Latvian WordNet

Pēteris Paikens, 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

Abstract

This paper describes the recently developed Latvian WordNet and the main linguistic principles used in its development. The inventory of words and senses is based on the Tēzaurs.lv online dictionary, restructuring the senses of the most frequently used words based on corpus evidence.

The semantic linking methodology adapts Princeton WordNet principles to fit the Latvian language usage and existing linguistic tradition. The semantic links include hyponymy, meronymy, antonymy, similarity, conceptual connection and gradation. We also measure inter-annotator agreement for different types of semantic links.

The dataset consists of 7609 words linked in 6515 synsets. 1266 of these words are considered fully completed as they have all the outgoing semantic links annotated, corpus examples assigned for each sense, as well as links to the English Princeton WordNet formed. The data is available to the public on Tēzaurs.lv as an addition to the general dictionary data, and is also published as a downloadable dataset.

1 Introduction

A wordnet (Fellbaum, 1998) is a lexico-semantic resource, which links an inventory of senses in synonym sets and other semantic relations, making it a valuable resource for NLP applications that can benefit from a formal structure of language semantics and relationships between specific word meanings.

Over the last three years we have been developing the first wordnet for Latvian, which is finally being formally released. We have chosen to form this resource based on corpus evidence and existing Latvian lexical resources, similar to the approach taken by plWordNet(Maziarz et al., 2016) and PolNet (Vetulani et al., 2010), instead of extending or translating word senses of some existing resource from other languages, such as the English Princeton WordNet.

The key tasks for forming Latvian WordNet were reviewing the sense inventory of the most frequently used Latvian words based on corpus evidence, annotating corpus examples to specific word senses, determining the members of synsets (synonym sets) and annotating outgoing semantic links, as well as later forming interlingual links to the English Princeton WordNet where applicable. The annotation work was performed with a custom lexicographic tool used for Tēzaurs.lv online dictionary, as described in (Paikens et al., 2022a).

The resulting manually curated resource consists of 7609 words linked in 6515 synsets. In addition we have an ongoing manual review of automatically obtained candidate links to Princeton WordNet (Strankale and Stāde, 2022). The consistency of semantic links was evaluated in an inter-annotator agreement experiment with three annotators on a limited subset of this data.

The following section describes the linguistic principles used in the development of Latvian WordNet, followed by a discussion of semantic links and an evaluation of their inter-annotator agreement in sections 3 and 4 respectively. After that, the article discusses the process of linking Latvian synsets to the Princeton WordNet in section 5 and the evaluation of these links in section 6. The concluding part consists of a discussion of the public availability of the resource in section 7 and conclusions and future work in section 8.

2 Linguistic Principles of Latvian WordNet

The decision was made to develop Latvian WordNet based on the inventory of Latvian word senses instead of adapting semantic hierarchy and relations from another language. It was also decided to build semantic relations between synsets from bottom up, allowing the hierarchy of word senses

to grow and develop on its own. All word sense relations are made between synsets.

To choose the initial set of senses to work with, a list of 2000 most frequently used words was created based on The Balanced Corpus of Modern Latvian (Levane-Petrova, 2019). The list was then revised, leaving only words from four main word classes - nouns (except proper nouns), verbs, adjectives and adverbs. The resulting list of words and their senses are the core of Latvian WordNet. The senses of these words were taken from from the Explanatory Latvian Dictionary Tezaurs.lv (Spektors et al., 2016), after which an additional sense revision was carried out, as the first attempts of semantic linking showed many outdated word senses, as well as inconsistent sense granularity. We chose to look for corpus evidence if the senses are still currently relevant, whether any new senses have appeared or whether specific uses of a word demonstrate the validity of word sense distinction, in a manner similar to how sense distinctions and definitions were done in Estonian WordNet (Kerner et al., 2010).

The sense revision was primarily based on data from The Balanced Corpus of Modern Latvian, however, for rare meanings that are only used in colloquial language or other specific language genres we looked for additional corpus evidence from several different corpora from Latvian National Corpora Collection (Saulite et al., 2022). The specific principles of distinguishing word senses were developed for the convenience of both annotators and target users of the dictionary (Lokmane et al., 2021). Given that the most frequently used words are also often polysemous, the lexicographic work of processing them proved time-consuming, but also resulted in a thorough, high-quality inventory for the core wordnet.

Regarding other linguistic principles, the semantic relations of Latvian WordNet are usually annotated between synsets of the same word class, with only rare, well-argumented exceptions when such a link is allowed between the senses of different word classes. For example, participles are considered as verb forms but can also be related to synsets of adjectives. Additionally, some meanings of definite adjectives can be linked to synsets of nouns. Such cases are often characteristic of partial word conversion, when a separate form of a word begins to perform the function of another word class and therefore has a separate meaning while still belonging to the same word entry in a dictionary. Word class boundaries are a separate research issue that was not addressed within the scope of our task.

3 Semantic Links in Latvian WordNet

The most common and better studied semantic relations traditionally included in wordnets of various languages are synonymy, antonymy, hyponymy and meronymy (Jurafsky and Martin, 2022, Chapter 18, pp. 4-5)

In addition to these four major relations, Latvian WordNet is enriched with gradation relations which are not included in most wordnets, an exception being e.g. plWordNet (Maziarz et al., 2012, 2015).

The basic unit of a wordnet is a **synset**. The opinions of language users about the synonymy of certain senses may differ, so the following set of criteria is used to determine a set of synonyms. Firstly, dictionary definitions, namely, semantic features are compared: if most of them match, the senses are considered synonymous. Secondly, a substitution criterion is applied: if the words are interchangeable in most contexts, the senses are considered synonymous. It should be noted that a synset may include both neutral and expressive meanings. The fact that the subtler semantic distinctions among the elements of a synset are beyond the scope of description might be considered one of the most serious shortcomings of wordnets (Geeraerts, 2009, p. 160). In Latvian WordNet, this is compensated by the representation of data in Tezaurs.lv which includes the definitions and stylistic nuances of the specific sense for each word, not a single definition for the whole synset. One of the sources used in annotating synset relations was an existing Latvian synonym dictionary (Grīnberga and Kalnciems, 1998), however, its application was limited as it lists synonyms on a word (not sense) level and includes many words that are related but not strictly synonymous.

As the degree of synonymy between senses may be different, Latvian WordNet also includes a **similarity** link for senses which do not fall under the category of a full synonym. Firstly, the similarity link is established between senses if the semantic differences are too significant to be considered synonymous, e.g. the synset (*diskusija, pārrunas*) ‘discussion, treatment, discourse’ is considered as similar to the synset (*apspriede, sanāksme, sēde, sapulce, konference, saruna*) ‘meeting, group meeting’. Secondly, the similarity link is estbalished between words which cannot be substituted for

each other in the context due to grammatical peculiarities, e.g. the sense of the word *kontekst* 'circumstance, setting' is characteristic only to the locative case form and can not be substituted with the locative forms of seemingly synonymous (*apstāklis, situācija, stāvoklis*) 'situation, state of affairs'. Thirdly, verbs with distributional differences are also considered similar, e.g. the transitive *spēlēt* 'to play' can not be substituted with the intransitive *rotaļāties* 'to play' despite their semantic closeness.

Hyponymy is mainly observed in nouns and verbs. The hyponymy between verbs is widespread (Cruse, 2004, p. 148), and wordnets tend to include a special subtype of verb hyponymy, namely, troponymy (Fellbaum, 1998, p. 80), in which case the hypernym denotes a more general action or process whereas the hyponyms differ in the manner of how the action or process happens or is carried out. Since the concept of troponymy is not known in Latvian linguistics so far, the Latvian WordNet does not differentiate any subtypes of hyponymy. Hyponymic relations are not established between adjectives and adverbs.

Meronymy is characteristic mainly of nouns especially those having a concrete meaning. In some cases, meronymy borders on hyponymy. This type of semantic relation can be applied mainly to physical objects, as well as to other more abstract ones, such as institutional units, e. g. the meronyms of (*uzņēmums*) 'enterprise' are (*filiāle, nodaļa*) 'subsidiary company'.

Antonymy is a relationship between semantic opposites. However, there are several subtypes of opposition and not all of them are considered antonymic. A prototypical group of antonyms consists of words denoting gradable notions (Löbner, 2002, pp. 88-90), e.g. for the synset (*brangs, dižens, dižs, ievērojams, liels, pamatīgs, prāvs*) 'large, big, great' the antonym is (*mačs, mazs*) 'small, little'. In Latvian WordNet, a wide understanding of antonymy is adopted, including other types of opposites as well. They are, firstly, complementaries, e. g. (*klātbūtne, klātiene*) 'presence' vs. (*trūkums*) 'absence', secondly, reversives, e.g. (*ārā*) 'outside' vs. (*iekšā*) 'inside', thirdly, converses, e.g. (*pārdot*) 'to sell' vs. (*pirkt*) 'to buy'. Other words that are often contrasted in language use are also considered antonyms, e.g. (*praktisks*) 'practical' vs. (*teorētisks*) 'theoretical' and (*sekas*) 'effect' vs. (*cēlonis*) 'cause'.

Words and synsets in one **gradation set** express different values of the same attribute. The relation of gradation is mainly seen between adjectives, however, it also occasionally occurs in nouns and verbs. In gradation sets, other semantic links may exist as well, e.g., if the gradable values cover the whole scale, antonymic relations are also included. On the other hand, gradation sets of verbs and nouns may include hyponymy, e.g. the word *līt* 'to rain' has a series of semantically linked verbs denoting raining of various intensity, which can also be considered types of raining and, thus, hyponyms. In the future, it is planned to develop a system of simultaneous marking for gradation and hyponymy where necessary.

In addition to the semantic links mentioned above, we also annotate conceptual connections (as **"see also"**), as a category for words that are semantically related, but not by any of the mentioned semantic relations.

4 Evaluation of Semantic Linking

In order to assess how consistently the linking principles developed during the project are applied, a three-person inter-annotator agreement (IAA) evaluation was conducted on 15 words (5 nouns, 5 verbs and 5 adjectives) with 85 senses altogether. Adverbs were excluded from the experiment as they are poorly represented in the dictionary due to the lexicographic tradition. The words were chosen from the core list of the most frequently used words by selecting words with a moderate number of senses (2-6 superordinate senses and possible subsenses). Revision of the sense inventory was not included in the scope of this experiment, so it was ascertained beforehand that the words selected from the dictionary already had comparatively suitable senses for wordnet linking.

The experiment was carried out in three stages.

1. In the given list of words, each linguist offered possible semantic links (including synonymy to form a synset); they could pick any sense or synset in the dictionary to form the link with.

2. All linked synset pairs (324 in total; 96 pairs for initial 5 nouns; 105 - for verbs, 123 - for adjectives) that appeared in the first step of the experiment (even if only one linguist suggested it) were collected into a list, and each linguist repeatedly considered what kind of a

	R1 All	R1 N	R1 V	R1 ADJ	R2 All	R2 N	R2 V	R2 ADJ
Given synset pairs	∞	∞	∞	∞	324	96	105	123
Overall annotated links†	535	160	166	209	833	252	262	319
Any link: 3 people	75	23	22	30	221	70	68	83
Any link: 2 people	60	18	17	25	69	16	23	30
Any link: 1 person	190	55	66	69	32	10	12	10
No link	-	-	-	-	6	0	2	4
Matching linking: 3 people	47	15	17	15	129	43	49	37
Matching linking: 2 people	295	90	98	107	277	85	92	100
No matching links	30	6	7	17	51	11	13	27

Table 1: Results of the first two stages of the experiment (R1 and R2).
† The total number of links annotated in the IAA experiment, i.e., if three annotators provide the same link, it is counted in this sum thrice.

semantic link (if any) was necessary in each case.

3. In the third stage, the results of the second round were compared and discussed by all three linguists. In this stage, differing answers were discussed, as well as the possibility to agree on one answer (a specific relation or the absence of it between the senses); the linguists also had the option of leaving their decision unchanged.

We are using Fleiss' kappa measurement to judge inter-annotator agreement between multiple annotators. It is interesting to note that most evaluations of wordnet quality in literature only rarely (e.g Ehsani et al. (2018)) attempt to make such estimates for the semantic links within the wordnet,

The results of the first stage (see *R1* part of Table 1) showed that the endpoints of the selected links were sufficiently different; at this point, Fleiss' kappa measurement was 0.55 (CI95% 0.48 - 0.63), i.e., moderate agreement. Out of 324 different linkable synset pairs which were proposed by annotators, only 47 had exact matching links for all three annotators. This was partially due to each annotator choosing different potential senses to link or not thinking of other possibly corresponding senses at all. Thus, it was concluded that additional automatic solutions for offering potential candidates would prove useful in the future; the identification of such candidates could be based, for example, on similarity of sense definitions. It should also be noted that data from a synonym dictionary were also available during the experiment. However, the coverage of such data is incomplete, as only some words from the experiment have synonym dictionary suggestions, and such a resource does not provide recommendations for any of the other types of semantic relations. This stage also demonstrated the differences in each annotator's individual approach: as seen from the data, one annotator connects synsets comparatively cautiously and less often, another much more freely, which also affects the inter-annotator agreement. Given the low number of matches in the chosen sense pairs themselves, it would be difficult to distinguish an actual agreement on semantic link creation. For this reason, the second stage of experiment was organised, with a prepared list of potential sense pairs to be linked.

The results of the second round where annotators got a pre-made list of potential sense are given in *R2* part of Table 1. Surprisingly the inter-annotator agreement showed by Fleiss' kappa was lower but still in the range of moderate agreement – 0.46 (CI95% 0.40 - 0.46), however this might also be due to the relatively small size of this experiment. As it was suspected before the experiment, the overall amount of proposed links increased dramatically – from 535 to 833. It seems that when annotators are provided a large quantity of proposed candidates, more links are made but inter-annotator agreement decreases as annotators are forced to make a choice about words they did not consider themselves.

The level of agreement on adjective links is lower than the agreement on noun and verb links, which indicates that the methodology of marking adjectival links should be further expanded and clarified. When looking at separate link types, a precise agreement also appeared in antonyms and gradation sets, suggesting that when such candidates are presented, the semantic relation is recognized.

The results of the third round were also used for making the McNemar's test, resulting in a p-value of 2.51×10^{-23} indicating that consultations made statistically significant changes to the data. In 15% of the discussed cases disagreements still remained even after consultations. From this it can be concluded that the linguists' seminars organized regularly during the project to solve various labeling and annotation dilemmas for specific words are notably beneficial for the creation of a more consistent system. At the same time, it can be seen that even after a unified theoretical base, a developed methodology and regular discussions, there are cases when annotators have differing opinions.

Some of the cases of disagreement are as follows. Firstly, there were varying opinions as to whether the synset (*vebkamera*) 'webcam' is a hyponym of synset (*kamera*) 'photografic or video camera', considering that a webcam carries out an additional function of transmitting an image instantly, which a regular camera does not. This raised speculations about whether a webcam is a new type of camera or they both are types of some more general meaning of camera that is not represented in the dictionary. Secondly, there were discussions regarding the synsets (*inspekcija*) 'inspection' and (*apskate*) 'examination'. Opinions differed as to whether they are members of the same synset or whether the 'inspection' includes 'examination', but 'examination' can exist without 'inspection'. Both of the given examples show a different understanding of the importance of one seme to distinguish a new meaning or a new semantic relation. The difference of opinion also occurred in situations where the linguist feels a close semantic connection between the senses, but is unable to define it in the currently available relation set, or in moments, when each linguist indicated a different type of relation, although most likely none of the currently available relations fully corresponds to it in its general sense. The synset (*tonis*) 'tone - a quality of a given color that differs slightly from another color' and synset (*krāsa*) 'color' serves as an illustrative example for this. One linguist suggested that color consists of various tones and therefore a meronymy/holonymy link could be used; at the same time, another linguist believed that tone is an attribute of color and therefore the appropriate link type is "See also". It should also be noted that none of the linguists suggested a relation to a hierarchy in this case, although that is exactly the type of link used in Princeton WordNet between these synsets.

In order to obtain a gold standard, it may be necessary to assign an authoritative linguist who will determine the final opinion in such cases.

The qualitative analysis of the data gave sufficient grounds for the additional conclusion that link formation can successfully highlight cases, when sense revision is necessary during the process of annotation. There were cases when it was agreed that the reason for disagreement was the vague definition of certain word senses, which, in turn, complicated the possibility if agreement, as there was too much space for interpretation.

In short, the experiment has demonstrated the complexity of the given problem, but also provides an opportunity to evaluate the consistency of annotated data. A more detailed analysis of separate semantic link types is planned in future, to further improve our methodology.

5 Linking Latvian WordNet to Princeton WordNet

As a part of the project, Latvian WordNet to Princeton WordNet sense mapping is carried out to identify English equivalents for Latvian word meanings. Currently, only a manual mapping has been implemented for the 2000 most frequently used Latvian words. However, the manually generated data are being used to develop and train the algorithm for automated sense linking, which will be carried out for a significantly broader scope of word meanings. The version that the Latvian word meanings are presently being mapped to is Princeton WordNet 3.0.

Currently, the project implements wordnet to wordnet interlinking on the level of synsets, as opposed to linking individual word senses as seen, for example, in plWordNet (Rudnicka et al., 2019). Such choice of approach is motivated by the need to primarily secure a foundation of optimal interlingual equivalence based on meaning, that would later potentially serve as a basis for more intricate equivalence structures based on stylistic register, dialect, gender and other aspects, which can be linked sense to sense.

The project's main theoretical base for creating interlingual links and word sense equivalence is taken from translation theories that offer various perspectives on equivalence (e.g. natural vs. directional) (Pym, 2014; Venuti and Baker, 2000; Chesterman, 2016), to better understand the poten-

tial asymmetry between two or more languages. Thus, not only full or direct equivalence is taken into account, but also such types as functional, formal, stylistic, situational and semantic equivalence (Venuti and Baker, 2000; Chesterman, 2016). This provides additional context for each decision to minimise inconsistency or artificially rigid or symmetrical interlingual structures.

The current process of interlinking is facilitated by automatic suggestions of possible equivalents for each word, based on bilingual dictionaries and machine translation. This feature is integrated in the editing tool, but the linguist may also freely choose and select other English word meanings if the suggestions do not seem to fit the specific meaning in Latvian. Therefore, both automated and manual methods are already combined in this step of the process. So far, 3139 interlingual links of various types have been created between Latvian WordNet and Priceton WordNet.

However, during the early stages of wordnet to wordnet linking it was concluded that direct links alone cannot fully convey the various cases of interlingual hyponymy, namely, cases when a synset in the source language conveys a broader or narrower scope of meanings than its closest equivalent in the target language. Consequently, three types of interlingual links were created, enabling the editors to mark a Latvian synset as a full equivalent, as well as being broader or narrower than its English counterpart. If an equivalent synset can be identified, links of narrower or wider meanings are not allowed. If an equivalent synset can not be identified, multiple links of narrower and wider meanings are allowed.

Full equivalence may be seen in the Latvian synset (*jautājums, prasījums, vaicājums*) and the Princeton WordNet synset (question, interrogation, interrogative, interrogative sentence): the meanings describe a sufficiently similar concept with the same level of semantisation. This type of direct link is the most often used – it constitutes 1891 of all interlingual links. But, for example, considering the Latvian synset (*pārmest*), roughly translated as 'reprimand' or 'reprove', it can be concluded that there is no single equivalent for it in the Princeton WordNet; instead, several, broader synsets, such as (reproach, upbraid) and (admonish, reprove, reproof) are linked to it through interlingual hyponymy links, each denoting a part of its full, comparatively broader range of meanings. There are currently 545 such links.

Conversely, there are also certain cases, when Princeton WordNet synsets have a broader set of meanings than their Latvian counterparts. For example, the synset (sibling), which includes both brothers and sisters, does not have a direct equivalent in Latvian[1]. Therefore two separate hyponymy links need to be made with the more specific (*bāleliņš, bāliņš, brālis*) 'brother', and (*māsa*) 'sister' to convey the full meaning of the concept of a sibling. 703 such links have been created in Latvian WordNet so far. Interlingual hyponymy links not only help in the previously described cases, but also in linking cultural realia to more general meanings in the other language. Thus, the data that would otherwise be left unmarked can be involved in forming the interlingual hierarchies between Latvian WordNet and Princeton WordNet.

A notably problematic aspect in the formation of interlingual links are word meaning definitions, which in some cases have become outdated over the course of time or have been left unnecessary broad or narrow. For example, Princeton WordNet lists only the general meaning of 'dispute' (disagreement), without separating the meaning of a legal dispute, which exists in Latvian WordNet. Similar cases have been observed in Latvian WordNet, especially in instances when meaning definitions list two aspects separated by a semicolon. Such ambiguous cases automatically involve selective use of annotators' personal knowledge or additional research to discern the true level of meaning equivalence; such cases are discussed in greater detail during the weekly project linguist seminars to reach the most objective solution.

So far, distinguishing three types of interlingual links has proved useful to bridge the gaps and differences between Latvian and English. It is expected, that this approach will also facilitate the future aspirations of incorporating Latvian WordNet into Open Multilingual Wordnet (Bond and Foster, 2013), as a working mechanism will already be established to deal with any potential inconsistencies or language differences.

6 Evaluating Interlingual Links

To evaluate our process of automatic interlingual link creation, another IAA experiment was car-

[1]In Latvian, *brālis* 'brother' refers exclusively to males. There is an English calque 'sibs' used as a term in genetics, but it is not understood or used by non-specialists.

ried out. In the experiment, annotators evaluated the machine-translated suggestions, taking into account the opinions of three annotators. The proposed links were separated in the following four categories:

1. link corresponds perfectly;

2. the proposed link points to a semantically wider or narrower sense than the Latvian word sense;

3. more information is needed to make a decision, as it is clear that there is some semantic relation but not obvious what type of relation;

4. the proposed link does not correspond at all.

The IAA experiment was performed using words from common vocabulary with only one sense in Latvian (including homonyms), excluding regional words, slang etc. There are up to five possible candidates of English equivalents offered by the system which the annotators can choose from.

Three linguists annotated 684 instances in total. On 272 corresponding outputs all annotators agreed that the proposed interlingual link should be approved, and in 94 cases all annotators decided that the suggested links definitely do not match the Latvian meaning. In 57% cases annotators fully agree, and out of all the automatically provided candidate links 40% are undisputed interlingual matches.

In cases when all three annotators chose to select the "wider/narrower meaning" option, several links were proposed. For example, *apnikums* (a mental state when a person is bored and tired of everything) had four suggested links: (boredom, ennui, tedium), (depression), (fatigue, weariness, tiredness), (tediousness, tedium, tiresomeness). All of suggested links are somehow semantically connected to *apnikums*, but none of them corresponds completely. From this it can be concluded that the automated system has already noted the absence of complete equivalence in this case.

The main reason of annotators' disagreement with automatic suggestions was the occasional inability of MT to correctly interpret the meaning of derived words. For example, *apgaismniecība* 'Enlightenment' (derived from *gaisma* 'light') had the automatic MT suggestion of "lighting" (the craft of providing artificial light).

Another reason for disagreement was based on grammatical differences between Latvian and English, especially in the use of genitive case. In Latvian, a noun in genitive case is often used to name a quality, taking the place of adjective. For example, inflexible genitive noun *apļveida* (derived form *aplis* 'circle') is used only in this (genitive) case and implies quality (circular, round). Because it is a noun, MT suggests a link to the noun synset (circle, round).

Differences in word meaning definitions between wordnets may occur for seemingly similar concepts. In that case answers between annotators may vary. For instance, *apašs* 'apache' is defined in Latvian a "a French gangster", whereas Princeton WordNet suggests that it is "a Parisian gangster". Two annotators considered this as a direct link, one viewed this as wider/narrower case. Thus, the annotators had to look at each case individually and decide whether to base their decision on their knowledge of the subject or to stick to the given definitions, leading to the conclusion that the result in this case cannot be completely objective. It also brings to attention the difference which even a minimal manual control can make in automatically created data.

Disagreement based on annotators' personal opinion frequently appeared on words that name state, condition, sensation and other abstract concepts. These differences are mainly based on annotators' personal understanding of the concept in Latvian. Personal opinion also may vary on how we perceive translation quality and which semantic differences are essential when choosing between direct, wider/narrower or no link. For example the Latvian meaning *asthma* "a fit of loss of breath, shortness of breath" and the English synset (asthma, asthma attack, bronchial asthma) "respiratory disorder characterized by wheezing; usually of allergic origin" has a different answer from each annotator: 1 "corresponds", 1 "wider/narrower" and 1 "needs more information".

The IAA results for interlingual links not only have helped reinforce the importance of multiple link types, but also aided in the future the development of clearer strategies and criteria for annotating ambiguous, more complicated meanings.

7 Publishing Results

The main access point for this resource to the general public is through the Tēzaurs.lv (https://

tezaurs.lv) online dictionary, which is widely used in Latvia. However, for the purposes of the research community we also publish this data in various formats and in multiple repositories. Latvian WordNet is developed and maintained in the Tezaurs.lv lexicographic platform with a PostgreSQL database custom data structure, which then can be exported in multiple widely recognised data formats.

Currently we provide an Open Multilingual Wordnet compatible LMF XML[2] export for the wordnet data, and a more detailed TEI 5 (Text Encoding Initiative) Dictionary chapter XML[3] which contains both Tezaurs.lv dictionary data and Latvian WordNet synsets and links. The TEI format also contains information about gradation sets, which is not available in LMF due to format restrictions.

All the latest version data (including a full database dump) are available on the project homepage[4], where we also provide a list of Latvian Wordnet core words. The TEI XML dataset is also regularly published in the CLARIN-LV repository[5] (Skadina et al., 2020). Our intent is to publish LMF export both via CLARIN-LV and OMW infrastructure. We do quarterly releases for all our dictionary and wordnet data.

8 Conclusions and future work

To summarize, we are happy to present the first major release of Latvian WordNet, providing a manually curated resource of a reasonable size, based on Latvian corpus evidence and linguistic tradition that can be a solid basis for future research work.

The current Latvian WordNet consists of 7609 words linked in 6515 synsets, out of which 1266 synsets are considered completed as they have all the outgoing semantic links annotated, corpus examples assigned for at least one word in the synset, as well as links to the English Princeton WordNet formed, and the remainder being less frequently used words that have been joined by outgoing semantic links from the 'core' synsets. 70826 corpus examples were linked to specific word senses and subsenses. This information is available to public as an integrated part of the Tēzaurs.lv online dictionary and has received positive user feedback regarding its usefulness.

From the perspective of linguistic principles, we are satisfied with our choice to form a wordnet from scratch. Even if bootstrapping from English resources would have taken less effort, our experience with linking to the Princeton WordNet has indicated many interlingual differences, which through automatic means would have imposed an artificial, English-derived structure upon this resource.

It is interesting to note that sense granularity is still an issue open for debate among the annotators, with no clear consensus despite the fact that it was one of the primary drivers for restructuring the existing sense inventory and a key part of the methodology discussion over the last three years. Developing an adequate sense inventory takes a large amount of time and effort compared to forming synonym sets and other semantic links.

Our approach of word sense selection based on corpus evidence has also resulted in a large quantity of corpus examples aligned to the specific word senses, which forms a useful dataset for training word sense disambiguation systems (Paikens et al., 2022b). Ongoing future work in this direction is annotating a gold standard text - the first two chapters of The Little Prince - with specific word senses from Latvian WordNet.

The results of our inter-annotator agreement experiments for semantic links within Latvian WordNet indicate the difficulty and the subjective nature of semantic linking. A relevant observation is that providing automatically generated candidates improves the linking coverage, as annotators often agree that the link should be made if they are aware of the option, but might not come up with the related word on their own. It seems that when annotators are provided a large quantity of proposed candidates, more links are made but inter-annotator agreement decreases as annotators are forced to make a choice about words they did not consider themselves. It also indicates that annotator discussions improve consistency, so the differences apparently involve also a different understanding of methodology, not a fundamental disagreement about the discussed words.

In 57% cases annotators fully agree, and out of all the automatically provided candidate links 40% are undisputed interlingual matches.

[2] https://globalwordnet.github.io/schemas/#xml

[3] https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

[4] https://wordnet.ailab.lv/data/

[5] https://repository.clarin.lv/repository/

For Latvian-English links we observe 57% exact match IAA between all three annotators, with some disagreement whether a certain sense is the same or broader in one of the languages. We observe less agreement over abstract concepts, as their perception seems to be more subjective, and it is difficult to decide on the most appropriate interlingual link. In general, the generation of automatically provided candidates were very helpful in rapidly creating links, as the 40% of candidates were clearly proper links, but they do need manual review.

For further improvement of Latvian WordNet the planned future tasks involve adding links for word derivation, extending the automatic link candidate derivation also for intra-language semantic links based on existing word definitions and language models from large corpora, and also continuing the manual review of proposed Latvian-English links which could then enable a transfer of semantic relations from Princeton WordNet to Latvian WordNet.

It would be interesting to apply this resource for cross-lingual research on semantic alignment and differences between Latvian and Lithuanian WordNet (Garabík and Pileckytė, 2013), as well as going beyond current semantic links to word derivation and etymology.

Continued extension of the manually developed Latvian WordNet is also an obvious direction of future work, but is highly contingent on funding opportunities. We are also considering a specific project to integrate idiomatic expressions and other multiword entities in the Latvian WordNet.

Acknowledgements

This work was supported by the Latvian Council of Science, project "Latvian WordNet and Word Sense Disambiguation", project No. LZP-2019/1-0464. We also thank the anonymous reviewers for their input in improving this paper.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Andrew Chesterman. 2016. *Memes of translation: The spread of ideas in translation theory*. John Benjamins Publishing Company.

Alan Cruse. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(3).

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

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to lithuanian wordnet. *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80.

Dirk Geeraerts. 2009. *Theories of lexical semantics*. OUP Oxford.

E. Grīnberga and O. Kalnciems, editors. 1998. *Latviešu valodas sinonīmu vārdnīca*. Avots, Rīga. 3. papildinātais un pārstrādātais izdevums.

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

Kadri Kerner, Heili Orav, and Sirli Parm. 2010. Growth and revision of Estonian WordNet. *Principles, Construction and Application of Multilingual Wordnets*, pages 198–202.

Kristīne Levane-Petrova. 2019. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Language: Meaning and Form*, 10:131–146. The Balanced Corpus of Modern Latvian, its role in grammar studies.

Sebastian Löbner. 2002. *Understanding Semantics*. UK: Hodder Arnold.

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

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0–a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268.

Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz, and Joanna Rabiega-Wiśniewska. 2015. Semantic relations among nouns in polish wordnet grounded in lexicographic and semantic tradition. *Cognitive Studies| Études cognitives*, (11):161–181.

Marek Maziarz, Stanisław Szpakowicz, and Maciej Piasecki. 2012. Semantic relations among adjectives in Polish WordNet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies| Études cognitives*, (12):149–179.

Peteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2022a. Towards Latvian WordNet. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2808–2815, Marseille, France. European Language Resources Association.

Pēteris Paikens, Laura Rituma, and Lauma Pretkalniņa. 2022b. Towards word sense disambiguation for latvian. *Baltic journal of modern computing*, 10(3):402–408.

Anthony Pym. 2014. *Exploring translation theories*. Routledge.

Ewa Rudnicka, Maciej Piasecki, Francis Bond, Łukasz Grabowski, and Tadeusz Piotrowski. 2019. Sense equivalence in plwordnet to princeton wordnet mapping. *International Journal of Lexicography*, 32(3):296–325.

B. Saulite, R. Dargis, N. Gruzitis, I. Auzina, K. Levane-Petrova, L. Pretkalnina, L. Rituma, P. Paikens, A. Znotins, L. Strankale, K. Pokratniece, I. Poikans, G. Barzdins, I. Skadina, A. Baklane, V. Saulespurens, and J. Ziedins. 2022. Latvian national corpora collection – korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

Inguna Skadina, Ilze Auzina, Normunds Gruzitis, and Arturs Znotins. 2020. Clarin in latvia: From the preparatory phase to the construction phase and operation. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pages 342–350.

Andrejs Spektors, Ilze Auzina, Roberts Dargis, Normunds Gruzitis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, and Baiba Saulite. 2016. Tēzaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2568–2571, Portorož, Slovenia. European Language Resources Association (ELRA).

Laine Strankale and Madara Stāde. 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.

Lawrence Venuti and Mona Baker, editors. 2000. *The translation studies reader*. Routledge London.

Zygmunt Vetulani, Marek Kubis, and Tomasz Obrębski. 2010. PolNet — Polish WordNet: Data and tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).