Building The Mongolian WordNet

Khuyagbaatar Batsuren[§†*], **Amarsanaa Ganbold**[†], **Altangerel Chagnaa**[†], and **Fausto Giunchiglia**[§]
[§]KnowDive Group, DISI, University of Trento, Italy
[†]Machine Intelligence Laboratory, DICS, National University of Mongolia, Mongolia
{k.batsuren, fausto.giunchiglia}@unitn.it
{amarsanaag, altangerel}@num.edu.mn

Abstract

This paper presents the Mongolian Wordnet (MOW), and a general methodology of how to construct it from various sources e.g. lexical resources and expert translations. As of today, the MOW contains 23,665 synsets, 26,875 words, 2,979 glosses, and 213 examples. The manual evaluation of the resource[1] estimated its quality at 96.4%.

1 Introduction

Language resources are crucial in the research of computational linguistics e.g., information retrieval, document classification, query answering. In recent years, world languages are divided in two groups: highly-resourced languages (e.g., English or Chinese) and under-resourced languages (e.g., Kazakh or Uyghur). Due to the lack of language resources, the second group of languages displays more mediocre performance than the first group. Mongolian was one of the under-resourced languages.

This paper describes a general methodology by which we built the Mongolian WordNet (MOW), a high-precision wordnet-like lexical resource. Our main technical contributions are (1) a general method to extract high-precision wordnet translations from a bilingual dictionary, (2) a medium-scale lexical resource for the Mongolian language.

The paper is organized as follows. Section 2 presents state-of-the-art methods. Section 3 provides the main methodology how the MOW is built, and Section 5 describes the automatic algorithm to extract the wordnet translations from a bilingual dictionary. We evaluated the results of this method in section 6. Finally, section 7 concludes the paper.

2 State of the Art

Princeton WordNet (PWN) has been a primary lexical resource for most researches involved in lexical semantics, from Computational Linguistics to Semantic Web. Examples of particular applications are word sense disambiguation (Navigli, 2009) and ontology research (Oltramari et al., 2002). This successful case for English inspired many researchers to build wordnets for other languages. Given the awareness of the structural and semantic diversity across languages (Giunchiglia et al., 2017), mono-lingual wordnets have been developed in two ways: the expansion method from PWN and the merge method with PWN.

- The *expansion method* – researchers first accept that the semantic structure of PWN should be more or less similar to their language's semantic network, and translate English synsets to that of a target language.

- The *merge method* – researchers first create a semantic network for their language, and develop its synsets by adding words and definitions. In a final round, they merge their semantic network with PWN by linking[2] synsets with PWN.

To our knowledge, a vast majority of the wordnets have been developed by using the expansion method (Bond and Paik, 2012), while very few wordnets including Open Dutch

[*] This work has been done during internship at National University of Mongolia

[1] https://milab.num.edu.mn/research/monwordnet/

[2] Hereby, a linking is a manual finding of an equivalent meaning between synsets of two resources.

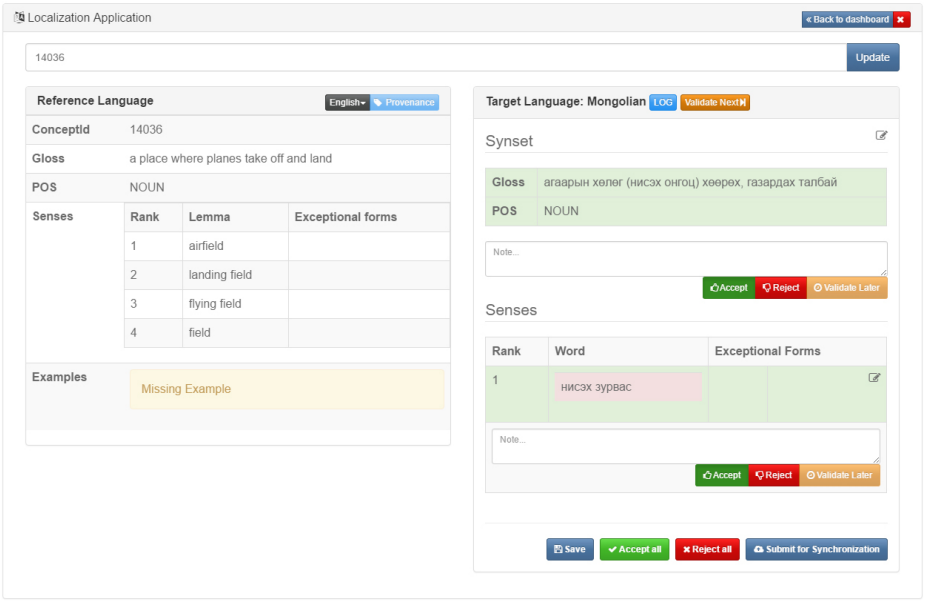

Figure 1: A screenshot of validator user interface

WordNet (Postma et al., 2016), Hindi WordNet(Bhattacharyya, 2017), Polish WordNet have used the merge method. The obvious obstacle is the cost of human labor and the deep expertise of several different domains and cultures, needed in the development of a semantic network.

Researchers in comparative linguistics state that the semantic space of languages are vast and very differential from one another (Von Fintel and Matthewson, 2008) (Giunchiglia et al., 2018). This is because of the differences between speakers of languages, e.g., culture, geographic environment. This is the primary condition underlying the actual choice of the *merge* method because of the importance of individual culture is a fundamental to their wordnet-like lexical resource.

Early linguists (Youn et al., 2016) revealed that an universal structure of lexical semantics exists across all languages at least between basic concepts, and it is why the majority of wordnet developers selected intuitively the *expand* method. Later on, the Global WordNet Association recommended that the monolingual semantic network should be extended by adding cultural synsets under the coordinated usage of the global wordnet grid between wordnets(Vossen et al., 2016).

3 Methodology

In terms of *Wordnet development*, we adopted the expansion method. In the future, we are planning to change and expand the core semantic structure by adding more cultural concepts under the coordination of the global wordnet grid (Vossen et al., 2016). Our wordnet project has two main stages of development: (1) expert translation and (2) automatic translation.

In the *expert translation*, the project has been running since 2016 by employing only expert linguists to translate PWN to Mongolian (Section 4). In the *automatic translation*, we have used a freely, available bilingual Mongolian dictionary to translate PWN to Mongolian (Section 5).

4 Expert Translation

The expert translation method generally follows ontology localization (Espinoza et al., 2009) (Das and Giunchiglia, 2016) which adapts an existing ontology in a language to another by using translation of terms. In this method (Ganbold et al., 2014) (Giunchiglia et al., 2015) (Huertas-Migueláñez et al., 2018), recruited linguistic experts and asked them to provide synsets, in the target language that properly represent a concept denoted by a synset in the source language. The main idea is to find out the most suitable words for

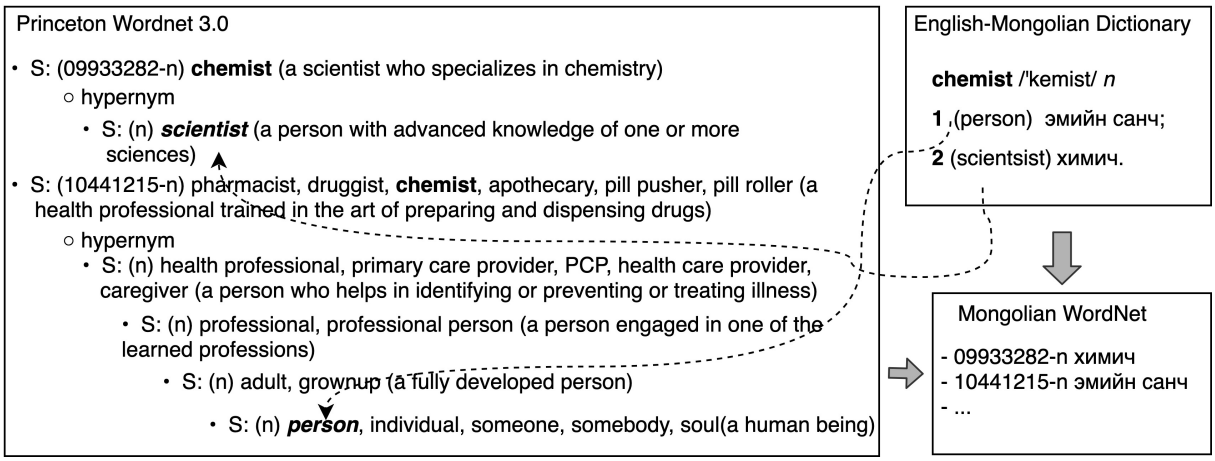

Figure 2: The hypernym-based translation between Princeton WordNet and Bilingual dictionary on a given word "chemist"

the concept in terms of linguistic context use rather than word-for-word translation between synsets.

This method consists of two main tasks: a) translation and b) validation. In the translation task, a language translator provides synset words, its gloss, and example sentences in the target language after she fully understands the meaning of a given synset to localize. If the translator assumes the concept does not exist in the target language, she should mark it as a lexical gap, which means a free combination of words represents the concepts. In this way, we avoid literal translations which may produce a wrong or unwanted result. In the validation task, a language validator evaluates all the elements of the given synset, provided by the translator. The validator either confirms each element or rejects elements one by one with feedback. In the case of a lexical gap, she can accept as it is or suggest word(s) for the synset where she denies it as a gap. When the translator receives feedback, he/she accommodates comments if she agrees with the validator. Alternatively, she can reject the evaluation with comments. Upon reaching an agreement between the translator and the validator, we believe this process produce target language synset with high-quality at the end.

Tasks for translators and validators are assigned by a language manager who manages overall translation activity. Tasks are grouped into a subset of wordnet hierarchy, called subtree, which allows the linguistic experts to understand what they translate/validate. It helps to differentiate concepts by exploring their hyponym/hypernym or sibling relations. The walk-through of tasks is breadth-first.

The linguistic experts use an expert sourcing tool whose screenshot of a validation process is shown in Figure 1. Several volunteered (Ganbold and Chagnaa, 2015) (Ganbold et al., 2018) and paid experts with this tool produced 12,141 synsets, 24,277 senses, and 12,830 words so far.

5 Automatic Translation

Given the two resources PWN and bilingual dictionary below, the main task is to find automatically a set of pairs of $<c, s>$ where c is a synset id from PWN and s is a sense instance of the dictionary. Our method in Algorithm 1 is based on the multiple intuitive criteria:

- if a collocate noun of the sense s maps into one of hypernyms of the synset c then s can express the meaning of the synset c. The example of hypernym-based translations is shown in Figure 2.

- if a given word w has one sense for both dictionary and PWN, the dictionary sense is equivalent to the PWN synset. For example, for the noun word 'mimic,' both PWN and dictionary has only one sense. This intuition of *monosemy translation* has been used to build a French WordNet (Sagot and Fišer, 2008) and Thai WordNet (Sathapornrungkij and Pluempitiwiriyawej, 2005).

The algorithm is structured with three main steps as follows.

Algorithm 1: WordNet Retrieval Algorithm

```
   Input  : w, an english word
   Input  : R, a lexical resource PWN
   Input  : D, a bilingual dictionary
   Output : M, a set of pairs of <id_R, w_D>
 1 C ← Synsets(R, w);
 2 S ← Senses(D, w);
 3 M ← ∅;
 4 if |C| == 1 and |S| == 1 then
 5     for one synset c ∈ C and one sense s ∈ S do
 6         if pos(c) ≠ pos(s) then
 7             continue;
 8         M ← M ∪ <c, words(s)>;
 9 else
10     for each synset c ∈ C do
11         for each sense s ∈ S do
12             if pos(c) ≠ pos(s) then
13                 continue;
14             if μ(collocate(s), c) then
15                 M ← M ∪ <c, words(s)>;
16 return M;
```

Step 1: Initialization (Lines 1–3). C is initialized with a list of synsets which are expressed by the input word w in the lexical resource $\mathcal{R}$ as PWN (line 1). S is initialized with a list of the Mongolian senses which are contained by the input word w in the bilingual dictionary $\mathcal{D}$ (line 2).

Step 2: Monosemy translation (Lines 4–8). In this step, it first checks if the lexical resource R and the bilingual dictionary D have one-to-one mapping between them for the input word w (line 4). if so, in the line 5, it assigns the corresponding one synset from $\mathcal{R}$ into c and the corresponding one sense from $\mathcal{D}$ into a sense instance s (line 5). Then it checks if the synset and the sense share same part of speech (line 6). Then if it succeeds it adds $<c,\ words(s)>$ into the answer set M where $words(s)$ returns only words of the sense s in the bilingual dictionary $\mathcal{D}$.

Step 3: Hypernym-based translation (lines 10–15). In this step, the algorithm iterates each possible pair of a synset c from C and a sense s from S. Then for each pair, if the synset c and the sense s share same part of speech (line 12). If so, the function μ checks if the collocate noun of the dictionary sense s is a hypernym of the synset c in the lexical resource $\mathcal{R}$. If it succeeds it adds $<c,\ words(s)>$ into the answer set M where $words(s)$ returns only words of the sense s in the bilingual dictionary $\mathcal{D}$.

Finally, in Line 16, the algorithm returns the answer set M.

5.1 English-Mongolian Bilingual Dictionary

This bilingual dictionary between English and Mongolian contains over 43,442 English headwords (including compound words) that are translated into 79,299 Mongolian words (or senses). For each english word, the dictionary provides its related senses with their mongolian words. For example, given a word "chemist", the dictionary stores an information as follows:

chemist /'kemist/ *n* **1**. (person) **эмийн санч**; **2**. (scientist) **химич**.

where the numbers represent each meaning and it is followed by the collocates (e.g. person or scientist) that are used to distinguish the meanings. Let the 3-tuple $a = <w,\ p,\ S>$ be the headword instance where w represents a head word, p represents a part of speech of the word w, S is a set of senses expressed by the word w. Let the sense instance, s, is the three tuple of $<id,$ col, $w_m>$ where id represents a sense number of s, col is a collocate noun to distinguish s from other meanings, and w_m is a mongolian translation word.

For the above example, the headword instance h is $<$'*chemist*', '*noun*', $S>$ where $S = \{<1,$ '*person*', '**эмийн санч**'$>$; $<2,$ '*scientist*', '**химич**'$>\}$.

6 Results and Evaluation

PWN has 133974 English words and then given in input to the algorithm 1, which, in turn, generated two sets of 3652 synsets and 7872 synsets from the two automatic methods of *hypernym* translation and *monosemy* translation respectively. For each of the three translations, 200 cases were randomly selected, which were equally selected across four parts of speech. Three linguists were selected to evaluate the samples. They were also provided with the corresponding English glosses and words for the synsets involved, and they were asked the following question: "Do you think meanings of the English synset s_e and the Mongolian synset s_m are equivalent?", and they had to provide a yes/no answer.

Table 1: The results of the three translations: *expert*, *monosemy*, and *hypernym-based* translations.

#	Method	Synsets	Senses	Words	Core Coverage	Accuracy
1	Expert translation	12141	24277	12830	41.1	99.0
2	+ monosemy translation	7872	11038	10235	8.1	98.2
3	+ hypernym-based translation	3652	5629	3792	12.4	92.1
Total	Mongolian Open WordNet	23665	40944	26857	61.6	Avg. 96.4

Table 2: The best twenty wordnets ranked by a number of synsets (Note: we only consider the wordnets that are publicly available and linked to PWN)

#	Language	Synsets	Senses	Words	Examples	Glosses	References
1	English	109942	191523	133974	48459	109942	(Miller, 1995)
2	Finnish	107989	172755	115259	0	0	(Lindén and Carlson, 2010)
3	Chinese	98324	123397	91898	17	541	(Wang and Bond, 2013)
4	Thailand	65664	83818	71760	0	0	(Thoongsup et al., 2009)
5	French	53588	90520	44485	0	0	(Sagot and Fišer, 2008)
6	Romanian	52716	80001	45656	0	0	(Tufiş et al., 2008)
7	Japanese	51366	151262	86574	28978	51363	(Bond et al., 2009)
8	Catalan	42256	66357	42444	2477	6576	(Gonzalez-Agirre et al., 2012)
9	Slovene	40233	67866	37522	0	0	(Fišer et al., 2012)
10	Portuguese	38609	60530	40619	0	0	(de Paiva et al., 2012)
11	Spanish	35232	53140	32129	651	17256	(Gonzalez-Agirre et al., 2012)
12	Polish	35083	87065	59882	0	0	(Piasecki et al., 2009)
13	Italian	33560	42381	29964	1934	2403	(Emanuele et al., 2002)
14	Indonesian	31541	92390	24081	0	3380	(Noor et al., 2011)
15	Malay	31093	93293	23645	0	0	(Noor et al., 2011)
16	Basque	28848	48264	25676	0	0	(Pociello et al., 2011)
17	Dutch	28253	57706	40726	0	0	(Postma et al., 2016)
18	**Mongolian**	23665	40944	26857	213	2976	our resource
18	Croatian	21302	45929	27161	0	0	(Oliver et al., 2016)
19	Persian	17705	30365	17544	0	0	(Montazery and Faili, 2010)
20	Greek	17302	23117	17278	0	0	(Stamou et al., 2004)

Table 1 provides accuracy values for the three translations. The average accuracy for all the translations is 96.4, and the inter-annotator agreement between three annotators was 98.1.

The Mongolian WordNet now contains 23665 synsets, 40944 senses, and 26857 words as a result of the combination of all the above methods. As can be seen from Table 1, the resource is covering the 61.6 percents of 4960 "core" synsets derived from (Boyd-Graber et al., 2006).

7 Conclusion

We described how Mongolian WordNet is created by using three types of translation: *expert*, *monosemy*, and *hypernym-based* translations under the expansion method of PWN. Our main goal was to create a high-quality lexical resource, so that in automatic translations, we only selected the intuitive patterns (*monosemy* and *hypernym*) which are ensuring high quality in principles.

Mongolian WordNet contains 23665 synsets, 40944 senses, and 26857 words. There are 15976 nouns, 3791 verbs, 601 adverbs, and 3037 adjectives. In addition, it has 213 examples and 2976 glosses. The average polysemy is 1.52. The resource is delivered in the tab-separated format (Bond and Foster, 2013) under the CC BY-NC-SA 4.0 license[3].

8 Acknowledgements

The research has received funding from the Mongolian Science and Technology Fund under grant agreement SSA_024/2016. The result described in this paper is part of the Mongolian Local Knowledge Core project, partially supported by the National University of Mongolia under grant agreement P2017-2383. The first author is supported by the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105.

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

References

Pushpak Bhattacharyya. 2017. Indowordnet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th workshop on Asian language resources*, pages 1–8. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36. Citeseer.

Subhashis Das and Fausto Giunchiglia. 2016. Geoetypes: Harmonizing diversity in geospatial data (short paper). In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 643–653. Springer.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Pianta Emanuele, Bentivogli Luisa, and Girardi Christian. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Mauricio Espinoza, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2009. Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, pages 33–40.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. slownet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117.

Amarsanaa Ganbold and Altangerel Chagnaa. 2015. Crowdsourcing Localization of Ontology and Geographical Names. In *The Eighth International Conference on Frontiers of Information Technology*, pages 120–124, Jilin, China.

Amarsanaa Ganbold, Feroz Farazi, Moaz Reyad, Oyundari Nyamdavaa, and Fausto Giunchiglia. 2014. Managing language diversity across cultures: The english-mongolian case study. *International Journal on Advances in Life Sciences*, 6(3-4).

Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using Crowd Agreement for Wordnet Localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, and Khuyagbaatar Batsuren. 2015. Crowdsourcing a large scale multilingual lexico-semantic resource. In *The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15), San Diego, CA*.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4009–4017.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One world—seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Mercedes Huertas-Migueláñez, Natascia Leonardi, and Fausto Giunchiglia. 2018. Building a lexico-semantic resource collaboratively. In *The XVIII EURALEX International Congress*, page 148.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

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

Mortaza Montazery and Heshaam Faili. 2010. Automatic persian wordnet construction. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 846–850. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

Nurril Hirfana Bte 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*.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2016. Automatic expansion of croatian wordnet. In *Međunarodni znanstveni skup Hrvatskoga društva za primijenjenu lingvistiku*.

Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. 2002. Restructuring wordnet's top-level: The ontoclean approach. *LREC2002, Las Palmas, Spain*, 49.

Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the Eighth Global WordNet Conference*, page 300.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of the Ontolex 2008 Workshop*.

Patanakul Sathapornrungkij and Charnyote Pluempitiwiriyawej. 2005. Construction of thai wordnet lexical database from machine readable dictionaries. *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring balkanet shared ontology for multilingual conceptual indexing. In *LREC*.

Sareewan Thoongsup, Kergrit Robkop, Chumpol Mokarat, Tan Sinthurahat, Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of the 7th workshop on Asian language resources*, pages 139–144. Association for Computational Linguistics.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of 4th Global WordNet Conference, GWC*, pages 441–452.

Kai Von Fintel and Lisa Matthewson. 2008. Universals in semantics. *The linguistic review*, 25(1-2):139–201.

Piek Vossen, Francis Bond, and J McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.