Towards Latvian WordNet

Pēteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, Laine Strankale
University of Latvia, Institute of Mathematics and Computer Science
Raina bulvaris 29, Riga, LV-1459, Latvia
{peteris.paikens, mikus.grasmanis, agute.klints, ilze.lokmane,
lauma.pretkalnina, laura.rituma, madara.stade, laine.strankale}@lumii.lv

Abstract

In this paper we describe our current work on creating a WordNet for Latvian based on the principles of the Princeton WordNet. The chosen methodology for word sense definition and sense linking is based on corpus evidence and the existing Tezaurs.lv online dictionary, ensuring a foundation that fits the Latvian language usage and existing linguistic tradition. We cover a wide set of semantic relations, including gradation sets. Currently the dataset consists of 6432 words linked in 5528 synsets, out of which 2717 synsets are considered fully completed as they have all the outgoing semantic links annotated, annotated with corpus examples for each sense and links to the English Princeton WordNet.

Keywords: Wordnet, Latvian, semantics, lexicography

1. Introduction

The intent of this paper is to share our experience in developing Latvian WordNet and to describe our concerns regarding the many aspects of resource development where the decisions were complex and different between projects building and maintaining wordnets for various languages.

Many natural language processing solutions, especially in the field of natural language understanding and semantic parsing, presume the availability of a lexical resource similar to Princeton WordNet (Fellbaum, 1998). This is a valid assumption for many languages including almost all EU languages, facilitated by multilingual efforts such as EuroWordNet (Vossen, 1998), but as of now Latvian has been lacking such a resource. This need for a structured lexical-semantic resource and a sense inventory for Latvian word sense disambiguation has been a primary motivation for initiating this project at the beginning of 2020, and we have now released the initial portion of the data.

When developing a wordnet from scratch, there are two possible methodology approaches: start with a bottom-up analysis of the language and its lexicography, or adopt and transfer the semantic hierarchy from another language (usually English or a very closely related one). In our initial experience and looking at other projects like plWordNet (Maziarz et al., 2012), we observed that the former approach is more labor intensive but it also results in a resource that is more accurate from a linguistic perspective. Therefore, we have chosen to build this resource based on corpus evidence and linguistic analysis of Latvian, and only afterwards attempt to link it with other language wordnets, without the expectation that there would necessarily be a one-to-one alignment for specific synsets. The existing Latvian digital dictionary Tēzaurs.lv (Spektors et al., 2020) was used as the basis for this project, providing the initial dataset and technological platform. However, we quickly identified a need to restructure word sense separation for most of the dictionary entries as described in more detail in section 2.1.

The development of this resource is largely motivated by the necessity for multilingual solutions and integration of Latvian in existing multilingual systems. Therefore we have also focused on annotating links between synsets of Latvian WordNet and English Princeton WordNet. Since many other languages also have resources linked to Princeton WordNet, for example as aggregated in Open Multilingual WordNet (Bond and Foster, 2013), this also enables us to retrieve matching word senses and translations from other languages using English WordNet as a pivot. In section 2 we describe our linguistic methodology used for building Latvian WordNet. Section 3 covers the custom technical platform developed for maintaining this resource. Section 4 describes the properties of the currently published dataset. The final section describes potential applications of this resource and ongoing future work.

2. Methodology

Latvian Wordnet is developed for the most commonly used Latvian words. The list of words is compiled by frequency of use in The Balanced Corpus of Modern Latvian (Levāne-Petrova and Darģis, 2018), after dropping particles, function words and proper words. It is planned that during this project the first 2000 words from the list will undergo full processing, which consists of several steps. The first step is a sense inventory review for each word from the list in the Tēzaurs.lv platform. Semantic links are then established between the revised senses and other senses in Tēzaurs.lv. In the first step of word processing, the greatest difficulty was establishing criteria for distinguishing senses and ensuring that they were used consistently within at least

one word class. It was difficult to determine how detailed the meanings should be so that the word sense granularity would be similar across all labeled entries. For this reason, we try to consider semantically related words from the list simultaneously to make sense distinctions similar at least within one semantic group (for a similar implementation of systematic polysemy into Estonian WordNet see (Kerner et al., 2010)). Our main objective is to improve the inventory of word senses in Tēzaurs.lv, make them clearer and more comprehensible to the user of the dictionary, which, in turn, leads to more coarse-grained senses. On the other hand we aim to link the word senses with as many synonyms and Princeton synsets as possible which leads to more fine-grained word senses.[1] At the same time, examples from corpora are added to each word sense, creating a training data set that can subsequently be used in automatic word sense disambiguation. It is the selection of examples that guarantees the quality of the distinguished word senses, as the linguists immediately examine them in practice and varifies that it is possible for a language user to distinguish between the various word senses in a text. In the second step of word sense processing, sets of synonyms are created and other semantic links are added to word senses, including external links to English Princeton WordNet. The approach chosen for this project is centered around processing basic relations - synonymy, hyponymy, meronymy, antonymy, but in practice it proved necessary to also use the links "Similarity" and "See also" in cases when we want to show the semantic connection, but it does not fit in the mentioned basic relations. Unlike Princeton and other wordnets, we do not create a separate gloss for the synset as we base our work on the existing Tēzaurs.lv dictionary which has separate glosses for each word. In Latvian WordNet we do not enforce a rule that a synset can contain only words of the same word class as words sometimes have senses which act as a different part of speech, for example, a sense of an adjective can be in the same synset as participles, adverbs linked with a certain noun inflection form, and multi-word expressions can be added to a synset as well - the principal factor is the shared meaning. A more detailed understanding of semantic relations is described in Section 2.2, and the creation of interlingual links is described in Section 2.3. Developing a wordnet is based on live data in the database. A system has been set up to ensure that the linguists do not select the same entry, and the history of all work within a section is also preserved, to enable contact with the annotator of a particular entry to clarify and discuss certain details or decisions. Complex cases are thoroughly discussed during linguist seminars in order to achieve a more consistent approach and to reduce the subjectivity of a linguist's individual sense of language.

[1]Regarding the computational linguistics needs for clustering senses, see also (Jurafsky and Martin, 2022, Ch. 18)

2.1. Word sense distinction in Tēzaurs.lv

The need to manually revise word senses has arisen due to a number of reasons. Firstly, the existing entries in the dictionary Tēzaurs.lv more often than not reflect an earlier stage of language use and, consequently, many outdated word senses. A corpus-based approach enables us to arrive at a more adequate reflection of relevant word senses in Modern Latvian. Secondly, traditional Latvian lexicography tends to be of the so-called 'splitting' type, with the sense splitting appearing to be rather subjective due to not being based on any clearly stated criteria. Thirdly, in traditional Latvian dictionaries, the word senses belonging to the same thematic group happen to be distinguished according to different principles, which often makes identification of semantic relations between these words impossible. In addition, it would be preferable if sense granularity would be similar to Princeton WordNet as it would simplify creating interlingual links (discussed in more detail in Section 2.3). Last but not least, such an approach would contribute to the improvement of word sense inventory of Tēzaurs.lv.

The process of revising individual word senses and definitions is based on a set of criteria that form a certain hierarchy. This hierarchy is not uniform for all parts of speech. In the case of verbs, differences in **semantic and syntactic distribution** is the dominating criteria for sense distinction (Lokmane et al., 2021; Lokmane and Rituma, 2021). Syntactic distribution implies the verbs' arguments and their coding (Williams, 2015, p. 47-61). For example, the verb *noteikt* in its basic sense 'to state; to say' is used in direct speech constructions, whereas one of its secondary senses 'to determine; to shape' attaches an object in the accusative. Semantic distribution, in turn, includes semantic or thematic roles (such as agent, patient, experiencer, beneficiary etc.) (Saeed, 2016, p. 150-156) and general or more specific semantic features (such as animate / inanimate, abstract / concrete etc.) of the arguments. For example, the senses of the word *iet* 'to go' are distinguished according to semantic groups (persons, animals, vehicles, mechanisms etc.) of the agent.

The method of **lexical decomposition** postulates that a word's sense may be broken down into smaller semantic components or features (Cruse, 2004, p. 235) and is relevant for some groups of verbs, as well as nouns. For example, the basic sense of the verb *redzēt* 'to perceive by sight; to see' differs from its secondary senses 'to perceive mentally, to understand' and 'to witness, to be contemporaneous with' by semantic features. Similarly, the basic sense of the noun *laiks*, namely, 'the continuum of experience in which events pass from the future through the present to the past; time' is distinguished from its secondary sense 'a suitable moment' according to semantic features which is clearly illustrated by the above-mentioned sense definitions.

Substitution with a synonym is a relevant and commonly used method not only for sense distinction, but

also for defining word senses (Jackson, 2002, p. 94) and is applicable to various parts of speech. However, it should be noted that the synonymous word may be polysemantic as well and thus correspond to multiple senses, for example, the verb *tapt* 'to become; to turn' is synonymous with *kļūt* 'to become; to turn' in at least four senses which are mutually distinguished according to the criteria mentioned above. This example illustrates also that words belonging to the same thematic group should be treated similarly to ensure a systemic approach to the vocabulary as a whole.

One of the greatest challenges of word sense distinction is the interrelation of **superordinate senses and subsenses**. Although the concept of subsense has not been clearly defined yet, it has been proven necessary for creating paradigmatic semantic relations in Latvian WordNet. In most cases, a subsense is a way of displaying metonymic (and less often metaphorical) shifts, which cannot be given the status of a separate sense. A vivid example is the noun *pašvaldība* with its basic sense 'local government' having two metonymically motivated subsenses: 'an urban district having corporate status and powers of self-government; municipality' and 'a municipal building'. Thus, a paradigmatic sense relation may link a sense of one word with a subsense of another.

Being fully aware that absolutely uniform and consistent word sense distinction is not likely to be possible, we aim at creating a compromise, which would allow both to improve sense distinctions and definitions in Tēzaurs.lv, and create wordnet links. In cases of uncertainty, the decisions are made in favour of the needs for Latvian WordNet development.

2.2. Semantic relations

The set of semantic relations of Latvian WordNet for now is restricted to the most common wordnet relations (synonymy, antonymy, hyponymy, meronymy) and three more categories that are not that common among WordNets – gradation, similarity, and "see also". The two latter categories are seen by linguists as subjective, given that such relations don't have a clear definition yet. Sense relations are identified by semantic and syntactic criteria and formed between word senses. A brief review of semantic relations is presented below.

Synonymy is crucial for creating a wordnet as a network of synsets linked by semantic relations. Synset is a basic unit of a wordnet, so it is primarily necessary to establish the definition of synonymy and to form synonymous relations between senses (Lokmane et al., 2021). Absolute synonymy, when synonyms share the same semantic components and can always substitute each other, is very rare. However, partial (near) synonyms are more common – they are semantically identical, but have limited substitutability. In Latvian WordNet, the category of synonymy contains both absolute and partial synonyms.

Words that are semantically close, but for some reason do not fall under the synonym category, are joined by the **similarity** link. The reasons why senses may not be synonymous include syntactic criteria, valency, definition features that leads to partial synonymy when the context of particular meaning differs too much to put in one synset. One example of semantically close word senses is adjective synset *labs, vērtīgs* 'valuable, dear' that is linked to noun $zelts_{3.1}$ 'gold' that is used only in genitive case and expressing a particular quality, feature. Semantics of these synsets are identical, but noun can replace adjective synset only in poetic context, it is rarely used. By studying this category we come to a clearer understanding of synonym definition itself.

Hyponymy forms a hierarchical network. It is formed if one term is more general, and the other is more specific. In Latvian WordNet, hyponymy links can mainly be formed between nouns and verbs (although other wordnets use troponymy and other categories for verb hyponymy), but in rare cases they can also be formed between adjectives and adverbs. A hyponym is a more specific term which has more semantic components than its hypernym (Löbner, 2013). For example, *mēnesis* 'month, calendar month' is a hypernym for its hyponyms – *janvāris, februāris, marts* 'January, February, March' etc., because all of them contain the semantic element 'month' as well as an additional element naming more a specific kind of month. In addition, Latvian verb hyponymy is sometimes formed through affixes, for example verb *dzert* 'drink, imbibe' has many possible derivations that contain a more specific semantical element – *iedzert* 'take (a sip)', *izdzert* 'drink (all of one's drink)', *nodzert* 'drink off (top layer or part of drink)', *uzdzert* 'drink after eating, wash (something) down' and so on. All derivations usually are hyponyms of the main verb.

There are various ways to express semantic opposition. They all involve words that are related in meaning yet incompatible or contrasting. S. Löbner mentions **antonymy** only as one of the opposition types, but in Latvian linguistics antonymy usually combines different types of opposition: simple antonyms (also called complementary pairs / binary pairs), gradable antonyms, reverses, and converses (Skujiņa, 2007). This approach is also used in the creation of Latvian WordNet, where antonyms are listed e. g. adjective pair *neparasts* 'extraordinary' and *parasts* 'ordinary, daily', noun synsets *jautājums, vaicājums* 'question, interrogation' and *atbilde* 'answer, reply', and also verbs *pirkt* 'buy, purchase' and *pārdot* 'sell'.

Meronymy is used to show the relationship between a part and a whole. Thus *koks* 'tree' is a meronym of *mežs* 'forest, wood, woods', and *sakne* 'root' and *zars* 'branch' are meronyms of *koks* 'tree'. Only nouns are joined by meronymy links. Meronymy in other wordnets may be further divided into three subrelations, but Latvian WordNet currently assumes only one basic meronymy relation.

Instead of pairing synsets, **gradation** forms groups or sets. Words in one gradation set express different values (intensity, speed etc.) of the same attribute. This relation is typically associated with adjectives (Saeed, 2016) and represents a transition between synonymy and antonymy (Veidemane, 1970). Although many gradable adjectives can be named in Latvian, other wordnets may not have this category as there are not many gradables that are lexicalized, for example, in English, so Princeton WordNet does not show this relation (Miller, 1998). One example is the gradation group that contains the synsets *milzīgs* 'huge'; *gigantisks* 'gigantic'; *liels* 'big'; *mazs* 'small'; *sīks* 'tiny' etc. – they all describe size, but the meanings of these words vary in the extent of size. In Latvian WordNet the gradation set units are interlinked with each other, but there can also be a superordinate that is linked with the entire set, similar to attribute relations in other wordnets. For example, the superordinate for the gradation group mentioned earlier is *lielums* 'size'. Linking gradation group elements with each other proves useful in cases when a superordinate term in Latvian is not lexicalized, but it is still possible to show connection between these gradation set elements, for example, if a colour differs in intensity, it is not marked as a hierarchy relation but as a gradation set instead: *zils* 'blue'; *zilgans* 'bluish'; *iezilgans* 'a little bluish'.

"See also" is a category for words that are semantically related, but not by any of the mentioned semantic relations. For example all of the colour names (*red, blue, green, gold*) are linked with the word *krāsa* 'colour' by relation "see also", because Latvian WordNet doesn't have relation "attribute" yet that is used in these kind of situations in other wordnets (Maziarz et al., 2012; Miller, 1998), so "see also" includes these cases. One more example in this category - verb *uzstāties* 'perform' is related to such verb synsets as *dejot* 'dance', *dziedāt* 'sing', *spēlēt, tēlot* 'act, play', but none of them are hyponym to other – *uzstāties* is semantically wider, it covers all of the ways anyone can perform, but its meaning contains condition 'in front of the public', however words that name the ways of performing are semantically wider in sense that a performer can dance, sing and act with or without public. As long as we don't have a precise semantic relation for linking semantically connected word senses like mentioned above, there is link "see also" between them. This category is a source for future research and further improvements of Latvian WordNet.

2.3. Wordnet to Wordnet Sense Mapping

Linking the created Latvian WordNet to Princeton WordNet presents another crucial stage of the project implementation.

In this project, mapping is carried out on the level of synsets and the process of interlingual sense-linking consists of two main stages: manual and automatic linking. In the first stage, synsets are linked manually to the closest possible equivalent in Princeton WordNet by the project's team of linguists. This stage concentrates on the interlingual sense-linking of 2000 most frequently used Latvian words. So far, 1920 links have been manually created between both wordnets. In the second stage, the manually created inventory of interlingual links is used for developing the algorithm responsible for automatic sense-linking. The manually created data is also used to verify and control the level of accuracy of the automatically generated links. Currently, automatic sense disambiguation and interlinking with Princeton WordNet is undergoing testing. Preliminary results show that the automatic links of adjectives and verbs have been the most difficult to map accurately, probably due to more specific and distinct meanings that are less interchangeable and more situationally used than other parts of speech (Strankale and Stāde, 2022).

As mentioned above, equivalence between source and target language plays a crucial role in the process of sense-linking. However, as preferable as direct equivalents are for clear and comprehensible links between the two wordnets, the natural asymmetry between languages can make it impossible to achieve such level of equivalence. Although there is a number of roughly universal concepts in English and Latvian that can be directly linked, more often than not, a linguist must choose to break equivalence down into various subtypes (e.g. formal, functional, semantic, stylistic) (Chesterman, 2016; Pym, 2017) and choose between them to select the most precise match. This means introducing more than one type of inter-wordnet links. Currently we are using three link types between the Latvian and Princeton WordNets: a direct link, a hyponymy link and a hypernymy link. Such an approach helps account for the interlingual hyper/hyponymy between Latvian and English, but also address other issues of sense asymmetry that lead to different types of equivalence. This proves convenient when linking such Princeton WordNet sense as 'trip', which includes both, a trip by transport and by foot, to possible equivalents in Latvian, which is more specific in distinguishing between these concepts. Therefore, the English 'trip' can be linked to both *brauciens*$_2$ ('trip by transport') and *gājiens*$_{1.2}$ ('trip by foot') with a hyponymy link, as both Latvian senses fit the one in English on a situational basis and are therefore its functional equivalents.

The method and links described above ensure successful sense mapping in most cases. Even when a sense in the source language is highly culture specific, the target language contains at least one sufficiently broad sense to form a hyponymy link. However, there have still been some cases when the exact or partial equivalent for a sense or synset in the Latvian WordNet does not exist in Princeton WordNet, even though the notion itself is present in the English language; one such example is 'science' in the sense of 'scientific thought'.

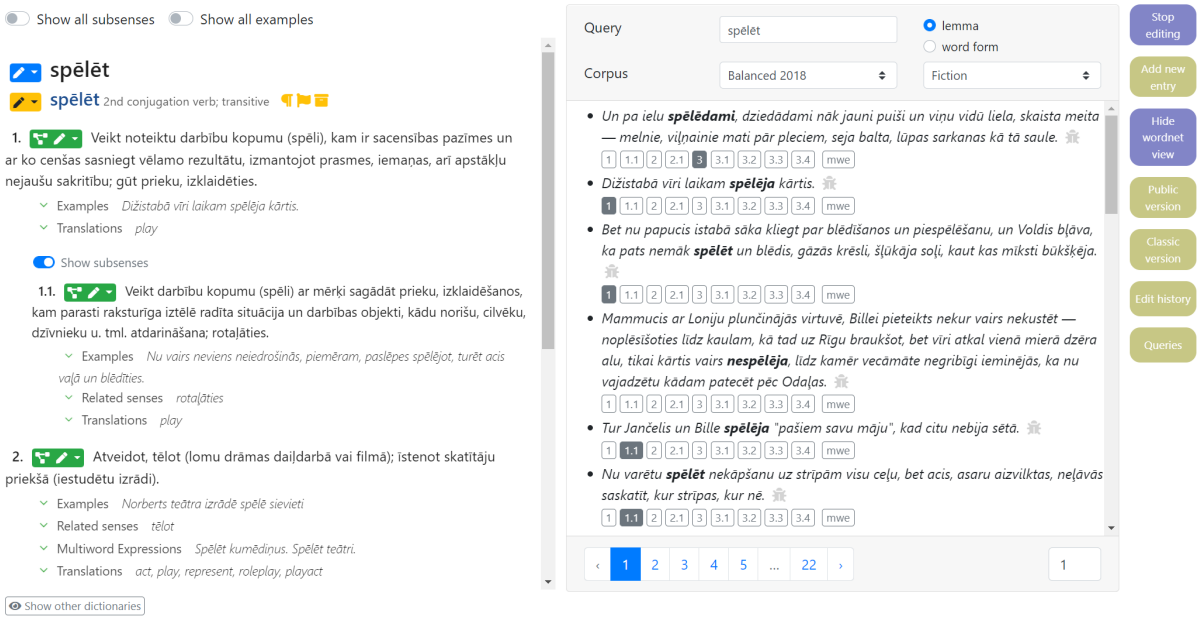

Figure 1: Editor's view of a Tēzaurs.lv entry. The upper left side shows the list of word senses. The newly added example browser is on the right. Extracts from other Latvian dictionaries for the same entry (used during sense revisions) can be accessed on the lower left side.

3. Technical platform

As Latvian WordNet uses data from Tēzaurs.lv, we decided to extend the existing Tēzaurs.lv platform (Spektors et al., 2016), which already had a custom-made web interface for editing entries built on Vue.js and Node.js, with the software tools needed for the construction of Latvian WordNet.

For this project, we required the following functionality: (1) to create and update word senses and subsenses; (2) to browse through different corpora and add usage examples to particular word senses; (3) to create synsets, i.e. to find and link synonymous senses; (4) to interlink the new synsets with semantic relations; (5) to link Latvian WordNet synsets to corresponding Princeton WordNet synsets; (6) to visually inspect Latvian WordNet; (7) to keep notes and see statistics.

The existing Tēzaurs.lv editor already supported editing core items of the lexicographical entry – word senses and subsenses. However, in regards to other requirements, we needed to make substantial updates to the editor. A major addition was a system for adding usage examples, seen in the entry view on Figure 1, which allows searching within five different corpora, notably, the balanced Latvian corpus and a large web corpus from Common Crawl data.

The functionality of (3), (4), and (5) is combined into a single view seen in Figure 2. Each word sense has such a synset view. Firstly, it displays synset information for the particular sense: a list of senses in the synset, a list of links to other Latvian and Princeton WordNet synsets, an excerpt from a Latvian dictionary of synonyms, and English translations for the current word. Secondly, it allows to search for a sense or a synset both in Latvian WordNet and Princeton WordNet and link them to the current synset.

Each synset has two graph views: one for displaying all links within a distance of two links (shown in Figure 3, and another for showing the hyponym-hypernym hierarchy which is useful to verify consistency.

4. Current data set

At the moment of submitting this paper there are 6432 words incorporated within Latvian WordNet (9489 senses), forming 5528 synsets. Of these synsets, most (3683) are one-member synsets, while 1845 synsets consist of two and more members. The average number of members of the multimember synsets is 3.15, while the average number of members of all synsets is 1.7. It should be noted that a sense starts to function as a synset when there is at least one link attached to it. If a sense is not involved in any synset and does not have any wordnet links with any other sense, it is not counted in the synset statistics.

In total, there are currently 3712 semantic relation links formed in Latvian WordNet shown in Table 1, along with 24 gradation sets in which 74 synsets are included. Currently we consider 1055 words (2717 senses) as fully processed - they have reviewed sense separation, added corpora examples for all the senses, senses linked in synsets and other relevant semantic links, including external links with Princeton WordNet synsets. 62 428 examples have been added to the processed senses (58 300 examples for general word senses and 4128 examples for multi-word expressions).

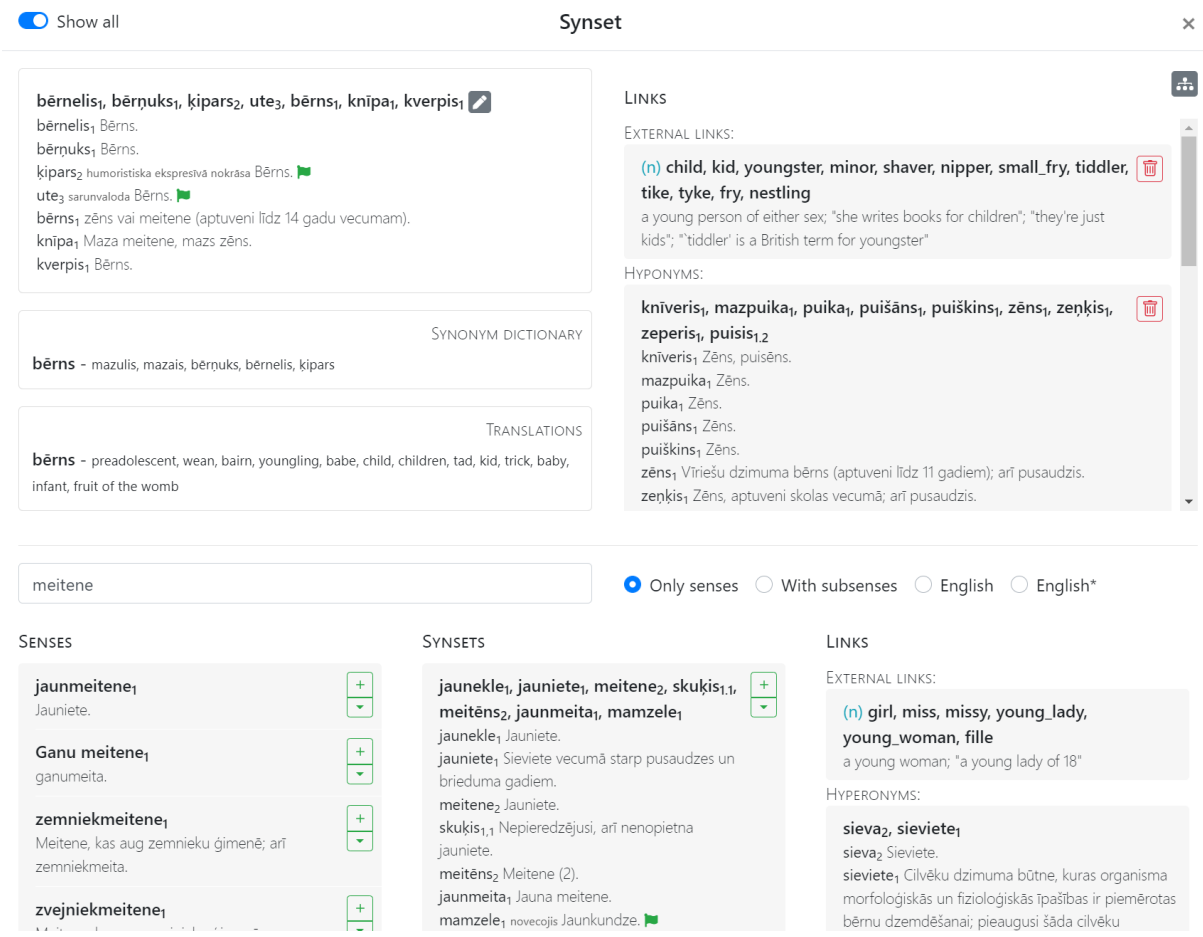

Figure 2: Synset edit view for a sense of the word *bērns* 'child'. The upper section contains synset information: included senses, synonyms and word translations on the left, a list of linked synsets by type on the right. The lower section allows to add new links by searching within Tēzaurs.lv or Princeton WordNet. The image shows Tēzaurs.lv results for a search query *meitene* 'girl'. The first result column contains a list of senses that are not yet in Latvian WordNet. The second column contains a list of Latvian WordNet synsets. By clicking on any of these synsets, a list of all its links is displayed in the third column.

Relation	Verbs	Nouns	Adj	Adv
Hyponymy	905	1758	53	50
Meronymy	-	247	-	-
Similar	93	116	30	11
Antonymy	28	37	25	5
See also	61	159	24	11

Table 1: Semantic relations by type and part of speech

Where possible, synsets of Princeton WordNet were interlinked with the corresponding Latvian synsets. Out of the total 3061 senses in the completed words, 2471 have been linked with Princeton WordNet on a synset level. Of these interlingual links, 1721 are equivalence links, 427 are links denoting that the meaning of the Latvian synset is wider than the meaning of the English synset, and the remaining 515 synsets have links where the meaning of the Latvian synset is narrower than the meaning of the English synset. Some synsets have multiple Princeton links, for example, if both wider and narrower senses are linked. As a result, 81 % of the senses processed have a corresponding Princeton link, and 70 % of those are equivalence links.

Latvian WordNet data is made publicly available in two ways - through the web platform at `https://tezaurs.lv` for everyday needs and through work-in-progress machine-readable TEI XML[2] format at `https://wordnet.ailab.lv/data/`. Currently this data format provides an overview for the underlying Tēzaurs.lv dictionary entries, including all the senses and lexemes for each entry. Senses contain identifiers for gradation sets and synsets they belong to and sysnset relations.

[2]TEI Consortium, Chapter 9: Dictionaries. `https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`

Figure 3: Graph for a sense of the word *minēt* 'guess'. The green boxes are all the senses of the word, including all the other synset members in the box. The blue boxes are related synsets in Latvian WordNet. The red boxes are corresponding synsets in Princeton WordNet.

5. Applications, ongoing and future work

For general public the main application of this resource is through the Tēzaurs.lv platform, which already serves users around 100 000 entries per day and is relatively well known among the Latvian general public (e.g, students, teachers, translators and editors). WordNet data effectively transforms Tēzaurs.lv into a 3-in-1 solution, combining an explanatory dictionary, a thesaurus and a translation dictionary into a single free online tool, and makes WordNet data accessible to many users.

Currently the platform displays WordNet data as additional information for each word sense involved in any synset, as seen in Figure 4. The resulting dictionary provides definitions of word meanings and nuances as well as usage examples. For the words which have extended with WordNet data it also lists the synonyms, antonyms, hyponyms and other semantic links of the word. This is useful for creating an accurate translation and ensuring the diversity of the text and helps both language learners and professional users. The dictionary also lists the closest word equivalents in Princeton WordNet, helping translators find the best match.

An important design goal for this resource was also to apply it as the word sense inventory for natural language understanding solutions. Because of this, when using corpus evidence for separating word senses, we chose not only a few illustrative examples for dictionary readers but annotate many usage samples to be used as training data for the word sense disambiguation systems which we are currently developing.

The immediate ongoing work is to continue extending the WordNet size, by continuing the current effort and also by exploring options of using Princeton WordNet data as source for candidate links. Current experiments (Strankale and Stāde, 2022) show that this would be plausible but require significant manual validation.

We also plan to use the currently annotated links between Latvian and Princeton WordNets to integrate Latvian WordNet into the Open Multilingual Wordnet resource. We are using compatible data formats and names of identifiers to facilitate this integration.

With respect to the public user interface, relevant future work is to add navigation between different words using the WordNet link graph like shown in Figure 3 and `https://wordnet.ailab.lv/demo`.

Future NLP work planned with this resource includes development of a word sense disambiguation system and applying it for tagging existing corpora of Latvian in order to enable searching of specific word senses. We also expect to apply the Latvian WordNet for semantic parsing, such as AMR (Banarescu et al., 2013) which was previously difficult for Latvian due to lack of a good sense lexicon.

6. Acknowledgements

This research work is supported by the Latvian Council of Science, project "Latvian WordNet and Word Sense Disambiguation" (project No. LZP-2019/1-0464) and National Research Programme "Letonika – Fostering a Latvian and European Society" project "Research on Modern Latvian Language and Development of Language Technology" (grant agreement No. VPP-LETONIKA-2021/1-0006).

7. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual WordNet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Chesterman, A. (2016). *Memes of translation: The spread of ideas in translation theory*, volume 123. John Benjamins Publishing Company.

Cruse, A. (2004). *Meaning in Language: An Introduction to Semantics and Pragmatics*.

Christiane Fellbaum, editor. (1998). *WordNet: An electronic lexical database*. MIT Press.

Jackson, H. (2002). *Lexicography: An Introduction*. Taylor & Francis Routledge.

Jurafsky, D. and Martin, J. H. (2022). Speech and language processing (3rd edition draft). Available from: `https://web.stanford.edu/~jurafsky/slp3/` [cited 2022 Jan 13].

Figure 4: Tēzaurs.lv public interface for entry *baznīca* 'church': 1 – search; 2 – entry header with lexemes; 3 – nearby entries; 4 – first gloss; 5 – corps examples (expandable); 6 – related senses, including 7 – synonyms, 8 – hyponyms, 9 – hypernyms; 10 – MWE block (expandable); 11 – translations obtained via WordNet mappings

Kerner, K., Orav, H., and Parm, S. (2010). Growth and revision of Estonian WordNet. *Principles, Construction and Application of Multilingual Wordnets*, pages 198–202.

Löbner, S. (2013). *Understanding semantics*. Routledge.

Lokmane, I. and Rituma, L. (2021). Verba nozīmju nošķiršana: teorija un prakse; verb sense distinction: theory and practice. *Valoda: Nozīme un forma*, 12:142–162.

Lokmane, I., Rituma, L., Stāde, M., and Klints, A. (2021). The Latvian WordNet and word sense disambiguation: Challenges and findings. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, pages 232–246.

Maziarz, M., Szpakowicz, S., and Piasecki, M. (2012). Semantic relations among adjectives in Polish WordNet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies*, (12).

Miller, K. J. (1998). Modifiers in WordNet. In *WordNet: an electronic lexical database*, pages 47–67. MIT Press.

Pym, A. (2017). *Exploring translation theories*. Routledge.

Saeed, J. (2016). *Semantics*. Wiley Blackwell.

Valentīna Skujiņa, editor. (2007). *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga: LU Latviešu valodas institūts.

Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L., and Saulite, B. (2016). Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Strankale, L. and Stāde, M. (2022). Automatic word sense mapping from Princeton WordNet to Latvian WordNet. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, volume 1, pages 478–485.

Ruta Veidemane, editor. (1970). *Latviešu valodas leksiskā sinonīmija*. Zinātne.

Piek Vossen, editor. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*.

Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge University Press.

8. Language Resource References

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.

Spektors, Andrejs and Pretkalniņa, Lauma and Grūzītis, Normunds and Paikens, Pēteris and Rituma, Laura and Saulīte, Baiba. (2020). *Tēzaurs.lv 2020*.