An Abstract Multilingual WordNet

Krasimir Angelov
Chalmers University and University of Gothenburg / Sweden
krasimir@chalmers.se

Abstract

We present a variant of WordNet for 265 languages where the primary constituents of the synsets are abstract identifiers, rather than language specific lexemes. The identifiers are then verbalized to each language through a grammar. Currently, for most of the languages, the grammar only provides lemmas, but for 28 of them, there is also, full morphology and syntax. We review the bootstrapping methodology, evaluate the quality, and show-case applications.

1 Introduction

WordNets exist and are being created for several languages, but more importantly, equivalent synsets across languages are kept linked together. This makes WordNets not just into valuable language-intrinsic resources but also into useful cross-lingual translation dictionaries.

The latter, however, is not without problems. To start with, by going from a particular word sense, in the synset for one language to a synset in another language we lose information and introduce unnecessary ambiguity. For example, for most plant and animal species, the corresponding synset contains both the colloquial name as well as the Latin name of the species. This means that, if one starts from the colloquial word "apple" and goes to another language, they will be forced to accept that "Malus pumila" is a possible translation, despite that most languages have a more reasonable translation.

The above is an extreme example but there are plenty of more subtle cases. For instance, in English, the words "marihuana" and "cannabis" share the same synset, and both words have their cognates in many other languages. If we just go through the linked synsets, then one could freely translate "marihuana" to "cannabis" and vice versa. In this case the semantic difference is not so clear, but if we want to stay close to the original text, it is preferable to use cognates when they exist.

Other synonyms exist for historical reasons, e.g. Thailand vs Siam or Cambodia vs Kampuchea, and using one or the other depends on the sociopolitical context.

In the next section we present the design of an Abstract WordNet, where in addition to synsets with semantic relations between them, we also preserve the translation relation. Obviously, this can be done in many ways, but the particular choice makes the resource compatible with Grammatical Framework (Ranta, 2011). The model has also been tried for a smaller set of languages already in Angelov (2020).

Grammatical Framework is a formalism centered around the concept of abstract syntax. In short, the abstract syntax is a collection of functions with fixed types, where the functions themselves are implemented in different ways for different languages. For example adjectival modification on an abstract level can be:

$$\mathrm{AdjCN} : \mathrm{AP} \text{ -> } \mathrm{CN} \text{ -> } \mathrm{CN}$$

i.e. it is a function which takes an adjective phrase and a common noun and returns a new common noun. The word order, the agreement and the possible inflections are defined separately for each concrete language.

Words are functions with no arguments, e.g.:

$$\mathrm{apple\_1\_N} : \mathrm{N}$$
$$\mathrm{apple\_2\_N} : \mathrm{N}$$

may represent the two senses of "apple" (the fruit and the tree).

2 Design for an Abstract WordNet

While the synsets of a traditional WordNet (Princeton, 2006) are language specific and contain lemmas, we choose to make them abstract – each synset contains names of functions which are defined separately in the grammar for each language.

By design, the words that we choose to represent a function in each language should be as close as

possible, meaning that they must share as many senses as possible. This also makes the choice robust in translation – when going from one language to another, the chosen translations will remain valid for as many senses as possible. This is advantageous since even if the model cannot pick the exact sense, the translation is still likely to remain correct. When relevant, these should be cognates or borrowed words, but only if these have not changed meanings or register.

For example, "house" and "hus" (Swedish) are good translation equivalents regardless of the sense. On the other hand, "familj" (Swedish) is not a good translation despite that "house" and "family" share one sense in English (a social unit living together). The word "family" should rather translate as "familj". As it happens both examples are also cognates.

Obviously, a tight one-to-one alignment between languages will be problematic. For instance, there are words which simply cannot be translated to another language. In that case we represent the gap by just leaving the corresponding function undefined. In other cases, a word may translate to a multiword expression. That is supported, since the abstract functions can also generate complex phrases. Finally, some languages provide more ways to express a concept than others. In that case, we either use the same word to represent several functions, or we leave some functions undefined.

Since all the data is initially automatically generated, we also store the current status per language. First of all, a definition might be missing either because there is a lexical gap, or because we simply do not have enough data. We cannot know the difference without consulting a native speaker.

When we do know the definition, it is either already **validated**, **unchecked** or **guessed**. The difference between the last two is that for an unchecked definition we know that it represents the right sense but maybe it is not the best translation. On the contrary, guessed definitions are possible translations but not necessarily for the same sense. The unchecked definitions come from existing monolingual WordNets, e.g. we used the Open Multilingual WordNet (Bond and Foster, 2013) extensively. On the other hand, guessed definitions are extracted from translation dictionaries.

Another notable difference is that while a typical WordNet contains only lemmas, in our implementation each function normally computes a full inflection table. This is necessary, if we want the definitions to be usable in combination with the existing syntactic combinators provided by the framework. Currently full morphology is only provided for 28 languages, but we aim at extending that to all of the languages.

3 Data Collection

The creation of the initial Abstract WordNet was reported in Angelov (2020), where the focus was on Swedish, Bulgarian and English, and the compatibility with Grammatical Framework. In the process existing resources for the two languages (Borin et al., 2013; Viberg et al., 2002; Borin et al., 2008; Kann and Hollman, 2011; Simov and Osenova, 2010) were also absorbed. At the time, lexicons for 11 other languages were also bootstrapped from PanLex (Kamholz et al., 2014) by using the method in Angelov and Lobanov (2016). Since then, 10 more languages were added, and the data has been incrementally cleaned up. As we will see in Section 5, the status per language is still varying.

Here we want to extend the lexicon with as many languages as possible, and we even include some low-resource languages. Since we rely on translation dictionaries, the quality of the existing data affects the quality of the new translations. The cleanup for the original languages is therefore beneficial when adding new languages.

The main source for both the old languages and the new ones is PanLex - a collection of translation dictionaries for thousands of language pairs. The problem with PanLex is that although we get translation pairs, we cannot know in what sense the translation is appropriate. This is partly resolved by using Wikidata - a collaborative taxonomy created by the Wikimedia community which has labels in many languages and a partial mapping to WordNet.

While the previous work relied on self-made links from WordNet synsets to English articles in Wikipedia, here we switched to Wikidata. The advantages are several. First of all Wikidata contains structural information rather than plain text. Furthermore, it is a direct hub between different Wikipedia editions which makes it easier to query for labels in different languages. Finally, Wikidata has entities with no corresponding English articles, and those can still be linked with WordNet.

Wikidata has the property P8814 which is its way to link to WordNet. To these links, we added new ones by projecting the Wikipedia links created in Angelov (2020). In the process the number of links

almost doubled and many mistakes were fixed on both sides. Currently there are 30416 links. Unfortunately, since Wikidata only consists of concepts, the links point only to nouns in WordNet.

To compensate for that we also used Wiktionary, which we accessed through the Wiktextract tool (Ylonen, 2022) which reads the raw markup and extracts structured data. The data contains both morphology and translations, but like for PanLex, the senses for which the translations apply are not linked to WordNet. We created our own mapping.

Each abstract function, we verbalized in English and matched it to an entry in the Wiktextract data with the same lemma and part of speech tag. So far, this means that different WordNet senses will map to the same entry. After that we check the glosses. For example, for "apple" in Wiktionary we have four glosses: *fruit*, *tree*, *wood* and *apples and pears*. Obviously the first two must correspond to the identifiers $\mathrm{apple\_1\_N}$ and $\mathrm{apple\_2\_N}$ above (Page 1). To find the best match, we computed the SBERT similarity (Reimers and Gurevych, 2019) between the possible WordNet and Wiktionary glosses.

When we printed the candidate matches in the order of the descending score, the candidates at the top matched very well, while the candidates at the bottom had nothing in common. Unfortunately, there is no clear cut-off point. We started looking at the matches from bottom to top. At score 0.3148 we observed that 3 out 5 candidates are correctly matched. At score 0.4 we got 4 out of 5. Finally, at level 0.45, 10 out of 10 candidates were correctly matched. We used that as a cut-off point and discarded everything with a lower score. We realize that we discarded a lot of useful data in this way but at the same time we are pretty sure that what we retained is very well matched.

In the rest of the data, it is still possible that an abstract function maps to several Wiktionary glosses. In that case, we just retained the highest scoring match. The exception is when the two top-most scores are very similar, in that case SBERT cannot make good distinctions. For that purpose, if the difference is less than 0.005 points, we looked manually at the two candidates. There were only 312 of those. At the end we got 40652 relations from abstract functions to Wikitionary glosses which unlike Wikidata contain different parts of speech.

4 Finding New Translations

From the content of the existing WordNet, PanLex and Wikidata, we construct the matrix T which represents the abstract lexicon together with the translations in all languages. Here for every function f and every language l, the element T_{fl} is a set of items of the form:

$$\langle \text{lemma}, s, w, l, c, d \rangle$$

where we have the language specific lemma followed by a number of scores. For the definition of the scores, it is also useful to define the set:

$$C_{fl} = \{\text{lemma} \mid \langle \text{lemma}, s, w, l, c, d \rangle \in T_{fl'}, l' \neq l\}$$

i.e. all lemmas for the same function but for a different language l'. The scores are:

s (status) is 0 if the verbalization is already validated, 1 if unchecked and 2 if guessed

w (wiki) is 0 if the verbalization appears in Wikidata and 1 otherwise

l (languages) is the number of sources in PanLex which claim that the current lemma is a translation for one of the lemmas in C_{fl}.

c (co-occurences) is the number of pairs of linked synsets for two different languages in which the current lemma co-occurs together with a lemma in C_{fl}

d (distance) is the shortest Levenshtein distance between the current lemma and any other lemma in C_{fl}.

The matrix is constructed by first inserting all existing translations in the current WordNet. For those we retrieve the current status s. If Open Multilingual WordNet has data for a language, we insert that as well with $s = 1$. After that we add translations from Wikidata, for these w is always 0, and $s = 2$ unless the translation was already added in the previous step with a different status. Finally we add translations from PanLex with $s = 2, w = 1$ unless the translation was already added with different scores.

Once we have collected all the data, we compute the scores l, c and d for all lemmas by looking at the lemmas in C_{fl}. Since here we compare the candidate lemmas for a new language with the already existing ones, it pays off if there are already many existing and well cleaned up languages.

Here it is also crucial that we work with abstract functions and not complete synsets. If we were using the synsets, this would add too many ambiguities when considering all possible synonyms in all possible languages.

Score l helps us to select translations which are recommended in most dictionaries. By looking at several languages and intersecting their dictionaries, we narrow down the right new lemma, even if the existing lemmas in some of the languages are ambigous. The efficiency of this criteria is obviously limited by the existance of multiple dictionaries and the number of already present lemmas.

Score c is higher for robust translations like the examples "house-hus" and "family-familj" in Section 2. Finally d helps us to put together cognates which often look similar. The later is especially relevant for closely related languages.

The best translation for each function f and language l is selected by sorting the set T_{fl} by the key $(s, -c, -l, w, d)$ in descending order. In other words, we prefer translations that are possibly validated (the s score), and robust (the $-c$ score), are cited by more translation dictionaries (the $-l$ score), and if possible are mentioned in Wikidata. Finally, since we also sort by the d score, we select candidates which look the most like a translation in another language if all of the previous criteria are the same.

From the sorted list we always pick the first lemma and we compute its new status as follows:

- if $s = 0$, the lemma retains its status as validated. This rule is useful when a language is regenerated to take into account new external data. In that case we want to keep already validated translations.

- if $s = 1$ and $l > 1$, the lemma is marked as validated. The intuition is that previously, we only knew that a lemma had the right sense but now we also found more than one translation dictionaries which also list it as a good translation.

- if $w = 0$ and $l > 1$ then the lemma is also marked as validated. This happens when the lemma is used as a label for a Wikidata entity. Since those are linked on the sense level, $w = 0$ has the same semantics as $s = 1$, i.e. the lemmas share the same sense but maybe are not good translations. If on the other hand the label is also confirmed by more than one translation dictionary, then we can accept it.

- if $w = 0$ and $s = 1$ then the lemma is confirmed to have the right sense by both Wikidata and an existing WordNet for the language. Since here we have a synergy of two independent semantic sources, we mark those entries as validated although we cannot be completely sure that this is the best translation.

- if $s = 1$ the lemma remains unchecked.

- in all other cases, we treat the lemma as guessed.

5 Evaluation

We started with the evaluation of the existing languages by comparison with Wiktionary. The statistics are on Table 1.

We first looked at already validated abstract identifiers and compared them with existing Wiktionary lemmas. The first two columns on the table show the number of such cases as well as how often the two definitions match. As we expected, there is no perfect agreement but nevertheless it is quite high.

When we looked manually at cases where there is a disagreement, we observed examples where there is more than one possible translation and none is a better choice than another. This means that naturally, when the grammars and the lexicon are used for translation or natural language generation, the choices that we made will carry a particular style. This is not dissimilar from human translators who tend to use certain words more frequently than others.

After that we focused on abstract identifiers that are not validated yet. Again, we counted how often the two lemmas match, and we show the statistics in the columns "Confirmed" and "Not Confirmed". As we can see there are many WordNet definitions which have not been checked yet, but by comparing with Wiktionary we can confirm that they were indeed correct. In other cases, we cannot confirm the validity of the translations yet. They may be wrong, but they may also be just alternatives choices. In any case, after the evaluation we changed the status of all confirmed definitions to validated, but choose not to remove the unconfirmed yet.

Finally, there are cases where Wiktionary has a translation, but we do not. In that case we just inserted those in our data set. The number of cases is

shown in the last column. There we have the absolute number of insertions as well as the percentage of holes that we filled in in that way.

Figure 1 shows the overall status of the lexicon before and after the update. There, green color corresponds to validated definitions, yellow unchecked, and red guessed. The height of the column shows the relative number of definitions. For each language there are two bars. The first one represents the status before the comparison with Wiktionary, and the second the status after the update.

Now we turn our attention to the newly added languages. All languages and their absolute sizes are shown on Table 3. The already existing languages are marked with an asterisk after the name. For them the percentage of validated items is generally much higher.

For all other languages the entries are simply extracted by using the algorithm from Section 4, and we have not done any manual validation yet. For those languages, we only have the lemma, the morphology and syntax are not integrated yet.

As we can see the sizes of the lexicons vary widely, and it is dependant on how much data we can find. In the collection, we only included languages for which we can find at least 5000 lemmas. The percentage after every language shows how many of the translations that we selected match the entries in Wikidata or Wiktionary. The number depends on both the quality of the selection and the actual size of Wikidata and Wiktionary for a particular language. In general, we expect that many more translations are correct but at the moment we have nothing else to compare those against.

6 Applications

Apart from semantic tasks where the main issue is the semantic interrelations between words inside the same language, the Abstract WordNet can also be used in translation and natural language generation. The key ingredient here is the integration with syntax. A recent example is Angelov et al. (2024) where Wikipedia articles for countries were automatically generated from information in Wikidata.

This is done by automatic generation of abstract syntax trees which are then verbalized to each language. The lexicon comes from the WordNet while the syntax from the libraries.

Since there are still mistakes in the lexicons, the initial draft of each article contained errors. On the other hand, we get a rapid prototype for multiple languages with no extra cost. The evaluation of the prototype is reflected on Table 2, which we copied from Angelov et al. (2024). As we can see the BLEU scores vary and roughly reflect the age of the resources. Swedish and Bulgarian have the highest scores, but they are also the oldest languages. On the other hand, Russian is only a recent addition.

After only fixing a few words the BLEU scores rappidly go to over 80%. Changing only the lexicon is often not enough to go to 100% since there are also idiomatic uses of the language which are not captured in the shared abstract syntax. The only exception here is English but this is only because the initial program was made to generate correct English to start with. The final gap is closed only by producing slightly different abstract expressions for every language.

By incorporating more languages, we aim to make this kind of rapid prototyping of language applications, accessible for all languages, even for low-resource ones. This also aligns with the goal of Abstract Wikipedia (Vrandečić, 2020) which aims to make Wikipedia more widely accessible.

7 Conclusion

We showed an alternative design for a WordNet which is integrated with a multilingual grammar, and which can be used for translation and natural language generation. We also show how the lexicon can be extended to hundreds of languages.

For the new languages we used Wiktionary only for evaluation, but it can be utilized better if we also used it during the selection of translations just like we used Wikidata. This will potentially increase the lexicon sizes for some languages and will validate more entries. On the other hand, if we do this then we will have nothing to evaluate against. We leave that therefore as a future work.

Only 28 languages are currently integrated with the corresponding grammars although the framework provides a library with more than 40 languages. For the integration it is crucial that we add morphology as well, which is necessary for the syntactic combinators that must inflect the words in the right way.

The existing grammars provide a morphological API which given one or more forms constructs the rest of the inflection table. The accuracy of the API depends on the language and on how irregular a particular word is (Détrez and Ranta, 2012). A key

	Existing								Inserted	
	Matching		Conflicting		Confirmed		Not Confirmed			
Africans	699	(91.97%)	61	(8.03%)	1627	(71.80%)	639	(28.20%)	248	(7.57%)
Bulgarian	9730	(62.60%)	5814	(37.40%)	1976	(34.67%)	3723	(65.33%)	1518	(6.67%)
Catalan	8783	(86.26%)	1399	(13.74%)	3318	(46.58%)	3805	(53.42%)	333	(1.89%)
Chinese	216	(67.71%)	103	(32.29%)	53	(0.78%)	6770	(99.22%)	8826	(55.27%)
Dutch	7624	(84.33%)	1417	(15.67%)	4274	(43.71%)	5505	(56.29%)	277	(1.45%)
Estonian	2729	(86.91%)	411	(13.09%)	1942	(54.50%)	1621	(45.50%)	100	(1.47%)
Finnish	17737	(61.12%)	11285	(38.88%)	428	(25.99%)	1219	(74.01%)	654	(2.09%)
French	13151	(74.73%)	4446	(25.27%)	3468	(42.56%)	4680	(57.44%)	532	(2.02%)
German	7304	(78.54%)	1996	(21.46%)	9205	(50.98%)	8850	(49.02%)	472	(1.70%)
Italian	9477	(83.85%)	1825	(16.15%)	4391	(42.68%)	5896	(57.32%)	394	(1.79%)
Korean	1865	(93.02%)	140	(6.98%)	4276	(51.46%)	4034	(48.54%)	180	(1.72%)
Maltese	153	(80.10%)	38	(19.90%)	1330	(73.64%)	476	(26.36%)	142	(6.64%)
Polish	7621	(86.60%)	1179	(13.40%)	4307	(48.93%)	4496	(51.07%)	2951	(14.36%)
Portuguese	12526	(69.00%)	5627	(31.00%)	3114	(53.08%)	2753	(46.92%)	402	(1.65%)
Russian	2958	(87.98%)	404	(12.02%)	16537	(71.59%)	6561	(28.41%)	345	(1.29%)
Slovenian	3924	(64.64%)	2147	(35.36%)	542	(50.89%)	523	(49.11%)	163	(2.23%)
Somali	69	(84.15%)	13	(15.85%)	139	(39.60%)	212	(60.40%)	102	(19.07%)
Spanish	10062	(81.05%)	2353	(18.95%)	6458	(50.96%)	6215	(49.04%)	361	(1.42%)
Swahili	39	(76.47%)	12	(23.53%)	539	(67.46%)	260	(32.54%)	3366	(79.84%)
Swedish	10182	(78.94%)	2717	(21.06%)	2693	(56.83%)	2046	(43.17%)	1416	(7.43%)
Thai	2087	(70.48%)	874	(29.52%)	475	(19.02%)	2022	(80.98%)	866	(13.69%)
Turkish	3388	(85.45%)	577	(14.55%)	3709	(47.78%)	4054	(52.22%)	332	(2.75%)

Table 1: Evaluation on Existing Languages

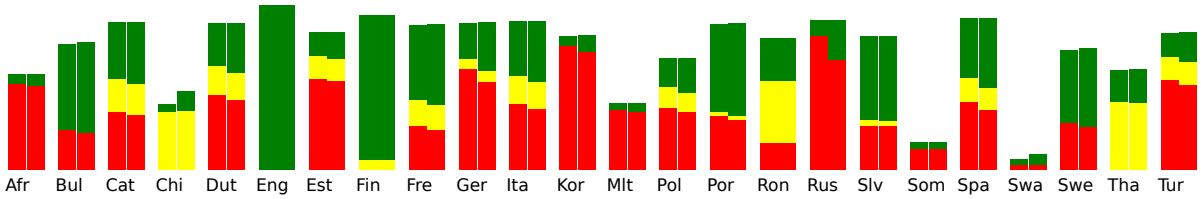

Figure 1: Status of the different languages before and after the validation with Wiktionary

	Initial Draft				Improved Draft			
	1-gram	2-gram	3-gram	4-gram	1-gram	2-gram	3-gram	4-gram
Bulgarian	80.04	72.94	67.05	**61.12**	99.17	98.88	98.65	**98.41**
English	94.08	93.13	92.19	**91.19**	100.00	100.00	100.00	**100.00**
French	64.43	53.94	44.80	**38.01**	95.60	93.45	91.10	**88.75**
Russian	43.21	26.01	17.28	**11.77**	93.12	88.82	85.68	**83.28**
Spanish	73.57	62.48	52.82	**44.75**	96.13	93.75	91.10	**88.31**
Swedish	81.82	76.44	71.67	**67.01**	99.26	98.94	98.75	**98.57**

Table 2: BLEU scores for the generated articles after each phase.

aar	Afar	9846	7%	abk	Abkhazian	13057	10%	ace	Achinese	6002	22%
ady	Adyghe	9304	12%	afr	Afrikaans*	70320	13%	als	Albanian (Tosk)	67430	4%
alt	Southern Altai	13788	3%	amh	Amharic	18565	13%	ang	Anglo-Saxon	42773	6%
arc	Aramaic	6906	19%	arg	Aragonese	44459	6%	ary	Arabic (Moroccan)	7380	21%
arz	Arabic (Egyptian)	31153	14%	asm	Assamese	23934	6%	ast	Asturian	60580	13%
ava	Avaric	26089	5%	aym	Aymara	22969	6%	azb	South Azerbaijani	10482	24%
azj	North Azerbaijani	52518	11%	bak	Bashkir	34063	12%	bam	Bambara	30442	4%
ban	Balinese	11438	16%	bar	Bavarian	14240	14%	bcl	Central Bicolano	10551	15%
bel	Belarusian	73977	13%	ben	Bengali	69110	9%	bul	Bulgarian*	94691	67%
bis	Bislama	12827	9%	bjn	Banjar	5443	19%	bod	Tibetan	22449	8%
bre	Breton	63208	9%	bos	Bosnian	45965	9%	bxr	Buriat	12127	12%
cat	Catalan*	97751	44%	cha	Chamorro	36070	3%	che	Chechen	31269	10%
chu	Old Church Slavonic	17400	7%	chv	Chuvash	26044	11%	ceb	Cebuano	36381	13%
ces	Czech	103625	12%	cho	Choctaw	6634	11%	chr	Cherokee	18943	6%
chy	Cheyenne	17770	6%	ckb	Central Kurdish	46370	6%	cmn	Chinese*	110165	13%
cor	Cornish	45052	6%	cos	Corsican	26614	8%	csb	Kashubian	18794	8%
crh	Crimean Tatar	27799	6%	cym	Welsh	73184	13%	dan	Danish	86194	24%
deu	German*	105914	33%	diq	Dimli	13296	16%	div	Divehi	8122	16%
dsb	Lower Sorbian	38371	6%	dzo	Dzongkha	31461	3%	ell	Greek	98997	19%
eng	English*	114563	100%	epo	Esperanto	93748	20%	est	Estonian*	98839	19%
eus	Basque	88130	26%	ewe	Ewe	17656	6%	ext	Extremaduran	8183	18%
fas	Persian	90365	14%	fao	Faroese	56735	8%	fin	Finnish*	112469	86%
fij	Fijian	11680	10%	fra	French*	106546	53%	frc	French, Cajun	9057	9%
frp	Franco-Provençal	25660	7%	frr	Northern Frisian	13301	19%	fry	Western Frisian	53486	8%
fur	Friulian	39397	6%	gle	Irish	78437	17%	gag	Gagauz	8292	16%
gan	Gan	9734	11%	gcr	Guianese	37803	1%	gla	Gaelic	62484	13%
glg	Galician	86233	22%	got	Gothic	12563	11%	grc	Ancient Greek	37803	6%
gsw	Alemannic	14891	12%	guj	Gujarati	71884	5%	glv	Manx	78691	5%
hau	Hausa	36862	6%	hak	Hakka Chinese	37624	5%	hat	Haitian Creole	39155	8%
haw	Hawaiian	52008	3%	heb	Hebrew	76553	13%	hin	Hindi	92061	9%
hrv	Croatian	90951	21%	hsb	Upper Sorbian	30878	9%	hun	Hungarian	96184	21%
hye	Armenian	80531	15%	ibo	Igbo	22518	5%	ido	Ido	74231	12%
ile	Interlingue	52695	4%	ina	Interlingua	74901	7%	ind	Indonesian	82157	23%
iii	Sichuan Yi	22938	3%	iku	Inuktitut	20087	6%	ilo	Iloko	24153	7%
inh	Ingush	12935	9%	isl	Icelandic	79210	11%	ita	Italian*	103916	41%
jam	Jamaican Creole	24284	5%	jav	Javanese	22267	12%	jbo	Lojban	24144	5%
jpn	Japanese	102801	35%	kaa	Karakalpak	7765	19%	kat	Georgian	73174	15%
kab	Kabyle	28291	5%	kal	Kalaallisut	6912	19%	kan	Kannada	23651	17%
kau	Kanuri	6282	11%	kaz	Kazakh	63718	10%	kbd	Kabardian	7346	13%
kcg	Tyap	5083	5%	kik	Kikuyu	8854	11%	kin	Kinyarwanda	26479	4%
khm	Khmer	43396	8%	kor	Korean*	95411	13%	krc	Karachay-Balkar	17052	6%
ksh	Colognian	19661	5%	koi	Komi-Permyak	11219	9%	kur	Kurdish	52224	6%
kpv	Komi-Zyrian	10880	12%	kir	Kyrgyz	45678	10%	lad	Ladino	6514	21%
lao	Lao	39200	6%	lat	Latin	74951	8%	lav	Latvian	71512	10%
lbe	Lak	25298	4%	lez	Lezghian	28201	5%	lfn	Lingua Franca Nova	58802	3%
lin	Lingala	19252	8%	lit	Lithuanian	71903	16%	lim	Limburgish	31220	7%
ltz	Luxembourgish	44917	10%	lug	Luganda	24644	3%	lij	Ligurian	35554	5%
lld	Ladin	26685	8%	lmo	Lombard	11674	19%	ltg	Latgalian	13349	9%
lzz	Laz	5558	13%	mah	Marshall	28233	3%	mal	Malayalam	41816	10%
mar	Marathi	65326	6%	mcn	Masana	7245	0%	mdf	Moksha	40323	5%
mhr	Eastern Mari	59020	3%	min	Minangkabau	6051	24%	mkd	Macedonian	70062	15%
mlg	Malagasy	18860	16%	mnw	Mon	7650	1%	mon	Mongolian	71589	6%
mrj	Western Mari	5838	24%	mlt	Maltese*	52795	12%	mwl	Mirandese	9496	15%
mya	Burmese	33043	10%	myv	Erzya	41797	4%	mzn	Mazanderani	7821	24%
nan	Min Nan Chinese	17795	19%	nav	Navajo	29752	9%	nap	Neapolitan	31818	5%
nau	Nauru	6692	18%	nds	Low German	18514	11%	nep	Nepali	21344	10%
nld	Dutch*	101817	34%	nno	Norwegian Nynorsk	53446	18%	nob	Norwegian Bokmål	90040	17%
nov	Novial	26373	5%	nya	Nyanja	11727	8%	oci	Occitan	11891	30%
ori	Oriya	9334	15%	oss	Ossetian	33339	9%	pag	Pangasinan	10459	9%
pap	Papiamento	47770	4%	pam	Pampanga	12181	12%	pcd	Picard	37787	4%
pli	Pali	12311	9%	pms	Piedmontese	14957	15%	pnb	Western Panjabi	10070	29%
pol	Polish*	99392	25%	por	Portuguese*	101483	64%	prg	Prussian	10926	7%
pus	Pushto	8651	20%	que	Quechua	11782	18%	rmy	Vlax Romani	12100	10%
roh	Raeto-Romance	34172	4%	ron	Romanian*	96712	32%	rue	Rusyn	7171	21%
run	Rundi	10561	7%	rup	Aromanian	26719	7%	rus	Russian*	109088	26%
sag	Sango	6360	13%	sah	Sakha	29653	8%	san	Sanskrit	47117	4%
sco	Scots	30370	14%	scn	Sicilian	51852	8%	sgs	Samogitian	9997	18%
shi	Tachelhit	10449	9%	shn	Shan	9061	12%	sin	Sinhalese	22048	13%
srd	Sardinian	31002	9%	snd	Sindhi	10598	15%	sma	Southern Sami	15527	6%
sme	Northern Sami	52119	6%	sms	Skolt Sami	11825	11%	slk	Slovak	91588	20%
slv	Slovenian*	87577	68%	smo	Samoan	12791	10%	smn	Inari Sami	13721	11%
sna	Shona	22913	7%	som	Somali*	22279	21%	spa	Spanish*	105949	46%
sqi	Albanian	54189	11%	srp	Serbian	59483	12%	srn	Sranan	29389	4%
sot	Southern Sotho	9226	10%	stq	Saterland Frisian	12528	12%	sun	Sundanese	18926	10%
swa	Swahili*	10349	82%	swe	Swedish*	99623	55%	szl	Silesian	9065	28%
tam	Tamil	69361	8%	tah	Tahitian	14139	7%	tat	Tatar	33533	8%
tel	Telugu	70048	9%	tet	Tetum	27917	4%	tgk	Tajik	39981	12%
tha	Thai*	102483	23%	tgl	Tagalog	60739	10%	tir	Tigrinya	26006	4%
tsn	Tswana	26531	3%	ton	Tonga	25382	4%	tpi	Tok Pisin	29355	5%
tur	Turkish*	95333	22%	tuk	Turkmen	52711	6%	tyv	Tuvinian	32380	4%
udm	Udmurt	21349	8%	uig	Uighur	23474	8%	ukr	Ukrainian	80760	16%
urd	Urdu	71483	9%	uzb	Uzbek	42173	14%	vec	Venetian	36120	10%
vep	Veps	14417	14%	vie	Vietnamese	97153	11%	vls	West Flemish	9407	17%
ven	Venda	14057	6%	vol	Volapük	38529	14%	vro	Võro	12109	13%
wln	Walloon	32361	9%	war	Waray	16651	22%	wol	Wolof	19317	8%
wuu	Wu Chinese	11813	27%	xal	Kalmyk	18634	9%	xho	Xhosa	11767	10%
xmf	Mingrelian	5658	39%	yid	Yiddish	53512	8%	yor	Yoruba	46988	4%
yue	Yue Chinese	54050	9%	zsm	Standard Malay	80181	24%	zul	Zulu	29660	1%

Table 3: List of languages with lexicon size and per cent of validated word senses

feature of the API is that it can always produce the right inflection but for irregular cases it will require more forms as input. By using it in combination with the inflection tables that Wiktionary provides, we can find how to use the API in the best way.

There are still more than 200 languages for which we do not have any grammars. We started looking into how the morphology can be learned automatically by using the inflection tables in Wiktionary as examples. Albanian, Kazakh and Macedonian are three pilot languages where we first attempted that. At least some of the syntactic combinators are easy to learn as well. This is definitely future work that we want to pursue.

The extracted lexicons are available on GitHub: https://github.com/GrammaticalFramework/gf-wordnet and can be queried through the search interface here: https://cloud.grammaticalframework.org/wordnet/

References

Krasimir Angelov. 2020. A parallel WordNet for English, Swedish and Bulgarian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3008–3015, Marseille, France. European Language Resources Association.

Krasimir Angelov, Andrea Carrión del Fresno, Ekaterina Voloshina, and Aarne Ranta. 2024. Leveraging grammatical framework and wordnet for natural language generation from wikidata.

Krasimir Angelov and Gleb Lobanov. 2016. Predicting translation equivalents in linked wordnets. In *The 26th International Conference on Computational Linguistics (COLING 2016)*, page 26.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *In 51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362, Sofia.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Grégoire Détrez and Aarne Ranta. 2012. Smart paradigms and the predictability and complexity of inflectional morphology. In *EACL*, pages 645–653. The Association for Computer Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Viggo Kann and Joachim Hollman. 2011. Slutrapport för projektet Folkets engelsk-svenska lexikon.

Princeton. 2006. WordNet 3.0 Reference Manual. http://wordnet.princeton.edu/doc.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Åke Viberg, Kerstin Lindmark, Ann Lindvall, and Ingmarie Mellenius. 2002. The Swedish WordNet project. In *Proceedings of Euralex*, pages 407–412.

Denny Vrandečić. 2020. Architecture for a multilingual Wikipedia. *Preprint*, arXiv:2004.04733.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 1317–1325.