Open Multilingual WordNets - Ideas for Multilingually Linked Language Resources

I. Arzamastseva, M. Siegel

Ulyanovsk State Technical University, Ulyanovsk, Russia,
Darmstadt University of Applied Sciences, Germany
e-mail: lingua@ulstu.ru; melanie.siegel@h-da.de

Abstract. In the Open Multilingual WordNet (OMW) initiative, WordNets are built for many different languages and made available under an open source license. These resources share the XML format and its DTD. At the same time, the concepts of multilingual WordNets are linked via an interlingual ID, such that semantic concepts can be accessed in different languages. The paper reports on WordNet developments for the German and Russian languages. Our focus is on the automatic conversion of existing resources into the OMW WordNet format and the linking of concepts.

1 Introduction

WordNets are well-established lexical resources with a wide range of applications. For more than twenty years they have been elaborately set up and maintained manually, especially the original Princeton WordNet of English (PWN) [6]. In recent years, there have been increasing activities, in which open WordNets for different languages have been automatically extracted from other resources and enriched with lexical semantics information, building the so-called Open Multilingual WordNet [4]. These WordNets were linked to PWN via shared synset IDs [3], [5]. In this context a German lexical semantics resource with the name Open German WordNet (OdeNet) is being developed with the aim to be included as the first open German WordNet into the Open Multilingual WordNet. OdeNet is automatically created from different information sources.

There are already three such resources for the Russian language: RuWordNet, WordNet.ru and RussNet; but none of them is complete. At present, they are not part of the OMW environment.

We analyse the process of production and structure of OdeNet and the Russian WordNets (RuWordNet, WordNet.ru and RussNet). We are looking for ideas on how to create a Russian resource in the OMW context, how to use the existing Russian resources, and how to extend OdeNet.

In the following we first give an overview of the state of the art of research on the development of WordNets. Afterwards we present the structure and the construction process of OdeNet. Then we describe the WordNets for Russian, RuWordNet, WordNet.ru and RussNet, and analyse which steps would be necessary to integrate

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

them into the OMW context. We also examine whether processes from OdeNet development are transferable to the Russian WordNet development.

2 State of the Art

In the Open Multilingual WordNet initiative [4], [3], several WordNets for multiple languages were developed and linked.

A manually well-designed WordNet resource for German is Germanet [7]. Germanet was developed over 20 years now and is very stable and precise. The problem is that it is not under an open-source license and is therefore not broadly used in language technology applications. Further, the restricted license makes it unable to include Germanet in the Open Multilingual WordNet initiative. This is the reason why we decided to build up a new resource.

Vossen [8] describes two basic approaches to develop new WordNet resources: In the first case (called “expand”), existing PWN synsets are taken and lexical entries added for the specific language. In the second case (called “merge”), language-specific resources are built and then linked to the PWN. The approach of the OdeNet initiative is “merge”. We use an existing synonym dictionary and try to link the synsets to PWN.

There are three main resources for the Russian language: RuWordNet [14], WordNet.ru [15] and RussNet [16].

WordNet.ru [15] is an automatically generated database that has been translated only half into Russian. There are no Russian definitions of synsets, all remain in English.

RuWordNet [14] is a thesaurus that was created based on the automated transformation of the RuThes Thesaurus into WordNet format. The relationships between synsets in this resource are the same as in WordNet. Between synsets relating to different parts of speech, but expressing the same meaning, there are established relations of part-to-synonym synonymy, connecting the separated synsets.

The RuWordNet search interface is located at ruwordnet.ru. For non-commercial use you can get xml files with RuWordNet thesaurus data.

The RussNet [16] is similar to Princeton WordNet and EuroWordNet. The thesaurus consists of four interconnected files containing words of the main parts of speech: nouns, verbs, adjectives and adverbs. The basic unit of RussNet is a synsets, combining words with similar meanings. Synsets are connected by various paradigmatic and syntagmatic relationships.

None of these resources is complete or fully translated. Moreover, only RuWordNet is currently being developed, and is the only Russian WordNet that can be used for non-commercial purposes.

Therefore, there is a need to create such a resource that would be compatible with open resources for German and English.

3 WordNet for German Language: Developing OdeNet

The Open-de-WordNet (OdeNet) initiative is aiming at including a crowd-sourced German resource in a multilingual WordNet context, where the concepts (the synsets) of the languages are linked, and where the resources are under an open-source license, being eventually included in the NLTK language processing package (Bird 2009), such that the resource can be further developed by researchers while using the resource for their Natural Language Processing projects. OdeNet is combining existing resources: The OpenThesaurus German synonym lexicon [9] and the Open Multilingual WordNet English resource [10], with the aim to link the resulting WordNet-compliant German resource to the Princeton WordNet of English (PWN) [6]. The OpenThesaurus is a great chance of using a large resource, generated and updated by the crowd. The PWN resource is a well-developed resource for English concepts. It includes many relations between the concepts and is linked to resources for multiple languages.

The first version of OdeNet was created as an experimental project at Darmstadt University of Applied Sciences. It was completely automatically created. In the following, manual corrections were made in different topic domains. German definitions were introduced, relations were corrected and supplemented and ili links (links to the multilingual concepts) were added. Further, we worked on the syntactic categories. The main focus was on correcting the POS tags of the multiword lexemes. The next step was the annotation of the most common German words, as listed in http://pcai056.informatik.uni-leipzig.de/downloads/etc/legacy/Papers/top1000de.txt. We annotated all lexical entries (except for function words) of this list with *dc:type="basic_German"* in OdeNet, added missing entries and corrected synsets manually. Then, we implemented an analysis of German nominal compounds and used this information for the addition of hypernym relations.

The OpenThesaurus data can be downloaded as txt from https://www.openthesaurus.de/about/download. The text file contains one synset per line, such that the lexical items in each synset are divided by semicolons, e.g.:

Mobilität;Unabhängigkeit;Beweglichkeit

We compiled it to a synset and lexical entries:

```
<Synset id="de-9784-n" ili="i62097" partOfSpeech="n"
      dc:description="the quality of moving freely">
      <SynsetRelation targets='odenet-23172-n' relType='hypernym'/>
</Synset>
<LexicalEntry id="w39185">
      <Lemma writtenForm="Mobilität" partOfSpeech="n"/>
      <Sense id="w39185_9784-n" synset="odenet-9784-n"></Sense>
</LexicalEntry>
<LexicalEntry id="w33556">
      <Lemma writtenForm="Beweglichkeit" partOfSpeech="n"/>
      <Sense id="w33556_8203-n" synset="odenet-8203-n"></Sense>
      <Sense id="w33556_9784-n" synset="odenet-9784-n"></Sense>
      <Sense id="w33556_11420-n" synset="odenet-11420-n"></Sense>
      <Sense id="w33556_19087-n" synset="odenet-19087-n"></Sense>
```

```
</LexicalEntry>
<LexicalEntry id="w35624">
    <Lemma writtenForm="Unabhängigkeit" partOfSpeech="n"/>
    <Sense id="w35624_8795-n" synset="odenet-8795-n"></Sense>
    <Sense id="w35624_9784-n" synset="odenet-9784-n"></Sense>
    <Sense id="w35624_28976-n" synset="odenet-28976-n"></Sense>
</LexicalEntry>
```

The first task was to find POS information. We used the Python library TextBlob for POS annotation [11]. The second task was to find an English synset that can be linked. We translated the words in the synset to English using google-translate [12]. Using a statistical machine translation system instead of a dictionary has the advantage that the translation is based on the context. In case of ambiguous words, the decision is context-based, with the context being the other words in the synset. We could link 19,845 German synsets to synsets in the PWN, about 55 % of the German synsets.

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

Hyponymy relations were added using a German compound analyser. In German regular compounds, the leftmost part – the head – links to its hyperonym, such as "Apfelsaft" (apple juice) has the hypernym "Saft" (juice).

4 WordNets for Russian Language

For the Russian language, we will analyse the vocabulary articles and databases of the above three resources. This will serve as the basis for creating our own WordNet type project for the Russian language.

The RussNet word base today consists only of verbs. In addition to the database, there are synsets for these verbs, and their descriptions. From the RussNet resource database, we select verbs and write dictionary entries for these, synsets are indicated, and relations between them are added.

From the WordNet.ru resource database, we select nouns, adjectives and adverbs and also write dictionary entries, synsets are indicated and relations between them added.

In order to determine which model of construction of dictionary entries to adhere to, and in order to bring this model as close as possible to the OMW, it is necessary to compare the most popular resources not only in terms of construction of a dictionary entry, but also in terms of the relationship between synsets, as well as the format.

But first we need to determine, which of the resources of the Russian language is as close as possible to the original WordNet and to OdeNet.

On the RussNet resource, we cannot search for words, because only one database is published, which, in addition to being exemplary, is read only in one program. Of course, the whole methodology and structure of such a resource is described, but using RussNet as a thesaurus is not possible.

RuWordNet, in our opinion, is closest to the format of OMW, because it is based on a thesaurus translated into the OMW format.

In this resource, one can search for words using the search form. Also, for a given word all existing meanings, synsets and relationships are displayed. At the moment, the resource RuWordNet is the most complete and newest. This project is completely Russian; definitions are given in Russian

To make our resource as useful and multilingually linked as possible, we need to translate words and their dictionary entries into English and German, taking into account all the features of these languages. To do this manually will take a lot of time. Therefore, in order to save time and resources, we tried to use machine translation for this task (Google translate from https://translate.google.com, or Yandex services акщь https://translate.yandex.ru/).

After analyzing the machine translation results of some dictionary entries, it can be concluded that they are of poor quality. Unfortunately, even the most modern technologies are still not able to convey all the semantic, stylistic and grammatical features of different languages. Below are the tables with the results of machine translation from Russian-speaking resources into English and German, and vice versa.

Table 1. Google machine translation of Russian language entries

Original entry	**English translation (Google translate)**	**German translation (Google translate)**
Любовь существительное Найдено несколько значений: Любовь 1 (любовь, сердечное влечение) Любовь 2 (склонность (постоянное влечение)) Любовь 3 (любовь (привязанность))	**Love** noun Found several values: Love 1 (love, heart attraction) Love 2 (addiction (constant attraction)) Love 3 (love (affection))	**Liebe** Nomen Mehrere Bedeutungen gefunden: Liebe 1 (Liebe, Herzattraktion) Liebe 2 (Sucht (ständige Attraktion)) Liebe 3 (Liebe (Zuneigung))
Любовь 1 любовь, сердечное влечение *Синсет* ЛЮБОВНОЕ ВЛЕЧЕНИЕ, ЛЮБОВНЫЙ ПОЖАР, **ЛЮБОВЬ**, ПОЖАР ЛЮБВИ, СЕРДЕЧНОЕ ВЛЕЧЕНИЕ *однокоренные слова*	**Love 1** love heart attraction *Synset* LOVE TREATMENT, LOVE FIRE, **LOVE** FIRE LOVE, HEART IMPACT *common words* LOVE	**Liebe 1 Liebesherzanziehungskraft** *Synset* Liebesbehandlung, Liebesfeuer, **LIEBE** FEUERLIEBE, HERZ-VERBESSERUNG *verwandte Wörter* LIEBE

ЛЮБОВНЫЙ		
гипероним ВНУТРЕННЕЕ ОЩУЩЕНИЕ, ДВИЖЕНИЕ ДУШИ, ДУШЕВНОЕ ПЕРЕЖИВАНИЕ, ОЩУЩЕНИЕ, ПЕРЕЖИВАНИЕ, ЧУВСТВО, ЭМОЦИЯ ЛЮБОВЬ	*hyperonym* INTERNAL FEEL, MOVEMENT OF THE SOUL, Soul experience, FEEL, EXPERIENCE, FEELING, EMOTION LOVE	*Hyperonym* Inneres Gefühl, BEWEGUNG DER SEELE Seelenerfahrung, FÜHLEN, LEBEN, Gefühl, EMOTION Liebe
домен ВНУТРЕННЕЕ ОЩУЩЕНИЕ, ДВИЖЕНИЕ ДУШИ, ДУШЕВНОЕ ПЕРЕЖИВАНИЕ, ОЩУЩЕНИЕ, ПЕРЕЖИВАНИЕ, ЧУВСТВО, ЭМОЦИЯ	*domain name* INTERNAL FEEL, MOVEMENT OF THE SOUL, Soul experience, FEEL, EXPERIENCE, FEELING, EMOTION	*Domänenname* Inneres Gefühl, BEWEGUNG DER SEELE Seelenerfahrung, FÜHLEN, LEBEN, Gefühl, EMOTION
гипоним ВЗАИМНОСТЬ, ВЗАИМНОСТЬ В ЛЮБВИ, ВЗАИМНОСТЬ ЧУВСТВА, ЛЮБОВНАЯ ВЗАИМНОСТЬ, ОТВЕТНАЯ ЛЮБОВЬ	*Hyponym* RECIPROCITY, MUTUAL IN LOVE, Reciprocity of feeling LOVE RELATIONSHIP, RESPONSE LOVE	*Hyponym* Gegenseitigkeit MUTUAL IN LIEBE, Reziprozität des Gefühls Liebesbeziehung, ANTWORT LIEBE
частеречная синонимия ЛЮБОВНЫЙ	*partial synonymy* LOVE	*Redeteilsynonymie* LIEBE
гипоним ЛЮБОВНОЕ ВЛЕЧЕНИЕ, ЛЮБОВНЫЙ ПОЖАР, ЛЮБОВЬ, ПОЖАР ЛЮБВИ, СЕРДЕЧНОЕ ВЛЕЧЕНИЕ	*hyponym* LOVE TREATMENT, LOVE FIRE, LOVE, FIRE LOVE, HEART IMPACT	*Hyponym* LIEBESBEHANDLUNG, LIEBE FEUER, LIEBE FEUER LIEBE, HERZVERBESSERUNG

There are too many errors in this translation. For example, "сердечное влечение" (heart attraction) has been translated into English as “heart impact” and into German as “Herzverbesserung” (heart improvement). But such a machine translation can be used at the first stage in order to facilitate the task. Then it is necessary that all examples be edited by a translation specialist.

Our dictionary entries will be based on the resource RuWordNet. Further, some properties and relationships from the RussNet resource will be added. Our dictionary entries for the Russian language will be compiled according to the format of the Open Multilingual WordNet (OMW) resources. This will provide an opportunity to use our resource with other foreign language resources in OMW.

To describe a word in a dictionary entry, information from the RuWordNet resource is used, such as meaning, synset, single-rooted words, hyperonym, domain, hyponym, partial synonymy. Additionally, we use information from the RussNet resource: antonyms / conversion, meronymy, complex relationships, sub-existence, causality, ingenuity and terminality.

For a more detailed description, examples for the use of a specific word will be added from the http://ruscorpora.ru resource, paradigmatic features that show all forms of the word and the areas of use of the given word.

This is an example of a dictionary entry for the abstract noun “love” with all additions:

ЛЮБОВЬ / LOVE

Значения / Values

- Любовь 1 (любовь, сердечное влечение) / Love 1 (love, heart attraction)
- Любовь 2 (склонность, постоянное влечение) / Love 2 (addiction, constant attraction)
- Любовь 3 (привязанность) / Love 3 (affection)

Любовь 1 / Love 1

Синсет / Sinset

- Любовное влечение
- Любовный пожар
- Любовь
- Сердечное влечение

Однокоренные слова / Root words

- Любовный

Гипероним / Hyperonym

- Внутреннее ощущение
- Движение души
- Душевное переживание
- Ощущение
- Переживание
- Чувство
- Эмоция

Домен / Domain

- Внутреннее ощущение
- Движение души
- Душевное переживание

- Ощущение
- Переживание
- Чувство
- Эмоция

Гипоним / Hyponym

- Взаимность
- Взаимность в любви
- Взаимность чувства
- Любовная взаимность
- Ответная любовь

Частеречная синонимия / Synonymy in parts of speech

- Любовный

Антонимия/конверсия / Antonymy/Conversion

- Ненависть

Меронимия / Meronymy

- MEMBER_OF (чувство - любовь)
- PART_OF (любовь - чувство)

Подсобытие / Sub event

- Увидеть
- Встретить

Причинность / Causality

- Жениться

Начинательность / Initiative

- Полюбить
- Возлюбить

Терминативность / Terminativity

- Разлюбить
- Охладеть

Сложные отношения и зависимости / Complex relationships and dependencies

- INVOLVED_AGENT Человек, мужчина, женщина, лицо
- INVOLVED_OBJECT Человек, мужчина, женщина, лицо

5 Conclusion

In this paper, we first described the process of compiling and extending a German WordNet (OdeNet) from an existing synonym dictionary, in a way that it can be used in a database of multilingual WordNets, OMW. Then, we explored methods and developed ideas to do the same thing for Russian. The next step will be a compliation of a Russian WordNet from the described resources, and extension with information.

References:

[1] Bird S., Klein E., Loper E. 2009. Natural language processing with Python. – 504pp.

[2] Bond, Francis, 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.

[3] Bond, Francis 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.

[4] Bond, Francis and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In Proceedings of the Global WordNet Conference.

[5] Bond, Francis, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In Proceedings of the Global WordNet Conference, vol. 2016.

[6] Fellbaum, Christiane, ed. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[7] Hamp, Birgit, 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.

[8] Vossen, Piek, ed. 1998. Euro WordNet. Kluwer

[9] https://www.openthesaurus.de/

[10] http://compling.hss.ntu.edu.sg/omw/

[11] https://textblob.readthedocs.io/en/dev/

[12] https://translate.google.de/

[13] http://ruscorpora.ru

[14] https://ruwordnet.ru/ru

[15] http://wordnet.ru/

[16] http://project.phil.spbu.ru/RussNet/index_ru.shtml