Towards a Derivational Semantics Resource for Latvian

Ilze Lokmane, Mikus Grasmanis, Agute Klints, Gunta Nešpore-Bērzkalne,
Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, Evelīna Tauriņa
Institute of Mathematics and Computer Science
University of Latvia
Raiņa bulvāris 29, Riga, Latvia
ilze.lokmane@lu.lv, (mikus.grasmanis, agute.klints,
gunta.nespore, peteris.paikens, lauma.pretkalnina,
laura.rituma, madara.stade)@lumii.lv, taurina.evelina@gmail.com

Abstract

In this paper, we describe the implementation of the first structured resource of semantic derivational links for Latvian, basing it on the largest online dictionary Tēzaurs.lv and linking it to the Latvian WordNet. We separate two kinds of derivational links: semantic derivation links between senses and morphological derivation links between lexemes. Semantic links between senses are defined as a pair of semantic labels assigned to both ends of the link. The process of semantic linking involves revising the sense inventory of both the base word and the derivative, defining semantic labels for lexemes of four basic word classes – nouns, verbs, adjectives, and adverbs, and adding the appropriate labels to the corresponding senses. We exemplify our findings with a detailed representation of the sense relations between a base verb and its nominal derivatives.

Keywords: morphosemantic relations, derivational semantics, polysemous words, WordNet, Latvian

1 Introduction

So far, no derivational semantics resource has been created for the Latvian language. The idea for its creation grew out of the desire to extend the Latvian WordNet (Paikens et al., 2023) because regular derivatives are an essential part of the lexicon, and they also have semantic relations both with their base words and with each other, for example, two derivatives can be synonyms. Latvian WordNet is planned to be supplemented with derivational links, similar to what Princeton WordNet (Mititelu et al., 2021) and others have implemented (e.g. Turkish (Bilgin et al., 2004), Bulgarian (Dimitrova et al., 2014), Romanian (Mititelu, 2012), Czech (Rambousek et al., 2018), Polish (Piasecki et al., 2012)). We consider derivational semantics resources relevant for NLP applications because the behavior of current large language model chat agents for less resourced languages like Latvian shows a misunderstanding of meaning of derived words, so the application of lexical resources has value even in the era of large pre-trained models.

Latvian WordNet has been developed manually for the past four years (Paikens et al., 2023). As of Autumn 2024, Latvian WordNet contains 8756 synsets which cover the meanings of the 2000 most frequently used words in The Balanced Corpus of Modern Latvian (Levāne-Petrova and Darģis, 2018) and their related synsets. The inventory of words and senses is based on the Tēzaurs.lv online dictionary (Spektors et al., 2023; Grasmanis et al., 2023), which is a large (approximately 405000 entries in the last release in September 2024) digital compilation of legacy dictionaries. Latvian WordNet is developed and maintained on the Tēzaurs.lv lexicographic platform, and the data are available in dictionary entries of words whose senses are included in WordNet. This lexical resource also contains links between Latvian WordNet and Princeton WordNet (Fellbaum, 1998). The important thing is that the Latvian WordNet is created between separate word senses, and we also want to create the semantics of derivatives separately for each word sense, so we think these resources will be well integrated. Currently, the semantics of derivatives is a network parallel to Latvian WordNet, the word sense inventory being the unifying element which is involved in both networks. In the future, that will help to integrate one resource into the other.

Up until now, according to the traditions of lexicography, the regular derivatives listed in the

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 359–366
March 3-4, 2025 ©2025 University of Tartu Library

Tēzaurs.lv dictionary had their own entries only if they had a specific sense which was far removed from the senses of the base word. In order to represent the diversity of derivational relations, we are currently creating new entries for the most frequently used regular derivatives.

In Latvian linguistics little or no attention has been paid to semantic relations between the senses of a polysemous base word and the senses of its derivatives, as only general semantics of derivational formatives has been studied and described referring to the basic sense of the base word (Kalnača and Lokmane, 2021; Soida, 2009). In order to improve Tēzaurs.lv and Latvian WordNet, it should be verified whether these relations exist between all senses of the base word and the derivative (in more detail in Chapter 3.3). Therefore, we have chosen to employ two kinds of derivational links: morphological derivation links between lexemes and semantic derivation links between exact word senses (described in more detail in Chapters 2 and 3). A morphological derivation link contains information about the formatives used in word formation, while a semantic link is formed as a pair of semantic labels that describe both linked senses.

The choice of word pairs for annotating is determined by their frequency of use in The Balanced Corpus of Modern Latvian (Levāne-Petrova and Darģis, 2018). First, the derivatives of the most frequently used verbs, which are already included in Latvian WordNet, are marked to enrich the lexical information of these words as much as possible. Second, the most frequently used derivations in each derivation group are selected, for example, the most frequently used adjectives derived from nouns. The following word pairs are annotated in this phase of the project: a) verbs – deverbal nouns, b) nouns – denominal verbs, c) nouns – denominal adjectives, d) adjectives – deadjectival adverbs. Such groups were chosen to cover the four main word classes of the Latvian language involved in word formation processes. Other patterns of derivational links will be annotated as the project progresses, including patterns when a derivative is of the same word class as the base word. The processed data set currently includes 1000 morphological links and 1600 semantic links.

To ensure a reliable resource for future research, the dataset is developed manually. However, we assume that in future some semi-automatic methods could also be applied to unambiguous words to ensure a larger coverage, which is essential for NLP applications of this dataset.

2 Morphological Derivational Links

A morphological derivation link between lexemes connects the base word entry to the derived word entry. This link contains two attributes: a derivational stem base and a derivational formative. The stem indicates which part and form of the base word the derivative is formed from. The formative is the means by which a new word is made; it can be a single morpheme, such as a prefix or a suffix, or a combination of morphemes, such as a suffix and an ending, that together form a complex formative. For example, the noun *skrējējs* 'runner' is formed by adding formatives *-ēj-* and *-s* to the past tense stem of the verb *skriet* 'to run'; and the adjective *mākoņains* 'cloudy' is formed by adding formatives *-ain-* and *-s* to the plural stem of the noun *mākonis* 'cloud'.

Since the Latvian language has an extremely rich inflectional and derivational morphology (Kalnača and Lokmane, 2021), new words can be made from various stems, e.g., the present, past, infinitive or participle stems of verbs and singular or plural stems of nouns, using prefixes, suffixes, endings, and interfixes. Therefore, information about the derivational stem seems to be crucial in describing Latvian derivational morphology.

In addition, this information will help in further studies regarding the semantic properties that derivatives obtain with certain derivative formatives. Although Latvian grammars (e.g., (Kalnača and Lokmane, 2021; Soida, 2009)) provide general information of the semantic aspects of such formatives, wider language material could potentially lead to new insights, assist in determining previously undescribed peculiarities of derivative senses, and specify derivational stem bases.

However, our aim does not include dividing the entire word into morphemes; the internal composition of Latvian words is the objective of another project, "Database of Latvian Morphemes and Derivational Models" (see `https://www.dlmdm.lu.lv`). Instead, we only indicate the morphemes involved in the derivative process.

In most cases, the derivational direction between two words is clear, i.e., the base word and the derivative can be discerned by consulting the already described models of word formation.

However, there are derivational relations in which it is not obvious which of the two is the base word and which is the derivative (e. g., *kontrolēt* 'to control' – *kontrole* 'control'; *spēlēt* 'to play (a game)' – *spēle* 'a game'). This problem arises mainly (but not exclusively) in pairs of loan words where it is not possible to establish which of the words was introduced into Latvian first; this means that both derivational paths are possible in such cases, as both models of word formation are possible in Latvian. A noun can be derived from a verb (e.g., *atsaukties* 'to refer' – *atsauce* 'a reference'; *aizstāvēt* 'to defend' – *aizstāvis* 'a defender'), and a verb can be derived from a noun (e.g., *skaips* 'Skype' – *skaipot* 'to communicate via Skype', *balva* 'an award' – *apbalvot* 'to reward'). There are also more recent loan word pairs that are clearly derivationally linked, but are probably not derived from each other (e.g., *bioloģija* 'biology' – *bioloģisks* 'biological'; *demokrātija* 'democracy' – *demokrātisks* 'democratic'). In such instances, the solution is to label the link between lexemes as 'derivationally related' without specifying which is the base word and which is the derivative; information on the stem base and the formatives is also not provided.

3 Semantic Derivational Links

Due to the fact that semantic relations between the senses of a polysemous base word and the senses of its derivatives are yet to be studied in depth in Latvian linguistics, a new system for annotating such instances had to be devised. This chapter describes the process of preparing entries for linking, creating semantic derivation links between the senses of the base word and its derivatives, semantic labels for each word class combination and more detailed observations of the relations between the senses of polysemous words.

3.1 Revising the Senses of the Base Word and the Derivative

First step for derivational link creation is revising dictionary entries and word senses. The Tēzaurs.lv entries come from various dictionaries, therefore, the criteria for dividing meanings may vary across different entries. We strive to standardize them according to the current criteria for distinguishing senses in the Tēzaurs.lv (see (Lokmane et al., 2021)) and based on the current situation in the language.

Derivatives mostly do not have entries in the Tēzaurs.lv because regular derivatives have not been included in the dictionary until now. Therefore, they need to be created anew. We strive to align the derivative's entry with the entry of the base word (sequence of senses, their granularity), but we try to not create "artificial" meanings for derivatives just to align the entry symmetrically with the base word entry. The verification of the sense is based on corpora data mentioned below. If the word is used in corpora in a particular sense, the sense has to be created and added to the word entry.

Usage examples from several corpora of the Latvian National Corpora Collection (Saulite et al., 2022) are added to the senses of base words and derivatives (examples must be short, clear, of simple syntactic constructions, in examples the word appears in various constructions). The examples also guide the creation and distinction of senses – if in many examples it is not possible to determine in which meaning the word is used, the division of senses should be reconsidered. We add several examples for each sense, but one example is enough to conclude that the sense is being used, therefore it is relevant to entry.

3.2 Semantic Labels

Semantic links between senses are formed as a pair of semantic labels, which are given to both ends of the link. It seems important to record not only the semantics of the derivative, as most grammars do, but also the semantic characteristics of the base word. For example, the sense 'to be lying down' of the base verb *gulēt* labeled as *toBeInState* is linked to the sense 'sleeping place' of the derived noun *guļa* labeled as *location*. Similarly, the sense 'group' of the base noun *kopa* labeled as *abstract notion* is linked to the sense 'used by several or many' of the derived adjective *kopējs* labeled as *related to*. Such an approach will allow future studies of word-formation processes not only from the perspective of the derivative, but also from the perspective of the base word.

Each of the four word classes discussed so far has a different number of semantic labels (see Table 1). Choosing and defining semantic labels is a labor-intensive process, because there are no ready-made samples that can be used without improvements. It should also be emphasized that

Word class	Semantic label	Description
verb	toBeInProcess	to undergo a change of a condition or a state
	toBeInState	to experience a state or a condition
	toDo	to perform an action
noun	abstract notion	a non-concrete concept or idea
	action	something that the verb argument does or performs
	agent	participant who initiates and carries out an action
	animal	a living being except humans
	body part	any part of an organism such as an organ or extremity
	cause	the non-volitional causer of the event
	device	an object or machine used to perform an action
	experiencer	participant experiencing some state or process
	feature	property of an entity
	instrument	the entity that is manipulated by the agent and with which an action is performed
	location	the place in which something is situated or takes place
	member of a profession	a person who works in a specified professional activity
	mythical creature	a supernatural creature that does not exist in real life
	natural phenomenon	a physical event that occurs in atmosphere or on the ground
	patient	participant undergoing the effect of some action
	person	a human being
	physical phenomenon	a natural phenomenon involving the physics of matter and energy
	process	a change in condition or state of the argument
	resource	the entity by which an action is performed and which is used up during the action
	result	entity that comes into existence through the event
	state	the state or condition of the argument
	thing	an inanimate material object
	time (noun)	the period or moment during which something exists or continues
adjective	evaluative	based on or relating to an assessment
	property	expressing a general property like colour, shape etc.
	including	including the entity named by base word
	measurable	expressing a measurable property
	possessing	possessing the entity named by base word
	similar to	similar to the entity named by base word
	related to	related to the entity named by base word
adverb	degree	specifying the degree to which a property applies
	frequency	describing how often something happens
	manner	describing how something happens
	place	describing location in which something is situated or takes place
	time (adv.)	describing when or how long something takes place

Table 1: Semantic labels for senses linked by a relation

the list of labels can be linguistically specific, although some labels are, of course, universal (Mititelu et al., 2021). Thus, the selection of semantic labels takes into account the experience of creating electronic resources of other languages (Bilgin et al., 2004; Piasecki et al., 2012) and linguistic studies of both word class semantics and derivational semantics (Wierzbicka, 1988; Raskin and

Nirenburg, 1995; Soida, 2009; Kalnača and Lokmane, 2021).

The noun has most semantic labels. Firstly, this is due to the fact that nouns are included in several pairs of derivationally linked word classes - as derivatives of verbs and as base words of both denominal verbs and adjectives. Secondly, the semantics of nouns are generally more specific and easier to classify than, for example, the semantics of adjectives (Wierzbicka, 1988). Verbs have only three semantic labels despite being both base words and derivatives in relation to nouns. However, the list of semantic labels is being constantly enriched as we proceed with new lexical groups. In the future, there might be a need for a more detailed semantic division of verbs, e.g., names of motion, communication, cognition etc. Adverbs have been assigned five semantic labels traditionally described in grammars.

One of the most difficult problems so far has been the semantic classification of adjectives, since they, being attributes, derive at least part of their semantics from the noun they are attached to. We have chosen to assign three rather general semantic labels to qualitative (descriptive) adjectives and four semantic labels to relational (denominal) adjectives (for a similar solution, see (Raskin and Nirenburg, 1995)) Qualitative adjectives are morphologically simpler than relational ones. The latter, being more complex formally, derive their semantics from their base words.

In each pair of word classes considered so far, the set of semantic labels is different, to best capture the specific semantics of derivative in relation to the base word.

3.3 Relations between Senses of Polysemous Words

Even within the boundaries of one word and its derivatives, there can be a large variety of semantic relations between them, especially when all the senses of a word are considered. This is exemplified by the verb *atgādināt*, which has 4 senses and 4 noun derivatives (see Figure 1).

The first sense, *atgādināt*$_1$ 'to prompt, to remind, to cue (something forgotten or imperfectly learned)', has two narrower subsenses, *atgādināt*$_{1.1}$ 'to give a reminder (by device)' and *atgādināt*$_{1.2}$ 'to bring back a memory (of something)'. The second sense, *atgādināt*$_2$ 'to resemble' has no subsenses. The 4 derivatives that have been linked to the verb (i.e., the base word) through morphological and semantic links are examined in more detail in the following paragraphs; the links between the base word and these derivatives are visualised in Figure 1. ***Atgādināšana*** names the action derived from the verb "to remind". It was created by one of the linguists of the project as it did not previously exist as a separate dictionary entry. This derivative contains two senses: *atgādināšana*$_1$ 'the act of reminding' and *atgādināšana*$_{1.1}$ 'the act of reminding (by device)' which are linked symmetrically to the base word senses 1 and 1.1 using the semantic link "toDo – action", where "toDo" and "action" are roles for both ends of the link from Table 1.

Atgādinājums denotes the result of the act of reminding; it has three senses: *atgādinājums*$_1$ 'a reminder (written or spoken)', *atgādinājums*$_{1.1}$ 'a written reminder (incl. by device)', and *atgādinājums*$_{1.2}$ 'a reminder of a fact, event'. All three semantic derivation links to these senses are of the "toDo – result" type, however, they are not symmetrical (see Figure 1). E.g., first sense of the derivative and its subsense are both linked to *atgādināt*$_1$. The reason can be both word meaning peculiarities and previous reviewing and amendment of the entries.

Atgādne 'a reminder (usually written)' is a more specific term for a general reminder. The entry only has one sense, which is linked to the first sense of the base word by the "toDo – instrument" semantic link type.

Atgādinātājs 'someone/something that reminds' is another three-sense derivative, but in this case, the semantic link distribution with the base word is symmetrical. It is the variety of semantic derivation links that stands out in this case: each derivative sense is ascribed a different role ("agent", "device", "cause"), whilst the roles of the base word are either "toDo" or "toBeInState", demonstrating the wide range of meanings that even relatively simple derivatives may contain.

It is worth noting that the second sense of the base word *atgādināt*$_2$ is not linked to any of the senses of the derivatives, which further highlights the complex, irregular semantic link structures between the senses of derived words.

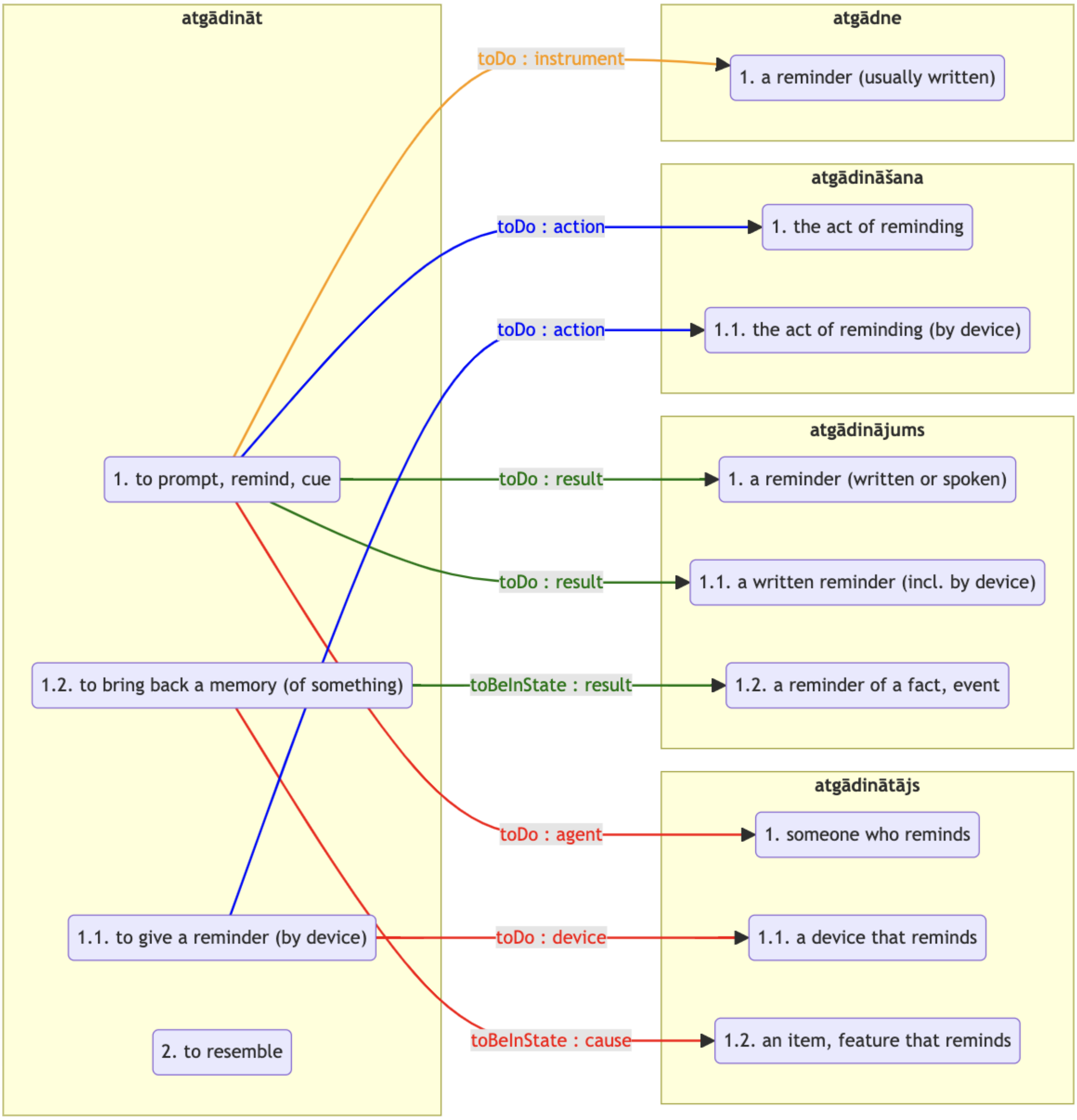
atgādināt
1. to prompt, remind, cue
1.2. to bring back a memory (of something)
1.1. to give a reminder (by device)
2. to resemble
atgādne
1. a reminder (usually written)
atgādināšana
1. the act of reminding
1.1. the act of reminding (by device)
atgādinājums
1. a reminder (written or spoken)
1.1. a written reminder (incl. by device)
1.2. a reminder of a fact, event
atgādinātājs
1. someone who reminds
1.1. a device that reminds
1.2. an item, feature that reminds
toDo : instrument
toDo : action
toDo : action
toDo : result
toDo : result
toBeInState : result
toDo : agent
toDo : device
toBeInState : cause

Figure 1: Relations between the senses of the verb *atgādināt* and its derivatives

3.4 Problematic Cases and Solutions

Polysemous derivatives can sometimes pose a challenge for annotation due to their gradual shifts in meaning. There are certain cases when the basic and usually most general sense of a derivative may be lost or rarely used, as the derivative has developed more specific senses over time. This is illustrated by the noun *laidiens* 'a release' derived from the verb *laist* 'to let' or the noun *darījums* 'a transaction' derived from *darīt* 'to do'. The solution for annotating such cases may be twofold depending on corpus data – either to include the basic sense in the entry with a tag 'rarely', or not to include it at all. In the latter case, the general derivational semantics exist only as a potential and remain unrevealed in semantic derivational links.

Due to diverse sense granularity of the base word and the derivative, attempts to obtain symmetry between the two might lead to an unnecessarily fine-grained distinction of senses. Instead, two following linking patterns can additionally be employed: (a) one sense of the base word is linked to several senses of the derivative (*plānot* 'to plan' is linked to two senses of the derivative *plānotājs* 'a planner': those of an agent and of a device), (b) several senses of the base word are linked to a single sense of the derivative (two senses 'to know (how to)' and 'to be able to' of the base word *mācēt* are linked to the single sense of the derivative *māka* 'a skill') (on a similar asymmetry between word senses in English see (Mititelu, 2018)).

4 Conclusions and Future Work

The creation of derivational semantics resource has been started, the first such open-access resource for the Latvian language. To reflect the possible difference in derivational semantics between the senses of one polysemous word, two types of links are created in the resource - a morphological link between lexemes and a semantic link between word meanings. A semantic link is formed by a pair of labels assigned to each linked sense. This results in a more informative resource than the general models of derivational semantics described in grammar alone. The first processed data consist of approx. 1000 morphological links and 1600 semantic links and the data is available in the autumn release of Tēzaurs.lv, and from the winter release, it will also be available in the public version of Tēzaurs.lv in the entries of the processed words.

In the future, first of all, it is planned to cover other pairs of word classes involved in Latvian derivation, including derivation pairs within the same word class. Secondly, it is planned to automate part of the process – to find the existing entries of derivatives in the dictionary according to templates, to check in the corpus what kind of derivatives are used for a certain base word and compare with the dictionary data to create the missing entries. Thirdly, it is planned to create a good search system in the data, so that we can further study which derivatives form which semantics. We would like to pay special attention to the semantic relations of polysemous words with their derivatives. Plans for further work also include the integration of the derivational links within Latvian WordNet, as there is a difference between synset-to-synset WordNet links and the derivational links that apply to specific words within that synset, and more study is needed to determine the proper representation for that interaction.

Acknowledgments

This research was funded by Latvian Council of Science project "Advancing Latvian computational lexical resources for natural language understanding and generation" (LZP2022/1-0443).

References

Orhan Bilgin, Ozlem Cetinoglu, and Kemal Oflazer. 2004. Morphosemantic relations in and across wordnets: A study based on Turkish. In *Proceedings of the Global Wordnet Conference*.

Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. Coping with derivation in the Bulgarian Wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 109–117, Tartu, Estonia. University of Tartu Press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

Mikus Grasmanis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, Laine Strankale, Arturs Znotins, and Normunds Gruzitis. 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.

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

Kristīne Levāne-Petrova and Roberts Darģis. 2018. Balanced corpus of modern Latvian (LVK2018).

Ilze Lokmane, Laura Rituma, Madara Stade, and Agute Klints. 2021. The Latvian WordNet and word sense disambiguation: Challenges and findings. In *7th Biennial Conference on Electronic Lexicography (eLex)*, pages 232–246.

Verginica Mititelu, Svetlozara Leseva, and Ivelina Stoyanova. 2021. Semantic analysis of verb-noun derivation in Princeton WordNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 108–117, University of South Africa (UNISA). Global Wordnet Association.

Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the Romanian Wordnet. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2596–2601, Istanbul, Turkey. European Language Resources Association (ELRA).

Verginica Barbu Mititelu. 2018. Investigating English affixes and their productivity with Princeton WordNet. In *Global WordNet Conference*.

Pēteris Paikens, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2023. Latvian WordNet. In *Proceedings of the Twelfth Global Wordnet Conference*, pages 187–196, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Maciej Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012. Recognition of Polish derivational relations based on supervised learning scheme. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 916–922, Istanbul, Turkey. European Language Resources Association (ELRA).

Adam Rambousek, Aleš Horák, and Karel Pala. 2018. Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive Studies — Études cognitive*, 18:75–81.

Victor Raskin and Sergei Nirenburg. 1995. Lexical semantics of adjectives: A microtheory of adjectival meaning. *MCCS report 95*, 288.

Baiba Saulite, Roberts Darģis, Normunds Gruzitis, Ilze Auzina, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Peteris Paikens, Arturs Znotins, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Barzdins, Inguna Skadiņaa, Anda Baklāne, Valdis Saulespurāns, and Jānis Ziediņš. 2022. Latvian national corpora collection – Korpuss.lv. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5123–5129, Marseille, France. European Language Resources Association.

Emīlija Soida. 2009. *Vārddarināšana*. University of Latvia Press, Riga.

Andrejs Spektors, Lauma Pretkalniņa, Normunds Grūzītis, Pēteris Paikens, Laura Rituma, Baiba Saulīte, Gunta Nešpore-Bērzkalne, Ilze Lokmane, Agute Klints, Madara Stāde, Mikus Grasmanis, Laine Strankale, Ilze Auziņa, Artūrs Znotiņš, Roberts Darģis, and Guntis Bārzdiņš. 2023. Tēzaurs.lv 2023 (summer edition). CLARIN-LV digital library at IMCS, University of Latvia.

Anna Wierzbicka. 1988. *The Semantics of Grammar*. Companion series. J. Benjamins Publishing Company.