Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3181–3188
Marseille, 11–16 May 2020

Linking the TUFS Basic Vocabulary to the Open Multilingual Wordnet

Francis Bond[♠], **Hiroki Nomoto**[♢], **Luis Morgado da Costa**[♠], **Arthur Bond**[♡]
[♠] Nanyang Technological University (NTU), Singapore
[♢] Tokyo University of Foreign Studies (TUFS), Japan
[♡] Brisbane School of Distance Education (BSDE), Australia
bond@ieee.org,nomoto@tufs.ac.jp,luis.passos.morgado@gmail.com,artcbond@gmail.com

Abstract

We describe the linking of the TUFS Basic Vocabulary Modules, created for online language learning, with the Open Multilingual Wordnet. The TUFS modules have roughly 500 lexical entries in 30 languages, each with the lemma, a link across the languages, an example sentence, usage notes and sound files. The Open Multilingual Wordnet has 34 languages (11 shared with TUFS) organized into synsets linked by semantic relations, with examples and definitions for some languages. The links can be used to (i) evaluate existing wordnets, (ii) add data to these wordnets and (iii) create new open wordnets for Khmer, Korean, Lao, Mongolian, Russian, Tagalog, Urdu and Vietnamese.

Keywords: basic vocabulary, wordnet, language teaching, multilingual lexicon

1 Introduction

In this paper we describe linking two complementary linguistic resources — the TUFS Basic Vocabulary Modules, created for online language learning (Kawaguchi et al., 2007), with the Open Multilingual Wordnet (OMW: Bond and Foster, 2013). Multilingual lexicons are still quite rare, with most created by linking various bilingual lexicons. The TUFS Basic Vocabulary Modules are hand created, using commonly occurring vocabulary (designed for beginning language learners). It includes some languages with few online language resources, like Lao and Khmer. Because the lexicons are used by learners of single languages, the current interfaces do not allow easy access to them as a single multilingual resource.

The first product of this paper is an easily accessible multilingual lexicon, based on the TUFS vocabulary, fully documented and available from github.[1] The second is a high quality mapping between the TUFS vocabulary and the open multilingual wordnet.

The structure of the paper is as follows. We start off by introducing the TUFS database, an open language resource, to provide context for the entries in § 1.1. Next, in § 1.2, we briefly describe the OMW. Then, in § 2, we created machine tractable versions of the data. In § 3, we mapped the OMW and TUFS using the wordnet synsets and the translation sets. In § 3.1 we show the mapping process. § 3.2 shows special cases such as when hyponym relations were assigned or when a new entry was recommended for words with two distinct meanings that needed to be separated. § 4 discusses TUFS can be used to evaluate the languages in the OMW.

1.1 TUFS Open Language Resources

The TUFS Open Language Resources includes basic vocabulary words and example sentences in 24 languages (Kawaguchi et al., 2007). These have also been used as a basis for the TUFS Asian Language Parallel Corpus, or TALPCo (Nomoto et al., 2018). The data is available as PostgreSQL backup files (`pg_dump` files), under an open license (CC BY 4.0).[2]

The Japanese component consists of 799 basic vocabulary words with example sentences for them.[3] These basic vocabulary words were selected in accordance with the lowest level of the Japanese Language Proficiency Test: N5. For this level, one must be able to:

1. read and understand typical expressions and sentences written in hiragana, katakana and basic kanji

2. listen and comprehend conversations about topics regularly encountered in daily life and classroom situations, and pick up necessary information from short conversations spoken slowly.

Some of the vocabulary is old-fashioned (words that we would not consider basic vocabulary today). Some examples are 万年筆 *mannenhitsu* "fountain pen" and フィルム *firumu* "(photographic) film".

Each vocabulary module is made for a learner of a language who speaks another language (mainly Japanese, but sometimes English, Malay, Indonesian or Burmese). A screenshot of an entry for *Lehrer* "teacher" shown to a Japanese student of German is given in Figure 1. It has the word itself (in red), a link to the pronunciation, possible translations, with meaning notes, and then example sentences (with glosses in Japanese).

This information is repeated for all the vocabulary modules, although there is no way to go between languages on the TUFS website. However, in the database storing the vocabulary, there is an id shared across all languages (`classified_id`) which we will call the TUFS ID (`tid`). This can be used to link the translations. We call the set of translations linked by the `tid` the TUFS translation set.

[1] https://github.com/fcbond/tufs

[2] https://malindo.aa-ken.jp/TUFSOpenLgResources.html

[3] https://www.jlpt.jp/e/about/levelsummary.html

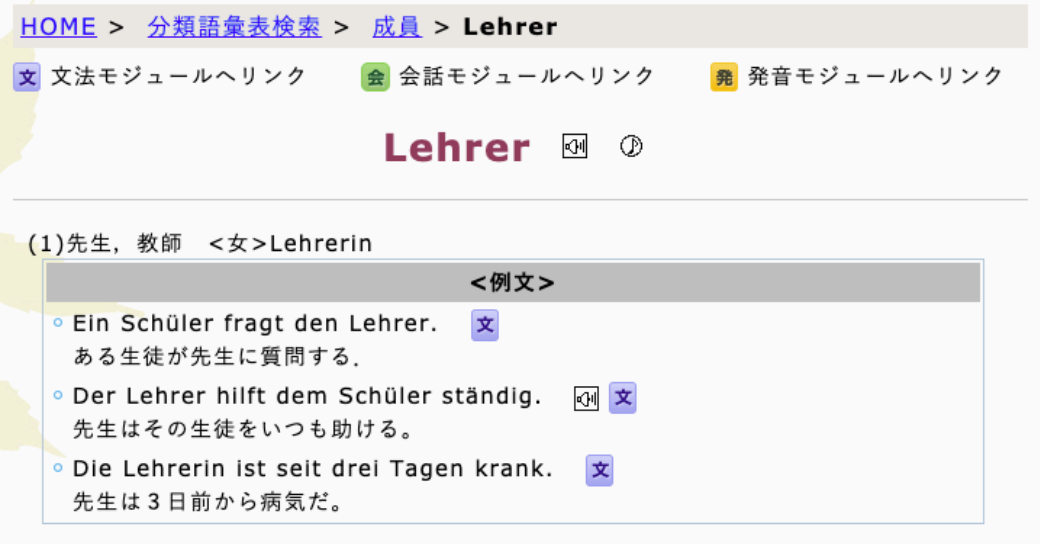

Figure 1: Teacher in German

Not all words are used in the teaching materials for all languages. We show the distribution of translations for words in Table 1. Many translation sets only have entries in one language (with a Japanese gloss). 505 have translations in more than fifteen, so we will focus on linking them to wordnet due to the diminishing returns for the rest.

Number of Translations	Translation Sets
$0 < x \leq 5$	2,004
$5 < x \leq 10$	214
$10 < x \leq 15$	2
$15 < x \leq 20$	80
$20 < x \leq 25$	425

Table 1: Distribution of Translations

We give the translation set for 先生 *sensei* "teacher" in Table 2. We see here that different languages offer extra information as free text (the meaning column): German, French and Vietnamese have male and female versions; Russian differentiates school and university teachers. This is in Japanese in the TUFS database, but we give a translation in English in the Table. For example, "先生，教師 ＜女＞Lehrerin" becomes "(classroom) teacher <woman>Lehrerin". Some languages have extra information in the lemma: Arabic offers irregular plurals and Chinese gives pinyin transliterations.

The vocabulary also comes with example sentences, all glossed in Japanese (as the material is aimed at Japanese learners). The sentences are not the same across all languages: e.g., (1–3). There is a revised corpus which is parallel in seven languages: Japanese, Burmese (Myanmar), Malay, Indonesian, Thai, Vietnamese and English (TALPCO: Nomoto et al., 2018). This has not been fed back into the original data.

(1) Mrs. McDonald is my English teacher. [en]

(2) Ein Schüler fragt den Lehrer. [de]
"A student asks the teacher a question."

(3) Akiko, Pak Yanto adalah dosen tata bahasa. [id]
"Akiko, Mr Yanto is a grammar teacher."

1.2 The Open Multilingual Wordnet

The Open Multilingual Wordnet, version 1.2, is a collection of wordnets linked through the Princeton wordnet. We use the extended version produced at Nanyang Technological University that includes some extra vocabulary, including Japanese and Chinese lexicalized time expressions, pronouns, exclamations and more (Bond et al., 2016). The individual wordnets used are the Princeton Wordnet, PWN (Fellbaum, 1998), the Japanese Wordnet (Isahara et al., 2008), the Chinese Open Wordnet (Wang and Bond, 2013), the Wordnet Bahasa (Nurril Hirfana, Suerya and Bond 2011 , Malay and Indonesian), the Wordnet Libre du Français (Fišer and Sagot, 2008, French), the Arabic Wordnet (AWN) (Elkateb et al., 2006), the Multilingual Central Repository (MSR) (Gonzalez-Agirre et al., 2012, Spanish), OpenWordNet-PT (de Paiva et al., 2012, Portuguese) and the Thai Wordnet (Thoongsup et al., 2009).

Wordnets are organized into synsets linked to the PWN entries. Each wordnet may have one or more lemmas, definitions and examples. There are also semantic links to other synsets. We give a simplified example of the entry for one sense of *teacher* in Figure 2: note that the hypernyms and hyponyms link to other synsets. We omit languages in OMW but not TUFS (31 wordnets had translations for teacher). We also do not show sense level links, sentiment, and links to other resources. The full entry can be seen at `http://compling.hss.ntu.edu.sg/omw/cgi-bin/wn-gridx.cgi?synset=10694258-n`.

Some languages, especially those semi-automatically constructed (like Malay, Japanese, French and Indonesian) have many entries, preferring to err on the side of inclusivity.

2 Creating the TUFS Translation Sets

The first task we did was to analyze the vocabulary for each language and link them into the translation sets. As part

Language	Translation	Meaning
Arabic (ar)	أُسْتَاذْ	[noun] professor, teacher
Arabic in Syria (as)	إستاذ (brk:أساتذة)	teacher
German (de)	Lehrer	(classrom) teacher <woman>Lehrerin
English (en)	teacher	teacher
Spanish (es)	profesor	teacher
French (fr)	instituteur	teacher. for primary school; feminine is institutrice
French (fr)	professeur	teacher. professor; same for masculine and feminine.
French (fr)	enseignant	teacher. teaching staff. teacher as a job.
Indonesian (id)	guru	teacher
Indonesian (id)	dosen	teacher
Japanese (ja)	先生	teacher; instructor; not used when referring to one's own job; doctor; ※ used when addressing a medical doctor in stead of "Dr.〜"
Central Khmer (km)	គ្រូ	teacher
Korean (ko)	선생님	teacher
Lao (lo)	ອາຈານ	teacher
Mongolian (mn)	багш	teacher
Malay (ms)	guru;cikgu	teacher; instructor; not used when referring to one's own job; doctor; ※ used when addressing a medical doctor in stead of "Dr.〜"
Burmese (my)	ဆရာ;ဆရာမ; ကျောင်းဆရာ;ဆရာဝန်	teacher; instructor; not used when referring to one's own job; doctor; ※ used when addressing a medical doctor in stead of "Dr.〜"
Por. in Brazil (pb)	professor	teacher
Portuguese (pt)	professor	teacher
Russian (ru)	учитель	primary·middle·high school teacher; (преподаватель is used for university teachers)
Thai (th)	อาจารย์	teacher
Tagalog (tl)	titser	teacher
Turkish (tr)	öğretmen	teacher
Urdu (ur)	استاد	teacher
Vietnamese (vi)	cô giáo	teacher (indicates a female teacher)
Vietnamese (vi)	thầy giáo	teacher (indicates a male teacher)
Mandarin Chinese (zh)	老师 (py:lǎoshī)	teacher

Table 2: Entry for 先生 *sensei* "teacher": 19186

of this we cleaned the lemmas (to make them more computationally tractable). The data is made available both as tab-seperated data `tabs-vocab.tsv` and as an html view, both availble from the github site. We show the html view for teacher in Figure 3.

3 Linking OMW and TUFS

In this section we describe the mapping between the TUFs translation sets and wordnet synsets. This mapping will allow several tasks:

(i) for languages with existing wordnets, we can do a general evaluation of their coverage (§ 4). It is hard to evaluate large multilingual lexicons as there is no one who can speak all languages. Further, pure size is not a perfect metric — coverage of basic vocabulary is very important. Like all selections, the TUFS vocabulary has its quirks, but is a welcome addition to the current widely used testset of 5,000 **core** concepts (Boyd-Graber et al., 2006) based on the British National Corpus. This is not multilingual — we can test for coverage in terms of synsets, but we have no idea if the language specific realizations are reasonable.

(ii) the basic vocabulary can be used to fill in gaps for existing wordnets, or to seed new wordnets. None of the existing wordnets have all of the vocabulary, and the vast majority of wordnets do not have example sentences — adding a few hundred high quality example sentences to basic vocabulary for 30 languages will be a great extension. To do this, we will (a) add it as high confidence entries and (b) send it to upstream projects. When it has been incorporated into new versions of the wordnet we will remove the automatic entries. This will (we hope) allow the projects to also take advantage of the usage notes.

In addition, TUFS has pronunciation for all of the vocabulary, we intend to add links to this from OMW.

(iii) The wordnets contain (or link to) information not in the TUFS vocab, that may be useful in teaching. This includes definitions, pictures (from Imagenet: Deng et al., 2009), synonyms, hyponyms, and for some languages, further examples and more.

Looking purely in terms of words, coverage of the TUFS vocabulary varies very widely for wordnets in the OMW,

Synset ID	10694258-n	
English (en)	lemma	**teacher**, instructor
	def	a person whose occupation is teaching
Arabic (ar)	lemma	مُعلِّم
Mandarin Chinese (zh)	lemma	**教师**, 教练, 老师
French (fr)	lemma	instructeur, **professeur**, **instituteur**, **enseignant**, maître
Indonesian (id)	lemma	pengajar, pendidik, pengasuh, instruktur, **guru**
Japanese (ja)	lemma	イントラ, 先公, **先生**, 指南番, 指導員, 指南役, インストラクター, ティーチャー, ティーチャ, 師匠, 師, 師資, 師範, 師家, 教官, 教師, 教員, 老師
	def	教職の人
Portuguese (pt)	lemma	Magistério, **professor**, instrutor, docente, mestre , mestra, magistério
Spanish (es)	lemma	maestro, **profesor**
Thai (th)	lemma	**อาจารย์**, ผู้ให้ความรู้, ผู้สอน, อ, ครู, ครูบาอาจารย์
Malay (ms)	lemma	jurutunjuk, **cikgu**, pengajar, pendidik, jurulatih, pengasuh, **guru**
Hypernym	educator	
Hyponym	art teacher, bahai catechist, coach, dancing-master, demonstrator, docent, …	

Figure 2: Simplified Wordnet Entry for *teacher*
Lemmas that also appear in the TUFS vocabulary are shown in bold.

先生 (1.2410: Membership) id=19186, pos=n

Language	Lemma	Cleaned	Meaning
Arabic (ar)	أُسْتَاذٌ	أُسْتَاذٌ	［名詞］教授（職名）、先生（呼称）أساتذةأستاذ\n أستاذ,استاذ,أساتذة,اساتذة
Arabic in Syria (as)	إستاذ (أساتذة)	إستاذ (brk:أساتذة)	先生\n
German (de)	Lehrer	Lehrer	先生，教師 <女>Lehrerin
English (en)	teacher	teacher	先生
Spanish (es)	profesor	profesor	先生
French (fr)	instituteur	instituteur	先生。とくに小学校の先生。\n＊女性形はinstitutrice。
French (fr)	professeur	professeur	先生。教授。\n＊男女とも同じ形。
French (fr)	enseignant	enseignant	先生。教員。職業としての先生。
Indonesian (id)	guru	guru	先生
Indonesian (id)	dosen	dosen	先生
Japanese (ja)	先生	先生	【よみ】\n せんせい\n\n【意味】\n teacher\n instructor\n ※ not used when referring to one's own job\n\n doctor\n ※ used when addressing a medical doctor in stead of "Dr. ～"
Central Khmer (km)	គ្រូ	គ្រូ	先生
Korean (ko)	선생님	선생님	先生
Lao (lo)	ອາຈານ	ອາຈານ	先生
Mongolian (mn)	багш	багш	先生
Malay (ms)	guru/cikgu	guru;cikgu	【よみ】\n せんせい\n\n【意味】\n teacher\n instructor\n ※ not used when referring to one's own job\n\n doctor\n ※ used when addressing a medical doctor in stead of "Dr. ～"
Burmese (my)	ဆရာ/ဆရာဝန်/ကျောင်းဆရာ/ဆရာဝန်	ဆရာ;ဆရာဝန်;ကျောင်းဆရာ;ဆရာဝန်	【よみ】\n せんせい\n\n【意味】\n teacher\n instructor\n ※ not used when referring to one's own job\n\n doctor\n ※ used when addressing a medical doctor in stead of "Dr. ～"
Por. in Brazil (pb)	professor	professor	先生
Portuguese (pt)	professor	professor	先生
Russian (ru)	учитель	учитель	小・中・高校の先生\n（大学の先生にはпреподавательを用いる。）
Thai (th)	อาจารย์	อาจารย์	先生
Tagalog (tl)	titser	titser	先生
Turkish (tr)	öğretmen	öğretmen	先生
Urdu (ur)	استاد	استاد	先生
Vietnamese (vi)	cô giáo	cô giáo	先生　（女性の先生を指します）
Vietnamese (vi)	thầy giáo	thầy giáo	先生　（男性の先生を指します）
Mandarin Chinese (zh)	老师 lǎoshī	老师 (py:lǎoshī)	先生

Figure 3: TUFS Translation Set for Teacher

shown in Table 3.[4]

[4]New open wordnets have been released for Turkish, Burmese, Russian and German. We hope to include their results in a follow up study.

3.1 Mapping Process

The mapping process was done through a naive algorithm inspired by multilingual sense intersection (Bond and Foster, 2013; Bond and Bonansinga, 2015; Bonansinga and Bond, 2016). The abstract idea behind multilingual sense intersection has a simple logical foundation: the semantic

Language		Concepts	Words	% in WN
Arabic	ar	2,043	2,928	3
Arabic in Syria	as	1,184	1,053	30
German	de	521	1,221	0
English	en	525	1,255	59
Spanish	es	587	1,207	49
French	fr	969	2,076	73
Indonesian	id	621	899	84
Japanese	ja	829	1,945	47
Central Khmer	km	517	465	0
Korean	ko	527	1,011	0
Lao	lo	526	492	0
Mongolian	mn	553	497	0
Malay	ms	829	829	61
Burmese	my	829	829	0
Por. in Brazil	pb	527	1,238	72
Portuguese	pt	519	477	84
Russian	ru	556	1,102	0
Thai	th	520	469	80
Tagalog	tl	525	1,195	0
Turkish	tr	537	1,959	0
Urdu	ur	755	669	0
Vietnamese	vi	629	723	0
Mandarin Chinese	zh	596	556	66
Total		16,224	25,095	

Table 3: Wordnet coverage of the basic vocabulary
0 values in % in WN means there is no wordnet for this language in OMW (1.2)

space of a polysemous word in any language can be constrained by aligned translations of the same word in other languages. This technique can be used to perform Word Sense Disambiguation (WSD) when parallel text is available – including aligned dictionaries.

Even though a detailed discussion of this algorithm falls outside the scope of this paper, in summary, it leverages multiple levels of information (e.g. language alignments, part-of-speech, number of overlaps per concept, etc.) to generate a ranked list of candidate senses. Naturally, links to entries that have been constrained by a larger number of languages score higher. Concepts constrained by the same number of languages can be ranked by other more nuanced metrics such as: the number of individual lemmas matched in each language, part-of-speech congruency, and ambiguity of each lemma.

In our case, we tried to intersect the parallel data provided by TUFS Open Language Resources with the OMW. This generated, for each entry in TUFS Open Language Resources, a ranked set of possible concept links in the OMW.

The wordnets were linked through the vocabulary to all possible synsets, and then scored.

These links were then evaluated by two students from Tokyo University of Foreign Studies, native speakers of Japanese, with some knowledge of English and Malay. They were asked to match things as: good, bad or questionable.

Their results were then checked by a third bilingual Japanese-English speaker (the last author) who consulted with the first author when unsure. The goal was to link the **basic** sense (as used in the example sentence) to the appropriate wordnet sense (or senses) or, when there was no appropriate sense to link to, a suitable hypernym in wordnet.

An example of a straighforward link is the translation set for 箸 *hashi* “chopstick”, which links to single synset: 03025755-n.

If a word had two common meanings, but only one was shown, we also noted this: for example バス *basu* “bass, bus” - bass, bus. The TUFS examples only talk about the “bus” meaning, although “bass” is also a common interpretation.

The distribution of the final mappings is given in Table 4. The most common mapping type (n=340) is from one translation set to one synset, but there are some that have zero (the hyponym links), many that have two (n=107) and the synset for いい *ii* “good” has twelve. Translation sets that did not link directly are discussed further below.

Links	Number	Example (ja)	Example (en)
0	10	かぶる	wear [on head]
1	340	男の子	boy
2	107	医者	medical doctor
3	32	近い	close [near in time or place or relationship]
4	10	言う	to say
5	1	弱い	weak
7	1	書く	to write
9	1	新しい	new
12	1	いい	good
Total	503		

Table 4: Number of links to OMW per translatoin group

3.2 Non-standard Mappings

Not all concepts mapped directly. There were ten cases where the Japanese concept was not in OMW, but a more general concept existed. In this case, we used a hyponym link. For example, Japanese differentiates between temperature in general and the temperature of objects. This means the English synset 01251128-a - *cold* "having a low or inadequate temperature or feeling a sensation of coldness or having been made cold by e.g. ice or refrigeration - links to both 寒い *samui* "low temperature" and 冷たい *tsumetai* "cold to touch" in TUFS. Neither has exactly the same meaning, they are both more specific than English cold.

The same was true for hot: 暑い *atsui* "hot temperature" and 熱い *atsui* "hot to touch". Therefore, they were linked with a hyponym relation. Interestingly, warm in Japanese also has two variants, but they are merged into the same entry in TUFS as あたたかい(暖・温)*atatakai* with both characters shown. These should probably be split for consistency.

Further, Japanese distinguishes the concept of ***wear*** depending on the clothing's destination on the body. The English synset has the more general concept of wear, which applies to all articles of clothing. Lemmas like 穿く *haku* "to wear on the lower body", かぶる *kaburu* "to put on (a hat)" and 着る *kiru* "to wear on the upper body" were all linked with a hyponym relation.

In addition, the closest English synsets that linked to the Japanese noun 茶わん *chawan* "teacup; rice bowl" are ***cup*** and ***bowl***. These concepts aren't specific enough, as it denotes the bowl used for eating rice out of or a cup for drinking tea out of. Therefore, this Japanese lemma was also linked using a hyponym relation.

Similar to the previous example, 交番 *koubann* "police box" is specific to the Japan. It is a small police office which was originally found only in Japan, although now variations exist elsewhere, such as in Singapore, where it is called a *neighbourhood police post*. Therefore, it needed a hyponym relation to the English synset for ***police station***: 03977678-n.

Surprisingly, the English synset for 来月 *raigetsu* "next month" was not in the OMW, even though the English synset "last month" was: 80000079-n.

In all these cases, new entries should be added to the Japanese wordnet. We will prepare entries for the next release and submit them to the Japanese wordnet. In this way, it will better cover basic Japanese concepts.

Another problem was when the English senses were finer than the Japanese ones. A basic word in the TUFS vocabulary could have multiple meanings. In most cases these were linked to the corresponding synsets as part of one entry.

In two cases, TUFS vocabulary had multiple basic meanings, some of which were not linked to any synset:

- The word 語 *go* "word" can be used to mean word, e.g. 単語 *tango* "word". However, 語 is also a suffix. Commonly, the name of a country is placed before the character, 語, to create the term for its language, e.g. フランス *furansu* "France" + 語 = フランス語 *furansugo* "French". TUFS (and the JLPT) did not make this distinction in the vocabulary database. The best solution is to create another entry in the TUFS database.
- Similarly, あの *ano* "excuse me; umm" is commonly used before a sentence to be polite: *excuse me, where is the bathroom located?*. But, it is also the pause in between words to think. The English equivalent is *umm*. These are two different senses. OMW has the first but not the second. The solution is again to create another entry for あの *ano* "umm" both in OMW and the TUFS database.

Because it is not easy to change the TUFS database, we currently link the above two entries as 'multiple' links, and warn the user that these do not link cleanly.

Another case of possibly problematic links was those that linked accross parts of speech. The translation set for 好き *suki* "to like" was linked as both a verb and an adjective. When it is translated to English, the synset 01777210-v is a correct definition: e.g. "I like jogging". But in Japanese this is an adjective. In another case, the synset 00409709-r described a correct link for the concept 近く *chikaku* "near in time or place or relationship". Lastly, いかが *ikaga* "phrase used when suggesting or recommending something" was correctly linked to two different synsets: 80001331-x and 77000090-n. The synsets described a link that was similar enough that they should be merged in the OMW.

The links are made available on the github site as a tab separated file: tufs-omw-map.tsv. We give a sample below:

translation set id	link type	synset
0577	synonym	07557434-n
99489	synonym	02395115-a
99489	synonym	02396720-a
30521	multi	06286395-n
100468	multi	80000672-x
9206	hyponym	15209413-n

Finally, there were clearly cases where the usage example indicated that some languages had more specific concepts that could be linked. For example, German, French and Vietnamese had different words for male and female teachers. French had different words for primary and university teachers, and Russian for university teachers. This suggests the need for finer divisions of concepts in these languages.

4 OMW Evaluation

In this section, we give the results of measuring the coverage of the wordnets in OMW (1.2), for the 503 translation sets. The results are shown in Table 5. The first column shows the language. Note that we linked Arabic and Arabic in Syria to the Arabic wordnet (Elkateb et al., 2006) and both Portuguese and Portuguese in Brazil to the Portuguese wordnet (de Paiva et al., 2012).

The second column contains the number of translations that exist of the 503 that were linked to OMW. For example, 13 entries were missing English translations even if the synset was linked. The third column shows how many of the lemmas were in the wordnet for this language in OMW. If a translation set linked to more than one lemma, then the score is spread amongst them (so if two out of three had lemmas it gets a score of $\frac{2}{3}$). For Arabic, we noted that TUFS shows vowels but OMW doesn't, so we stripped the vowels before we compared the lemmas.

Language	Number of Translations	% of synsets in OMW	% of lemmas in OMW	overlap of lemmas
Arabic	443	53.9	73.6	39.0
Arabic in Syria	346	58.3	53.3	27.7
English	490	**92.3**	78.5	49.5
Spanish	477	70.6	74.3	**57.5**
French	484	81.7	**78.8**	47.1
Indonesian	483	80.0	71.7	35.5
Japanese	433	81.8	67.4	19.0
Malay	433	78.6	66.1	32.6
Por. in Brazil	457	77.9	78.0	43.7
Portuguese	491	76.5	72.5	40.4
Thai	491	81.2	61.3	36.8
Mandarin Chinese	448	61.7	76.4	36.1

Table 5: Evaluation

For the evaluation, we only look at synsets in the PWN, not the extended OMW. The fact that English only matches for 92.3% of the synsets is because PWN does not include pronouns or exclamations. This shows how important they are for basic vocabulary.

The next column shows the percentage of synsets that had at least one lemma from the TUFS translation set in the OMW, if the synset exists. This shows how well the basic vocabulary is covered. This is analogous to precision. Unsurprisingly English has the highest score here. Most wordnets do fairly well — the word selected for teaching beginners is included over 70% of the time.

Finally, the fifth column indicates how well the two resources match in terms of lemmas, if the synset exists. This is analogous to recall. For example, in Figure 2 Japanese has many lemmas that link to the synset for teacher. A low overlap of lemmas in the TUFS vocab and the OMW shows that perhaps the wordnet has too many entries: it is overcompensating to increase its accuracy. In reality, this makes it more confusing for a beginner learner of a language or to anyone using a dictionary. Here Japanese is by far the worst, but this partly can be explained by the orthographic variation: the wordnet lists multiple variants while the TUFS picks just one. Similarly for Malay and Indonesian, the wordnets list both root and derived forms for verbs, while TUFS only lists the derived forms. Note that a low score for overlap does not necessarily mean all the words are wrong: for example for the synset 00081591-r, TUFS has only *every week*, but OMW has *each week, every week, hebdomadally, weekly*. These are all good synonyms for the concept.

The results were unexpected — we thought the Japanese wordnet would be better than French and Thai, but although it did well in precision, it clearly needs more work on recall. This shows the value of the evaluation.

4.1 Future Work

The work-in-progress OMW 2.0 also has wordnets for Burmese, German, Turkish, with potential wordnets for Russian and Mongolian, as well as new releases for some other wordnets: when it is ready we will redo the evaluation with the new lexicons.

For the time being, however, since many of these languages are unlikely to be become full wordnet projects in the near future, we also plan to convert the linked portion of the TUFS Basic Vocabulary Modules into a small multilingual wordnet project. Using the Global Wordnet Association's new Wordnet LMF format proposed in Vossen et al. (2016), it is now possible to create a wordnet with data for multiple languages. We believe that this would be the ideal format for this lexicon, and it would mean that it would become immediately ready to be exploited by other projects (or individuals) through the the OMW.

5 Conclusions

The TUFS Basic Vocabulary Modules were linked with the Open Multilingual Wordnet. The TUFS Modules were created by hand, which provided basic vocabulary for languages with scarce online language resources like Lao and Khmer. The TUFS interface provides modules which are bilingual resources for international students who wish to study Japanese. Fortunately, each module was connected by an ID (`tid`) that corresponds to a basic Japanese lemma. To make all languages accessible, a multilingual lexicon was produced, combining all translations and links to the respective modules under its `tid`. Additionally, three bilingual speakers mapped the TUFS data to the OMW synsets — two students from TUFS evaluated the algorithm by assigning "confidence scores" and the final speaker (also the last author) further refined the mapping by making sure the links were correct. For example, *warm* should only be linked to the basic definition of it and not, say, an emotional connotation of kindness. This ensured a high quality mapping of the TUFS basic vocabulary to the OMW synsets.

Acknowledgements

This research is supported by the JSPS grant *A collaborative network for usage-based research on lesser-studied languages*. We thank the creators of the individual wordnets and the TUFS vocabulary modules.

References

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proc. of the 8th Global WordNet Conference*, pages 44–49.

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel

corpus. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 56–61. Trento.

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)*, pages 1352–1362. Association for Computational Linguistics, Sofia, Bulgaria. URL http://aclweb.org/anthology/P13-1133.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Global WordNet Conference (GWC2016)*, pages 50–57. Bucharest, Romania.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*. Jeju.

Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMAp technical report, Escola de Matemática Aplicada, FGV, Brazil.

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)*.

Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

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

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue*, LNCS 2546:61–68.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 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.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori, and Yoichiro Tsuruga, editors. 2007. *Corpus-Based Perspectives in Linguistics*, volume 6 of *Usage-Based Linguistic Informatics*. John Benjamins Publishing Company, Amsterdam. URL http://www.jbe-platform.com/content/books/9789027292384.

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.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. TUFS Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, pages 436–439. URL https://www.anlp.jp/proceedings/annual_meeting/2018/pdf_dir/C3-5.pdf.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference 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.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. 419–426.

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, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.