OdeNet: Compiling a German Wordnet from other Resources

Melanie Siegel
Darmstadt University
of Applied Sciences
melanie.siegel@h-da.de

Francis Bond
School of Humanities
Nanyang Technological University
bond@ieee.org

Abstract

The Princeton WordNet for the English language has been used worldwide in NLP projects for many years. With the OMW initiative, wordnets for different languages of the world are being linked via identifiers. The parallel development and linking allows new multilingual application perspectives. The development of a wordnet for the German language is also in this context. To save development time, existing resources were combined and recompiled. The result was then evaluated and improved. In a relatively short time a resource was created that can be used in projects and continuously improved and extended.

1 Introduction

The goal of this initiative is to have a German resource in a multilingual wordnet initiative, where the concepts (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 et al., 2009)) and/or the spaCy package.[1]

Wordnet resources are largely used in NLP projects all over the world. Our idea is to create a German resource that starts from a crowd-developed thesaurus; is open; and included in the NLTK package. Then it can be further developed by researchers while using the resource for their NLP projects.

For the first version, we combined existing resources: The OpenThesaurus German synonym lexicon,[2] the Open Multilingual Wordnet[3] (OMW: Bond and Foster, 2013) the English resource, the Princeton WordNet of English (PWN: Fellbaum, 1998)). The OMW data (Bond and Foster, 2013) was made by matching multiple linked wordnets to Wiktionary (Wikimedia, 2013) and the Unicode Common Locale Data Repository (Unicode, 2012). The OpenThesaurus is 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 synsets from the OMW data have an estimated accuracy of 90%. We call our new resource "OdeNet", from "Offenes deutsches Wordnet - open German wordnet". The first version of OdeNet was automatically compiled. We also describe the efforts to extend and correct the entries.

2 Related Work

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

A manually well-designed wordnet resource for German is GermaNet (Hamp and Feldweg, 1997). 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 impossible to include GermaNet in the Open Multilingual Wordnet initiative. This is the reason why we decided to build up a new resource. In order not to violate the license terms, we do not use anything from GermaNet in OdeNet.

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

[1] https://spacy.io/
[2] https://www.openthesaurus.de/
[3] http://compling.hss.ntu.edu.sg/omw/

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 fully automatically built: 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.

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.

The approach of the OdeNet initiative is merge. We use an existing synonym dictionary and try to link the synsets to PWN.

Braslavski et al. (2016) describe the creation of a large thesaurus for Russian by means of crowd sourcing. The data is directly collected in a wordnet style, but synsets are not linked to the OMW. The basic data for OdeNet is also generated in a crowd sourcing style, in the OpenThesaurus project. The OpenThesaurus project (Naber, 2004) is a crowd initiative to set up a German synonym lexicon. The version we downloaded in April 2017 has about 120,000 lexical entries in about 36,000 synsets.

2.1 German

The establishment of an ontology for the lexical information of a language requires an in-depth study of ambiguities and multi-word lexemes. In German, compounds are also an issue. There are many examples of lexical ambiguities in German, such as *Mutter* "mother, nut" or *umfahren* "bypass, to knock over". These are in many cases not parallel to English ambiguities, which makes the translation more difficult (for the purpose of linking in OMW). In most cases, ambiguities remain within a syntactic category (POS). The capitalization of German nouns prevents ambiguities between nouns and other syntactic categories, as is often the case in English (e.g. *change* "money" or "transform"). Morpho-syntactic ambiguities, which occur frequently in German, are not relevant for OdeNet because only lemmata are included. There are some words that can be used both as verbs and adjectives, such as *verlegen* "to place, to relocate, to publish - embarrassed". Other POS ambiguities are not relevant for this work because they refer to finer POS distributions than we can provide at the moment (particles - prepositions, demonstrative pronouns - articles).

In the area of multi-word lexemes we are concerned with support verb constructions, such as *Abschied nehmen* "to say goodbye" or *in Rechnung stellen* "to invoice". In addition, there are idioms such as *das geht auf keine Kuhhaut* "it beggars description". Especially for idioms it is difficult to automatically determine the syntactic category.

However, complex nouns are not realized - as in English - by means of multi-word expressions, but with compounds. Nominal compounds are very productive in German. They can be very long, like the well-known example *Donaudampfschifffahrtskapitänsmütze* "Danube steamship captain's cap". They can constantly be newly created. Automatic extraction and analysis from text data is complex because there are ambiguities here too.

In the case of regular German compounds, there is a hyponymy relationship between the head and the compound. For example, *Wassereis* is an ice that consists of water, while *Eiswasser* is water that is ice-cold. Different relations can exist to the modifier. The regularity of the hyponymy relationship to the head of German compounds is used to add relations to OdeNet.

3 Process of Creating OdeNet

The first version of OdeNet was completely automatically created by compilation from OpenThesaurus. In the following, manual corrections were made in the domains of project management and business reports. German definitions were introduced, relations were corrected and supplemented and CILI links (links to the multilingual concepts in OMW) were added. Then we worked on the syntactic categories. The main focus was on correcting the POS tags of multi-word lexemes.The next step was the annotation of basic 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"
```

We then 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.

3.1 Linking OpenThesaurus Synsets with the Multilingual Wordnet

The OpenThesaurus data can be downloaded as txt. 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
```

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:

```
<Synset id="de-9784-n"
  ili="i62097"
  partOfSpeech="n"
  dc:description="the quality of moving
                              freely">
  <SynsetRelation
                targets='odenet-23172-n'
                relType='hypernym'/>
</Synset>
```

The synset has a unique synset ID, a link to the international wordnet IDs in "ili", a POS, a definition, and relations to other synsets.

The first task is to find POS information. POS information is not included in the OpenThesaurus download data. We use the Python library TextBlob for POS annotation.[4] OdeNet just uses "n", "v" and "a" as POS tags, such that we map the Penn Treebank POS tags that TextBlob gives to these. In the case of multi-word expressions, such as *moralische Werte* "ethical values", we take the POS value of the last word in the expression, which is the head word in most cases.

The second task is to find an English synset that can be linked. We translate the words in the synset to English using google-translate.[5] 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. Using the NLTK wordnet API, we then search for synsets with these English words in the PWN and access their synset ID.

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

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

```
(id="de-39-n",pwn="in-05890249-n"),
```

We could link 19,845 German synsets to synsets in the PWN, about 55 % of the German synsets. Synsets that could not be linked were often multi-word expressions and metaphorical, such as: *es kann Gott weiß was passieren; für nichts garantieren können; mit allem rechnen müssen* "God knows what can happen; can't guarantee anything; have to count on everything". The link gives direct access to the definition in PWN, such that we could copy these into OdeNet in `dc:description`. Thus, we have an English definition as long as German definitions are still missing. The synset relations in PWN link to English synsets. We searched for German synsets with the `ili` that links to the target of a relation in the PWN and added these as targets.

3.2 Lexical Entries and Senses

These are the lexical entries for the words in the synset above:

```
<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 id="w33556_9784-n"
              synset="odenet-9784-n"/>
        <Sense id="w33556_11420-n"
            synset="odenet-11420-n"/>
         <Sense id="w33556_19087-n"
             synset="odenet-19087-n"/>
</LexicalEntry>

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

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).

When a synset gets its ID and link to PWN, all words in the synset are added to a tuple with this ID, as for example:

```
("Beweglichkeit",
"odenet-8203-n", "in-05003850-n"),
("Beweglichkeit",
"odenet-9784-n", "in-04773351-n"),
("Beweglichkeit",
"odenet-11420-n", "in-04875728-n"),
("Backlogged",
"odenet-19087-n", "in-05003850-n"),
```

The sense relations (antonym and pertainym) again are taken from the PWN and linked back to German.

4 Corrections and Extensions

4.1 POS Corrections

In a first evaluation, we found that POS information in OdeNet was only correct in 77% of the cases. With many multi word expressions in OdeNet, standard procedures to POS assignment do not seem to be sufficient. The basic idea for corrections was that a synset should in principle contain only lexical items of the same syntactic category. Therefore, we extracted all synsets containing lexical items with different POS information and manually corrected them. The evaluation showed an increase of correct POS to 90%. The next idea was to look at endings of lexemes. In German, words ending in *-ung*, *-heit*, and *-keit* are always nouns, while words ending in *-lich* are adjectives. Further, nouns are capitalized. We used this information to automatically correct further POS assignments. The evaluation showed an increase of correct POS to 93.3%.

4.2 Using German Compounds for Hyponymy Relations

Regular German nominal compounds have a hyponymy relation to their head, as explained above. A large part of the German compounds are regular and many synsets contain compounds. We decided to make use of these facts in order to add relations to OdeNet.

The idea is to use the regularity of German compounds to automatically generate hypernym relations for OdeNet. For this purpose, we have implemented a compound analysis tool that recognizes the head of the compound. Using this tool, we then analyzed all lexical items that are not multi-word expressions in OdeNet and extracted compounds and their heads.

Basis for the compound analysis is a list of nouns extracted from the TIGER tree bank (Brants et al., 2004). If the word to analyze consists of less than three letters, it is not a compound. If there are hyphens in the word (such as *Lehr-Lern-Forschung*, teaching-learning-research), the compound is split at these.

Using the pyphen module,[6] we split the compound into syllables. If the word to analyze consists only of one syllable (as in the case of *Stuhl* "chair"), it is not a compound. If the word consists of two noun components with one syllable each, as in the case of *Haustür* "front door", then both components are searched for in the TIGER lexicon. If they exist as entries, then the result of the analysis is a list with both components, such as *([Haus],[Tür])*. If the two syllables do not exist as words, then an attempt is made to delete a linking element from the first syllable and then look it up again. This is e.g. the case with *Wirtshaus* "pub", consisting of *Wirt* + *s* + *Haus*. If there are more than two syllables, different combinations of syllables are tested, as in the case of *Herstellungskosten* "production costs", until it can be split into parts that can be found in the noun list:

```
("Herstellungskosten")
SYLLABLES:
 ['Her', 'stel', 'lungs', 'kos', 'ten']
SYLLABLE COMBINATIONS:
  ['Herstel', 'Stellungs', 'Lungskos',
  'Kosten', 'Herstellungs',
  'Stellungskos', 'Lungskosten',
  'Herstellungskos', 'Stellungskosten']
COMPONENTS:
 ['Herstellungs', 'Kosten']
```

If the analysis with syllables does not lead to a result, we look up all combinations of n-grams in the word, considering fugen elements.

We ran our compound analyzer on all lexical entries that are not multiword entries and could identify 3,630 compounds. In case that the head has a singular sense in OdeNet, we added a hypernym relation to that synset and a hyponym relation backwards. Using synsets instead of lexical entries results in relations not only between single words, but also between groups of words. For example, because of the analysis of the word *Butterbrot* "sandwich" as consisting of *Butter* "butter" and *Brot* "bread", we added a hyponym relation between the synsets 11770-n ['Brotlaib', 'Wecken', 'Brot'] and 10073-

[6] https://pyphen.org/

n ['Knifte', 'belegtes Brot', 'Scheibe', 'Butterbrot', 'Schnitte', 'Bemme', 'Stulle'].

There are some exceptions to the hyponymy relation of compound and compound head. In some cases, the compound is synonym to its head, as in the case of *Fachterminus* "technical term" and *Terminus* "term". In these cases, both appear in the same synset and could therefore be automatically excluded.

More complicated are negations in compounds. A *Nichtraucher* "non-smoker" is not hyponym to *Raucher* "smoker", but antonym. On the other hand, *Nichteisenmetall* "non-ferrous metal" is a kind of *Metall* "metal". Thus, we manually checked all compounds with negations. Another problem are expressions with *Pseudo* "pseudo" or *Schein* "phantom". Is a pseudo-documentation a documentation? Is a *Scheinschwangerschaft* "phantom pregnancy" a pregnancy? We decided to not treat these as hyponyms. The compound analysis found 19,115 nominal compounds in OdeNet. In 12,132 cases, the found head was ambiguous between multiple senses and did not get a relation entry. In 1,810 cases, there was no entry for the head in OdeNet, such that these were also ignored.

For all hypernym relations that we added, we added the backward hyponym relation as well. 10,346 relations were added to the OdeNet synsets by this method. OdeNet contains around 35,000 synsets, such that we could add information for 29% of all synsets.

For the evaluation we randomly extracted 100 compounds from OdeNet. The compound analysis found 83 of these. Only one of the 83 analyzed compounds got a wrong analysis: *Blockdiagramm* (block diagram) was analyzed as ['Block', 'Dia', 'Gramm'] (block - slide - grams). This analysis is syntactically fine, but semantically nonsense. Thus, the precision of the compound analysis is very high (0.99), while the recall is moderate (0.83). For our purpose, extending OdeNet, precision is highly important, while a moderate recall is fine.

The 100 entries had 41 hypernym relation entries that originated from compound analyses. One of the relation entries was wrong: in the case of *Fleischsaft* "meat juice", the compound analysis was correct (['Fleisch', 'Saft']), but the hypernym relation led to the synset ['Strom', 'Saft', 'Elektrizität'] (electricity). The German word *Saft* is ambiguous between *juice* and *electricity*, but had only the electricity entry in OdeNet, which is wrong. If there was more than one sense for a word, there was no hypernym relation added to avoid such errors.

Synset relation	Number
hypernym relations	9,907
hyponym relations	10,101
member holonyms	84
part holonyms	647
member meronyms	74
part meronyms	282

Table 1: Number of synset relations

Therefore, for 100 synsets that had compounds, we could add 40 good hypernym relations by this method, and one wrong relation, which is a precision of 98%.

5 Current State of OdeNet

The resulting wordnet resource (v1.3) contains about 120,000 lexical entries in about 36,000 synsets. About 20,000 of these synsets are linked to synsets in the English PWN and then to the multilingual CILI numbers. There are 2,664 antonym relations and 1,053 pertainym relations linking lexical entities. The number of synset relations can be seen in table 1.

For evaluation of preciseness, we randomly chose 90 lexical entries, 30 with POS "n", "v" and "a" respectively, and evaluated them manually, see Table 2.

The **POS information** was correct in 93.3% of the cases. In 5 cases of 6 wrong POS assignments, the lemma was a multi-word lexeme, such as *nicht unumstößlich* "not unalterable". POS tagging of multi-word lexemes needs more sophisticated procedures than the ones we used here, as standard POS taggers do not tag multi-word expressions. A good part of this problem could be solved with POS corrections in synsets that had lexical items with different POS. The linked English synsets could also give a hint that there might be a problem, as they have POS assigned, which often would be the same for German. A further attempt to improve OdeNet could therefore be to search for cases where the synsets are linked and the POS tags of the English and German synsets do not match.

The German synsets that are linked to English ones, contain the **definitions** from the correspond-

Tested	Correct	Comment
POS	93%	many multi-word lexemes
DEFINITIONS	82%	in cases of errors, POS of the English words are often different
RELATIONS	61%	in cases of errors, definitions are also wrong

Table 2: Precision of 90 randomly chosen lexical entries

ing English synsets. We checked if the definitions are correct (and therefore the synsets are correctly linked). 55 of the 90 cases had a link to an English synset, and therefore a definition. In 45 of the 55 cases (82%), these definitions were correct.

There were 41 cases, where **relations** on the lexical or the synset level were assigned (34%). 12 of these cases had wrongly assigned relations (39%). In 5 of these cases, the link to PWN was also wrong, and the relation was taken over from the English synset. In one case, the relation from the English synset was wrong, while the relation that was automatically added by the compound analysis was correct. The next correction step will have to address the linking.

We have annotated the entries with a default confidence of **0.6**, with entries that have been manually validated given a confidence of **1.0** and those from the extended OMW a confidence of **0.85**.

Release

The wordnet is released through GitHub, as a compressed tar file containing the wordnet itself, its license (CC-BY-SA 4.0)[7] and canonical citation.[8] This can 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 CILI.

6 Discussion and Future Plans

It has been possible to set up a wordnet for the German language in a couple of years. We have benefited both from OpenThesaurus and the knowledge in the OMW. In this way, we were able to build a very large resource, with the synsets being created manually in the OpenThesaurus project, and therefore very precise. We have used NLP techniques to add more information, namely POS and the relations to OMW and CILI. We have used the knowledge in the OMW to supplement relations between the German synsets - parallel to the relations in the other wordnets.

The Open Multilingual Wordnet initiative is a great chance to get highly linked and standardized language resources for multiple languages. The standardization makes it possible to include these resources in NLP packages, such as NLTK or spaCy.

We have shown that it is possible using NLP techniques to combine language resources such as the OpenThesaurus and the English PWN to gain a new resource in this standardized multilingual context, with a reasonable precision.

The next step will be to further work on the quality of OdeNet. We have already started to implement methods that allow the semi-manual correction and extension:

- A tool for adding more hyponym relations in case of compounds that shows the user different synsets for a compound head and asks which one to set the relation to. It then adds the relation to OdeNet automatically.
- A tool that shows the user all information for a word and gives her multiple possibilities to correct and extend it.
- A tool that allows to search for a word in PWN and give the corresponding CILI(s) and allows the user to add the CILI to OdeNet.

Further, it will again be compared to the English PWN, such that cases where linked synsets differ in their POS assignment will be further investigated. Another source of problems is multiword lexemes, where we will have to search for better POS tagging methods.

We started to work on the basic German words, adding and correcting information. This will be a valuable information source for simplified language projects.

Through the Wn library (Goodman and Bond, 2021) the resource will be available to NLTK, such that it can be used in NLP projects. The open-source idea will help to let researchers working

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

[8] https://github.com/hdaSprachtechnologie/odenet/releases/tag/v1.3

on German language further improve and expand OdeNet. We ourselves plan to use it in information extraction in the business domain and sentiment analysis projects. By this approach, we will add synsets from the business domain and sentiment polarity for many words.

We will add a user interface to make crowd development possible, in order to extend and correct OdeNet.

We would also like to tag some German texts.

The resource is available on GitHub under an open-source license: https://github.com/hdaSprachtechnologie/odenet.

Acknowledgments

We would like to thank Michael Wayne Goodman for his help in preparing the GitHub release.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., Beijing.

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. URL http://aclweb.org/anthology/P13-1133.

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.

S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on language and computation*, 2(4):597–620.

P. Braslavski, D. Ustalov, M.l Mukhin, and Y. Kiselev. 2016. Yarn: Spinning-in-progress. In *Proceedings of the 8th Global WordNet Conference, GWC 2016*, pages 58–65.

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

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

B. Hamp and H. Feldweg. 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*. Madrid.

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.

Daniel Naber. 2004. Openthesaurus: Building a thesaurus with a web community. *Retrieved January*, 3:2005. URL https://www.openthesaurus.de/download/openthesaurus.pdf.

Inc. Unicode. 2012. Unicode, inc. license agreement - data files and software. URL http://www.unicode.org/copyright.html.

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

Wikimedia. 2013. List of wiktionaries. (accessed on 2013-09-12).