Bringing together over- and under-represented languages: Linking Wordnet to the SIL Semantic Domains

Muhammad Zulhelmy bin Mohd Rosman
Francis Bond and **František Kratochvíl**
Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore
muhammad20@e.ntu.edu.sg, bond@ieee.org, fkratochvil@ntu.edu.sg

Abstract

We have created an open-source mapping between the SIL's semantic domains (used for rapid lexicon building and organization for under-resourced languages) and WordNet, the standard resource for lexical semantics in natural language processing. We show that the resources complement each other, and suggest ways in which the mapping can be improved even further. The semantic domains give more general domain and associative links, which wordnet still has few of, while wordnet gives explicit semantic relations between senses, which the domains lack.

1 Introduction

In this paper we compare, and semi-automatically link using Python with NLTK (Bird et al., 2009), two very different approaches to organizing lexical knowledge. The first is the **Semantic Domains** (**SD**) from SIL International.[1] **SD** is a tool designed to aid in the rapid construction and subsequent organization of lexicons for languages which may have no dictionary at all. The second is the linked concepts from the **wordnet** (**WN**) lexical databases, largely based on the Princeton WordNet of English (Fellbaum, 1998). This lexical database was designed to be consistent with models of how human beings process language and is now widely used in natural language processing.

SD is a standard tool in development of dictionaries for under-resourced languages. Wordnets on the other hand, are primarily built for languages that already have many lexical resources, such as English, Japanese and Finnish (Bond and Paik, 2012).

SD is designed for rapid construction and intuitive organization of lexicons, not primarily for the analysis of the resulting data. As a result, many potentially interesting relationships are only implicitly realized. By linking **SD** to **WN** we can take advantage of the relationships modeled in **WN** to make more of these explicit. For example, the semantic relations in **WN** would be a useful input into **SD** while the domains hierarchy would enforce the existing **WN** relations. This will allow more quantitative computational modeling of under-resourced languages.

It is currently an exciting time for field lexicography with better tools and hardware allowing for rapid digitization of lexical resources. Typically, linguists tag text soon after they collect it. As semantic tags are integrated into the workflow, the new words are instantly linked to structured data. We will make it possible to then link them to languages with fuller descriptions and formal ontologies.

In the following sections we introduce the resources in more detail (Section 2), then describe the automatic mapping (Section 3). The results of the mapping are presented (Section 4) and discussed (Section 5). Both **SD** and **WN** are freely available under open licenses, and we release our mapping in the same way (licensed with the Creative Commons Attribution License (CC-BY).[2]

2 Resources

In this section we introduce the resources. As WordNet is more established in the field of computational linguistics, we will mainly describe the semantic domains.

[1] "SIL International is a [Christian] faith-based nonprofit organization committed to serving language communities worldwide as they build capacity for sustainable language development." http://sil.org

[2] See http://creativecommons.org/licenses/by/3.0/

2.1 Semantic Domains (SD)

SD is a standard tool in descriptive linguistics aiding in dictionary building and organization. It comprises of nine major headings where similar domains are placed close to each other. We show the two upper levels in Figure 2.[3] There are several versions in circulation for various regional languages, the latest version is DDP.v4, on which **SD** is built. **SD** draws on a number of thesauri developed as tools for historical linguists (enabling them to track words despite sound change or meaning shift). An excellent example of such approach is Buck (1949), which is a dictionary of synonyms in principal Indo-European languages. It contains more than 1,100 clusters of synonyms grouped into 172 domains, listing related words and reviewing their etymology and semantic history. It allows to detect changes in meaning and replacement of older forms by newer forms, of colloquial or foreign origin. **SD** are also informed by English lexicographic resources, including the 20,000 most frequent words from the Corpus of Contemporary American English (450m words).

Multilingual versions of **SD** are available, covering currently besides English also Chinese, French, Hindi, Indonesian, Khmer, Nepali, Russian, Spanish, Telugu, Thai, and Urdu.

SD has been built into several standard software tools for language documentation and description such as SIL Toolbox, SIL FieldWorks, and WeSay (Moe, 2013).[4]

Each domain includes:

- a number for sorting purposes
- a domain label (consisting of a word or short phrase that captures the basic idea of the domain)
- a short description of the domain
- a series of questions designed to help people think of the words that belong to the domain
- a short list of words under each question that belong to the domain.

We show examples of the domains in Figure 1. The semantic domains are released under an open source license — Creative Commons Attribution-ShareAlike license (CC-BY-SA).

There are no explicit relational links between the domains, although the most common tool used with it (FLEX[5]) allows for the addition of **hypernym/hyponym**, **meronym/holonym**, **antonym/synonym** and **calendar** relations. We show more detailed of a group of domains in Figure 1. The relations between super and sub domains is generally random. Within each domain questions are designed to elicit words associated with the domain, and these can be related in almost any way.

2.1.1 Users

We took a survey among the users of SIL Toolbox and SIL Fieldworks on the respective online fora. Among the 12 respondents, DDP is mainly used to build dictionaries (72%), organize them (63%), and let native speakers enrich them (54%). The option to produce language materials is also valued. Most respondents would appreciate an increased compatibility with other systems such as WordNet (Fellbaum, 1998) and were planning to make their dictionaries available online in the future. The DDP tool has been used in several projects aimed to crowd-source the vocabulary documentation. The RapidWords project explores rapid vocabulary building where within 2 weeks a substantial dictionary can be compiled, counting up to 15,000 entries [6].

In our recent experience with Abui[7] we were able to triple the size of the corpus-based lexicon (about 2,500 entries which took around 10 years to compile) in just four days, during a workshop with just 15 Abui speakers. We expect to easily go over 15,000 words, when we continue for another ten days next year. The structured intuitive interface of **SD** is extremely easy to grasp even for native speakers of under-resourced languages who only have a basic literacy and received limited or no formal training. It is a great resource to substantially increase the amount of information on the lexicons of under-resourced languages.

The **SD** method opens up new possibilities for refining linguistic analysis. As an example of such

[3] The list of domains was developed by Ron Moe, a linguist working with SIL International, and originally called The Dictionary Development Process (DDP).

[4] See `http://www.sil.org/computing/toolbox`; `http://fieldworks.sil.org/`;`http://wesay.palaso.org/`

[5] FieldWorks Language Explorer (FLEx) is a tool for language documentation and analysis `http://fieldworks.sil.org/flex/`.

[6] See `http://rapidwords.net/`

[7] ISO 639-3 abz: a language spoken by approximately 16,000 speakers in the central part of the Alor Island in Eastern Indonesia.

1 **Universe, creation**

Use this domain for general words referring to the physical universe. Some languages may not have a single word for the universe and may have to use a phrase such as 'rain, soil, and things of the sky' or 'sky, land, and water' or a descriptive phrase such as 'everything you can see' or 'everything that exists'.

Q What words refer to everything we can see?
– ***universe, creation, cosmos, heaven and earth, macrocosm, everything that exists***

1.1 **Sky**

Use this domain for words related to the sky.

Q1 What words are used to refer to the sky?
– ***sky, firmament, canopy, vault***
Q2 What words refer to the air around the earth?
– ***air, atmosphere, airspace, stratosphere, ozone layer***
Q3 What words are used to refer to the place or area beyond the sky?
heaven, space, outer space, ether, void, solar system
...

1.1.1 **Sun**

Use this domain for words related to the sun. [...]

• Related domains: 8.3.3 Light, 8.3.3.2.1 Shadow, 8.4.1.2.3 Time of the day

Q1 What words refer to the sun?
– ***sun, solar, sol, daystar, our star***
Q2 What words refer to how the sun moves?
– ***rise, set, cross the sky, come up, go down, sink***
Q3 What words refer to the time when the sun rises?
– ***dawn, sunrise, sunup, daybreak, cockcrow***
...

1.1.1.1 **Moon**

Use this domain for words related to the Moon. [...]

1.1.1.2 **Star** [...]

1.1.1.3 **Planet** [...]

Figure 1: Depth First View of **Universe**

new step is the study of verbal semantics. Abui is a language with a complex alignment system described most recently in Kratochvíl (2011). There are multiple parameters determining the realization of arguments. **SD** method enables us to map the verbal inventory in great detail, map the **SD** for Abui onto **WN** and use computational tools to test the predictions outlined in Kratochvíl (2011). Linguistic description and the accuracy of linguistic analysis will be improved by the compatibility with **WN**, a standard resource in natural language processing.

2.1.2 Access to Lexical Resources

The structure of the **SD** further opens a possibility to create useful and refined lexical resources for the language community, such as dictionaries and language teaching materials.

Dictionaries using DDP have already been made available online in projects such as Webonary[8] or E-kamus2.org for languages of Eastern Indonesia.[9] There are many dictionaries in informal circulation, because there is no easy way to publish them online. By linking **SD** and **WN**, we open a possibility for small dictionaries to be published in the multilingual WN environment, which is better established and supported.

2.2 Wordnet (WN)

A wordnet is a semantic lexicon modeled on the Princeton WordNet (Fellbaum, 1998). Groups of similar words[10] are grouped together into synonym sets (or **synsets**) which are roughly equivalent to concepts. A combination of a word and synset defines a **sense**. Synsets are linked together by semantic relations, predominantly **hyperonymy** and **meronymy**, but including many others. Relations can also link senses to senses or synsets. Wordnets have been built for many languages, in this research we use the Princeton WordNet and the Wordnet Bahasa: a wordnet with Malay and Indonesian words linked to the Princeton WordNet structure (Nurril Hirfana at al., 2011). Over twenty wordnets have been linked together as the Open Multilingual Wordnet[11] and there is data for many, many more (Bond and Foster, 2013). Almost all wordnets have been built for established languages: building a wordnet from scratch is a considerable undertaking. The Princeton WordNet is released under an open source license that allows reuse with attribution , and most new wordnets (including the Wordnet Bahasa we use here) are released under a similar license.

The Princeton WordNet has been linked to

[8] See `http://webonary.org/`
[9] See `http://e-kamus2.org/`
[10] More properly, **lemmas**, which may be multiword expressions.
[11] See `http://compling.hss.ntu.edu.sg/omw/`

No.	Domain
1	**Universe, creation**
1.1	Sky
1.2	World
1.3	Water
1.4	Living things
1.5	Plant
1.6	Animal
1.7	Nature, environment
2	**Person**
2.1	Body
2.2	Body functions
2.3	Sense, perceive
2.4	Body condition
2.5	Healthy
2.6	Life
3	**Language and thought**
3.1	Soul, spirit
3.2	Think
3.3	Want
3.4	Emotion
3.5	Communication
3.6	Teach
4	**Social behavior**
4.1	Relationships
4.2	Social activity
4.3	Behavior
4.4	Prosperity, trouble

No.	Domain
4	**Social behavior** (cont)
4.5	Authority
4.6	Government
4.7	Law
4.8	Conflict
4.9	Religion
5	**Daily life**
5.1	Household equipment
5.2	Food
5.3	Clothing
5.4	Adornment
5.5	Fire
5.6	Cleaning
5.7	Sleep
5.8	Manage a house
5.9	Live, stay
6	**Work and occupation**
6.1	Work
6.2	Agriculture
6.3	Animal husbandry
6.4	Hunt and fish
6.5	Working with buildings
6.6	Occupation
6.7	Tool
6.8	Finance
6.9	Business organization

No.	Domain
7	**Physical actions**
7.1	Posture
7.2	Move
7.3	Move something
7.4	Have, be with
7.5	Arrange
7.6	Hide
7.7	Physical impact
7.8	Divide into pieces
7.9	Break, wear out
8	**States**
8.1	Quantity
8.2	Big
8.3	Quality
8.4	Time
8.5	Location
8.6	Parts of things
9	**Grammar**
9.1	General words
9.2	Part of speech
9.3	Very
9.4	Semantic constituents related to verbs
9.5	Case
9.6	Connected with, related
9.7	Name

Figure 2: Top two levels of the Semantic Domains

many other useful resources, including corpora (Landes et al., 1998), images (Bond et al., 2008; Deng et al., 2009), geographical locations, verb frames (Baker et al., 1998), Wiktionary and Wikipedia (de Melo and Weikum, 2010; Bond and Foster, 2013), many NLP tools (Bird et al., 2009) and ontologies (Niles and Pease, 2001; Gangemi et al., 2003). Allowing under-resourced languages to access these is an important goal for this project.

2.3 Comparison

As can be seen from Figures 1 and 2, the relations between domains are not as strongly typed as in WordNet, or at all uniform: for example **bodily functions** are related to **person**, but not as **synonyms**, **hyponyms** or **meronyms**. These somewhat looser relations are not captured well by WordNet: the so-called **tennis problem** (Wordnet does not link clearly related words such as ***racket***, ***ball***, ***net***: Fellbaum, 1998). The general associations of the **SD**s can go some way to providing these kinds of links.

3 Mapping

The objective of this task is to map the **SD** files to the **WN** files. Both the Indonesian and English versions of **SD** and **WN** were used. For Wordnet Bahasa, only the words tagged under Indonesian were taken. As such, mapping was done for the same language file (i.e. English **SD** to English **WN**) while across the two languages these two mappings were merged. As both files are in different formats, they were normalized first. This is to ensure that words from both the **SD** and **WN** file will be able to match each other during mapping.

To make the mappings more specific, we treat each question as a **class**: so we build for example: **1.1.s1** "What words are used to refer to the sky?" which contains the words: ***sky, firmament, canopy, vault***. We remove any meta information in brackets, part of speech information and so forth. We thus try to link both domains and classes (we will use the terms interchangeably from here on).

For both the English and Indonesian **WN** words, the underscore character was replaced with a space to harmonize with the **SD** words: ***outer_space*** becomes ***outer space***.

3.1 Initial Mapping

For each class in **SD**, the class name and each word was looked up in **WN**, and any matching synsets recorded (examples are given in Table 1). It was possible for **SD** classes to match to **WN** synsets through multiple paths: through more than one word (in either English or Indonesian). Of

SD ID (class)	**WN** Synset	Word
6.5.2.4.s3	01202651-v	bolt
8.3.1.5.1.s2	00124854-v	scroll
7.4.1	05021151-n	give
2.1.2.s2	05578911-n	girdle
1.6.4.2.s1	01181166-v	feed

Table 1: **SD**-**WN** ID mapping

course, many of these mappings would be inappropriate, due to the ambiguity of the word used as a pivot, so we need to further constrain the mapping.

We give some examples of words that did not match in Table 2. Typically the **SD** title is more informal than the **WN** synset entries. For example **SD**'s something used to see should map to **WN**'s optical instrument "an instrument designed to aid vision". The automatic mapping is very much a lower bound on the number of possible mappings.

3.2 Confirming the mappings

We looked at a variety of sources of information to improve confidence in the mappings: the structure of the domains and WordNet, the degree of polysemy, and the cross-lingual reliability.

3.2.1 Extracting Relations

We compared classes that were in a hierarchical relation to see if we could identify it with one of the relations used in WordNet. We used the following semantic relations from WordNet (**hypernym, part meronym, member meronym, substance meronym , part holonym, member holonym, substance holonym, entailment, attribute, cause, also see, verb group, similar to**). As the objective of **WN** and **SD** is to map semantic relationships of languages, we did not used formal relationships such as derivational links.

Some examples of classes linked in this way are given in Table 3. In general, if we could find a link, it was good evidence that the synset used in the link was the correct mapping to the domain. For example, in Wordnet, dry (**SD** ID:1.3.3.1) is a hyponym of sear (**SD** ID:1.3.3.1.s4). As the relations exist in Wordnet and these two words occurs under the same ID (1.3.3.1). We consider the Wordnet mapping to be applicable to Semantic Domains.

Table 4 shows another good example of mapping for the **SD** labels using the WordNet semantic relations. 75% of the related **SD** words were mapped to the main words (8.4.1: period of time/ janka waktu). For **SD** word 8.4.1.8 (Special days/hari-hari khusus), it was unable to be mapped under 8.4.1 as the expression, for both English and Indonesian, does not exist in WordNet. While for 8.4.1.1 (Calendar/Kalender), there is no direct semantic relation between the words available WordNet synsets and the main word 8.4.1. As such, 8.4.1.8 could not be mapped using WordNet relations (2nd level mapping) even though the word was mapped with WordNet synsets (1st level mapping).

SD ID	Word
1.3.3.1:	dry
Hypernym of:	
1.3.3.1.s5:	sear
1.3.3.1.s4:	wither
Cause:	
1.3.3.1.s2:	thirsty
1.3.3.1.s1:	dehydrated, desiccated, dried
1.3.3.1.s4:	wither
Similar to:	
1.3.3.1.s2:	thirsty
1.3.3.1.s1:	dehydrated, desiccated, dried
1.3.3.1.s5:	sear

Table 3: Classes linked with Semantic Relations

3.2.2 Monosemous Words

If a word is monosemous (that is it only appears in one synset) then we can assume it links a class to a synset unambiguously. We give some examples of such mappings in Table 5. In this case, there is no ambiguity, so the mapping is good.

3.2.3 Translation

Lexical ambiguity is often language specific and multiple languages can thus be used to disambiguate meanings (Bond and Ogura, 2007). If we can find matching synsets through pivots in two languages (in our case English and Indonesian) then we consider it a good mapping. We give an example in Table 6.

4 Results

We produced three kinds of mappings:

- **class**↔ **synset**: classified as related; monosemous; translated. (monosemous, e.g. 1.3.1.3↔ 09411430-n ***river***)
- **class**↔ **class**: classified with the WordNet relation. (hypernym↔ hyponym,

	English		Indonesian
8.3.3.3.4:	colors of the spectrum	8.3.3.3.4:	rentetan warna yang diuraikan oleh cahaya
3:	language and thought	3:	bahasa dan pikiran
9.4:	semantic constituents related to verbs	9.4:	konstituen atau unsur semantik yang berkaitan dengan
1.3.5:	solutions of water	1.3.5:	larutan air
2.3.1.9:	something used to see	2.3.1.9:	sesuatu yang digunakan untuk melihat

Table 2: **SD** main words not mapped to **WN**

English			Indonesian		
8.4.1	15113229-n	period of time	8.4.1	15115926-n	jangka waktu
	Hyponym			Hyponym	
8.4.1.2	14484516-n	day	8.4.1.2	14484516-n	hari
8.4.1.3	15135996-n	week	8.4.1.3	15135996-n	minggu
8.4.1.4	15206296-n	month	8.4.1.4	09358226-n	bulan
8.4.1.5	00294884-v	season	8.4.1.5	15239292-n	musim
8.4.1.6	15201505-n	year	8.4.1.6	15201505-n	tahun
8.4.1.7	15248564-n	era	8.4.1.7	15248564-n	zaman
	not mapped			not mapped	
8.4.1.1	08266849-n; 06487395-n; 15173479-n	Calendar	8.4.1.1	15173479-n	Kalender
8.4.1.8	NIL	Special days	8.4.1.8	NIL	Hari-hari khusus

Table 4: Example of a good 2nd level mapping

e.g. 8.4.1↔ 8.4.1.2)

- **sense**↔ **sense**: this is the direct word level, sense disambiguated mapping (class+lemma↔ synset+lemma, e.g. 7.4.1+give↔ 05021151-n+give).

The results of the mapping in terms of **class**↔ **synset** are summarized in Table 7 (which also shows the numbers of **class**↔ **class** mappings found). Potential mappings were found for 75% the domains, but confirmations were only found for around 21%.

The results for **class+lemma**↔ **synset+lemma** are shown in Table 8: about 69% of the English and 60% of the Indonesian **SD** words were mapped to entries in their respective wordnets. Out of the mapped **SD** label names, 27.92% and 31.92%, for English and Indonesian respectively, were confirmed using the **WN** semantic relations. Overall, about 20% of the **SD** label names were mapped to the second level.

Thus, the **class**↔ **synset** mapping improved as we go towards the lower levels as there is an increase in monosemous terms. However, the opposite occured for the SD-WN Main mapping because of the difference in word usage and structures in the two dictionaries. These weaknesses will be discussed in the following section

5 Discussion and Further Work

This is only the first step in the **SD**-**WN** mapping. The work that was done focuses on linking the **SD** words to the **WN** words before the **WN** semantic relationship is used to connect the words. As **WN** categorizes its words differently than **SD**, we expect some relations not to be mapped by the program: the cover should not be 100%, and is rarely one-to-one. In most cases, a single **SD** class links to multiple **WN** synsets.

When we started this process, full **SD** files were only available for English and Indonesian. There are now versions for Chinese and French which we intend to map to Chinese and French WordNets in the same way (Xu et al., 2008; Huang et al., 2010; Wang and Bond, 2013; Sagot and Fišer, 2012). This should increase the number of monosemous and translated mappings.

SD ID	Word	**WN** ID	Meaning
4.1.9.2.s3	intermarry	02490090-v	marry within the same ethnic, social, or family group
6.5.2.7.s4	kantor	08337324-n	an administrative unit of government

Table 5: Monosemous Words

SD ID	Word	**WN** ID	Meaning
2.s2	someone, somebody	00007846-n	a human being
2.s2	seseorang	00007846-n	a human being

Table 6: Classes that are Matched through Multiple Languages

Most of the **SD** words that were not mapped to **WN** synsets were not lemmas in **WN**. As shown in Table 2, these are mainly informal multi-word expressions, consisting of 4 or more words while the multi-words expression in wordnet are rarely of more than 3 words. As that mapping was done by matching both **SD** and **WN** expressions as a whole, these **SD** expressions were unable to be matched with **WN**. Having formal and informal names for the concepts (domains/synsets) could be useful for both resources.

Error analysis found some matches due to inconsistent structures, which suggest the resources themselves may need to be revised. For example, ***contact lens*** is under **SD** "something used to see" which we hand-mapped to **WN**'s optical instrument "an instrument designed to aid vision". However in **WN** it is a hyponym of optical device "a device for producing or controlling light" which puts it in the same grouping as camera lenses, not spectacles. It is possible it should inherit from both, but it should definitely inherit from optical instrument, as it is an aid to vision. In this case **SD** reveals a missing link in **WN**. The opposite case was also common.

We intend to use the mapping to generate a wordnet for the under-resourced language Abui (Kratochvíl, 2007). As a part of this process, we will correct and refine the mapping. We can then compare, for example, verb classes in Abui with those in wordnets for English and other well described languages. Linking descriptions of under-studied languages to well-studied languages makes it easier to leverage existing linguistic knowledge.

Even though most classes do not map one-to-one to WN synsets, the combination of class and lemma/gloss is generally enough to disambiguate. For example, consider the class 1.1.1.s2 "What words refer to how the sun moves". This links to at least four WordNet classes $\text{rise}_{\text{v:16}}$ "come up, of celestial bodies", $\text{sink}_{\text{v:6}}$ "appear to move downward", $\text{cross}_{\text{v:1}}$ "travel across or pass over" and $\text{set}_{\text{v:10}}$ "disappear beyond the horizon". Linking to these suggests several other possible entries for the class: ***go under*** [the horizon], ***traverse*** [the sky]. When we want to build a wordnet for, e.g., Abui, we would look at the Abui word with the gloss "go down" ***sei*** in the class 1.1.1.s2 and we know that this links to the synset $\text{sink}_{\text{v:6}}$. Even though the mapping is not one-to-one, the combination of mapping and gloss will generally lead to a specific synset. In addition, **WN** gives the information that $\text{rise}_{\text{v:16}}$ and $\text{set}_{\text{v:10}}$ are antonyms and this is true for the Abui equivalents ***marang*** and ***sei***.

The mapping can also be used to help translate the semantic domains into new languages (assuming there is a wordnet for the language) and to add new instances of the classes from the wordnets.

Finally, there has been a recent movement within the wordnet community to make the lexical resources more open (Bond and Paik, 2012; Bond and Foster, 2013). We hope to show the advantages of openness (more usable and accessible data) with the under-resourced language community and make the data open in the same way. The Wordnet-Semantic Domain Mappings themselves are available for download at the Open Multilingual Wordnet,[12] and linked in the search interface.

6 Conclusion

A simple **SD**-**WN** mapping was done using the **WN** semantic relationships. Even though the program was unable to cover all the semantic relationships that exist in both the English and Indonesian **SD** data, it provided a basis for further work in mapping the semantic relationships that are available in the **SD** file. The mapping is freely avail-

[12] See `http://compling.hss.ntu.edu.sg/omw/`

LVL	Example	# IDs	ID linked to WN		≥ 1 relation		≥ 2 relation		monosemous	
			eng	ind	eng	ind	eng	ind	eng	ind
1	1: universe	9	3	4	3	4	1	2	1	1
2	1.1: sky	68	54	46	27	27	6	13	7	7
3	1.1.1: sun	419	252	237	73	74	16	32	33	29
4	1.1.1.1: moon	985	702	605	90	69	8	8	86	65

Table 7: Summary of Mapping

	English (eng)			Indonesian (ind)		
	Word	Immediate (%)	Label (%)	Word	Immediate (%)	Label (%)
SD words	1,793			1,793		
1st level mapping	1,243	69.32		1,090	60.75	
2nd level mapping	347	27.92	19.35	384	31.92	21.42

Table 8: Coverage of **SD-WN** Main mapping

able, and we hope that it will provide a useful link between wordnet and the semantic domains.

Acknowledgments

This research was partially funded by the NTU SUG grant on *Documentation and Analysis of Endangered Papuan Languages of Alor-Pantar Archipleago, Southern Indonesia* (M4080390.100). We would like to thank Ronald Moe for producing and sharing with us the SIL semantic domains, as well as his constructive support.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*. Montreal, Canada.

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. `www.nltk.org/book`.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Francis Bond and Kentaro Ogura. 2007. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2):127–136. URL `http://dx.doi.org/10.1007/s10579-007-9038-4`, (Special issue on Asian language technology).

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

Carl Darling Buck. 1949. *A dictionary of selected synonyms in the principal Indo-European languages : a contribution to the history of ideas*. Chicago University Press, Chicago.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156. Narosa Publishing, New Delhi, India.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition (CVPR09)*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).

František Kratochvíl. 2007. *A Grammar of Abui: A Papuan Language of Indonesia*. LOT, Utrecht.

František Kratochvíl. 2011. Transitivity in Abui. *Studies in Language*, 35(3):588–635.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Fellbaum (1998), chapter 8, pages 199–216.

Ron Moe. 2013. Semantic domains. `http://semdom.org`. (Accessed 2013-04-01).

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Maine.

Benoît Sagot and Darja Fišer. 2012. Automatic Extension of WOLF. In *Proceedings of GWC2012 - 6th International Global Wordnet Conference*. Matsue, Japan.

Shan Wang and Francis Bond. 2013. Building a Chinese wordnet: Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*. Nagoya.

Renjie Xu, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.