A Multilingual Lexico-Semantic Database and Ontology

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh,
Chu-Ren Huang, Adam Pease and Piek Vossen

Abstract We discuss the development of a multilingual lexicon linked to the SUMO formal ontology. The ontology as well as the lexicon have been expressed in OWL, as well as their original formats, for use on the semantic web and in linked data. We describe the Open Multilingual Wordnet, a multilingual wordnet with twenty two languages and a rich structure of semantic relations. It is made by exploiting links from various monolingual wordnets to the English Wordnet. Currently, it contains 118,337 concepts expressed in 1,643,260 senses in 22 languages. It is available as simple tab separated files, wordnet-LMF or lemon and had been used by many projects including Babelnet and Google translate. We then discuss some issues in extending the wordnets and improving the multi-lingual representation to cover concepts not lexicalized in English and how concepts are stated in the formal ontology.

Key words: Semantic Lexicon, Multilingual, Wordnet, Open Data, Ontology

Francis Bond
Nanyang Technological University, Singapore e-mail: bond@ieee.org

Christiane Fellbaum
Princeton University, USA e-mail: fellbaum@princeton.edu

Shu-Kai Hsieh
National Taiwan University, Taipei e-mail: shukaihsieh@ntu.edu.tw

Chu-Ren Huang
Hong Kong Polytechnic University e-mail: churen.huang@polyu.edu.hk

Adam Pease
Articulate Software, USA e-mail: apease@articulatesoftware.com

Piek Vossen
Vrije Universiteit, Amsterdam e-mail: piek.vossen@vu.nl

Final version before submission: Appears in: Paul Buitelaar and Philipp Cimaniao (eds), 2014, *Towards the Multilingual Web*, Springer-Verlag, pp 243–258 DOI 10.1007/978-3-662-43585-4_15 (pagination is different in the final version)

1 Introduction

What do words mean and how are the words in different languages related? We make a start at answering these questions with a large multilingual lexical database and formal ontology. Each formalism captures knowledge about words and language in a different way. Linked together, they form a unified representation of knowledge suitable for language processing and logical reasoning.

An electronic lexicon is a fundamental resource for computational linguistics in any language, and Princeton's English WordNet (PWN) (Fellbaum, 1998) has become a de facto standard in English computational linguistics. WordNet represents meanings in terms of lexical and conceptual links between concepts and word senses. This allows us to model how concepts are represented in various languages. Ontologies offer a complementary representation where concepts are defined more axiomatically and can be formally reasoned with. The Suggested Upper Merged Ontology (SUMO) model of meaning (Pease, 2011) addresses language-independent concepts, formalized in first- and higher-order logic. Bringing these two models together (Niles & Pease, 2003) has resulted in a uniquely powerful resource for multi-lingual computational processes.

There have been a number of efforts to create wordnets in other languages than English. The EuroWordNet (EWN) project provided a first solution for also connecting these wordnets to each other by introducing a shared InterLingual Index (ILI), (Vossen, 1998). The ILI was based on the English Wordnet (mainly for pragmatic reasons) and was considered as an unstructured fund of concepts for linking synsets across wordnets.

Most wordnets developed since EWN have used PWN as a common pivot to which each new wordnet is linked. This has the drawback of making English a privileged language, and creating a certain linguistic bias. Since all languages have a different set of lexicalized concepts, it is not possible to have an interlingua where everything is lexicalized in all languages. A solution to this was proposed in the ILI using the union of synsets from all languages, arranged and related via the semantic links of PWN (Laparra et al., 2012). In this case, wordnets in the individual languages do not have to lexicalize all synsets but can still be linked together.

Another approach is to use a language-independent formal ontology – SUMO (Pease, 2006) – as the common hub, which allows for the creation of arbitrary new concepts that can eventually encompass the union of lexicalized concepts in all languages. This has additional advantages such as a logical language for creating definitions of concepts that can be checked automatically for logical consistency, and a much larger inventory of possible relations among concepts. Using the ILI as an intermediate approach collects and arranges synsets that are in need of formalization, while defering that effort to a later time. It is hoped that by cataloging these synsets it should be possible to have some of the benefits of a common hub, while speeding construction. This will likely be used as input to full SUMO-based formalizations in the future.

Currently we are exploring both approaches in parallel — creating an ILI (not yet released) and extending SUMO (which has been released and is regularly updated).

A key organizational challenge for a true multilingual lexico-semantic database has been the large-scale nature of the effort needed. Each wordnet project has generally had its own funding and processes, even when coordinated in a broad sense with the original PWN. A variety of formats have proliferated. Wordnets do not all link to one another or a central ontology. Another challenge has been that some wordnets have not been released under open licenses and thus cannot be legally redistributed. This has greatly improved since the initial survey in (Bond & Paik, 2012) with many more wordnets being made open (Bond & Foster, 2013). Some years ago, we introduced the idea of combining wordnets in a single resource [1] (Pease et al., 2008). This original vision has now been realized in the Open Multilingual Wordnet described in Section 4. At the time of this writing, there are 22 wordnets that have been put into a common database format and linked to SUMO.

In the next section we describe the Princeton Wordnet in more detail. We then introduce the linked ontology, SUMO (§ 3). In the next section we describe how we built and made accessible the open multilingual wordnet: the main new resource described here (§ 4). Finally we discuss how it can be extended to cover more languages better (§ 5).

2 Princeton English WordNet

Princeton WordNet (PWN: Fellbaum, 1998) is a large lexical database comprising nouns, verbs, adjectives and adverbs. Cognitively synonymous word forms are grouped into **synsets**, each expressing a distinct concept. Within each synset, words are linked by synonymy. Synsets are interlinked by means of lexical relations (among specific word forms) and conceptual relations (among synsets). Examples of the former are antonymy and the morphosemantic relation; examples of the latter are hyponymy, meronymy and a set of entailment relations. The resulting network can be navigated to explore semantic similarity among words and synsets. PWN's graph structure allows one to measure and quantify semantic similarity by simple edge counting; this makes PWN a useful tool for computational linguistics and natural language processing.

The main relation among words in PWN is synonymy, as between the words *shut* and *close* or *car* and *automobile*. A group of synonyms – words that denote the same concept and are interchangeable in many contexts – is grouped into an unordered set. Synsets are linked to other synsets by means of a small number of **conceptual relations**, such as **hyperonymy**, **meronymy** and **entailment**. Additionally, each synset contains a brief definition and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented by appearing in as many distinct synsets as there are meanings. Thus, each form-meaning pair (or **sense**) in PWN is unique.

[1] `http://www.globalwordnet.org/gwa/gwa_grid.html`

3 SUMO

The Suggested Upper Merged Ontology[2] (Niles & Pease, 2001; Pease, 2011) began as just an upper level ontology encoded in first order logic. The logic has expanded to include higher order elements. SUMO itself is now a bit of a misnomer as it refers to a combined set of theories: (1) the original upper level, consisting of roughly 1,000 terms, 4,000 axioms and some 750 rules. (2) A MId-Level Ontology (MILO) of several thousand additional terms and axioms that define them, covering knowledge that is less general than those in the upper level. We should note that there is no objective standard for what should be considered upper level or not. (3) There are also a few dozen domain ontologies on various topics including theories of economy, geography, finance and computing. Together, all ontologies total roughly 22,000 terms and 90,000 axioms. There are also an increasing group of ontologies which are theories that consist largely of ground facts, semi-automatically created from other sources and aligned with SUMO. These include YAGO (de Melo et al., 2008), which is the largest of these sorts of resources and has millions of facts.

SUMO is defined in the SUO-KIF language,[3] which is a derivative of the original KIF (Genesereth, 1991). It has been translated automatically, although in what is a necessarily very lossy translation into the W3C Web Ontology Language (OWL).[4] The translation also includes a version of PWN in OWL, [5] and the mappings between them.[6]

SUMO proper has a significant set of manually created language display templates that allow terms and definitions to be paraphrased in various natural languages. These include Arabic, French, English, Czech, Tagalog, German, Italian, Hindi, Romanian, and Chinese (traditional and simplified characters).

SUMO has been mapped by hand to the entire PWN lexicon (Niles & Pease, 2003). The mapping statistics are given in Table 1. There are a number of other approaches for mapping ontologies to wordnets (Fellbaum & Vossen, 2012; Vossen & Rigau, 2010). However these have not involved ontologies that are either comparable in size or degree of formalization to SUMO.

---

[2] www.ontologyportal.org

[3] http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/sigma/suo-kif.pdf

[4] http://www.ontologyportal.org/SUMO.owl

[5] http://www.ontologyportal.org/WordNet.owl

[6] http://sigma-01.cim3.net:8080/sigma/OWL.jsp?kb=SUMO also provides a ”live” generation of OWL one term at a time, where ”&term=name” can be appended to the URL and the desired term name substituted for ”name”.

	instance	equivalence	subsuming
noun	9,837	3,329	68,919
verb	0	600	13,150
adj	724	540	14,771
adverb	57	99	3,235
total	10,618	4,568	100075

Table 1 SUMO WordNet mappings (115,261 total)

4 Open Multilingual Wordnet

Wordnets have now been made for many languages. The Global Wordnet Association currently lists over 60 wordnets.[7] The individual wordnets are the result of many different projects and vary greatly in size and accuracy. The Open Multilingual Wordnet (OMW)[8] provides access to some of these, all linked to the PWN and SUMO. The goal is to make it easy to access lexical meaning in multiple languages. OMW has (i) extracted and normalized the data, (ii) linked it to PWN 3.0 and (iii) put it in one place. It includes a simple search interface that uses the SQL database developed by the Japanese Wordnet.

In order to make the wordnets more **accessible**, we have built a simple server with information from those wordnets whose licenses allows us to do so. It is based on a single shared database with all the languages in it. We only include data that is open: "anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike".[9]

The accessibility of the data means that it is becoming widely used. Babelnet 2.0,[10] a very large multilingual encyclopedic dictionary and semantic network, is made by combining the OMW, PWN, Wikipedia and Omegawiki (a large collaborative multilingual dictionary). Google Translate[11] also uses the OMW data.

The majority of freely available wordnets have been based on the **expand** approach, basically adding lemmas in new languages to existing PWN synsets (Vossen, 1998, p11). These wordnets can easily be combined by using the PWN as a pivot. We realize that this is an incomplete solution and a better one is discussed in Section 5.2. Some wordnets are based on the **merge** approach, where independent language specific structures are built first and then some synsets linked to the PWN. For those merged wordnets in the OMW (Danish and Polish), only a small subset are actually linked, due more to lack of resources to link them than semantic incompatibility.

Adding a new language to the OMW turned out to be difficult for two reasons. The first problem was that the wordnets were linked to various versions of PWN.

[7] `http://globalwordnet.org/`

[8] `http://compling.ntu.edu.sg/omw`

[9] Definition from the Open Knowledge Foundation: `http://opendefinition.org/`.

[10] `http://babelnet.org/about.jsp`

[11] `http://translate.google.com/about/intl/en_ALL/`

In order to combine them into a single multilingual structure, we had to map to a common version. The second problem was the incredible variety of formats that the wordnets are distributed in. Almost every project used a different format and thus required a new script to convert it. In fact, different releases from the same project often had slightly different formats. These two problems mean that, even if a wordnet is legally available, there is still a technical hurdle before it becomes easily accessible.

The first problem can largely be overcome using the mappings from Daude et al. (2003). Mapping introduces some distortions. In particular, when a synset is split, we chose to only map the translations to the most probable mapping, so some new synsets will have no translations. For example, the synset pwn16-***leg***$_{n:8}$ "a section or portion of a journey or course" in PWN 1.6 maps to two senses in PWN 3.0: pwn30-***leg***$_{n:9}$ "a section or portion of a journey or course" and pwn30-***leg***$_{n:8}$ "the distance traveled by a sailing vessel on a single tack". pwn16-***leg***$_{n:8}$ to pwn30-***leg***$_{n:9}$ is the most probable mapping, so any lemmas associated with pwn16-***leg***$_{n:8}$ will be associated only with pwn30-***leg***$_{n:9}$.

The second problem we have currently solved through brute force, writing a new script for every new wordnet we add. We discuss better possible solutions in Section 5.2. In the future, we hope people will move to a common standard for exchange, with Wordnet-LMF being the strongest contender (Vossen et al., 2013).

The server currently includes English (Fellbaum, 1998); Albanian (Ruci, 2008); Arabic (Black et al., 2006); Chinese (Huang et al., 2010; Wang & Bond, 2013);[12] Danish (Pedersen et al., 2009); Finnish (Lindén & Carlson., 2010); French (Sagot & Fišer, 2008); Hebrew (Ordan & Wintner, 2007); Indonesian and Malaysian (Nurril Hirfana et al., 2011); Italian (Pianta et al., 2002); Japanese (Isahara et al., 2008); Norwegian (Bokmål and Nynorsk: Lars Nygaard 2012, p.c.); Persian (Montazery & Faili, 2010); Polish (Piasecki et al., 2009); Portuguese (de Paiva & Rademaker, 2012); Thai (Thoongsup et al., 2009) and Basque, Catalan, Galician and Spanish from the Multilingual Common Repository (Gonzalez-Agirre et al., 2012).

The wordnets are all in a shared `sqlite` database with either `Python` or `PERL` cgi clients using the wordnet module produced by the Japanese Wordnet project (Isahara et al., 2008). The database is based on the logical structure of the PWN, with an additional language attribute for lemmas, examples, definitions and senses. It is thus effectively a single open multilingual resource. We summarize the size of the wordnets and their coverage of **core concepts** in Table 2. Core concepts are the 5,000 synsets proposed as a core lexicon based on the frequency of the word forms in the British National Corpus (Burnard, 2000) and an intuitive sense of salience (Boyd-Graber et al., 2006). That is, the core concepts are frequently occurring concepts (at least in British English).

We make available the synset-lemma pairs as tab separated files, where they can be used by the Natural Language Tool Kit[13] (Bird et al., 2009) as well as WordNet-

[12] Users see the union of the data from the two Chinese wordnets.

[13] With the extensions that were added with the Japanese translation by Masato Hagiwara (Bird et al., 2010).

Wordnet	Lang	Synsets	Words	Senses	Core	Licence
Albanet	als	4,676	5,990	9,602	31%	CC BY 3.0
Arabic WordNet (AWN)	arb	10,165	14,595	21,751	48%	CC BY SA 3.0
Chinese Wordnet (Taiwan)	cmn	4,913	3,206	8,069	28%	wordnet
Chinese Open Wordnet	cmn	42,316	61,536	79,812	99%	wordnet
DanNet	dan	4,476	4,468	5,859	81%	wordnet
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet
Persian Wordnet	fas	17,759	17,560	30,461	41%	Free to use
FinnWordNet	fin	116,763	129,839	189,227	100%	CC BY 3.0
WOLF	fra	59,091	55,373	102,671	92%	CeCILL-C
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet
MultiWordNet	ita	34,728	40,343	61,558	83%	CC BY 3.0
Japanese Wordnet	jpn	57,179	91,959	158,064	95%	wordnet
Multilingual	cat	45,826	46,531	70,622	81%	CC BY 3.0
Central	eus	29,413	26,240	48,934	71%	CC BY-NC-SA 3.0
Repository	glg	19,312	23,124	27,138	36%	CC BY 3.0
(MCR)	spa	38,512	36,681	57,764	76%	CC BY 3.0
Wordnet Bahasa	ind	51,755	64,948	142,488	99%	MIT
Wordnet Bahasa	zsm	42,615	51,339	119,152	99%	MIT
Norwegian Wordnet	nno	3,671	3,387	4,762	66%	wordnet
Norwegian Wordnet	nob	4,455	4,186	5,586	81%	wordnet
plWordNet	pol	14,008	18,860	21,001	30%	wordnet
OpenWN-PT	por	41,810	52,220	68,285	79%	CC by SA 3.0
Thai Wordnet	tha	73,350	82,504	95,517	81%	wordnet

Table 2 Available Wordnets

LMF (Lexical Markup Framework: Vossen et al., 2013) and lemon (McCrae et al., 2011).[14]

Finally, we also make the SQL database available (with all languages except French and Basque, whose licenses are incompatible with the others). We use a simple database schema extended from the schema for the Japanese wordnet (Bond et al., 2009). When we use the combined database in applications, we typically use the database directly, or through the Perl interface. Licenses that allow redistribution of derivative works allow people to make the entire lexicons available in any format, thus greatly improving their usefulness. There are also APIs for the database produced by other researchers in Python, Java, Ruby, Objective-C, Gauche and an alternative Perl module.[15]

There has been much research on making Wordnets available to the semantic web, including formatting as RDF (van Assem et al., 2006; Koide et al., 2006), serving LMF directly (Savas et al., 2010) or serving them through the lemon format (McCrae et al., 2011). Typically, these do not involve any changes in the actual content, the emphasis is instead on making it more easily accessible as Linked Open Data (Berners-Lee, 2009). The proliferation of these approaches suggests that there is still some way to go until we will have an agreed upon universal standard. Therefore, our approach has been to make our data open, clearly documented, well-

[14] Thanks to John P. McCrae for help in adding this.

[15] `http://nlpwww.nict.go.jp/wn-ja/index.en.html`

formatted and validated in a simple format we use ourselves (tab separated text) and some standard formats for exchange (LMF and lemon). This can then be straightforwardly converted to whatever format is desired by those who want it in that format. Currently, in most of our use scenarios (principally word sense disambiguation and semantic processing) the latency of a web interface is problematic — we expect that most of the users of our data will want to download the entire lexicon, and this is what we offer.

Possible Wordnet Structural Enhancements

In this section we will discuss some extensions people have suggested to the structure of the original PWN: these are not currently part of the open wordnet. One advantage of having many language specific projects loosely coordinated is that there can be a wide variety of experimentation.

Our conversion scripts basically reduce each wordnet to a list of synset-lemma pairs, plus frequency, definitions and examples if available. Everything is mapped to PWN 3.0 synsets. Therefore, the current version loses any synsets not in the English 3.0 wordnet. Many of the wordnets have such synsets, as well as meta-data, definitions, examples and other useful information. One of the ongoing goals of the OMW project is to make this information more easily accessible between projects.

We do not consider wordnets with licenses that do not allow redistribution, as we cannot legally include them. This includes some very well constructed wordnets with excellent coverage, such as the Dutch,[16] German and Korean wordnets (Vossen et al., 2008; Kunze & Lemnitzer, 2002; Yoon et al., 2009). It is unfortunate that they cannot be integrated into the Open Wordnet. Some wordnets are built with their own structure and do not link to the PWN. These also cannot be included. Finally, some wordnets were not included even though they were open as the quality was still too poor due to the fact that they had been automatically made, with very little quality control.

Many of the wordnet projects extend the PWN relations in some way. For example, EWN defined many cross-part-of-speech links: ***hammer***$_{n:1}$ is an involved-role of ***hammer***$_{v:1}$ (Vossen, 1998, pp97–110). Another instance of extensions is the Chinese Wordnet (Taiwan) which takes a different approach in representing lexical meanings. Unlike most models of lexical ambiguity resolution that assume only one meaning is chosen in a given context, it allows more than one (related) meanings to co-exist in the same context. A lexical item is **actively complex** if it allows simultaneous multiple readings.[17] Meaning extensions thus are proposed to be distin-

[16] We are delighted to see that an Open Dutch Wordnet will be released soon (Vossen & Postma, 2014) and will integrate it as soon the data is available.

[17] Note that according to psycholinguistic studies from Ahrens et al. (1998), there are two types of active complexity in natural language. The first is 'triggered complexity' initiated by the speaker that involve puns; the second is 'latent complexity' in which no pun or vagueness is intended. The Chinese Wordnet's model focuses only on latent complexity.

guished between two types: **sense** and **meaning facet** (Ahrens et al., 1998). These can be distinguished as follows: given multiple possible meanings of a lemma, if a sentence that allows co-existing multiple readings for that lemma can be found, the distinction of these meanings are recognized as **meaning facet** distinction, otherwise they are **sense** distinctions. The **co-existence test** for sense/meaning facet distinction can be illustrated in (1–4). The lemma *kànbìng* "seeing-sickness" in (1) allows two readings ("seeing the doctor" or "examining the patient"). The ambiguity can be resolved given more contextual information and we can not find a sentence that allows the co-existence of these two readings. Therefore, it is treated as two senses of that lemma. However, for the lemma *zázhì* "magazine", it can refer to the `physical object` in (2), or the `information contained` in (3), more specifically, we can find a sentence like (4) in which the meaning of the lemma can refer to both the `physical object` and the `information` contained in that object. We therefore consider this meaning distinction of *zazhi* "magazine" is a **meaning facet** rather than a **sense**. Interestingly, among the 5,890 meaning facets being identified in Chinese Wordnet, 9 regular systematic patterns are extracted, which are similar to the regular polysemy (Apresjan, 1973) (of complex types) proposed by Pustejovsky (1995). This fine grained distinction is implemented by extending the types of semantic relations within the Chinese wordnet. Many (perhaps most) of these relations are not specific to Chinese. One of the advantages of the OMW is that we can look at research like this being done for one language, and easily test its applicability to other languages.

(1) 他 正在 看病
tā zhèngzài kànbìng
he PROG seeing-sickness
'He is seeing the doctor./He is examining the patient'

(2) 他 手 上 拿 了 本 雜誌
tā shǒu shàng ná le běn zázhì
he hand on hold asp. CL magazine
'He is holding one magazine in his hand.'

(3) 他 在 讀 那 一 本 雜誌
tā zài dú nà yì běn zázhì.
he PROG read that one CL magazine.
'He is turning the pages of the magazine and reading it.'

(4) 他 拿 一 本 雜誌 給 我 看
tā ná yì běn zázhì gěi wǒ kàn
he takes one CL magazine give me read
'He passed me a magazine (to read).'

5 Extending the Multilingual Wordnet

In this section we discuss the immediate plans to extend the wordnets to deal with multilingual issues. As was demonstrated in EWN, we can expect most languages to have concepts that are not lexicalized in English. In addition, there are still many concepts lexicalized in English, but not in PWN. Thus different wordnets will have synsets that do not appear in most or even any other existing wordnet (this was the case for seven of the wordnets in the OMW). Consider the example of the Tagalog word *hilamos – to wash one's face* (Borra et al., 2010).

Words such as this form part of the motivation for using a formal ontology. While some wordnets have used English as an interlingua and created phrases to stand in the place of otherwise unlexicalized concepts, another approach is to use SUMO as an interlingua which can contain concepts which stand for the lexicalized concepts of any particular language.

Exactly what counts as lexicalized can be hard to determine. Consider the following example: ***foal*** is lexicalized in English so must be in the English Wordnet. In Malay, the closest equivalent is a phrase: *anak kuda* "horse child" which can be produced compositionally by fully productive syntactic rules. In Japanese it is *kouma* "child+horse" a word produced by a semi-productive process. So it is not clear whether the Malay wordnet should have an entry here. On the one hand, it is produced by a fully productive process. On the other, it is useful to have an entry, even if fully compositional, for completeness. We suggest that it should be entered but marked as syntagmatic using meta-data, following the example of Italian, Basque and Hungarian wordnets (Pianta et al., 2002; Pociello et al., 2011). Vincze & Almázi (2014) show how it is possible to exploit this meta-data to automatically make two versions of the monolingual wordnets — one showing translation equivalents and one only showing concepts lexicalized in particular language.

EWN distinguished a few types of non-universal lexicalisations and expressions, which call for different methods of handling:

cultural concepts concepts that exist in some cultures and not in others, e.g. Dutch *klunen=to walk on skates*

pragmatic lexicalisations concepts that are known in all/most cultures but are not considered lexicalised in all of these, e.g. we all know the concept of a small fish but Spanish happens to have a separate word for it *alevin*

morpho-syntactic mismatches concepts that are lexicalised through words with different morpho-syntatic properties across languages, e.g. Dutch has no equivalence for *like* but uses the adjective *aardig*

differences in perspective some languages distinguish things depending on who is doing what to whom in ways that other languages don't, e.g. *teach* and *learn* in English whereas French uses *apprendre* for both.

A pertinent question is what defines a word and what defines a concept. Commonly occurring collocations may have transparent, compositional semantics, yet we may still consider these words. For example, noun compounds such as *sailing*

boat are so common and ready made that we consider them to be one word. Another point is that the relation between the components cannot be predicted from the structure: who is doing the sailing, who has the sail and what is being sailed? A classical Dutch example is *kindermeel: meal for children* and *tarwemeel: flour made of oats*. From the structure, we cannot infer the relation. It needs to be learned or inferred but Dutch speakers are probably not deriving them over and over again.

We are also extending the wordnets in terms of their size and coverage both within individual projects and by exploiting the disambiguating power of multilingual data to link to other open resources such as Wiktionary (Bond & Foster, 2013). The core idea is that by looking at multiple translations of a concept, we can pin-point the meaning exactly: ***bat*** in English is ambiguous between the sporting equipment and the flying mammal, but adding, e.g. French removes the ambiguity (***batte*** vs ***chauve-souris***).

We are investigating two (compatible) methods of dealing with these new concepts. One is to create a concept in an external ontology and use this to link languages. In this approach, as *hilamos* is not lexicalized in English, it is not linked directly to English *wash* in the English wordnet. The fundamental value of the ontology is to define meaning using axioms in an expressive logic so that the meanings can then be manipulated without recourse to a human's intuition about the meaning of a word.

The second approach is to have a shared group of synsets for all languages, but not have them lexicalized in all languages. In this model English *wash* and Filipino *hugas* are both lexicalizations of the same synset, and the synset for *hilamos* "wash one's face" inherits from this, but would be marked as unlexicalized in English. Most **expand** style wordnets take this approach with non-lexicalized synsets being either just left blank, or explicitly marked as non-lexicalized (as in, for example the MCR (Gonzalez-Agirre et al., 2012)).

5.1 Wordnets linked to external Ontologies

Using ontologies[18] to link words (the first approach) is more labor intensive, but offers other advantages.

Consider the notion of ***earlier***. PWN has a synset for this word, but not a way to use it in temporal inference. SUMO however has a relation for earlier, and a formal rule (among others) that allows an automated inference system such as those available with Sigma (Pease & Benzmüller, 2013; Pease et al., 2010) to conclude that an interval that is earlier than another has an endpoint that precedes the start point of the following interval. This is a necessary and sufficient definition for `earlier` and uses the bi-implication or equivalence sign `<=>`.

[18] It would be possible to link ontologies other than SUMO. There are other ontologies with at least partial links to wordnet, including DOLCE (Gangemi et al., 2003) and the Kyoto Ontology (Laparra et al., 2012). We only discuss SUMO here, as it is both the largest ontology and the most fully integrated with the OMW.

```
(<=>
  (earlier ?INTERVAL1 ?INTERVAL2)
  (before
    (EndFn ?INTERVAL1)
    (BeginFn ?INTERVAL2)))
```

Another example is the SUMO-based content developed to represent Muslim cultural concepts in Arabic Wordnet (Black et al., 2006). The ***Udhiyah*** ritual is performed during the period of Eid-Aladha and involves slaughtering a lamb by a Muslim. If a lamb has the attribute of being `Udhiyah` then there necessarily exists an `UdhiyahRitual` in which it is the subject of the ritual.

```
(=>
  (instance ?UR UdhiyahRitual)
  (exists (?S ?EA ?P)
    (and
      (instance ?EA EidAladha)
      (during ?UR ?EA)
      (attribute ?S Udhiyah)
      (agent ?UR ?P)
      (attribute ?P Muslim)
      (patient ?UR ?S))))
```

```
(=>
  (attribute ?S Udhiyah)
  (exists (?UR)
    (and
      (instance ?S Lamb)
      (instance ?UR UdhiyahRitual)
      (patient ?UR ?S))))
```

Each of these symbols is further formalized, allowing them to be checked for logical consistency by automated theorem provers. This is also a key advantage for formal logic representation. The more expressive the representation, and the more extensive the set of formalizations for each concept, the more things that can be checked automatically. A conventional dictionary must be checked by humans to ensure correctness of definitions. This is true with a conventional data dictionary, in which concepts in a database are defined in natural language in hopes of ensuring their correct usage. But when such a corpus of definitions grows large, into the thousands or more, it is not likely that a human or even many humans will be able to find all inconsistencies. Automated means are needed. At that point, expressiveness also matters. In a taxonomy, the only error that can be caught automatically is the presence of a cycle in the graph. With a description logic, many more checks can be performed. In a higher-order language such as that used by SUMO, theorem proving (Benzmüller & Pease, 2010) can find much more deep and subtle errors, leading to definitions of considerable depth and consistency.

Because SUMO terms are mathematical symbols, with a semantics given solely by their logical axioms, and unlike taxonomies or semantic networks, the symbol names can be changed without altering their meaning. In fact, the current Sigma browser can display terms with their names in different languages, in order to emphasize this point, and make them more accessible to logicians who may not speak English.

5.2 InterLingual Index

The second approach is basically that of the **Interlingual Index** (ILI: Peters et al., 1998). The variety of approaches in the EWN initially resulted in wordnets that were mapped to very different sets of concepts in the ILI. Likewise, only a small set of synsets could be traced to other languages through the ILI. To harmonize the output, EWN took two measures: (i) the definition of a shared set of (1,000 up to 5,000) Base Concepts that were manually aligned, and (ii) the classification of these Base Concepts using a small top-ontology of 63 terms. Base Concepts (not to be confused with the ”Basic Level Categories” of Rosch (1978)) represent synsets that have the highest connectivity to the other synsets. The top-ontology classification of these synsets provided a shared semantic framework. Each wordnet made sure the Base Concepts were presented properly in their language and manually mapped to the ILI. The minimal intersection across these wordnets through the ILI is thus the set of Base Concepts but in practice the intersection is much larger. During the EWN project, it became clear that there are many problems with the ILI being based on PWN and that there are many possibilities to improve the ILI for linking wordnets (Vossen et al., 1999).

6 Conclusion

Several goals are being pursued in parallel: (i) research on building wordnets for individual languages; (ii) research on building a more formal upper ontology; (iii) research on linking wordnets in many languages to make a multilingual resource. The ontology as well as some of the lexicons have been expressed in OWL, as well as their original formats, for use on the semantic web and in linked data. This effort builds on WordNet, Global Wordnet, and SUMO to create a rich web of linguistic data and mathematically specified world knowledge.

References

Ahrens, K., Chang, L. L., Chen, K. J., & Huang, C.-R. (1998). Meaning representation and meaning instantiation for Chinese nominals. In *Int. J. Computational Linguistics and Chinese Language Processing*, vol. 3, (pp. 45–60).

Apresjan, J. (1973). Regular polysemy. *Linguistics*, *142*(5), 5–32.

Benzmüller, C., & Pease, A. (2010). Progress in automating higher-order ontology reasoning. In B. Konev, R. Schmidt, & S. Schulz (Eds.) *Workshop on Practical Aspects of Automated Reasoning (PAAR-2010)*. Edinburgh, UK: CEUR Workshop Proceedings.

Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly. www.nltk.org/book.

Bird, S., Klein, E., & Loper, E. (2010). *Nyumon Shizen Gengo Shori [Introduction to Natural Language Processing]*. O'Reilly. (translated by Hagiwara, Nakamura and Mizuno).

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic WordNet Project. In *Proceedings of the Third International WordNet Conference*. Choi, Fellbaum and Vossen: Sojka.

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, (pp. 1352–1362). Sofia.
URL http://aclweb.org/anthology/P13-1133

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009). Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, (pp. 1–8). Singapore: ACL-IJCNLP 2009.

Bond, F., & Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Borra, A., Pease, A., Roxas, R., & Dita, S. (2010). Introducing Filipino WordNet. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.) *Principles of Construction and Application of Multilingual Wordnets: Proceedings of the 5th Global WordNet Conference*, (pp. 306–310). Mumbai, India: Narosa pub.

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*. Jeju.

Burnard, L. (2000). *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Daude, J., Padro, L., & Rigau, G. (2003). Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria.

de Melo, G., Suchanek, F., & Pease, A. (2008). Integrating YAGO into the Suggested Upper Merged Ontology. *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence*.

de Paiva, V., & Rademaker, A. (2012). Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fellbaum, C., & Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, *46*(2), 313–326. Doi=10.1007/s10579-012-9186-z.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI Magazine*, *24*(3), 13–24.

Genesereth, M. (1991). Knowledge interchange format. In J. Allen, R. Fikes, & E. Sandewall (Eds.) *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, (pp. 238–249). Morgan Kaufman.

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., & Huang, S.-W. (2010). Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, *24*(2), 14–23. (in Chinese).

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Koide, S., Morita, T., Yamaguchi, T., Muljadi, H., & Takeda, H. (2006). OWL expressions on WordNet and EDR. In *AI society Semantic Web Ontology SIG 13*, SIG-SWO-A601-03. (in Japanese).
URL `http://www.jaist.ac.jp/ks/labs/kbs-lab/sig-swo/fpapers.%htm`

Kunze, C., & Lemnitzer, L. (2002). Germanet — representation, visualization, application. In *LREC*, (pp. 1485–1491).

Laparra, E., Rigau, G., & Vossen, P. (2012). Mapping wordnet to the kyoto ontology. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC2012)*, (pp. 2584–2589). Publ. European Language Resources Association (ELRA).

Lindén, K., & Carlson., L. (2010). Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, *17*, 119–140. In Swedish with an English abstract.

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, (pp. 245–259).

Mohamed Noor, N., Sapuan, S., & Bond, F. (2011). Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, (pp. 258–267). Singapore.

Montazery, M., & Faili, H. (2010). Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, (pp. 846–850).

Niles, I., & Pease, A. (2001). Toward a Standard Upper Ontology. In C. Welty, & B. Smith (Eds.) *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, (pp. 2–9).

Niles, I., & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, (pp. 412–416).

Ordan, N., & Wintner, S. (2007). Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, *19*(1), 39–58.

Pease, A. (2006). Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics. In *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. New York: Mouton de Gruyter.

Pease, A. (2011). *Ontology: A Practical Guide*. Angwin, CA: Articulate Software Press.

Pease, A., & Benzmüller, C. (2013). Sigma: An Integrated Development Environment for Logical Theories. *AI Communications, 26*, 9–97.

Pease, A., Fellbaum, C., & Vossen, P. (2008). Building the Global WordNet Grid. In *Proceedings of the CIL-18 Workshop on Linguistic Studies of Ontology*. Seoul, South Korea.

Pease, A., Sutcliffe, G., Siegel, N., & Trac, S. (2010). Large Theory Reasoning with SUMO at CASC. *AI Communications, Special issue on Practical Aspects of Automated Reasoning, 23*(2-3), 137–144.

Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.

Peters, W., Vossen, P., Díez-Orzas, P., & Adriens, G. (1998). Cross-linguistic alignment of wordnets with an inter-lingual-index. In Vossen (1998), (pp. 149–251).

Pianta, E., Bentivogli, L., & Girardi, C. (2002). Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, (pp. 293–302). Mysore, India.

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. (ISBN 978-83-7493-476-3).
URL `http://www.plwordnet.pwr.wroc.pl/main/content/files/pub%lications/A_Wordnet_from_the_Ground_Up.pdf`

Pociello, E., Agirre, E., & Aldezabal, I. (2011). Methodology and construction of the Basque wordnet. *Language Resources and Evaluation, 45*(2), 121–142.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.) *Cognition and Categorization*, (pp. 27–48). Hillsdale (NJ), USA: Lawrence Erlbaum Associates. Reprinted in *Readings in Cognitive Science. A Perspective from Psychology and Artificial Intelligence*, A. Collins and E.E. Smith, editors, Morgan Kaufmann Publishers, Los Altos (CA), USA, 1991.

Ruci, E. (2008). On the current state of Albanet and related applications. Tech. rep., University of Vlora. (`http://fjalnet.com/technicalreportalbanet.pdf`).

Sagot, B., & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In E. L. R. A. (ELRA) (Ed.) *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Savas, B., Hayashi, Y., Monachini, M., Soria, C., & Calzolari, N. (2010). An lmf-based web service for accessing wordnet-type semantic lexicons. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokarat, C., Sornlertlamvanich, V., & Isahara, H. (2009). Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint confer-*

ence of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),. Suntec, Singapore.

van Assem, M., Gangemi, A., & Schreiber, G. (2006). Conversion of wordnet to a standard RDF/OWL representation. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Vincze, V., & Almázi, A. (2014). Non-lexicalized concepts in wordnets: A case study of English and Hungarian. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, (pp. 118–126). Tartu.

Vossen, P. (Ed.) (1998). *Euro WordNet*. Kluwer.

Vossen, P., Maks, I., Segers, R., & Van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC 2008*. Marrakech, Morocco: European Language Resources Association (ELRA).

Vossen, P., Peters, W., & Gonzalo, J. (1999). Towards a universal index of meaning. In *Proceedings of ACL-99 Workshop, Siglex-99, Standardizing Lexical Resources*, (pp. 81–90). Maryland.

Vossen, P., & Postma, M. (2014). Open Dutch wordnet. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. (presentation only).

Vossen, P., & Rigau, G. (2010). Division of semantic labor in the global wordnet grid. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.) *5th Global Wordnet Conference: GWC-2010*. Mumbai: Narosa Pub.

Vossen, P., Soria, C., & Monachini, M. (2013). LMF - lexical markup framework. In G. Francopoulo (Ed.) *LMF - Lexical Markup Framework*, chap. 4. ISTE Ltd + John Wiley & sons, Inc.

Wang, S., & Bond, F. (2013). Building a Chinese wordnet: Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*. Nagoya.

Yoon, A., Hwang, S., Lee, E., & Kwon, H.-C. (2009). Construction of Korean wordnet KorLex 1.5. *Journal of KIISE: Software and Applications*, *36*(1), 92–108.