Building a BIOWORDNET by Using WORDNET's Data Formats and WORDNET's Software Infrastructure — A Failure Story

Michael Poprat **Elena Beisswanger** **Udo Hahn**

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
D-07743 Jena, Germany
{poprat,beisswanger,hahn}@coling-uni-jena.de

Abstract

In this paper, we describe our efforts to build on WORDNET resources, using WORDNET lexical data, the data format that it comes with and WORDNET's software infrastructure in order to generate a biomedical extension of WORDNET, the BIOWORDNET. We began our efforts on the assumption that the software resources were stable and reliable. In the course of our work, it turned out that this belief was far too optimistic. We discuss the stumbling blocks that we encountered, point out an error in the WORDNET software with implications for research based on it, and conclude that building on the legacy of WORDNET data structures and its associated software might preclude sustainable extensions that go beyond the domain of general English.

1 Introduction

WORDNET (Fellbaum, 1998) is one of the most authoritative lexical resources for the general English language. Due to its coverage – currently more than 150,000 lexical items – and its lexicological richness in terms of definitions (glosses) and semantic relations, synonymy via synsets in particular, it has become a *de facto* standard for all sorts of research that rely on lexical content for the English language.

Besides this perspective on rich lexicological data, over the years a software infrastructure has emerged around WORDNET that was equally approved by the NLP community. This included, e.g., a lexicographic file generator, various editors and visualization tools but also meta tools relying on properly formated WORDNET data such as a library of similarity measures (Pedersen et al., 2004). In numerous articles the usefulness of this data and software ensemble has been demonstrated (e.g., for word sense disambiguation (Patwardhan et al., 2003), the analysis of noun phrase conjuncts (Hogan, 2007), or the resolution of coreferences (Harabagiu et al., 2001)).

In our research on information extraction and text mining within the field of biomedical NLP, we similarly recognized an urgent need for a lexical resource comparable to WORDNET, both in scope and size. However, the direct usability of the original WORDNET for biomedical NLP is severely hampered by a (not so surprising) lack of coverage of the life sciences domain in the general-language English WORDNET as was clearly demonstrated by Burgun and Bodenreider (2001).

Rather than building a BIOWORDNET by hand, as was done for the general-language English WORDNET, our idea to set up a WORDNET-style lexical resource for the life sciences was different. We wanted to *link* the original WORDNET with various biomedical terminological resources vastly available in the life sciences domain. As an obvious candidate for this merger, we chose one of the major high-coverage umbrella systems for biomedical ontologies, the OPEN BIOMEDICAL ONTOLOGIES (OBO).[1] These (currently) over 60 OBO ontologies provide domain-specific knowledge in terms of hierarchies of classes that often come with synonyms and textual definitions for lots of biomedical subdomains (such as genes, proteins, cells, sequences,

[1] http://www.bioontology.org/repositories.html#obo

Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pages 31–39,
Columbus, Ohio, USA, June 2008.

etc.).[2] Given these resources and their software infrastructure, our plan was to create a biomedically focused lexicological resource, the BIOWORDNET, whose coverage would exceed that of any of its component resources in a so far unprecedented manner. Only then, given such a huge combined resource advanced NLP tasks such as anaphora resolution seem likely to be tackled in a feasible way (Hahn et al., 1999; Castaño et al., 2002; Poprat and Hahn, 2007). In particular, we wanted to make *direct* use of available software infrastructure such as the library of similarity metrics without the need for re-programming and hence foster the reuse of existing software *as is*.

We began our efforts on the assumption that the WORDNET software resources were stable and reliable. In the course of our work, it turned out that this belief was far too optimistic. We discuss the stumbling blocks that we encountered, point out an error in the WORDNET software with implications for research based on it, and conclude that building on the legacy of WORDNET data structures and its associated software might preclude sustainable extensions that go beyond the domain of general English. Hence, our report contains one of the rare failure stories (not only) in our field.

2 Software Around WORDNET Data

While the stock of lexical data assembled in the WORDNET lexicon was continuously growing over time,[3] its data format and storage structures, the so-called *lexicographic file*, by and large, remained unaltered (see Section 2.1). In Section 2.2, we will deal with two important software components with which the lexicographic file can be created and browsed. Over the years, together with the continuous extension of the WORDNET lexicon, a lot of software tools have been developed in various programming languages allowing browsing and accessing WORDNET as well as calculating semantic similarities on it. We will discuss the most relevant of these tools in Section 2.3.

[2] Bodenreider and Burgun (2002) point out that the structure of definitions in WORDNET differ to some degree from more domain-specialized sources such as medical dictionaries.

[3] The latest version 3.0 was released in December 2006

2.1 Lexicon Organization of WORDNET and Storage in Lexicographic Files

At the top level, WORDNET is organized according to four parts of speech, *viz.* noun, verb, adjective and adverb. The most recent version 3.0 covers more than 117,000 nouns, 11,500 verbs, 21,400 adjectives and 4,400 adverbs, interlinked by *lexical relations*, mostly derivations. The basic semantic unit for all parts of speech are sets of synonymous words, so-called *synsets*. These are connected by different semantic relations, imposing a thesaurus-like structure on WORDNET. In this paper, we discuss the organization of noun synsets in WORDNET only, because this is the relevant part of WORDNET for our work. There are two important *semantic* relation types linking noun synsets. The *hypernym* / *hyponym* relation on which the whole WORDNET noun sense hierarchy is built links more specific to more general synsets, while the *meronym* / *holonym* relation describes partonomic relations between synsets, such as part of the whole, member of the whole or substance of the whole.

From its very beginning, WORDNET was built and curated manually. Lexicon developing experts introduced new lexical entries into WORDNET, grouped them into synsets and defined appropriate semantic and lexical relations. Since WORDNET was intended to be an electronic lexicon, a data representation format had to be defined as well. When the WORDNET project started more than two decades ago, markup languages such as SGML or XML were unknown. Because of this reason, a rather idiosyncratic, fully text-based data structure for these lexicographic files was defined in a way to be readable and editable by humans — and survived until to-day. This can really be considered as an outdated legacy given the fact that the WORDNET community has been so active in the last years in terms of data collection, but has refrained from adapting its data formats in a comparable way to to-day's specification standards. Very basically,[4] each line in the lexicographic file holds one synset that is enclosed by curly brackets. Take as an example the synset for "monkey":

[4] A detailed description can be found in the WORDNET manual *wninput(5WN)*, available from `http://wordnet.princeton.edu/man/wninput.5WN`.

```
 { monkey, primate,@ (any of various
long-tailed primates (excluding the
prosimians)) }
```

Within the brackets at the first position synonyms are listed, separated by commas. In the example, there is only one synonym, namely "monkey". The synonyms are followed by semantic relations to other synsets, if available. In the example, there is only one hypernym relation (denoted by "@") pointing to the synset "primate". The final position is reserved for the gloss of the synset encapsulated in round brackets. It is important to notice that there are no identifiers for synsets in the lexicographic file. Rather, the string expressions themselves serve as identifiers. Given the fundamental idea of synsets – all words within a synset mean exactly the same in a certain context – it is sufficient to relate one word in the synset in order to refer to the whole synset. Still, there must be a way to deal with homonyms, i.e., lexical items which share the same string, but have different meanings. WORDNET's approach to distinguish different senses of a word is to add numbers from 0 to 15, called *lexical identifiers*. Hence, in WORDNET, a word cannot be more than 16-fold ambiguous. This must be kept in mind when one wants to build a WORDNET for highly ambiguous sublanguages such as the biomedical one.

2.2 Software Provided with WORDNET

To guarantee fast access to the entries and their relations, an optimized index file must be created. This is achieved through the easy-to-use GRIND software which comes with WORDNET. It simply consumes the lexicographic file(s) as input and creates two plain-text index files,[5] namely `data` and `index`. Furthermore, there is a command line tool, WN, and a graphical browser, WNB, for data visualization that require the specific index created by GRIND (as all the other tools that query the WORDNET data do as well). These tools are the most important (and only) means of software support for WORDNET creation by checking the syntax as well as allowing the (manual) inspection of the newly created index.

[5]Its syntax is described in `http://wordnet.princeton.edu/man/wndb.5WN`.

2.3 Third-Party WORDNET Tools

Due to the tremendous value of WORDNET for the NLP and IR community and its usefulness as a resource for coping with problems requiring massive amounts of lexico-semantic knowledge, the software-developing community was and continues to be quite active. Hence, in support of WORDNET several APIs and software tools were released that allow accessing, browsing and visualizing WORDNET data and measuring semantic similarity on the base of the WORDNET's lexical data structures.[6]

The majority of these APIs are maintained well and kept up to date, such as JAWS[7] and JWNL,[8] and enable connecting to the most recent version of WORDNET. For the calculation of various similarity measures, the PERL library WORDNET::SIMILARITY initiated and maintained by Ted Pedersen[9] can be considered as a *de facto* standard and has been used in various experimental settings and applications. This availability of well-documented and well-maintained software is definitely a strong argument to rely on WORDNET as a powerful lexico-semantic knowledge resource.

3 The BIOWORDNET Initiative

In this section, we describe our approach to extend WORDNET towards the biomedical domain by incorporating terminological resources from the OBO collection. The most obvious problems we faced were to define a common data format and to map non-compliant data formats to the chosen one.

3.1 OBO Ontologies

OBO is a collection of publicly accessible biomedical ontologies.[10] They cover terms from many biomedical subdomains and offer structured, domain-specific knowledge in terms of classes (which often come with synonyms and textual definitions) and class hierarchies. Besides the hierarchy-defining relation *is-a*, some OBO ontologies provide

[6]For a comprehensive overview of available WORDNET tools we refer to WORDNET's 'related project' website (`http://wordnet.princeton.edu/links`).

[7]`http://engr.smu.edu/~tspell/`

[8]`http://jwordnet.sourceforge.net/`

[9]`http://wn-similarity.sourceforge.net/`

[10]`http://www.bioontology.org/`

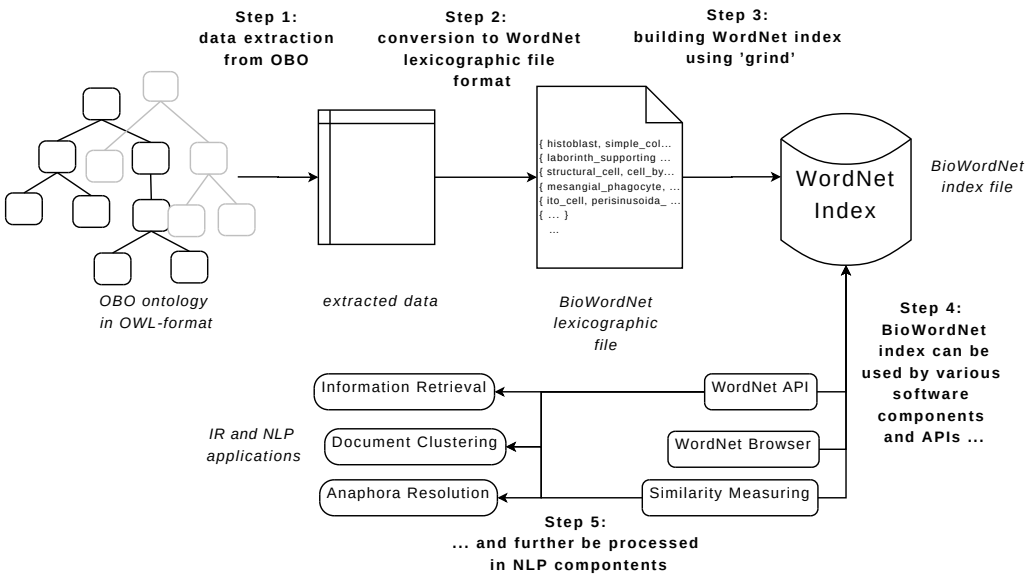

Figure 1: From OBO ontologies to BIOWORDNET— towards a domain-specific WORDNET for biomedicine

additional semantic relation types such as *sequence-of* or *develops-from* to express even more complex and finer-grained domain-specific knowledge. The ontologies vary significantly in size (up to 60,000 classes with more than 150,000 synonyms), the number of synonyms per term and the nature of terms.

The OBO ontologies are available in various formats including the OBO flat file format, XML and OWL. We chose to work with the OWL version for our purpose,[11] since for the OWL language also appropriate tools are available facilitating the extraction of particular information from the ontologies, such as taxonomic links, labels, synonyms and textual definitions of classes.

3.2 From OBO to BIOWORDNET

Our plan was to construct a BIOWORDNET by converting, in the first step, the OBO ontologies into a WORDNET hierarchy of synsets, while keeping to the WORDNET lexicographic file format, and building a WORDNET index. As a preparatory step, we defined a mapping from the ontology to WORDNET items as shown in Table 1.

The three-stage conversion approach is depicted in Figure 1. First, domain specific terms and taxonomic links between terms were extracted separately from each of the OBO ontologies. Then the extracted data was converted according to the syntax specifications of WORDNET's lexicographic file. Finally for each of the converted ontologies the WORDNET-specific index was built using GRIND.

OBO ontology	BIOWORDNET
ontology class	synset
class definition	synset gloss
class name	word in synset
synonym of class name	word in synset
C_i *is-a* C_j C_j *has-subclass* C_i	S_i *hyponym* of S_j S_j *hypernym* of S_i

Table 1: Mapping between items from OBO and from BIOWORDNET (C_i and C_j denote ontology classes, S_i and S_j the corresponding BIOWORDNET synsets)

Following this approach we ran into several problems, both regarding the WORDNET data structure and the WORDNET-related software that we used for the construction of the BIOWORDNET. Converting the OBO ontologies turned out to be cumbersome, especially the conversion of the CHEBI ontology[12] (long class names holding many special characters) and the NCI thesaurus[13] (large number

[11] http://www.w3.org/TR/owl-semantics/

[12] http://www.ebi.ac.uk/chebi/

[13] http://nciterms.nci.nih.gov/

of classes and some classes that also have a large number of subclasses). These and additional problems will be addressed in more detail in Section 4.

4 Problems with WordNet's Data Format and Software Infrastructure

We here discuss two types of problems we found for the data format underlying the WordNet lexicon and the software that helps building a WordNet file and creating an index for this file. First, WordNet's data structure puts several restrictions on what can be expressed in a WordNet lexicon. For example, it constrains lexical information to a fixed number of homonyms and a fixed set of relations. Second, the data structure imposes a number of restrictions on the string format level. If these restrictions are violated the WordNet processing software throws error messages which differ considerably in terms of informativeness for error tracing and detection or even do not surface at all at the lexicon builder's administration level.

4.1 Limitations of Expressiveness

The syntax on which the current WordNet lexicographic file is based imposes severe limitations on what can be expressed in WordNet. Although these limitations might be irrelevant for representing general-language terms, they do affect the construction of a WordNet-like resource for biomedicine. To give some examples, the WordNet format allows a 16-fold lexical ambiguity only (lexical IDs that are assigned to ambiguous words are restricted to the numbers 0-15, see Section 2). This forced us to neglect some of the OBO ontology class names and synonyms that were highly ambiguous.[14]

Furthermore, the OBO ontologies excel in a richer set of semantic relations than WordNet can offer. Thus, a general problem with the conversion of the OBO ontologies into WordNet format was that except from the taxonomic *is-a* relation (which corresponds to the WordNet *hyponym* relation) and the *part-of* relation (which corresponds to the WordNet *meronym* relation) all remaining OBO-specific relations (such as *develops-from*, *sequence-of*, *variant-of* and *position-of*) could not be represented in the BioWordNet. The structure of WordNet neither contains such relations nor is it flexible enough to include them so that we face a systematic loss of information in BioWordNet compared to the original OBO ontologies. Although these restrictions are well-known, their removal would require extending the current WordNet data structure fundamentally. This, in turn, would probably necessitate a full re-programming of all of WordNet-related software.

[14] This is a well-known limitation that is already mentioned in the WordNet documentation.

4.2 Limitations of Data Format and Software

When we tried to convert data extracted from the OBO ontologies into WordNet's lexicographic file format (preserving its syntactic idiosyncrasies for the sake of quick and straightforward reusability of software add-ons), we encountered several intricacies that took a lot of time prior to building a valid lexicographic file.

First, we had to replace 31 different characters with unique strings such as "(" with "-LRB-" and "+" with "-PLU-" before Grind was able to process the lexicographic file. The reason is that many of such special characters occurring in domain specific terms, especially in designators of chemical compounds such as *"methyl ester 2,10-dichloro-12H-dibenzo(d,g)(1,3)dioxocin-6-carboxylic acid"* (also known as *"treloxinate"* with the CAS registry number 30910-27-1), are reserved symbols in the WordNet data formatting syntax. If these characters are not properly replaced Grind throws an exact and useful error message (see Table 2, first row).

Second, we had to find out that we have to replace all empty glosses by at least one whitespace character. Otherwise, Grind informs the user in terms of a rather cryptic error message that mentions the position of the error though not its reason (see Table 2, second row).

Third, numbers at the end of a lexical item need to be escaped. In WordNet, the string representation of an item is used as its unique identifier. To distinguish homonyms (words with the same spelling but different meaning, such as *"cell"* as the functional unit of all organisms, on the one hand, and as small compartment, on the other hand) according to the WordNet format different numbers from 0 to 15 (so-called lexical IDs) have to be appended

Problem Description	Sample Error Message	Usefulness of Error Message	Problem Solution
illegal use of key characters	*noun.cell, line 7: Illegal character %*	high	replace illegal characters
empty gloss	*sanity error - actual pos 2145 != assigned pos 2143!*	moderate	add gloss consisting of at least one whitespace character
homonyms (different words with identical strings)	*noun.rex, line 5: Synonym "electrochemical_reaction" is not unique in file*	high	distinguish word senses by adding lexical identifiers (use the numbers 1-15)
lexical ID larger than 15	*noun.rex, line 4: ID must be less than 16: cd25*	high	quote trailing numbers of words, only assign lexical identifiers between 1-15, omit additional word senses
word with more than 425 characters	*Segmentation fault (core dumped)*	low	omit words that exceed the maximal length of 425 characters
synset with more than 998 direct hyponymous synsets	*Segmentation fault (core dumped)*	low	omit some hyponymous synsets or introduce intermediate synsets with a limited number of hyponymous synsets
no query result though the synset is in the index, access software crashes	none	–	not known

Table 2: Overview of the different kinds of problems that we encountered when creating a BIOWORDNET keeping to the WORDNET data structure and the corresponding software. Each problem description is followed by a sample error message that GRIND had thrown, a statement about how useful the error message was to detect the source of the error and a possible solution for the problems, if available. The last row documents a special experience with data viewers for data from the NCI thesaurus.

to the end of each homonym. If in a lexicographic file two identical strings occur that have not been assigned different lexical identifiers (it does not matter whether this happens within or across synsets) GRIND emits an error message that mentions both, the position and the lexical entry which caused this error (cf. Table 2, third row).

Numbers that appear at the end of a lexical item as an integral part of it (such as *"2"* in *"IL2"*, a special type of cytokine (protein)) have to be escaped in order to avoid their misinterpretation as lexical identifiers. This, again, is a well-documented shortcoming of WORDNET's data specification rules.

In case such numbers are not escaped prior to presenting the lexicographic file to GRIND the word closing numbers are always interpreted as lexical identifiers. Closing numbers that exceed the number 15 cause GRIND to throw an informative error message (see Table 2, fourth row).

4.3 Undocumented Restrictions and Insufficient Error Messages

In addition to the more or less documented restrictions of the WORDNET data format mentioned above we found additional restrictions that lack documentation up until now, to the best of our knowledge.

First, it seems that the length of a word is restricted to 425 characters. If a word in the lexicographic file exceeds this length, GRIND is not able to create an index and throws an empty error message, namely the memory error "segmentation fault" (cf. Table 2, fifth row). As a consequence of this restriction, some very long CHEBI class names could not have been included in the BIOWORDNET.

Second, it seems that synsets are only allowed to group up to 988 direct hyponymous synsets. Again, GRIND is not able to create an index, if this restriction is not obeyed and throws the null memory er-

ror message “segmentation fault” (cf. Table 2, sixth row). An NCI thesaurus class that had more than 998 direct subclasses thus could not have been included in the BIOWORDNET.

Due to insufficient documentation and utterly general error messages the only way to locate the problem causing the “segmentation fault” errors was to examine the lexicographic files manually. We had to reduce the number of synset entries in the lexicographic file, step by step, in a kind of trial and error approach until we could resolve the problem. This is, no doubt, a highly inefficient and time consuming procedure. More informative error messages of GRIND would have helped us a lot.

4.4 Deceptive Results from WORDNET Software and Third-Party Components

After getting rid of all previously mentioned errors, valid index files were compiled. It was possible to access these index files using the WORDNET querying tools WN and WNB, indicating the index files were ‘valid’. However, when we tried to query the index file that was generated by GRIND for the NCI thesaurus we got strange results. While WN did not return any query results, the browser WNB crashed without any error message (cf. Table 2, seventh row). The same holds for the Java APIs JAWS and JWNL.

Since a manual examination of the index file revealed that the entries that we were searching for, in fact, were included in the file, some other, up to this step unknown error must have prevented the software tools from finding the targeted entries. Hence, we want to point out that although we have examined this error for the NCI thesaurus only, the risk is high that this “no show” error is likely to bias any other application as well which makes use of the the same software that we grounded our experiments on. Since the NCI thesaurus is a very large resource, even worse, further manual error search is nearly impossible. At this point, we stopped our attempt building a WORDNET resource for biomedicine based on the WORDNET formatting and software framework.

5 Related Work

In the literature dealing with WORDNET and its structures from a resource perspective (rather than dealing with its applications), two directions can be distinguished. On the one hand, besides the original English WORDNET and the various variant WORDNETs for other languages (Vossen, 1998), extensions to particular domains have already been proposed (for the medical domain by Buitelaar and Sacaleanu (2002) and Fellbaum et al. (2006); for the architectural domain Bentivogli et al. (2004); and for the technical report domain by Vossen (2001)). However, none of these authors neither mentions implementation details of the WORDNETs or performance pitfalls we have encountered, nor is supplementary software pointed out that might be useful for our work.

On the other hand, there are suggestions concerning novel representation formats of next-generation WORDNETs. For instance in the BALKANET project (Tufiş et al., 2004), an XML schema plus a DTD was proposed (Smrž, 2004) and an editor called CISDIC with basic maintenance functionalities and consistency check was released (Horák and Smrž, 2004). The availability of APIs or software to measure similarity though remains an open issue.

So, our approach to reuse the structure and the software for building a BIOWORDNET was motivated by the fact that we could not find any alternatives coming with a software ensemble as described in Section 2. Against all expectations, we did not manage to reuse the WORDNET data structure. However, there are no publications that report on such difficulties and pitfalls we were confronted with.

6 Discussion and Conclusion

We learnt from our conversion attempt that the current WORDNET representation format of WORDNET suffers from several limitations and idiosyncrasies that cannot be by-passed by a simple, yet ad hoc work-around. Many of the limitations and pitfalls we found limiting (in the sense what can be expressed in WORDNET) are due to the fact that its data format is out-of-date and not really suitable for the biomedical sublanguage. In addition, though we do not take into doubt that the WORDNET software

works fine for the official WORDNET release, our experiences taught us that it fails or gives limited support in case of building and debugging a new WORDNET resource. Even worse, we have evidence from one large terminological resource (NCI) that WORDNET's software infrastructure (GRIND) renders deceptive results.

Although WORDNET might no longerbe the one and only lexical resource for NLP each year a continuously strong stream of publications on the use of WORDNET illustrates its importance for the community. On this account we find it remarkable that although improvements in content and structure of WORDNET have been proposed (e.g., Boyd-Graber et al. (2006) propose to add (weighted) connections between synsets, Oltramari et al. (2002) suggest to restructure WORDNET's taxonomical structure, and Mihalcea and Moldovan (2001) recommend to merge synsets that are too fine-grained) to the best of our knowledge, no explicit proposals have been made to improve the representation format of WORDNET in combination with the adaption of the WORDNET-related software.

According to our experiences the existing WORDNET software is hardly (re)usable due to insufficient error messages that the software throws and limited documentation. From our point of view it would be highly preferable if the software would be improved and made more user-supportive (more meaningful error messages would already improve the usefulness of the software). In terms of the actual representation format of WORDNET we found that using the current format is not only cumbersome and error-prone, but also limits what can be expressed in a WORDNET resource.

From our perspective this indicates the need for a major redesign of WORDNET's data structure foundations to keep up with the standards of today's meta data specification languages (e.g., based on RFD (Graves and Gutierrez, 2006), XML or OWL (Lüngen et al., 2007)). We encourage the reimplementation of WORDNET resources based on such a state-of-the-art markup language (for OWL in particular a representation of WORDNET is already available, cf. van Assem et al. (2006)). Of course, if a new representation format is used for a WORDNET resource also the software accessing the resource has to be adapted to the new format. This may require substantial implementation efforts that we think are worth to be spent, if the new format overcomes the major problems that are due to the original WORDNET format.

Acknowledgments

This work was funded by the German Ministry of Education and Research within the STEMNET project (01DS001A-C) and by the EC within the BOOTSTREP project (FP6-028099).

References

Luisa Bentivogli, Andrea Bocco, and Emanuele Pianta. 2004. ARCHIWORDNET: Integrating WORDNET with domain-specific knowledge. In Petr Sojka, Karel Pala, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2004 – Proceedings of the 2nd International Conference of the Global WordNet Association*, pages 39–46. Brno, Czech Republic, January 20-23, 2004.

Olivier Bodenreider and Anita Burgun. 2002. Characterizing the definitions of anatomical concepts in WORDNET and specialized sources. In *Proceedings of the 1st International Conference of the Global WordNet Association*, pages 223–230. Mysore, India, January 21-25, 2002.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WORDNET. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2006 – Proceedings of the 3rd International* WORDNET *Conference*, pages 29–35. South Jeju Island, Korea, January 22-26, 2006.

Paul Buitelaar and Bogdan Sacaleanu. 2002. Extending synsets with medical terms WORDNET and specialized sources. In *Proceedings of the 1st International Conference of the Global WordNet Association*. Mysore, India, January 21-25, 2002.

Anita Burgun and Olivier Bodenreider. 2001. Comparing terms, concepts and semantic classes in WORDNET and the UNIFIED MEDICAL LANGUAGE SYSTEM. In *Proceedings of the NAACL 2001 Workshop 'WORDNET and Other Lexical Resources: Applications, Extensions and Customizations'*, pages 77–82. Pittsburgh, PA, June 3-4, 2001. New Brunswick, NJ: Association for Computational Linguistics.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of The International Symposium on Reference Resolution for Natural Language Processing*. Alicante, Spain, June 3-4, 2002.

Christiane Fellbaum, Udo Hahn, and Barry Smith. 2006. Towards new information resources for public health:

From WORDNET to MEDICAL WORDNET. *Journal of Biomedical Informatics*, 39(3):321–332.

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

Alvaro Graves and Caludio Gutierrez. 2006. Data representations for WORDNET: A case for RDF. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2006 – Proceedings of the 3rd International* WORDNET *Conference*, pages 165–169. South Jeju Island, Korea, January 22-26, 2006.

Udo Hahn, Martin Romacker, and Stefan Schulz. 1999. Discourse structures in medical reports – watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system. *International Journal of Medical Informatics*, 53(1):1–28.

Sanda M. Harabagiu, Răzvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *NAACL'01, Language Technologies 2001 – Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8. Pittsburgh, PA, USA, June 2-7, 2001. San Francisco, CA: Morgan Kaufmann.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *ACL'07 – Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 680–687. Prague, Czech Republic, June 28-29, 2007. Stroudsburg, PA: Association for Computational Linguistics.

Aleş Horák and Pavel Smrž. 2004. New features of wordnet editor VisDic. *Romanian Journal of Information Science and Technology (Special Issue)*, 7(1-2):201–213.

Harald Lüngen, Claudia Kunze, Lothar Lemnitzer, and Angelika Storrer. 2007. Towards an integrated OWL model for domain-specific and general language WordNets. In Attila Tanács, Dorá Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2008 – Proceedings of the 4th Global* WORDNET *Conference*, pages 281–296. Szeged, Hungary, January 22-25, 2008.

Rada Mihalcea and Dan Moldovan. 2001. EZ.WORDNET: Principles for automatic generation of a coarse grained WORDNET. In *Proceedings of the 14th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pages 454–458.

Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Madolo. 2002. Restructuring WORDNET's top-level. In *Proceedings of ONTOLEX 2002 @ LREC 2002*.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In Alexander F. Gelbukh, editor, *CICLing 2003 – Computational Linguistics and Intelligent Text Processing. Proceedings of the 4th International Conference*, volume 2588 of *Lecture Notes in Computer Science*, pages 241–257. Mexico City, Mexico, February 16-22, 2003. Berlin etc.: Springer.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WORDNET::Similarity: Measuring the relatedness of concepts. In *AAAI'04 – Proceedings of the 19th National Conference on Artificial Intelligence & IAAI'04 – Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference*, pages 1024–1025. San José, CA, USA, July 25-29, 2004. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press.

Michael Poprat and Udo Hahn. 2007. Quantitative data on referring expressions in biomedical abstracts. In *BioNLP at ACL 2007 – Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*, pages 193–194. Prague, Czech Republic, June 29, 2007. Stroudsburg, PA: Association for Computational Liguistics.

Pavel Smrž. 2004. Quality control and checking for wordnets development: A case study of BALKANET. *Romanian Journal of Information Science and Technology (Special Issue)*, 7(1-2):173–181.

D. Tufiş, D. Christea, and S. Stamou. 2004. BALKANET: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology (Special Issue)*, 7(1-2):9–43.

Mark van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of WORDNET to a standard RDF/OWL representation. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, May 22-28, 2006. Paris: European Language Resources Association (ELRA), available on CD.

Piek Vossen, editor. 1998. EUROWORDNET*: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Piek Vossen. 2001. Extending, trimming and fusing WORDNET for technical documents. In *Proceedings of the NAACL 2001 Workshop* '*WORDNET and Other Lexical Resources: Applications, Extensions and Customizations'*. Pittsburgh, PA, June 3-4, 2001. New Brunswick, NJ: Association for Computational Linguistics.