Extracting WordNet links from dictionary glosses – Latvian Wordnet example

Elīza Gulbe and **Agute Klints** and **Gunta Nešpore-Bērzkalne** and
Laura Rituma and **Madara Stāde** and **Ilze Lokmane** and **Pēteris Paikens**
Institute of Mathematics and Computer Science, University of Latvia
Raiņa bulvāris 29, Riga, Latvia
Correspondence: peteris@ailab.lv

Abstract

This paper presents a project report on methods for extending the Latvian WordNet by automated extraction of candidate semantic links. We describe our experiments with neural network classifiers applied to extract, score and rank potential synonymy and hypernymy links between senses based on the sense glosses in the Tezaurs.lv online dictionary, and provide a manual evaluation of these candidates, demonstrating 72% and 67% accuracy for synonymy and hypernymy links respectively. As the methods used are language-independent, we hope that this research would be applicable to other wordnets as well.

Keywords **–** Latvian WordNet, machine learning, semantic link detection, hypernymy, synonymy

1 Introduction

In this paper we describe our experiments with automated WordNet link candidate extraction to support the enlargement of Latvian WordNet (Paikens et al., 2023). The current version of Latvian WordNet consists of 11 399 words that are linked in 8 768 synsets by manually curated links. However, potential NLP applications of this data require a high coverage and motivate a search for approaches that could significantly extend this resource without excessive amount of manual linguist labor. As the underlying lexical resource – Tezaurs.lv digital dictionary (Grasmanis et al., 2023) – contains more than 400 000 entries, their sense definition gloss data is a valuable potential source for further WordNet links. While the earlier development of Latvian WordNet focused on the most frequent core words, which often are highly polysemous and required careful restructuring of the sense inventory to ensure clear separation of senses and appropriate granularity, the senses for the less common words usually are usable as-is. Motivated by the recent advances in applying language models and embeddings to link extraction, as described in the next chapter, we investigated options to automatically identify the 'low hanging fruit' of potential new WordNet links based on this data.

2 Related Work

There are several strategies for detecting semantic links between words or synsets automatically. The first strategy is to create semantic relations based on Princeton WordNet. For example, (Bakay et al., 2021) mapped eight different semantic relations semi-automatically for Turkish WordNet KeNet by finding corresponding synset and its relation to other synsets in English WordNet PWN which were then checked by human annotators.

The second strategy for detecting semantic relations is to use word embeddings. For example, (Oliveira, 2023) applied masked patterns on BERT models to identify semantic relationship between words. (Tseng and Hsieh, 2019) attempted to identify hypernymy-hyponymy relationship in Chinese by building binary classifier based on the assumption that there is a semantic relationship between a word and its composing character. A similar approach was also taken by (Berend et al., 2018) where logistic regression model was trained for hypernymy discovery, which output the likelihood of two words being hypernyms for a given query. The highest ranked candidates were then selected as the most suitable hypernyms.

Whereas (Pocostales, 2016) approached the hypernymy-hyponomy detection problem by computing an average embedding offset of 200 known hypernym-hyponym pairs to then predict the hypernyms of new words. This approach was also studied by (Kafe, 2019) showing that none of the tested offset calculations methods were able to detect symetric relations (synonymy and antonymy), whereas for asymetric relations (hyperonymy/hyponymy and meronymy) this method was able to detect semantic relationship for Skip-Gram

and GloVe embeddings (Tan et al., 2020).

The third strategy is to use existing lexical resources, such as dictionaries or corpora, as the basis for the detection of semantic links. For example, earlier work on compiling lexical resources directly from monolingual dictionaries includes DanNet, a Danish WordNet, which reused sense distinctions and hypernym-hyponym relationships explicitly available in the Den Danske Ordbog to semi-automatically construct a WordNet, with missing information supplemented manually to ensure a consistent semantic hierarchy.(Pedersen et al., 2009) Whereas, in the construction of plWordNet for Polish, a hybrid method combining automated and semi-automated techniques was used.(Broda et al., 2009) Distributional methods, leveraging co-occurrence patterns in texts, and pattern-based methods, utilizing linguistic templates, were employed to extract semantic relations.

Many of the described methods focus solely on semantic relations at the word level. However, in WordNet, a synset encompasses both words and their specific senses, each linked to the same underlying concept. Building on existing approaches, we aim to leverage lexical resources, such as dictionaries, and pattern-based recognition methods to generate candidate word pairs with potential semantic relationships. To address the challenge of polysemy, this research introduces a mechanism that compares words along with their respective senses to other words and senses, enabling the detection of semantic relations at a sense level by training a neural network classifier trained on already existing Latvian WordNet dataset.

3 Candidate Extraction

Table 1 shows the unique counts of relationship types in the current Latvian WordNet dataset for nouns. For example, if a synset contains *n* records, the unique synonymy links within the synset is calculated by:

$$C(n, 2) = \frac{n!}{2!(n-2)!} \quad (1)$$

The sums of unique synonymy links are then accumulated for all synsets. For other links we calculate the unique count by multiplying the number of senses within two synsets - if $synset_1$ has *n* entries and $synset_2$ has *m* entries, we obtain **n * m** unique relationships in total which are later accumulated for the whole Latvian WordNet for nouns.

For the purpose of this research we focused only on hypernym and synonym detection because the sample count for other relation types was insufficient for the selected solution implementation as shown in Table 1.

Relation type	Unique count
synonymy	18 682
hypernymy	11 802
similar	919
holonym	811
see-also	659
antonym	274

Table 1: Unique counts of relationship types recorded in the Latvian WordNet dataset for nouns

3.1 Hypernymy

To get hypernym candidates we applied one of the rule-based extraction principles using Tēzaurs database previously studied in (Grūzītis et al., 2007) - if the principal clause consists of a noun in the nominative case, the noun is usually a hypernym of the word being explained. This approach does not map specific hypernym senses together, therefore, if $sense_1$ has a hypernym candidate $word_2$ with *n* senses, we generate *n* potential hypernym sense candidates. This approach was applied to 8000 most frequent words in the corpora for words with four or less senses thus generating 12000 hypernym candidate senses in total. The word frequency data was calculated based on the Latvian National Corpora Collection (Saulite et al., 2022)

For example, the word *lidaparāts* 'aircraft' has only one sense - *ierīce, transportlīdzeklis, kas spēj pārvietoties pa gaisu vai kosmosā* 'device, vehicle capable of moving through air or space'. In this case the extracted hypernymy candidate words are *ierīce* 'device' and *transportlīdzeklis* 'vehicle' because both of these words are included in the definition in the nominative case and belong to the principal clause of the sentence. Both *ierīce* 'device' and *transportlīdzeklis* 'vehicle' have only one meaning in their respective glosses, therefore, we generate two sense pairs as potential hypernym candidates that are later evaluated by our method.

3.2 Synonymy

For synonymy we had an unstructured data source available for the candidates - an older synonym dictionary (Grīnberga et al., 1972) that was digitized,

which included both absolute and near synonyms for 5839 words. This allowed to generate potential synonym candidates using Tēzaurs database. For example, if $word_1$ has *n* senses and $word_2$ has *m* senses, in total **n * m** synonym candidates are generated. As this is a first prototype for semantic link extraction, headings with four or less senses were selected. Additionally, as the selected dictionary consisted also of less popular words we chose 7000 synonym candidates with most popular words based on the Latvian National Corpora Collection (Saulite et al., 2022).

4 Relation Detection

To detect hypernym and synonym relations between two word senses, we trained a single hidden layer neural network. We created the vector embedding representation of the dataset using a pretrained monolingual encoder-only BERT model for Latvian, provided by the HPLT project (de Gibert et al., 2024). For each relationship type in the dataset (hypernymy, synonymy, or other) **we embedded both the sense and its respective word** for the training process. Therefore, when training the model we use the candidate pair senses, their respective word and class they belong to.

During the experimentation phase, we tested different model architectures, using 20% of the entire dataset as a validation set to compare performance. We explored variations in hidden layer sizes, activation functions, and optimizers. Additionally, we performed hyperparameter tuning on parameters such as the number of epochs, batch size, and learning rate to maximize model's performance. The tuning process involved systematic experimentation with different values for each parameter to identify the optimal configuration for our dataset.

The highest validation dataset results, shown in Table 2, were obtained by training a single hidden layer neural network with the following architecture:

- Input layer - a concatenation of $word_1$ embedding, $sense_1$ embedding, $word_2$ embedding, $sense_2$ embedding of size 3072;
- Hidden layer of size 512 followed by ReLU activation function;
- Output layer of size 3 (to predict if the given input layer is synonym, hypernym or other) followed by Softmax activation function to convert logits into probabilistic distribution.

The best performance was achieved with a hyperparameter configuration of 140 epochs, a batch size of 32, and a learning rate of 1.62×10^{-5}.

After training the model, we applied it to data retrieved from the candidate extraction phase, described in Section 3, to obtain probabilities for each candidate's relationship type: synonymy, hypernymy, or other. We specifically focused on candidates with potential synonym or hypernym relations, passing them through the model to identify the highest probability of synonymy or hypernymy for each word pair. In this task, we concentrated on the highest probability sense pair within each word pair, and even if the probabilities were close, only the sense pairs with the highest probability were considered as candidates for a potential semantic link that were later evaluated manually as described in Section 5.1.

4.1 Dataset

A subset of labeled nouns from the Latvian WordNet data was used as examples for training and evaluating the classifier. The dataset includes three classes: *synonyms* (18 682 samples), *hypernyms* (11 802 samples), and *negative examples* (40 000 samples). Negative examples are derived from the Latvian Wordnet data specifically for nouns and include **(1) random negatives**: senses not classified as synonyms or hypernyms; **(2) higher-level hypernyms**: those not falling under direct hypernyms; **(3) close embeddings without relations**: word pairs with small Euclidean distances but no labeled relations; **(4) unrelated senses of related words**: instances where a word has multiple senses, with only one being a hypernym or synonym of another word, while the unrelated senses are used as negatives; and **(5) similar/also/antonyms/holonyms**: pairs not qualifying as synonyms or hypernyms, included to differentiate other relations from hypernyms and synonyms.

5 Evaluation

The evaluation of the provided solution consists of automatic validation after the completion of training process and manual evaluation of generated semantic links analyzed by linguists followed by analysis of systematic errors produced by the selected implementation method.

5.1 Methodology for manual evaluation

We conducted a manual evaluation of 400 sense pairs identified as synonyms or hypernyms. Each

Class	Precision	Recall	F1-score
Hypernymy	0.87	0.68	0.77
Synonymy	0.91	0.93	0.92
Other	0.91	0.93	0.92

Table 2: Validation dataset results

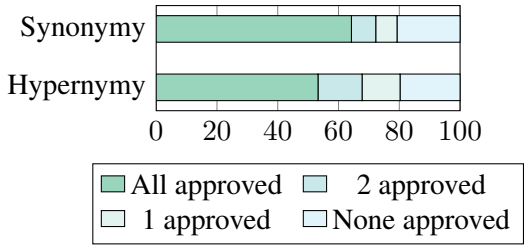

Figure 1: Approval rates based on number of approvals

dataset was independently evaluated by three linguists. Each rater was presented with an Excel spreadsheet where each row contained a word with its gloss, as well as a candidate link target word with its gloss, highlighting the proposed specific synonym or hypernym candidate sense. The raters were instructed to evaluate each pair based on whether the senses shared either an interchangeable (synonym) or a hierarchical (hypernym) relationship. They provided a definitive "yes" or "no" for each candidate pair. We define "complete rejection" as an instance when all the raters responded with "no" and "partial rejection" as an instance when 1 or 2 raters responded with "no". Whereas "complete approval" is defined as an instance when all three raters responded with "yes".

5.2 Results

Comparing the results of the validation dataset shown in Table 2 and from manual review in Table 3 we see a substantial discrepancy between the automatic validation results from the validation dataset and the manual evaluation results where the automatic validation shows a more optimistic outcome, - 87% precision in automatic validation versus average precision of 72% for hypernyms and 91% precision in automatic validation opposed to average precision of 67% for synonyms. This discrepancy likely arises because the validation dataset consists of preprocessed entries, where meanings have been refined to ensure similar granularity, facilitating accurate link prediction. In contrast, the manually reviewed data lacks this level of pre-processing, meaning definitions are less aligned in granularity, which makes it more challenging to confirm links and leads to lower precision.

5.3 Observations

As mentioned above, there were two types of negative results: complete rejection and partial rejection (Figure 1). Approximately 20 % of proposed candidates were rejected outright in both (synonymy and hypernymy) cases. 15% of synonym candidates and 27% of hypernym candidates were partially rejected - at least one linguist would have approved the candidate.

The first observation of the **complete rejection** of proposed synonym pairs is that the proposed words were correct, but the senses were mismatched. However, it can be inferred that the validation tool should enable linguists to assess alternative meaning pairs when a word has multiple interpretations.

Among the completely rejected hyperonym candidates, there were cases where one of the coordinated elements of the sentence was chosen from the definition as a candidate, for example, in definition *mutācijas rezultātā radies dzīvnieks, augs, tā pazīme* 'animal, plant or the feature as a result of mutation' the first conjunct "animal" is chosen as the candidate. That indicates that the principle mentioned in the section 3.1 can give negative results in the case of coordination.

Fully or partially rejected candidates also include senses of derivatives which in Tēzaurs are often explained with schematic definitions like *ļaunprātība – 'Vispārināta īpašība –> ļaunprātīgs'* ('malice - generalized quality –> malicious') with a reference to its base word. In these examples hypernym *īpašība* 'quality' was proposed. But *īpašība* 'quality' in this gloss indicates derivation's semantic role in relation to its base word rather than names a hypernym, therefore such schematic definitions should be excluded from candidate extraction.

The **partial rejection** of proposed hypernyms and synonyms can be attributed to several factors.

First of all, a different granularity of word sense and the resulting ambiguity of definitions - the word's meaning is not sufficiently detailed to encompass the proposed hypernym. In many cases some linguists could detect these details as a reason to reject proposed hypernym, some may not. For example, the case of *ceptuve* 'bakery' where the gloss includes both a company and a building where this company acts, had a proposed hypernym *uzņēmums* 'company'. Some approved this as hy-

	Rater 1	Rater 2	Rater 3	Average	Fleiss' Kappa
Synonymy	77.50%	65.75%	72.75%	72.00%	0.76
Hypernymy	69.00%	60.50%	71.75%	67.08%	0.59

Table 3: Acceptance rate of synonym and hypernym candidates, including average score and Fleiss' Kappa

pernym, but some didn't, because of the more narrow semantic element 'building' in the hyponym's definition.

The same goes for synonym candidates - the senses of one of the candidates are separated in more detail, while the other one is given only a synonym in the definition, for example, the synonym candidates *tracis*, *troksnis*, *strīds*, *kņada*, *drūzma* ('ruckus', 'argument', 'quarrel') all can be applied both to the noise made by several living beings together, often quarreling, and to the quarrel itself, which can be obtained through active actions, speeches that create noise. The meaning of noise is not synonymous with the meaning of arguing. However, only some of these dictionary entries explicitly separate these senses, and thus proper WordNet links can not be made without restructuring the senses to ensure the same granularity.

Another reason for the rejection of candidates can be that the proposed hypernym is at a higher level, and some of the linguists thought of a more accurate direct hypernym for the hyponym. For example for *vieglatlēts* '(track and field) athlete' the proposed hypernym was *sportists* 'sportsman'. Some accepted this as its hypernym, but one considered that direct hypernym for *vieglatlēts* is *atlēts* 'athlete', whereas its hypernym is *sportists* 'sportsman'. Additionally, some proposed hypernyms may seem overly broad — such as *auklīte* 'nanny' being categorized under *sieviete* 'woman' — leading to ambiguity about whether a lower-level hypernym, such as *speciāliste* 'specialist' or other, should apply in between.

The third reason is the linguist's subjective sense of language and knowledge of the world. This leads to differences in linguists' understanding of how well a hypernym fits. For example, there may be disagreement over whether *kājsargs* 'leg guard' and *rāvējslēdzējs* 'zipper' qualify as a form of a device (the proposed hypernym for both was *ierīce* 'device'), because they are not the prototypical devices. In some cases it may differ in how literally linguists interpret sense definitions, as seen in examples like *bļitka* which means certain very low-valued paper money that was used in World War I, and the proposed hypernym *sīknauda* 'small change' where the sense definition explicitly mentions coins and technically excludes paper money.

Likewise, the candidate's evaluation can be influenced by the linguist's knowledge of the meaning of words - if they do not know the word, they will rely only on the definition, but if the other rater has more specific personal knowledge of the essential nuances of the meaning of the word, the evaluations can differ.

A lot of disagreements were observed in pairs of abstract concepts. The perception and interpretation of such concepts are strongly influenced by individual linguistic intuition, which is why some linguists may feel that the candidates are appropriate in the case of abstract concepts, while others may feel they are not.

6 Conclusions

The observed accuracy of identified candidate links means that all the links do need manual review, however, they are useful to speed up the Latvian WordNet extension as the significant number of unambiguously acceptable links can be rapidly annotated, and the manual verification of the candidate sense pairs took much less effort than annotating a similar quantity of synsets from scratch.

We identified that the addition of targeted negative examples and exclusion o certain word groups was valuable in improving the accuracy of selected candidates.

It is relevant to note that even after putting significant effort in the evaluation process, we still observe a major accuracy difference between the automatic evaluation on the previously annotated WordNet links and the manual evaluation on truly new, unseen data. Apparently the selection of dictionary entries used for the initial core Latvian WordNet and also the manual restructuring of their sense inventory means that their 'linkability' is substantially different than the rest of the dictionary. Most of the disagreements between the raters about the proposed candidates arose due to inconsistencies and imperfections in the underlying dictionary data. In several cases, a single sense in the entry should

be divided into multiple senses. The manually reviewed data lacks the degree of pre-processing that is put in current Latvian WordNet development, resulting in definitions that are less consistent in granularity, thus making it more difficult to verify connections and reducing precision.

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

Özge Bakay, Özlem Ergelen, Elif Sarmış, Selin Yıldırım, Bilge Nas Arıcan, Atilla Kocabalcıoğlu, Merve Özçelik, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Begüm Avar, and Olcay Taner Yıldız. 2021. Turkish WordNet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174, University of South Africa (UNISA). Global Wordnet Association.

Gábor Berend, Márton Makrai, and Péter Földiák. 2018. 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 928–934, New Orleans, Louisiana. Association for Computational Linguistics.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. A wordnet from the ground up. *Oficyna Wydawnicza Politechniki Wroclawskiej*.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

M. Grasmanis, P. Paikens, L. Pretkalnina, L. Rituma, L. Strankale, A. Znotins, and N. Gruzitis. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex): Invisible Lexicography*, pages 400–418.

E. Grīnberga, O. Kalnciems, G. Lukstiņš, and J. Ozols, editors. 1972. *Latviešu valodas sinonīmu vārdnīca*. Liesma, Rīga.

Normunds Grūzītis, Gunta Nešpore, and Baiba Saulīte. 2007. Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas. *Vārds un tā pētīšanas aspekti*, 11:147–159.

Eric Kafe. 2019. Fitting semantic relations to word embeddings. In *Proceedings of the 10th Global Wordnet Conference*, pages 228–237, Wroclaw, Poland. Global Wordnet Association.

Hugo Gonçalo Oliveira. 2023. On the acquisition of WordNet relations in Portuguese from pretrained masked language models. In *Proceedings of the 12th Global Wordnet Conference*, pages 41–49, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Peteris Paikens, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2023. Latvian WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 187–196, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: The challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Joel Pocostales. 2016. NUIG-UNLP at SemEval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1298–1302, San Diego, California. Association for Computational Linguistics.

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 Skadiņa, Anda Baklāne, Valdis Saulespurēns, 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.

Yixin Tan, Xiaomeng Wang, and Tao Jia. 2020. From syntactic structure to semantic relationship: hypernym extraction from definitions by recurrent neural networks using the part of speech information. *CoRR*, abs/2012.03418.

Yu-Hsiang Tseng and Shu-Kai Hsieh. 2019. Augmenting Chinese WordNet semantic relations with contextualized embeddings. In *Proceedings of the 10th Global Wordnet Conference*, pages 151–159, Wroclaw, Poland. Global Wordnet Association.