Towards Ukrainian WordNet: Incorporation of an Existing Thesaurus in the Domain of Physics

Melanie Siegel and **Maksym Vakulenko** and **Jonathan Baum**
Darmstadt University of Applied Sciences

Abstract

In this paper, we represent the first version of the Ukrainian wordnet – Ukrajinet 1.0. It contains 3,360 sets of full synonyms in the field of physics, consisting of 8,700 words. This knowledge base will help incorporate the Ukrainian language into multilingual scenarios of Natural Language Processing that need information about lexical-semantic relations.

1 Introduction

Information about words and their meanings is traditionally stored in dictionaries. With the increasing importance of automatic processing of language, a need for machine-readable dictionaries arose. In this context, wordnets emerged to store lexical information in a format that can be used by language processing systems. A wordnet (WN) is a lexical database of semantic relations between words in a given language. The basis of wordnets are synsets: groups of synonyms in the language that stand for the concepts of meaning. The first wordnet was created for the English language at Princeton University (also known as Princeton WordNet, (Fellbaum, 1998)). As the usefulness of wordnets as lexical resources for a wide variety of language technology applications became clear, the Princeton WordNet (PWN) was expanded and wordnets in other languages were created. The Open Multilingual Wordnet (OMW) is an open-source project created with the goal of facilitating the use of wordnets in multiple languages with open source license (Bond and Foster, 2013). The OMW has the added benefit of connecting equivalent synsets in different languages (Bond et al., 2016). This connection is created by an Interlingual Index called "ILI". The English version of the OMW (Open English WordNet, OEWN) is basically a copy of the PWN, with some improvements and additions, most notably the addition of an interlingual index for each synset (McCrae et al., 2019); (McCrae et al., 2020). Many of the OMW wordnets in other languages were developed using existing translations in the Natural Language Toolkit (NLTK). These translations were extracted and packaged into new wordnets. Consequently, the corresponding synsets in the resulting wordnets were linked using the ILI. Goodman and Bond (2021) developed the Wordnet Python library that can be used to access the OMW project wordnets in Python. The OEWN is distributed in electronic form as part of the NLTK, among others, and can be used with a corresponding Python library. NLTK provides translations for synsets into different languages, although these translations are incomplete. This means that not every synset in English has an equivalent translation in another language. There are also wordnets in other languages that were developed independently of OMW, such as GermaNet (Hamp et al., 1997). Many of these wordnets contain high-quality data that is resource- and time-consuming to create manually. As a result, some of these wordnets are commercially licensed and not free to use (nor are they part of NLTK, for example).

Ukrainian is a language with still few linguistic resources that is not yet contained in OMW. Therefore, an initiative has been launched to create an open-source Ukrainian wordnet (Ukrajinet, Ukrainian pronunciation [ʊːkrəːjiːnət]), which is being developed as part of the OMW project. The Ukrainian wordnet, Ukrajinet, will help incorporate the Ukrainian language into multilingual scenarios of Natural Language Processing that need information about lexical-semantic relations. This paper presents the first version and demonstrates how this resource will be expanded.

We will present the related work and show, how other wordnets have been developed and how the development of Ukrajinet fits into it. We outline the process of developing the first version of Ukrajinet and show how we applied existing methods. Fi-

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 121–126
September 18–22, 2023.

nally, we discuss the initial results and demonstrate how we will proceed.

2 Related Work

In the Open Multilingual Wordnet initiative (OMW, Bond and Paik, 2012; Bond et al., 2015), wordnets for several languages were developed and linked with each other.

Vossen (1998, p11) describes two basic approaches to developing new wordnet resources: In the first case (*expand*), existing PWN synsets of other languages are taken, and lexical entries are added for the specific language. In the second case (*merge*), language-specific resources are built and then linked to the PWN.

An example of *expand* is the Japanese wordnet (Isahara et al., 2008). It is based on translations of PWN to Japanese. The Japanese wordnet is not built fully automatically: most translations are manually checked. The authors found that there are differences between concept structures in English and Japanese, such that several synsets could not be translated. Other examples of *expand* include the Finnish (Lindén and Carlson, 2010) and the French (Sagot and Fišer, 2008) wordnets.

The Russian wordnet (Alexeyevsky and Temchenko, 2016) is an example of the *merge* approach. It is based on a monolingual dictionary and the word definitions in these. The idea is that definitions contain hypernyms of the defined words, often in the form of WORD:HYPERNYM . . . , and that this information can be used to set up hierarchical structures in the wordnet. Other examples of *merge* with partly different ideas are the Polish Wordnet (Derwojedowa et al., 2008), the Norwegian Wordnet (Fjeld and Nygaard, 2009), the Danish Wordnet (Pedersen et al., 2009), and the Turkish Wordnet (Bakay et al., 2021).

There were previous attempts to create Ukrainian wordnets that, however, did not result in the release of an open Ukrainian wordnet. In particular, (Kuljchycjkyj et al., 2010) state that their earlier attempts to apply an expansion method to Ukrainian failed. The authors claim that in the next attempt, having used frequency dictionaries, they created the fragment of a wordnet-like dictionary of the Ukrainian language, in which 194 noun synsets were implemented, being connected by hypo-/hyperonymy links (183 examples), antonymy (14 examples), as well as additionally meronymy/holonymy connections (over 150 cases). However, the project was not continued, and the results were not made publicly available.

(Anisimov et al., 2013) report the main results of a project aimed to create the Ukrainian lexical-semantic knowledge base UkrWordNet (UWN), describing tools and results. The authors claim that they automatically created more than 82,000 noun synsets and have about 145,000 nouns in the lexicon. However, this wordnet cannot be accessed.

Nykonenko et al. (2013) describe a correction tool designed to create and modify the Ukrainian linguistic ontology in the UWN. However, the site of the mentioned project UWN (`http://lingvoworks.org.ua/`) is not accessible any more.

Thus, we may conclude that despite some efforts and announced results, a Ukrainian wordnet as part of the OMW effort under an open source license is still not available and remains an open field for research.

3 Method and Material

For Ukrajinet, we decided to use the same approach as for the (Siegel and Bond, 2021; Bergh and Siegel, 2023) wordnet. So, the approach of the Ukrajinet initiative is *merge*. We use an existing synonym dictionary and several methods to link the synsets to OMW. The methods from the development of the (Siegel and Bond, 2021; Bergh and Siegel, 2023) wordnet are reused for Ukrajinet.

The first version of the open Ukrainian wordnet, Ukrajinet 1.0, was created on a basis of a dictionary of physical synonymous terms (Vakulenko and Vakulenko, 2017).

As in other languages, the establishment of an ontology for the Ukrainian lexical information necessitates proper accounting of ambiguities resulting from homonymy and polysemy. These lexical semantic relations prevalently occur within the same syntactic category (Part of Speech, POS) but can also arise across different POS, e.g.

мати ‘mother’ (noun) – мати ‘have’ (verb)

In most cases, such ambiguities are not parallel to English ones, which results in difficulties in translation and linking Ukrajinet to OMW. For example, the Ukrainian term вал has three main meanings corresponding to different English terms: 1. (tech.) ’shaft’; 2. ’barrage’; 3. (arch.) ’torus’. In addition, Ukrainian verbal nouns stemming from the same verb, bear subtle semantic differences that cannot be reflected in other languages (Vakulenko

and Vakulenko, 2017).

4 Process of Creating Ukrajinet

Basic information on the wordnet idea can be found in (Fellbaum, 1998) and (Kunze and Lemnitzer, 2010), among others. The data structure of wordnets in OMW is an XML structure (which can be converted to a JSON format). A lexeme has a "LexEntry" with a unique ID, information about the written form, syntactic category, and meanings, with links to associated concepts.

The dictionary of physical terms that we use as a basis for Ukrajinet was not created primarily for NLP purposes. It is in Microsoft Word © format and has entries such as[1]

будова

будова атомного ядра, структура атомного ядра /+/ збудова атомного ядра

Therefore, the first step was to convert the dictionary entries into a machine-readable format. Then, existing methods could be used to compile this information into the OMW XML format (section 4.1). Furthermore, the information is extended with POS (section 4.2) and multilingual indexing information (section 4.3).

4.1 From the Dictionary of Physical Synonym Terms to Synsets

We extracted only the synonym information from the dictionary and ignored (for the time being) other information, such as subdomains (optics, molecular physics, quantum mechanics, etc.). This information will be added in future work. The output of the preprocessing was a file of synsets, with each synset on one line. An example of such a synset is:

аглютинація;склеювання;грудкування (agglutination, adhesion, clumping)

The target of the transfer process of this synset is to have three lexical entries and a synset entry. The format is described in Bond et al. (2016). We start with the synset and its basic information[2]:

<Synset id="ukrajinet-30-n" ili="i36192" partOfSpeech="n">
<Definition> міцне з'єднання між собою (strong connection between each other) </Definition>
</Synset>

The synset has a unique synset ID, a link to the interlingual wordnet IDs in "ili", a POS, and a definition. Further, it has relations to other synsets that we ignore for the moment.

4.2 Adding POS Information

The next task is to find information about the syntactic category (Part-of-Speech, POS). One option for the part-of-speech tagging of Ukrainian words is to use a tool such as VESUM[3]. However, a noticeable part of our terms is not present in VESUM, such as the words "видим ('antinode'), вогкомір ('psychrometer'), замичник ('relay'), іскриш ('pyrites'), etc. This is due to the fact that we have many very specific terms in the field of physics. Given this, we used the following heuristic approach, which showed better results.

As the dictionary contains only verbs and nouns (with rare exceptions), we recognize verbs by their endings. If a word ends with one of the verbal endings, then it is a verb in the infinitive form (with rare exceptions for " ти "), otherwise a noun:

- ти
- тися
- тись

As with other wordnets, we have some cases of multiword expressions. An example is ставати більшим (to grow larger). We use the POS of the first word in the expression, as these are (in this dictionary) mostly consisting of verb + adjective (POS V) or noun + noun (POS N). We manually checked and corrected the cases where a synset contained words with different assigned POS's.

A further task is to generate the lexical entries for the words, sharing the synset sense. This is what is aimed for:

<LexicalEntry id="w76">
<Lemma writtenForm=" аглютинація "[4]
partOfSpeech="n"/>
<Sense id="w76_30-n" synset="ukrajinet-30-n"/>
</LexicalEntry>

[1] structure, structure of the atomic nucleus, construction of a nuclear core

[2] The English translation is not part of the synset; the translation is given here only for better understanding

[3] https://github.com/brown-uk/nlp_uk

[4] agglutination

```
<LexicalEntry id="w77">
<Lemma writtenForm=" склеювання "[5]
partOfSpeech="n"/>
<Sense id="w77_30-n" synset="ukrajinet-30-n"/>
</LexicalEntry>

<LexicalEntry id="w78">
<Lemma writtenForm=" грудкування "[6]
partOfSpeech="n"/>
<Sense id="w78_30-n" synset="ukrajinet-30-n"/>
</LexicalEntry>
```

The lexical entries in a synset belong to one sense with the same synset ID. Further senses for lexical entries come from other synsets in the dictionary. Each lexical entry has a unique word ID, a lemma, and a part of speech (POS).

A validation process is implemented to ensure correctness of the wordnet. It checks for XML correctness, duplicate lexical entries (that are only allowed for homonyms), consistency of POS in LexEntries, synsets, duplicate ilis, synsets without words, words without synsets, and others.

4.3 Linking the Synsets with the Open Multilingual Wordnet

In order to create a useful resource in the OMW context, it is necessary to link the Ukrainian synsets by adding an interlingual index in "ili". We used the translation table that we had created for another wordnet (Bergh and Siegel, 2023). It contains the words and definitions for each English synset in OEWN. The idea behind using the definitions with the words is that these provide some context for the translation, such that lexical ambiguity is reduced. For our example above, we get:

i36192 bonding: fastening firmly together

The obtained list was automatically translated into Ukrainian through the DeepL tool and post-processed by a linguist to render precise meaning. Then we searched for the Ukrainian terms in our dictionary. Hence, we found the rows in the following form:

i36192 bonding: fastening firmly together
аглютинація;склеювання;грудкування

[5]adhesion
[6]clumping

In the non-ambiguous cases in which an ILI could be assigned exactly to one synset, we were able to transfer these words and definitions directly to Ukrajinet. We used the words and definitions from WordNet corresponding to those of Ukrajinet where 571 synsets were connected. We have also adopted the Ukrainian translation of the definition in these cases.

The ambiguous cases, where either one ILI is assigned to more than one synsets or a synset got more than one ILI assigned, are currently checked manually.

4.4 Results

So far, we have the first version of Ukrajinet with 8,700 lexical entries organized in 3,360 synsets, all in the physical domain. 571 of these synsets are connected to OMW via the ILI. We use a validation script for Ukrajinet that is based on the OMW validation, before submitting the wordnet to Github. Ukrajinet is released via GitHub, under a (CC-BY-SA 4.0)[7] license at https://github.com/hdaSprachtechnologie/ukrajinet. This can then be loaded directly into the WN Python library (Goodman and Bond, 2021), which allows easy use: either on its own or linked to other wordnets through the Collaborative Interlingual Index (CILI).

5 Discussion and Future Plans

We presented in this paper the process of creating the first version of the Ukrainian wordnet, Ukrajinet 1.0, which synsets and lexical entries in the field of physics.

It was possible to reuse methods that were developed for the creation of the German Wordnet OdeNet (Siegel and Bond, 2021) and therefore prove that this is an efficient way to create a wordnet for a new language.

Ukrajinet 1.0 is a starting point for future elaboration of this resource.

We are currently checking ambiguous translations, such that most of the terms in Ukrajinet 1.0 can be linked to OMW. Wordnets contain relations between synsets, such as hypernym, meronym, or antonym relations. Some relations are available in the dictionary that we use as the basis for our wordnet. Others can be taken over from OEWN, in cases where we have the ILI connection. Defini-

[7]https://creativecommons.org/licenses/by-sa/4.0/

tions for the terms in the domain of physics will be taken from the "Explanatory dictionary in physics" (Vakulenko and Vakulenko, 2008).

We have so far ignored information in the physics dictionary that we plan to include in the future: information about hierarchical relations and information about subdomains of physics.

Once the information for the terms we now have in Ukrajinet 1.0 is complete, we can begin to *expand* the wordnet. Various sources of information come into question for this: We can use the existing translation table to add general terms translated from OEWN to Ukrajinet. This will be done following the method described by (Bergh and Siegel, 2023). The domain information can be used to fine-tune the synsets. Further, we can look at the Wiktionary database of Ukrainian lemmata. We can also include the information in an academic dictionary of Ukrainian words, such as (Burjachok et al., 2001).

It is also planned to provide a Latinized version of Ukrajinet, Romanized according to the Ukrainian national standard 9112:2021[8] that yields isomorphic transliteration of Ukrainian texts (Vakulenko, 2022).

We are currently developing a user interface for manual work on Ukrajinet - corrections, edits, and additions.

Ukrajinet will be used in various multilingual scenarios of NLP requiring Ukrainian semantic and lexical resources, such as multilingual information retrieval, text analysis and comparison, machine translation, etc.

Limitations

The work described is work in progress. The results are promising, but not yet complete.

Acknowledgements

The described work was supported by Darmstadt University of Applied Sciences and the Hessian Ministry of Science and Art.

References

Daniil Alexeyevsky and Anastasiya V. Temchenko. 2016. WSD in monolingual dictionaries for Russian WordNet. In *Proceedings of the Eighth Global WordNet Conference*, Bucharest, Romania.

Anatoly Anisimov, Oleksandr Marchenko, Andrey Nikonenko, Elena Porkhun, and Volodymyr Taranukha. 2013. Ukrainian wordnet: creation and filling. In *Flexible Query Answering Systems: 10th International Conference, FQAS 2013, Granada, Spain, September 18-20, 2013. Proceedings 10*, pages 649–660. Springer. In Ukrainian.

Özge Bakay, Özlem Ergelen, Elif Sarmış, Selin Yıldırım, Bilge Nas Arıcan, Atilla Kocabalcıoğlu, Merve Özçelik, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Begüm Avar, et al. 2021. Turkish wordnet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174.

Johann Bergh and Melanie Siegel. 2023. Connecting multilingual wordnets: Strategies for improving ILI classification in OdeNet. In *Proceedings of the Global Wordnet Conference*, Donostia, Spain.

Francis Bond, Luis Morgado Da Costa, and Tuan Anh Le. 2015. Imi—a multilingual semantic annotation environment. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 7–12.

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*, pages 1352–1362, Sofia.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.

Andrij Burjachok, Andrij Ghnatjuk, Serghij Gholovashchuk, Ghalyna Ghorjushyna, Nina Lozova, Natalija Meljnyk, Oljgha Nechytajlo, Lidija Rodnina, Valentyna Taranenko, and Oleksandr Frydrak. 2001. *Slovnyk synonimiv ukrajinsjkoji movy: V dvokh tomakh (A dictionary of Ukrainian synonyms: In two volumes)*. Naukova dumka, Kyjiv. In Ukrainian.

Magdalena Derwojedowa, Maciej Piasecki, Stanislaw Szpakowicz, Magdalena Zawislawska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of Polish WordNet. *Proceedings of GWC 2008*, pages 162–177.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ruth Vatvedt Fjeld and Lars Nygaard. 2009. Nornet-a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources-between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7, pages 13–16.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The Wn Python library for wordnets. In *11th International Global Wordnet Conference (GWC2021)*.

[8] http://online.budstandart.com/ru/catalog/doc-page.html?id_doc=95601

Birgit Hamp, Helmut Feldweg, et al. 1997. GermaNet-a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese wordnet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

I.M. Kuljchycjkyj, A.B. Romanjuk, and K.B. Khariv. 2010. Rozroblennja wordnetpodibnogho slovnyka ukrajinsjkoji movy (development of a wordnetlike dictionary for the ukrainian language). *Visnyk Nacionaljnogho universytetu Ljvivsjka politekhnika. Informacijni systemy ta merezhi*, 673:306–318. In Ukrainian.

Claudia Kunze and Lothar Lemnitzer. 2010. Lexical-semantic and conceptual relations in germanet. *Lexical-semantic relations: Theoretical and practical perspectives*, pages 163–183.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet – finnish wordnet by translation. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.

John Philip McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English wordnet 2019–an open-source wordnet for english. In *Proceedings of the 10th Global WordNet Conference*, pages 245–252.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English wordnet 2020: Improving and extending a wordnet for english using an open-source methodology. In *proceedings of the LREC 2020 workshop on multimodal WordNets (MMW2020)*, pages 14–19.

A.O. Nykonenko, E.V. Lyman, K.S.and Zabelin, and B.O. Rybachok. 2013. Uwn: Ontocorrector as a tool for ukrainian language linguistic ontology creation. *Shtuchnyj intelekt*, 4. In Ukrainian.

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*, 43:269–299.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *OntoLex*.

Melanie Siegel and Francis Bond. 2021. Compiling a German wordnet from other resources. In *11th International Global Wordnet Conference (GWC2021)*.

M. O Vakulenko and O. V Vakulenko. 2008. Tlumachnyj slovnyk iz fizyky [explanatory dictionary on physics]. *Kyjiv: VPC Kyjivsjkyj universytet*.

Maksym Vakulenko. 2022. Deep contextual disambiguation of homonyms and polysemants. *Digital Scholarship in the Humanities*.

Maksym O. Vakulenko and Olegh V. Vakulenko. 2017. *Tlumachnyj slovnyk iz fizyky: [6644 statti] (Explanatory dictionary on physics: [6644 articles])*. Naukova dumka, Kyjiv. In Ukrainian.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.