Adapting GermaNet for the Semantic Web using OntoLex-Lemon

Claus Zinn
Department of Linguistics
University of Tuebingen
Germany
claus.zinn@uni-tuebingen.de

Marie Hinrichs
Department of Linguistics
University of Tuebingen
Germany
marie.hinrichs@uni-tuebingen.de

Erhard Hinrichs
Department of Linguistics
University of Tuebingen
Germany
erhard.hinrichs@uni-tuebingen.de

Abstract

GermaNet is a large lexical-semantic net that relates German nouns, verbs, and adjectives semantically. The word net has been manually constructed over the last 25 years and hence presents a high-quality, valuable resource for German. While GermaNet is maintained in a Postgres database, all its content can be exported as an XML-based serialisation. Recently, this XML representation has been converted into RDF, largely by staying close to GermaNet's principle of arrangement where *lexunits* that share the same meaning are grouped together into so-called *synsets*. With each lexical unit and synset now globally addressable via a unique resource identifier, it has become much easier to link together GermaNet entries with other lexical and semantic resources. In terms of semantic interoperability, however, the RDF variant of GermaNet leaves much to be desired. In this paper, we describe yet another conversion from GermaNet's XML representation to RDF. The new conversion makes use of the OntoLex-Lemon ontology, and therefore, presents a decisive step toward a GermaNet representation with a much higher level of semantic interoperability, and which makes it possible to use GermaNet with other wordnets that already support this conceptualisation of lexica.

1 Introduction

GermaNet was conceived in the mid-nineties (Hamp and Feldweg, 1997) and soon became the largest lexical-semantic wordnet for German. While it is still maintained as a relational database, it profited from quite a few format conversions in the meantime. With the wide adoption of the data interchange format XML, GermaNet's internal data representation – it is represented as a collection of relational database tables – was reformalized in terms of DTD-based document types. Four DTDs were defined: for synsets and their lexical unit children, for lexical and conceptual relations between them, for mapping GermaNet via the interlingual index to the Princeton Wordnet (Miller, 1995; Fellbaum, 1998), and for enriching GermaNet entries with Wiktionary paraphrases.[1] The current distribution of GermaNet provides both the database dump as well as an XML serialisation with XML documents that adhere to the DTD, and hence are syntactically valid. In total, the distribution encompasses 54 files (23 files for nouns, 15 files for verbs, and 16 files for adjectives). Each file name encodes the word category and the semantic class of the synsets they contain.[2] For each of the three word classes, there is also an XML file which encodes Wiktionary entries, and there is a single file for the XML encoding of the interlingual index and another file to encode the conceptual and lexical relations.

The single source of truth for GermaNet, however, is the Postgres-based database. A special-purpose tool called *GernEdit* is used to edit and extend the German wordnet (Henrich and Hinrichs, 2010a). Programming APIs in Java and Python are available to access all GermaNet information programmatically.[3] Users without a usage licence for GermaNet can use the web-based *Rover* application to explore GermaNet content. With Rover, users can also calculate and visualize the semantic relatedness between any two given synsets.

The latest version of GermaNet (release 17.0, April 2022) has about 205,000 lexical units and 159,514 synsets. There are 1,29 lexical units per

[1] In the EuroWordNet framework (Vossen, 1998) (see https://archive.illc.uva.nl/EuroWordNet), about 28,500 concepts from GermaNet have been linked to Princeton WordNet(R) 2.0. We have used mappings from WordNet(R) 2.0 to WordNet(R) 3.0 provided by the NLP group of the Universitat Politècnica de Catalunya to link GermaNet synsets to WordNet(R) 3.0. The mapping to WordNet(R) 3.0 was created automatically thus 100% accuracy of those mappings cannot be guaranteed.

[2] For instance, all nouns related to humans are given in the XML file nomen.Mensch.xml.

[3] https://uni-tuebingen.de/en/142806 (Applications & Tools).

```
<synset id="s25806" category="nomen" class="Tier">
  <lexUnit id="l35305" sense="3" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Ei</orthForm>
  </lexUnit>
</synset>
<synset id="s39427" category="nomen" class="Nahrung">
  <lexUnit id="l57850" sense="4" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Ei</orthForm>
  </lexUnit>
  <lexUnit id="l57851" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Hühnerei</orthForm>
    <compound>
      <modifier category="Nomen">Huhn</modifier>
      <head>Ei</head>
    </compound>
  </lexUnit>
</synset>
<synset id="s73239" category="nomen" class="Form">
  <lexUnit id="l100105" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Ei</orthForm>
  </lexUnit>
</synset>
<synset id="s25813" category="nomen" class="Koerper">
  <lexUnit id="l35317" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Eizelle</orthForm>
    <compound>
      <modifier category="Nomen">Ei</modifier>
      <head>Zelle</head>
    </compound>
  </lexUnit>
  <lexUnit id="l35318" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no"> [2 lines]
  <lexUnit id="l90270" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="yes"> [2 lines]
  <lexUnit id="l103438" sense="2" source="core" namedEntity="no" artificial="no" styleMarking="no">
    <orthForm>Ei</orthForm>
  </lexUnit>
</synset>
```

Figure 1: Four different entries for *Ei*.

.

synset. Moreover, GermaNet defines 173,742 conceptual relations between synsets, and 12,204 lexical relations between lexical units (excluding synonymy). In addition to conceptual relations known from Princeton WordNet, GermaNet also features a good number of lexical relations that have no correspondance in the Princeton Wordnet. The German language makes good use of compounds, and this is also reflected in the high number of compounds and their proper segmentation in subterms (115,366 compounds are represented).

GermaNet already has some substantial linking to external data sources such as 28,564 pointers to the interlingual index and 29,546 links to Wiktionary. Note that any linking to external data sources is established through "local" identifiers only so that contextual information (say, this is an identifier in Princeton Wordnet 2.0) is required to resolve or look-up such linkages.

It is worth pointing out that GermaNet has also been converted to the lexical markup framework (LMF, see (Vossen et al., 2013)), which is discussed in (Henrich and Hinrichs, 2010b).

Recently, we have converted GermaNet's XML serialisation of its database into RDF (Zinn et al., 2022). The conversion stayed close to GermaNet's conceptualisation of organising lexical-semantic nets (see Sect. 2). While our conversion of GermaNet into RDF comes with no information loss, it ignores the work of others that aim at defining a standard for the description of wordnets. One such standard for representing wordnets is the OntoLex-Lemon conceptualisation[4], which we briefly describe in Sect. 3. In Sect. 4, we show how we converted GermaNet's XML serialisation to the OntoLex-Lemon format and that most but not all of GermaNet content can be expressed in terms of this ontology. The conclusion and future work is discussed in Sect. 5.

2 Background

2.1 GermaNet Overview

In GermaNet, the meaning of a word, its word sense, is represented as a *lexical unit*. Word senses that express the same semantic concept are grouped together into *synsets*, a short form of synonym

[4] https://www.w3.org/2016/05/ontolex/

```
<relations>
  <con_rel name="has_hypernym" from="s68168" to="s25806" dir="revert" inv="has_hyponym"/> <!--- Vogelei -->
  <con_rel name="has_hypernym" from="s25806" to="s25809" dir="revert" inv="has_hyponym"/> <!--- Keim -->
  <lex_rel name="is_container_for" from="l115250" to="l57850" dir="one" />   <!--- Eierbecher -->
  <lex_rel name="has_ingredient" from="l58624" to="l57850" dir="one" />      <!--- Eierkuchen -->
</relations>

<interLingualIndex>
  <iliRecord lexUnitId="l103438" ewnRelation="synonym" pwnWord="egg cell" pwn20Id="ENG20-05144345-n"
    pwn30Id="ENG30-05457973-n" pwn20paraphrase="the female reproductive cell; the female gamete"
    source="initial"/>
  <iliRecord lexUnitId="l35305" ewnRelation="synonym" pwnWord="egg" pwn20Id="ENG20-01383930-n" [4 lines]
  <iliRecord lexUnitId="l57850" ewnRelation="synonym" pwnWord="egg" pwn20Id="ENG20-07367088-n"
    pwn30Id="ENG30-07840804-n"
    pwn20paraphrase="oval reproductive body of a fowl (especially a hen) used as food"
    source="initial"/>
</interLingualIndex>

<wiktionaryParaphrases>
    <wiktionaryParaphrase lexUnitId="l100105" wiktionaryId="w18622" wiktionarySenseId="3"
        wiktionarySense="ein ovales, dreidimensionales und entlang einer Achse symmetrisches Gebilde"
        edited="no"/>
    <wiktionaryParaphrase lexUnitId="l103438" wiktionaryId="w18622" wiktionarySenseId="0"
        wiktionarySense="eine Keimzelle" edited="no"/>
    <wiktionaryParaphrase lexUnitId="l35305" wiktionaryId="w18622" wiktionarySenseId="1" [2 lines]
    <wiktionaryParaphrase lexUnitId="l57850" wiktionaryId="w18622" wiktionarySenseId="2" [2 lines]
</wiktionaryParaphrases>
```

Figure 2: ILI and Wiktionary entries for *Ei*.

sets. To a large extent, GermaNet follows the design rationale of the Princeton WordNet for English, but there are, however, subtle differences that reflect the specifics of the German language. GermaNet's verbal frames, for instance, capture more detail than those represented in WordNet: reflexives, grammatical case, expletive subjects, and to-infinitives are explicitly encoded in GermaNet. With the German language making extensive use of compound constructions, GermaNet has rich descriptive means to describe them (see below).

2.2 GermaNet Example

GermaNet has four different lexical units with an orthographic form *Ei* (egg), which are distributed over the thematic domains *Form* (form), *Körper* (body), *Nahrung* (food), and *Tier* (animal), and therefore, distributed over four different files. Fig. 1 depicts the four lexical units, each of which is part of a different synset. Each synset comes with an identifier unique to GermaNet, a category encoding the part of speech of its lexUnits, and a class that marks their thematic domain. In GermaNet's XML representation, a lexical unit is always a child element of a synset. Each lexical unit also comes with a unique identifier, a sense identifier, and four other attributes: *namedEntity* specifies whether the lexical unit denotes a named entity or not; *styleMarking* is true if the lexical unit represents a stylistic variant; *artificial* is true if the lexical unit is used to represent an artificial node in the graph.[5] The source attribute is for internal use only. All attributes are mandatory. Each lexical unit must have a child *orthForm*. If the lexical unit is a compound, its head and modifier are also given. Fig. 1 also depicts two lexical units whose orthographic form is a compound, for instance, *Eizelle* (egg cell). In this case, GermaNet specifies the head of the compound *Zelle* and its modifier *Ei*. Note that GermaNet encodes eight different properties for compound constituents and seven modifier classes.

Fig. 2 depicts the three ILI records and the four entries into Wiktionary that GermaNet knows about the lemma *Ei*. An ILI record links a lexical unit of GermaNet via some relation to an entry into the Princeton Wordnet. It is worth to note that the target of the relation is (also) not an URI but an identifier locally unique to the wordnet. Note that *ewnRelation* encompasses not only *synonym* relationships but also *hypernym*, *hyponym*, *is_caused_by*, *causes* relationships, among others. Usually, a paraphrase from Princeton WordNet 2.0 is given.

Fig. 2 also shows four entries into Wiktionary paraphrases (again, some lines omitted), a useful addition to GermaNet data. – Note that both linkages were established more than 10 years ago and need to be updated and extended, where possible.

[5] GermaNet is a completely connected graph hierarchy without any dangling subgraphs, whereas WordNet consists of several distinct hierarchies – one for each semantic field.

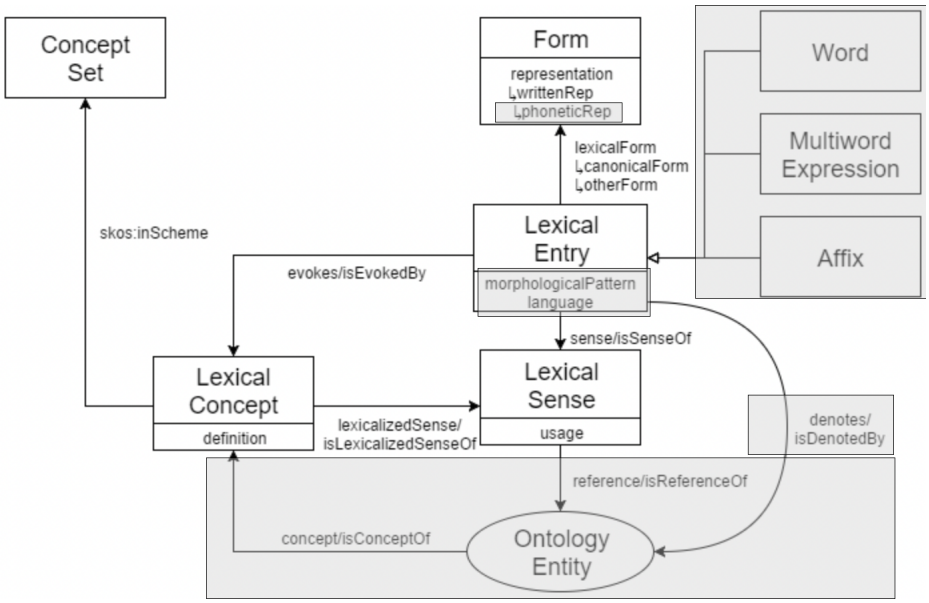

Figure 3: Core model of OntoLex.

3 OntoLex-Lemon's Design Principle

Fig. 3 depicts the core model of OntoLex, see also (McCrae et al., 2012).

In GermanNet's XML serialisation, a *synset* element contains one or more elements of type *lexUnit*, which in turn has a single obligatory child *orthForm* and optional children such as *compound*, or orthographic variants. Much information is encoded into XML attributes. The *category* attribute at the *synset* level encodes part-of-speech information whereas the *sense* attribute at the *lexUnit* level encodes a sense identifier marking a lemma (*orthForm*) as being part of several synsets.

In OntoLex, the structure of a lexical-semantic net is different with *Lexical Entry* encoding the entries of a lexicon. Each entry has a *Form* (the written representation, mirroring GermaNet's *orthForm* element), and a *Lexical Sense*, which in turn is the lexicalized sense of a *Lexical Concept*, the equivalent of a GermaNet *synset*.

4 Conversion and Extension

4.1 Conversion to OntoLex-Lemon

Our conversion makes use of all parts of OntoLex-Lemon apart from the items greyed out in Fig. 3.

Fig. 4 depicts a fragment of our conversion from GermaNet to the OntoLex-Lemon conceptualisation. Compared with the four occurrences shown in Fig. 1, there is now a single lexical entry having a *Form* with the written representation *Ei*.[6] In line with our example entries given in Fig. 1, this lexical entry has four different lexical senses, and hences evokes four different lexical concepts. Each sense inherits the *lexUnit* identifier of our XML representation, and each lexical concept inherits the respective *synset* identifier.

Note that our conversion failed to map information that in the GermaNet representation of *lexunit* is expressed in terms of attributes: *namedEntity*, *artificial*, and *styleMarking*. In these cases, we fall back to our initial conversion approach using our own *gn* vocabulary.

Fig. 4 also shows a number of conceptual relations between the lexical concept evoked by *Ei* and its super- and subclasses (hypernym and hyponym). Lexical relations are attached to the resources of type *LexicalSense*. At the time of writing, we still reuse our own vocabulary to express lexical relations.

Note that a lexical concept comes with two attributes that specify their semantic field: *skos:inScheme* carries the German name of the semantic field. and *dc:subject* carries its English translation.[7]

[6] For the sake of brevity, we have chosen to use a blank node to refer to something that has a written representation.

[7] As Henrich (2015) pointed out: "the semantic fields resemble the unique beginners in WordNet. However, mainly due to language specific differences of the two wordnets, the lists are not exactly identical: for instance, labels Verhalten and privativ are not available in WordNet, while act and process are not used in GermaNet".

Figure 4: OntoLex example encoding of GermaNet.

German Compounds. GermaNet has information about over 115,000 nominal compounds, splits them into their constituent parts, and labels them with linguistic information. In GermaNet, the constituents of compounds can have one of the following eight properties, see Fig. 5, also see (Henrich, 2015, Chapt. 3.6) and (Henrich and Hinrichs, 2011). This kind of information makes particular sense for German, where compounds are almost always spelled as one word.

Consider, for instance, the GermaNet's lexical unit *l36389* with orthographic form *Tollwut* (rabies). The modifier of the compound is *toll* of class *adjective* and its head is *Wut*.[8]

In our representation in Ontolex, we have chosen the following representation:

```
gn:Tollwut-n
        a ontolex:LexicalEntry ;
        lexinfo:partOfSpeech lexinfo:noun ;
        decomp:subterm gn:Wut-n ,
                       gn:toll-adj ;
        ontolex:canonicalForm [ ontolex:writtenRep "Tollwut"
                              ] ;
        ontolex:evokes gn:s26628 ;
        ontolex:sense gn:l36389 .
```

It consists of two *subterm* triples pointing to the lexical entries *Wut-n* and *toll-adj*. In this case, both lexical entries are part of GermaNet so that both subterms properly resolve. There are many other examples, however, where this is not the case, in particular in cases where modifiers are adverbs, particles, prepositions, or pronouns. Those word classes are not (yet) represented in GermaNet. This is an issue we have yet to resolve.

It is also clear that the *decomp:subterm* property does not distinguish between heads and modifiers, and cannot represent the information given in Fig. 5, so more work is required here.

GermaNet has a rich representation of verbal frames. For the representation of syntactic frames, we consider using the lexinfo ontology[9] (verb frame), but this is not done yet.

4.2 Processing ILI and Wiktionary Information

Fig. 4 has a number of triples whose properties have the namespace *gn:*, and hence do not make use of vocabularies such as *ontolex* or *lexinfo*. Consider, for instance, the information stemming from the Interlingual Index, which are all associated with the lexical sense of the lexical entry *Ei*:

```
gn:l35305
   gn:pwn20Id "ENG20-01383930-n" ;
   gn:pwn30Id "ENG30-01460457-n" ;
```

[8] The other classes are *adverb*, *noun*, *particle*, *preposition*, *pronoun*, and *verb*, see (Henrich, 2015, Chapt. 3.6).

[9] http://www.lexinfo.net/ontology/3.0/lexinfo.

Property	Example (and explanation, if needed)
abbreviation*	*IP* 'IP' in *IP-Paket* 'IP packet'
affixoid*	affixoids have a special grammatical status between bound and free morphemes; e.g., *haupt* 'main' in *Hauptbahnhof* 'main station'
foreign word*	*Offset* 'offset' in *Offsetdruck* 'offset printing'
combining form*§ (German: *konfix*)	bound morphemes which are borrowed from a foreign language and whose meaning stems from that particular language, e.g., *bio-* 'organic' in *Biosiegel* 'organic seal'
opaque morpheme*	*Him-* in *Himbeere* 'rasperry'
proper name†	*Valentin* 'Valentine' in *Valentinstag* 'Valentine's Day'
virtual word form‡	*Zieher* nominalization for 'to pull' (word does not exist in isolation) in *Schraubenzieher* 'screwdriver'
word group†	*drei Zimmer* 'three-room' in *Dreizimmerwohnung* 'three-room flat'

Figure 5: Properties for compound constituents, see (Henrich, 2015, Chapt. 3.6)

.

```
gn:pwn31Id pwn:01463098-n ;
gn:hasILIid ili:i42980 ;
```

Note that the first two triples stem from the mapping between GermaNet and the Interlingual Index.[10] Their objects make use of Princeton Wordnet (PWN) identifiers that do not resolve automatically. As part of the conversion, we have used a mapping from PWN 3.0 to PWN 3.1 to update the identifiers to the latest version of PWN (Zendel, 2019). The predicate `pwn31Id` now points to a resolvable URI into the RDF version of the Princeton WordNet.[11] Moreover, using the mapping between PWN 3.0 to the CILI (Bond et al., 2016) supplied by Francis Bond[12], `gn:hasILIid` points to the `http://globalwordnet.org/ili/` (namespace prefix `ili`).

The linkage of GermaNet with Wiktionary dates back to 2011 and made use of a large Wiktionary dump in order to automatically harvest sense definitions from the German Wiktionary for GermaNet senses (Henrich et al., 2014b). The *wiktionaryId* on Fig. 2 was introduced for purely technical reasons and cannot be used to lookup Wiktionary content in the current release.

During the conversion process, we downloaded a recent RDF version of Wiktionary and established a local SPARQL endpoint. We then queried the endpoint for all subjects that have a *skos:definition* to a node whose value is string-identical to the passphrase of the 2011 data linkage. The *gn:wiktionaryId* now points to a new resolvable URI.

4.3 Linkage to Other Lexical Resources

With GermaNet now being available in RDF, it is tempting to link its content to other resources in the Linked Data world. As a start, we have established links to Wikidata and the authority files of the German National Library.

GermaNet has a wealth of information on nouns with the semantic field *Ort* (*location*). Entries range from *Tagungshotel* (conference hotel) to 25 entries centered around the concept of *Gefängnis* (prison) such as *Frauengefängnis* and *Gefängnisinsel*. A substantial part of the information, however, represents names for cities, rivers, and mountains, and other geographic places. For this kind of information, a valuable subset of the *Integrated Authority File* (GND)[13] of the German National Library is available, namely, the subset holding *Geographika* with approximately 4.5 million triples.

The query for the geographical dataset is rather simple, searching for all entities where the *preferredNameForThePlaceOrGeographicName* of an entity is the location name, say *Potsdam*. As a result, the synset *s43887* with its lexical unit *l63714* and its orthographic form *Potsdam* was automatically linked to the entity `https://d-nb.`

[10] `https://tinyurl.com/y9znkzjz`
[11] `http://wordnet-rdf.princeton.edu/id`
[12] `https://github.com/globalwordnet/cili.git`
[13] `https://gnd.network`

info/gnd/4046948-7 of the GND dataset. The semantic linkage gives users access to a variety of information such as alternative names or lexicalisations (*e.g.*, Bostanium, Potestampium, Pozdam), the geographical coordinates in terms of latitude and longitude, and other information (*Hauptstadt vom Bundesland Brandenburg, kreisfreie Stadt, 993 als Poztupimi urkundl. erwähnt, 1317 Stadt*), hence demonstrating the potential of linked data. In this initial work, 1778 GermaNet entries were linked to entities in the subset of the GND dataset.

We have also queried Wikidata for location names. Here, the situation is more complicated, in part due to the crowd-sourcing approach of the platform, and because no geographical subset of Wikidata is readily available. We hence had to guide our search to only take into account entities whose type indicate their geographic nature. In Wikidata, there are a large amount of location types such as *big city*, *capital*, *city*, *state of the USA*, *river*, *commune of France*, *town*, *geographic region*, *country*, *historical country*, *inferior planet*, *peninsula*, *sea*, *ocean etc.* so that the query to Wikidata becomes quite complex.

At the time of writing, we were able to establish 2,564 links to Wikidata entries of type location. For *Potsdam*, two Wikidata entries were found: the wikidata entity *Q1711*, found via the location type *big city* (Q1549591), and the wikidata entity *Q1022943*, identified via the location type *town of the United States* (Q15127012). For GermaNet, it can be argued that only the first hit should be linked, but we decided to include all associations.

4.4 Implementation Details

Our conversion takes GermaNet's XML-based serialisation of its database content as a starting point. The conversion has been implemented in Prolog using SWI-Prolog, its built-in library `sgml` for XML parsing and its semantic web library `semweb/rdf11`. The conversion processes all main input files for nouns, verbs, and adjectives, the XML file that defines conceptual and lexical relations, and the ILI and Wiktionary files. While those files are being parsed, RDF triples are being asserted. At the end of the process, the triple store is written into a file resulting in approximately 3.5 million triples.

5 Conclusion and Future Work

There have been two prominent translations of Princeton Wordnet into RDF (Graves and Gutierrez, 2006; van Assem et al., 2006), but there is only one that uses the *lemon* vocabulary (McCrae et al., 2014). In this paper, we have described our conversion of GermaNet's XML format to a RDF representation that makes use of the OntoLex-Lemon conceptualisation, hence mirroring the work of McCrae and colleagues for GermaNet. This makes it possible for GermaNet to be part of a linked data cloud that combines rich linguistic information from various, high-quality resources.

In the near future, we will complement our conversion to include a more detailed representation of nominal compounds, and we still have to tackle the issue of representing syntactic frame information using the *lexinfo* vocabulary. The aim is to replace, whenever possible, our local vocabulary (in the namespace *gn*) with well-known terminology well-defined elsewhere. In this regard, GermaNet is monitoring recent developments in the Global Wordnet Formats (McCrae et al., 2021). Hence, our RDF version of GermaNet should not be considered final (yet) but open to change in the future.

Future work includes linking GermaNet with other RDF-based resources. In part, this is already done, as we have seen with the introduction of the ILI link into the RDF-based Princeton WordNet. At the time of writing, our GermaNet resource identifiers are not yet web-resolvable. In the future, an HTTP request to, say, `https://uni-tuebingen.de/germanet/v17/Ei-n`, will return the top left part of Fig. 4.

Rover, a web-based user interface for the exploration and visualization of GermaNet data (Hinrichs et al., 2020) is currently using the XML representation and the Java API in the back-end. In the future, we would like to experiment with using a back-end that executes SPARQL queries on the triple store.

GermaNet is free for academic users with a signed license.[14] For licence holders, both the database and the XML export are included in the data download.[15] In the future, licence holders will also be able to obtain the RDF export of GermaNet.

[14] `https://uni-tuebingen.de/en/142806` (Licenses).

[15] The mapping from GermaNet to Wiktionary and the ILI can be downloaded separately from GermaNet.

For accessing RDF-data via the Web, we will follow our technical solution taken for our web-based Rover application: a sign-in via the CLARIN Service Provider Federation will allow users to authenticate as academic user, and subsequently, make use of the SPARQL endpoint to GermaNet.

The main reason for having an RDF-based representation of GermaNet, however, is to unleash its potential when properly linked to other high-quality lexical sources. In the context of the Text+ project, it is our aim to link GermaNet with the DWDS dictionary of the German language[16] and also with the Leipzig Corpora Collection[17]. There are plans to convert both resources into RDF, which would allow the creation of a linked data cloud for the German language. In addition, we plan to link GermaNet to the lexicographical data of Wikidata[18].

Mapping location entities of one dataset to the locations of another dataset is relatively straightforward. In general, the main task to properly link together nodes from different RDF graphs is – essentially – a word disambiguation task. Our work will build upon Henrich et al. (2014b), where GermaNet senses were linked to wiktionary senses, and Henrich et al. (2014a), where word senses in GermaNet were linked with those in the DWDS Dictionary of the German Language. The linking task will be supported by the WebCAGe corpus (Henrich et al., 2012).

References

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Alvaro Graves and Claudio Gutierrez. 2006. Data representations for WordNet: A case for RDF. In *3rd International WordNet Conference, GWC 2006*. Masaryk University, Brno. South Jeju Island, Korea.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain.

[16] https://www.dwds.de
[17] https://corpora.uni-leipzig.de/
[18] https://wikidata.org

Verena Henrich. 2015. *Word Sense Disambiguation with GermaNet*. Ph.D. thesis, Univerisity of Tuebingen. http://dx.doi.org/10.15496/publikation-4706.

Verena Henrich and Erhard Hinrichs. 2010a. GernEdiT: A graphical tool for GermaNet development. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24, Uppsala, Sweden. Association for Computational Linguistics.

Verena Henrich and Erhard Hinrichs. 2010b. Standardizing wordnets in the ISO standard LMF: Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 456–464, Beijing, China. Coling 2010 Organizing Committee.

Verena Henrich and Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 420–426, Hissar, Bulgaria. Association for Computational Linguistics.

Verena Henrich, Erhard Hinrichs, and Reinhild Barkey. 2014a. Aligning Word Senses in GermaNet and the DWDS Dictionary of the German Language. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, pages 63–70. Tartu, Estonia.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2012. WebCAGe – a Web-Harvested Corpus Annotated with GermaNet Senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 387–396. Avignon, France.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2014b. Aligning GermaNet Senses with Wiktionary Sense Definitions. In *Human Language Technology: Challenges for Computer Science and Linguistics*, pages 329–342.

Marie Hinrichs, Richard Lawrence, and Erhard Hinrichs. 2020. Exploring and Visualizing Wordnet Data with GermaNet Rover. In *Proceedings of the CLARIN Annual Conference*, pages 32–36.

John P. McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Lang. Resour. Evaluation*, 46(4):701–719.

John P. McCrae, Christiane D. Fellbaum, and Philipp Cimiano. 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luís Morgado da Costa. 2021. The Global Wordnet Formats: Updates for 2020. In *Proceedings of*

the 11th Global Wordnet Conference, GWC 2021, University of South Africa (UNISA), Potchefstroom, South Africa, January 18-21, 2021, pages 91–99. Global Wordnet Association.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mark van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of wordnet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 237–242. European Language Resources Association (ELRA).

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer Dordrecht.

Piek Vossen, Claudia Soria, and Monica Monachini. 2013. *Wordnet-LMF: A Standard Representation for Multilingual Wordnets*, pages 51–66. Wiley.

Oliver Zendel. 2019. Wordnet v3.0 vs. v3.1 mapping. `https://github.com/ozendelait/wordnet-to-json`.

Claus Zinn, Marie Hinrichs, and Erhard Hinrichs. 2022. Adapting GermaNet for the Semantic Web. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 41–47, Potsdam, Germany.