism

The majority of Latvian roots have variants (root allomorphs) resulting from both historical and synchronic morphophonological processes (Kalnača, 2004; Kalnača and Lokmane, 2021; Nītiņa and Grigorjevs, 2013). Allomorphism is significant in inferring derivational relations between words. E.g., *ved, ves, ve, vez, vež, vad, vaz, važ* are all variants of the same root as in *vest* ‘to carry’, *vešana* ‘carrying’, *vedējs* ‘carrier’, *vadīt* ‘to lead’, etc.

Allomorphism also occurs in affixes, e.g., suffixes *-niek-, -niec-, -nieč-*, as in (15):

(15) *saim-niek-s* ‘owner, host’ (M), *saim-niec-e* (F), *saim-nieč-u* (GEN PL, F)

DLMDM encodes relations for all allomorphs occurring in the dataset, but not for all allomorphs that are, in principle, possible in Latvian.

4.3 Synchrony vs. diachrony

While most automated solutions for derivational morphology are synchronically oriented and focus on productive models, correct morphemic segmentation and word-family membership identification may sometimes require a diachronic stance, i.e. recognizing derivational models that are not synchronically productive, but are found in already established words, while retaining semantic motivation, e.g.:

(16) *(zag)-t* ‘to steal’ – *(zag)-l-is* ‘a thief’, *(bēg)-t* ‘to run away, to flee’ – *(bēg)-l-is* ‘a fugitive’, *(ie)-t* ‘to walk’ – *(ie)-l-a* ‘a street’
(17) *(sil)-t* ‘to warm’ – *(sil)-t-s* ‘warm’,

[6] https://mev.tezaurs.lv/sulainis
[7] https://mev.tezaurs.lv/dienests/
[8] https://mev.tezaurs.lv/karogs/

(sal)-t ‘to be cold, to freeze’ – *(sal)-t-s* ‘cold’

(18) *(bes)-t* (<*bed-t) ‘to dig’ – *(bed)-r-e* ‘a pit, a hole’, *(svīs)-t* (<*svīd-t) ‘to sweat’ – *(svied)-r-i* ‘sweat’

On the one hand, defining a synchronically unproductive word-formation model of this sort would probably lead to overgeneration (in generation tasks) and false positives (in analysis). On the other hand, not defining such models would lead to words like *zaglis, bēglis, iela* being segmented and marked as simplex, which would also entail loss of derivational semantic motivation and word-family membership.

In DLMDM, established complex words not corresponding to synchronically productive word-formation models are segmented from a diachronic perspective.

4.4 Non-straightforward derivational relations and semantic motivation

Defining a single directed derivational relation and a single base (i.e. a single base word for derivation or a single syntactic construction for compounds) for each derivationally complex word is not always possible. Some words, in Latvian, may be simultaneously motivated by more than one base, and the perceived motivation may even vary from speaker to speaker, e.g., *burvīgs* ‘charming, enchanting’ and *burvība* ‘charm, sorcery, magic, enchantment’ are both related to *burvis / burve* ‘wizard, sorcerer (M) and (F)’ and to each other, esp. when taking word senses into account. Certain kinds of words, often these are compounds, rather than having a single base tend to form clusters around concepts (or some would perhaps say, fill in paradigms of possible meanings and parts-of-speech), while also forming links to one another, e.g., *aitas kopt* ‘to farm sheep’ – *aitkopis / aitkope* ‘sheep farmer (M) and (F)’, *aitkopība* ‘sheep farming’; *gara aste* ‘a long tail’, *garaste, garastis* ‘someone having a long tail (F) and (M)’, *garastes* ‘long-tailed’ (a compound genitive noun), *garastains* ‘long-tailed’ (an adjective), *Garastene* (a proper noun in LVK2018); *lēkt ar izpletni* ‘to parachute’ – *izpletņlēkšana* ‘parachuting’, *izpletņlēcējs* ‘someone who parachutes’, etc. Another kind of examples are pairs of compound genitive nouns and adjectives related to one and the same concept, e.g., *starpnāciju, starpnacionāls* ‘international’; *pārreģionu, pārreģionāls* ‘transregional’, *bezgaršas, bezgaršīgs* ‘tasteless’, where a prior existence of an adjective that can fill the slot in the right-hand part of the compound seems to be a pre-requisite.

To summarize, a rooted tree does not seem to be able to accommodate all observable kinds of derivational relations in Latvian, therefore, word families in DLMDM are not designed to fit the rooted tree data structure.

4.5 Root hierarchies

Some roots or non-segmentable stems stand in a hierarchical relationship to one another. This is important for accurate morphemic segmentation and word-family membership:

- two or more inherited roots or an inherited and a borrowed root may be siblings with one common parent:

```
# zero-element
## dar
## darb

# zero-element
## dilb
## delm
## deln

# zero-element
## as
## aksi (borrowed)
## akson (borrowed)
```

- one inherited root may be a child of another inherited root when there is no sufficient basis for further segmentation of the former:

```
# aug, audz, audž
## augst
## augš

# av
## aun
## ait
```

Thus, a root (or a non-segmentable stem) in DLMDM may have allomorphs and also a parent root and siblings or a child root, which, in turn, may have allomorphs of their own. Lemmas are linked to a concrete root in a root hierarchy.

5 Types of data in DLMDM

DLMDM consists of co-indexed text files for lemma-level data, morpheme-level data and

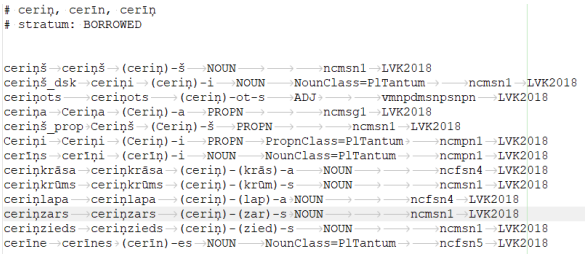

Figure 1: The word family # cerin̦, cerīn, cerī̦ 'lilacs' in a simplified format

source identifiers. To improve readability, manual revision is performed in a simplified format (see Figure 1). Upon completion, the files will be converted to a format compatible with CoNLL-U Plus to facilitate harmonization with other resources.

Each line in a DLMDM file contains data for one entry – a lemma, a morpheme or a source. Column values are tab-delimited.

The format of the database is largely inspired by DeriNet (Vidra et al., 2019) and Morpholex (Sánchez Gutiérrez et al., 2018), but, in terms of contents, DLMDM is different in many respects, the primary objective being to reflect the derivational morphology of Latvian as fully as possible. The major differences, apart from manual revision, include root hierarchies and morpheme-level data, as well as a different approach to marking derivational relations.

5.1 Lemma-level data

At the current stage, lemma data include the following columns:

Column	Description
LEMID	a unique case-sensitive lemma identifier coinciding with the original string extracted from the corpus
LEMMA	a manually validated base form of a lemma
SEGMENTATION	morphemic segmentation of a lemma
POS	part-of-speech tag in the UD format
FEATS	grammatical features
VARIANTS	lemma variants
MORPHTAG	an automatically generated morphological tag
SOURCE	a source identifier

Table 1: Lemma-level data

In addition, each lemma is linked to a concrete root or a non-segmentable stem in a root hierarchy through word-family membership.

Lemmas will be subsequently annotated for means of word-formation (e.g., syntactic: compounding, morphological: prefixation, suffixation), types of a derivational relationship (e.g., single base, multiple motivation) and participants of a derivational relationship.

Since DLMDM includes proper nouns, the LEMID, LEMMA and SEGMENTATION columns are case-sensitive. Two lemmas in the database can have identical values of the LEMMA and SEGMENTATION columns, but not of the LEMID column.

The parts-of-speech represented in DLMDM are shown in Table 2:

POS label	Description
NOUN	a noun
PROPN	a proper noun
ADJ	an adjective
ADV	an adverb
VERB	a verb, incl. participles
INTJ	an interjection
PRON	a pronoun
NUM	a numeral
ADP	an adposition
PART	a particle
CCONJ	a coordinating conjunction
SCONJ	a subordinating conjunction
OTHER	indeclinable words with a verbal motivation that do not fit any of the existing classes, e.g., *paslepu* 'secret', *piespiedu* 'compulsory'

Table 2: POS column values in DLMDM

Developing a unified approach to what is to be considered a valid base form of a lemma (the LEMMA column) has also required some conscious decision-making, e.g., what to do in cases when the corpus contains both a masculine and a feminine version of a derivative, e.g. *nosūtītājs* (M), *nosūtītāja* 'sender' (F), but the automatically generated lemma list only has one of them, as inflectional endings partly overlap; or what to do in cases when the lemma list contains a participle, but not the corresponding verb, although both exist in language.

The manually validated base forms of lemmas in DLMDM are given as follows:

POS	Base forms
NOUN, PROPN	nominative singular or nominative plural for pluralia tantum
ADJ	nominative singular masculine indefinite positive, unless an adjective is only used with the definite ending, e.g., *galvenais* 'principal'
VERB	the infinitive for verb tense forms and nominative singular masculine for declinable participles, except for the past participle active, which is given in masculine and feminine

Table 3: The base forms of lemmas for major declinable parts-of-speech in DLMDM

The FEATS column encodes several specific grammatical features that either cannot be reliably automatically inferred from base forms or are required for other reasons, e.g., because participles do not have a dedicated POS tag (see Table 4).

FEATS	POS
PlTantum – pluralia tantum	NOUN, PROPN, NUM
Gen – genitive nouns or numerals	NOUN, NUM
Indecl – indeclinable words	NOUN, ADJ, NUM
Part – participles	VERB

Table 4: Values of the FEATS column

The VARIANTS column is reserved for linking together different versions or variants, e.g., orthographic, dialectal, of the same word. The MORPHTAG column, which has been automatically generated for the purposes of automated pre-processing, incl. generating POS column values, will be removed in the final version of the database.

5.2 Morpheme-level data

DLMDM contains a separate file for morpheme data co-indexed with the lemma file. Morpheme-level data will include concrete morphemes with allomorphs and homonymy/ homography resolution through unique IDs, as well as information on morpheme types, morpheme strata (e.g., for borrowed roots or non-segmentable stems), hierarchical relationships between roots or non-segmentable stems in a root hierarchy, and, for roots, links to lemmas through word-family membership.

6 Summary

We hope that DLMDM will be useful as a reliable large-scale resource for further research on Latvian derivational morphology from various perspectives, incl. computational linguistics, corpus linguistics and linguistics. Future work might include a more in-depth analysis of the structure of borrowed words in Latvian, esp. international words, words of classical (Greek, Latin) origin, incl. neoclassical compounds.

Abbreviations

GEN – genitive
F – feminine
M – masculine
PL – plural

References

Gerhard Augst. 2009. *Wortfamilienwörterbuch der deutschen Gegenwartssprache*. Max Niemeyer Verlag, Berlin, New York.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Online. Association for Computational Linguistics.

Jānis Endzelīns. 1951. *Latviešu valodas gramatika*. Latvijas Valsts izdevniecība, Rīga.

Andra Kalnača. 2004. *Morfēmika un morfonoloģija*. Latvijas Universitātes Akadēmiskais apgāds, Rīga.

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

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Arya D. 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). European Language Resources Association (ELRA).

Kristīne Levāne-Petrova and Roberts Darģis. 2018. Balanced Corpus of Modern Latvian (LVK2018). CLARIN-LV digital library at IMCS.

Baiba Metuzāle-Kangere. 1985. *A Derivational Dictionary of Latvian*. Helmut Buske Verlag, Hamburg.

Daina Nītiņa and Juris Grigorjevs, editors. 2013. *Latviešu valodas gramatika*. Latvijas Universitātes Latviešu valodas institūts, Rīga.

Peteris Paikens, Lauma Pretkalniņa, and Laura Rituma. 2024. A computational model of Latvian morphology. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 221–232, Torino, Italia. ELRA and ICCL.

Erika Rimkutė, Asta Kazlauskienė, Gailius Raškinis, and Irena Markievicz. 2013. *Lietuviu kalbos morfemikos duomenu bazė*. Vytauto Didžiojo universitetas, Kaunas.

Pavel Rychly. 2007. Manatee/bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masaryk University.

Uģis Sarkans. 1996. Morphemic and morphological analysis of the latvian language. *Proceedings of the Fourth conference on Computational Lexicography and Text Research*, 28(1):219–225.

Valentīna Skujiņa. 1999. *Latīņu un grieķu cilmes vārddaļu vārdnīca*. Kamene, Rīga.

Eleonora Slavíčková. 1975. *Retrograde morphemic dictionary of Czech language*. Academia, Prague.

Miloslava Sokolová, Gustav Moško, František Šimon, and Vladimir Benko. 1999. *Morfematický slovník slovenčiny*. Náuka, Prešov.

Jochen Splett. 2009. *Deutsches Wortfamilienwörterbuch: Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes*. De Gruyter, Berlin, New York.

Claudia Sánchez Gutiérrez, Hugo Mailhot, Hélène Deacon, and Maximiliano Wilson. 2018. Morpholex: A derivational morphological database for 70,000 english words. *Behavior Research Methods*, http://link.springer.com/article/10.3758/s13428-017-0981-8:1–13.

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 2019. DeriNet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Zbyněk Šiška. 1998. *Bázový morfematický slovník češtiny*. Palacký University, Olomouc.