Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4211–4217
Marseille, 11–16 May 2020

Multilingual Dictionary-Based Construction of Core Vocabulary

Winston Wu, Garrett Nicolai, David Yarowsky
Center for Language and Speech Processing
Johns Hopkins University
(wswu, gnicola2, yarowsky)@jhu.edu

Abstract

We propose a new functional definition and construction method for core vocabulary sets for multiple applications based on the relative coverage of a target concept in thousands of bilingual dictionaries. Our newly developed core concept vocabulary list derived from these dictionary consensus methods achieves high overlap with existing widely utilized core vocabulary lists targeted at applications such as first and second language learning or field linguistics. Our in-depth analysis illustrates multiple desirable properties of our newly proposed core vocabulary set, including their non-compositionality. We employ a cognate prediction method to recover missing coverage of this core vocabulary in massively multilingual dictionary construction, and we argue that this core vocabulary should be prioritized for elicitation when creating new dictionaries for low-resource languages for multiple downstream tasks including machine translation and language learning.

Keywords: core vocabulary, Swadesh list, multilingual dictionaries

1. Introduction

Dictionaries are available for most of the world's languages, but coverage can be sparse for those with fewer resources. In sparse dictionaries, many entries are *core vocabulary* words from lists such as the Swadesh list (Swadesh, 1952; Swadesh, 1955), probably the most well-known formulation of a core vocabulary containing around 100–200 words, depending on the version. This list of basic words is used in historical comparative linguistics to determine the relationships between languages, and there have been many attempts to revise or expand these concept lists for this purpose. (See List et al. (2016) for a recent survey and compilation of such lists.)

Morris Swadesh chose the words in the Swadesh lists based on certain criteria: the words should be culturally universal, stable over time (not likely to change meaning), and not likely to be borrowed. Swadesh lists now exist in over 1000 languages and can be used as a dictionary to perform lexical translations. However, in a low-resource setting, the ability to translate a mere 100 concepts is insufficient for understanding in a language. In addition, the Swadesh list, like many other lists, was manually created and revised through years of experience and extensive fieldwork. Inspired by these shortcomings, we propose a novel data-driven criterion for a core vocabulary list: high coverage in dictionaries of different languages.

This paper presents the automatic creation of a core vocabulary list based on the number of entries a concept has in dictionaries. That is, the criterion for our inclusion in our list is the consensus of many lexicographers who deemed a word important enough for inclusion in a language's (possibly small) dictionary. The top entries of our list are presented in Table 1. We empirically find that roughly 3000 words is an adequate size for the list, which is on par with other major core vocabulary lists. In-depth analysis illustrates that due to substantial overlap with several established lists, our core vocabulary can serve well for downstream tasks such as language phylogenetics and language learning. In terms of low resource languages, our core vocabulary consists of words that should be prioritized for elicitation should they not exist in a dictionary. We also successfully experiment on the task of dictionary induction by generating these core words with cognate prediction models.

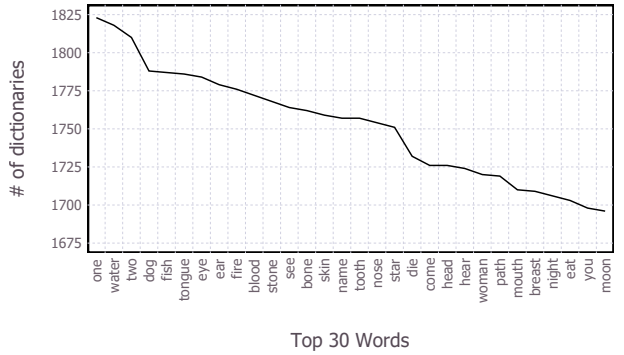


Figure 1: Coverage of the top 30 most common words across all languages in Panlex.

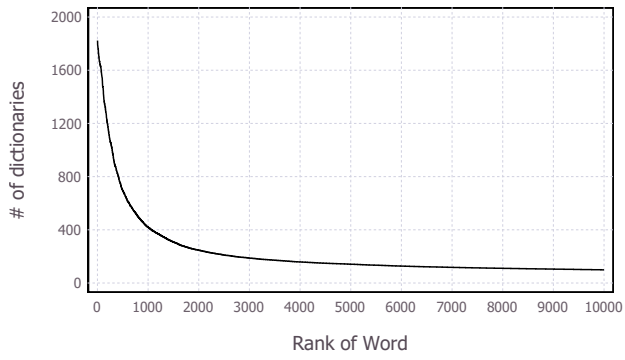


Figure 2: Coverage of the top 10,000 most common words across all languages. A zoomed out version of Figure 1.

2. Construction

For the construction of our core vocabulary, we utilize LanguageNet[1], a multilingual lexicon that is a subset of PanLex (Baldwin et al., 2010), a freely available multilingual dictionary. PanLex contains lexical translations across thousands of the world's languages and has recently garnered interest in the multilingual research community. Its lexical translations are sourced from existing dictionaries and thesauri such as Wiktionary and WordNet. LanguageNet, as of September 2019, contains 1895 languages.

We employ a simple procedure: using English as a pivot, we collect counts of how many languages have a translation for each English concept. (This dictionary pivoting strategy has previously been applied to model color terminology (McCarthy et al., 2019).) The concepts are then sorted in decreasing order by this count, resulting in our core vocabulary list. Up until recently, such a computational procedure would have been impossible without the computing resources and datasets available today.

Figure 1 shows the top 30 concepts along with the number of dictionaries that contain them.[2] The fact that so many languages' dictionaries contain these words is a strong indicator of the coreness of these words. This point is even more salient for dictionaries of low-resource languages: that so many lexicographers have included these words in their language's dictionary is a testament to the word's importance in the language and thus should be included in a list of core vocabulary. Figure 2 shows the rank of each concept (in the core vocabulary) and the number of languages containing the concept. The curve follows a typical exponential (Zipfian) decay, and we see that the top 1000 words are (at least) contained in roughly 500 languages. Using this curve, we see that around rank 3000 is when the curve begins to drastically flatten out, pointing to a reasonable number for the size of a core vocabulary list. For this work, we assume the top 3000 words as our core vocabulary list. Indeed, several other existing lists contain a similar number of words, affirming our choice of vocabulary size.

3. Analysis of Core Vocabulary

Linguists have always been interested in core vocabulary, and there have been many existing approaches for constructing sets of core words. Many of these lists share a substantial number of words, but the lists differ in the purpose of their construction. We examine two motivations: establishing linguistic relationships, and facilitating language acquisition. The former lists (*à la* Swadesh) are generally composed of words that are universal across cultures and are resistant to borrowing, so that a comparison across language of the words in these lists can help determine linguistic relationships. Words in the latter lists (for language learning) are often chosen for their frequency of use in written and spoken language as well as for their range of use across multiple genres or domains.

1. one	2. water	3. two
4. dog	5. fish	6. tongue
7. eye	8. ear	9. fire
10. blood	11. stone	12. see
13. bone	14. skin	15. name
16. tooth	17. nose	18. star
19. die	20. come	21. head
22. hear	23. woman	24. path
25. mouth	26. breast	27. night
28. eat	29. you	30. moon
31. smoke	32. hair	33. bird
34. black	35. fly	36. sleep
37. man	38. egg	39. new
40. three	41. white	42. I
43. liver	44. hand	45. rain
46. hide	47. tail	48. we
49. drink	50. louse	51. snake
52. good	53. say	54. small
55. fat	56. sun	57. tree
58. cloud	59. meat	60. rock
61. neck	62. sand	63. wind
64. cold	65. leaf	66. dry
67. earth	68. four	69. person
70. go	71. kill	72. bite
73. that	74. red	75. burn
76. mother	77. road	78. big
79. sit	80. father	81. long
82. five	83. mountain	84. male
85. what	86. knee	87. leg
88. root	89. soil	90. large
91. grind	92. ashes	93. fall
94. who	95. right	96. foot
97. house	98. all	99. heavy
100. back	101. stand	102. bad
103. little	104. child	105. hot
106. know	107. ten	108. give
109. short	110. walk	111. dead
112. female	113. heart	114. salt
115. old	116. hill	117. belly
118. sky	119. laugh	120. cut
121. ash	122. close	123. wing
124. six	125. shoulder	126. smell
127. stick	128. human being	129. green
130. dull	131. seven	132. single
133. eight	134. many	135. far
136. he	137. breasts	138. day
139. the	140. title	141. yellow
142. near	143. nine	144. full
145. this	146. lie	147. dig
148. where	149. rat	150. every

Table 1: Top 150 words from our core vocabulary list.

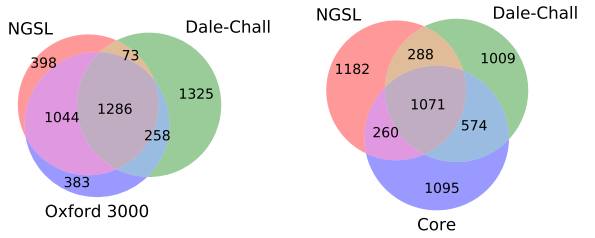
In this section, we show that our empirically derived, dictionary coverage–based lists have high overlap with several existing lists that were developed via these motivations and can indeed be used for such purposes. In addition, our core vocabulary list has high coverage over several well-known linguistic corpora which span multiple domains, making this list particularly suited for language learning.

[1] http://uakari.ling.washington.edu/languagenet

[2] Here, we use *dictionary* to mean *language*, i.e. every language in PanLex has one dictionary. Each dictionary is represented by a separate ISO 639-3 language code, so this number represents language variants.

List	Coverage	%
Swadesh	207/207	100
Dogolpolsky	15/15	100
Leipzig-Jakarta	100/100	100
Ogden	698/850	82
Dale–Chall	1669/2942	57
Oxford 3000	1525/2989	51
NGSL	1362/2801	49
Chinese	1518/2462	62
Russian	1243/1817	68

Table 2: Overlap with existing core vocabulary lists.



(a) (b)

Figure 3: Overlap in core vocabulary lists; (a) compares existing lists, (b) compares existing lists with our own Core Vocabulary list.

3.1. Comparison with Other Lists

We compare our 3000-word core vocabulary list with several well-known lists:

Linguistically Motivated Lists The Swadesh list (Swadesh, 1952) has already been extensively mentioned. The Dogolpolsky list (Trask, 2000) is a small set of 15 words that were chosen for their resistance to be replaced by other words over time. The Leipzig–Jakarta list (Tadmor, 2009) is a set of 100 words that are most resistant to borrowing from other languages.

We also investigate the following language-learning lists:

Ogden's Basic English (Ogden, 1932) A list of 850 words compiled by C. K. Ogden of simple concepts encountered in everyday life.

Oxford 3000 A list[3] of 3000 words (2989 unique lemmas) that were selected for their "importance and usefulness" for English language learners based on their frequency, range of domains, and familiarity in the English language.

New General Service List (NGSL) (Browne, 2014) A list of 2801 lemmas along with their inflected forms, billed as a list of general words for English language learners. It is based on the Cambridge English Corpus and seeks to improve upon an earlier list, the General Service List (West, 1953).

Dale–Chall (Dale and Chall, 1948) A list of 3000 words that a United States 4th grader would know. This list is used in readability metrics.

In addition, we compare against a couple language learning focused lists in other languages to evaluate the linguistic universality of our core vocabulary list:

Chinese We use a wordlist from the Hanyu Shuiping Kaoshi, also known as the Chinese Proficiency Exam. We use a total of 2500 words from levels 1–5, roughly corresponding to B1 or B2 proficiency level.

Russian We use a wordlist from OpenRussian.org containing 1819 words up to a B2 proficiency level.

In Table 2, we see that our list has complete coverage over three established core vocabulary lists for historical linguistics: the Swadesh list, Dogolpolsky list, and Leipzig–Jakarta list. This is not surprising: from Table 1, we see that many of these words are indeed Swadesh words. What is more interesting is how our list compares to similarly-sized lists for language learning. Figure 3a shows that the NGSL and Oxford 3000 lists have considerable overlap with each other, but less overlap with Dale–Chall. This is possibly because both the NGSL and Oxford 3000 are largely corpus-based, while Dale–Chall is manually curated. In Figure 3b, we see that our list covers a little over half of each of the other lists, meaning that there are roughly 1300 words that experts have deemed important for learners that are not commonly found in dictionaries. Conversely, there are roughly 1000 words that lexicographers have deemed important for entry into dictionaries but are not found in language learning lists. What kind of words are these?

In terms of words contained in our core vocabulary but excluded from other lists, we first examine the top ten words, along with their rank in our list, that are not present in any language learning list are: 129 *human being*, 181 *mosquito*, 210 *left hand*, 342 *urine*, 355 *crocodile*, 370 *vein*, 378 *buttock*, 401 *armpit*, 422 *buttocks*, 423 *excrement*. *Human being* shares translations with *human* and *man*, which occur higher in our core list; the same is for *left hand* and *left*. The other words are animals (mosquito, crocodile), and body parts or functions, which also occur in other core lists but might not be relevant for a language learner.

To examine the differences between our core vocabulary list and other lists, we first group our words into topics based on the topic dictionaries in the Oxford Learner's Dictionary.[4] Table 3 presents the top few topics whose words our list contains but other lists do not. These topic dictionaries are not comprehensive, so these counts are underestimates. Nevertheless they give an indication of the types of words missing from language learning lists.

Our core list notably contains roughly 160 country names and their adjectival forms (e.g. *Spain* and *Spanish*) not present in the other language learning lists. In our interconnected society, knowledge of such proper nouns is useful for reading or translating modern text, especially on the web. Many body parts, animals, and family words exist in

[3] https://www.oxfordlearnersdictionaries.com/us/about/oxford3000

[4] https://www.oxfordlearnersdictionaries.com/us/topic/

Topic	#	Example Words
Country	68	Europe, France, French, Spanish
Body	66	abdomen, belly, palm, wrist, nostril
Animal	55	beetle, mosquito, moth, louse, fowl
Family	42	sibling, stepfather, father-in-law, adolescent
Food	30	tasty, herb, acid, garlic
Other		wisdom, noble, merchant, murderer, funeral

Table 3: Examples of words in our Core Vocabulary that do not appear in other major core vocabulary lists.

our list but are missing from existing lists. One explanation is that these lists are mainly for English language learners. Other cultures may place more importance on such topics, and thus knowledge of these terms would be more important for learners of those languages. For example, familial relationships are an important part of Asian cultures, and Asian languages are known for having many specific kinship terms that do not exist as a single word in English.

Our list contains 112 multiword concepts not present in language learning lists. Along with their associated rank, these include

- multiword expressions and questions (2828 *a lot*, 512 *how many*)
- phrasal verbs (180 *lie down*, 391 *look for*)
- infinitival phrases (532 *be alive*, 1315 *be born*)
- kinship terms (575 *older brother*, 754 *mother-in-law*)
- other multiword nouns (129 *human being*, 1157 *day before yesterday*)

While almost all lists contain a MWE's constituent words (e.g. *day*, *before*, and *yesterday*), a language may not have a single word for the concept of *day before yesterday*. The presence of these MWE's in our core lists highlights the deficiencies of relying on English lists.

For the non-English language lists we examined, we see over 60% coverage over these lists (Table 2). As expected, a small number of concepts that our list missed are culture specific (e.g. for Chinese: *Chinese chess*, *tai chi*, *Beijing*; for Russian: *Leningrad*, *St. Petersburg*, *Soviet*). As observed with the other lists, a large portion of missed concepts (37% for Chinese, 15% for Russian) are multiword concepts (e.g. *can't help but*, *in total*, *of course*). We noticed that many of these phrasal concepts are not content words, which usually have high representation in dictionaries and thus rank highly in our core vocabulary. Anecdotally, proficient usage of adverbs can give the impression of fluency in a foreign language even when knowledge of nouns and verbs is lacking, which might have lead to their inclusion in these language learning lists.

4. Corpus Coverage

We also examine coverage of our core vocabulary list on various corpora which span a wide range of sizes and domains. Note that while these corpora are comprised of English text, we use them not as corpora of words but concepts that are universal across languages and cultures.

Bible The Bible is perhaps the most widely translated document in the world. Because of this fact, the Bible can be a useful resource for starting a dictionary in a low-resource language when other resources do not exist. We use the New Simplified English edition which contains both the Old and New Testament.

UDHR The Universal Declaration of Human Rights is also a widely translated document. It is considerably smaller than the (already small) Bible.

British National Corpus (BNC) (Leech et al., 2014) A multi-domain corpus of written and spoken British English from the late 20th century. We use words with a frequency above 800.

American National Corpus v2 (ANC) (Macleod et al., 2000) A similar multi-domain corpus. It also contains web-domain text like emails and tweets, which are not included in the British National Corpus. We remove words that occur only once.

Google N-Grams Corpus (GNG) (Michel et al., 2011) Google has scanned millions of books and computed frequency statistics per year. We use unigram frequencies from the 2012 version, accumulated over all years.

Coverage on a type and token basis are presented in Table 2. We compare against other lists by truncating our own list to match the size. We remove proper names using a heuristic if it does not appear in lowercase in the text. We also exclude hapaxes (words that appear only once) from the Bible, and truncate the frequency lists over the larger corpora, the sizes of which are shown in Table 4. To interpret Figure 4, we see for example that the top 2995 core vocabulary list gives 22% type and 57% token coverage over the Bible, using 1905 core vocabulary words. This means knowing roughly 2/3 of our core list allows one to read roughly 2/3 of the Bible, an impressive figure. While the NGSL and Oxford have higher coverage over these corpora, this is due to the fact that these lists were constructed in part based on frequency in such corpora. Nevertheless, our multilingual dictionary-based core list only trails slightly behind in coverage relative to other English core lists, indicating that over a thousand lexicographers' stamp of approval across languages tends to work well for specific languages, such as English.

If our core list has high coverage over existing corpora, why not use the corpora themselves as the basis? Large, diverse corpora are hard to find for low-resource languages. Using the Bible as the sole corpus for a language skews the vocabulary to a specific domain and limits the usefulness of the core vocabulary list. The intent of this project is to create a universally applicable core vocabulary list where knowledge of these concepts in any language will enable the comprehension of text across a variety of domains.

5. Experiments

We have argued that core vocabulary words have high priority for elicitation if they do not exist in a dictionary. While human annotation is ideal, in lieu of this, we can treat this elicitation as a lexicon induction task. Core vocabulary lists like the Swadesh lists are commonly used to determine

	Core-100		Swadesh 100		Core-8414		NGSL		Core-2995		Oxford	
	Type	Token	Type	Token	Type	Token	Type	Token	Type	Token	Type	Token
Bible	0.011	0.069	0.011	0.077	0.40	0.65	0.43	0.69	0.22	0.57	0.23	0.59
UDHR	0.025	0.034	0.036	0.026	0.68	0.62	0.78	0.69	0.43	0.51	0.67	0.63
BNC	0.017	0.055	0.017	0.067	0.71	0.92	0.56	0.94	0.34	0.73	0.51	0.94
ANC	0.010	0.048	0.009	0.053	0.35	0.58	0.51	0.66	0.17	0.45	0.27	0.56
GNG	0.010	0.049	0.010	0.059	0.41	0.78	0.54	0.89	0.19	0.61	0.28	0.75

Figure 4: Coverage of lists over various corpora. The number of types and tokens for each corpus is in Table 4. Comparisons are only valid between same size lists, i.e. between columns 1 and 2, 3 and 4, and 5 and 6.

Corpus	Types	Tokens
Bible	8,674	790K
UDHR	197	1,773
BNC	5,464	62M
ANC	10,000	20M
GNG	10,000	341B

Table 4: Corpus sizes

phylogenetic relationships between languages. Thus if two languages are related, their respective Swadesh words are likely to be cognates.

In this section, we expand on the work of Wu and Yarowsky (2018b), who devised a cognate translation method for the bilingual lexicon induction task. They discovered cognates from a multilingual dictionary in an unsupervised manner by using English as a pivot and then clustered these translations into cognate groups based on edit distance. Taking the Cartesian product of words in each cluster as word pairs, they run an aligner to extract character insertion, deletion, and substitution probabilities to be used as costs in a weighted edit distance in a second clustering iteration. The results of the second clustering were used to train character-based machine translation systems to predict missing cognates in each cluster.

As a motivating test case, we examine 18 Mayalo-Polynesian languages,[5] which is under the Austronesian language family. These are all low-resource languages with small dictionaries suitable for dictionary expansion. We first gather translations of our core vocabulary words in these languages. Following Wu and Yarowsky (2018b), we perform cognate clustering to separate translations into cognate groups. Then, following Wu and Yarowsky (2018a)'s multi-source approach for transliteration, we train a single neural machine translation system to predict held out cognate forms.

We train a single neural machine translation system to translate cognates. To prepare the training data, we preprocess the cognate clusters into bitext, following a procedure illustrated with the following example. Suppose there is a cognate group with the following cognates of the concept "man" in their respective languages: *mone* (bhp), *mone* (hvn), *mooni* (kje), *mane* (tet), *monu* (xbr). In this case, there are $\binom{5}{2} \times 2$ training pairs. The $\times 2$ is because we use each word as both as a source and target. When we hold out a pair for testing, e.g. (src=mone, tgt=mooni), we also remove the pair (src=mooni, tgt=mone) from the training data, so the system will not have encountered this pair. Words are split into characters, with the space character replaced by an underscore. We use an 80-10-10 train-dev-test split, and the architecture is a encoder-decoder network with 500-dimension word embedding size, with Adam optimizer with 0.001 learning rate.

We report results in Table 6, where the metric is top-N accuracy, i.e. does the gold appear in the top n predictions of the system. Each system generated ten predictions, but we found that the performance did not improve by looking further than the top 7 results. A sample of prediction results is shown in Table 5. While further analysis by native speakers might garner more insight, we see that when systems' 1-best predictions were incorrect, they were only off on average by 1 to 2 characters Table 7. One interesting phenomenon we noticed is the confusion between the glottal stop ʔ and the number 7. Apparently this artifact occurs in the PanLex data, possibly due to OCR errors. Nevertheless, our experiments show that the system can accurately generate missing cognates even in a low-resource setting by making use of information from related languages.

Src	Word	Tgt	Top 5 predictions
alp	buai	xbr	wua, **wo**, wue, bua, wu
xbr	wo	alp	bua, **buai**, bui, bu, buau
lti	sulu	mqy	**culu**, tulu, Culu, pulu, ulu
mqy	culu	lti	tulu, **sulu**, tulmu, mulu, culu
kje	iʔur	mqy	**iko**, éko, ca, ʔiko, Cko
mqy	iko	kje	**iʔur**, i7ur, ʔiʔur, ʔiu, iʔu
aoz	manu	ksx	**manuʔ**, manuk, manu7, manur
ksx	manuʔ	aoz	**manu**, tanu, manuʔ, manú
kje	haʔa	tet	**sa'e**, sa, sa'é, san, sade
tet	sa'e	kje	**haʔa**, ha7a, caʔa, ha, saʔa

Table 5: Sample of system predictions. Gold is bolded.

6. Compound Analysis

We further analyze the mechanisms of word formation in these core vocabulary words by employing the word com-

[5] As seen in Tables 6 and 7. These are ISO 693-3 language codes. alp is Alune, aoz is Uab Meto, bhp is Bima, hvn is Hawu, jmd is Yamdena, kei is Kei, kje is Kisar, ksx is Kedang, lti is Leti, mhs is Buru, mqy is Manggarai, nxg is Ngadha, plh is Paulohi, ski is Sika, slp is Lamaholot, slu is Selaru, tet is Tetum, and xbr is Kambera.

Lang	1	2	3	4	5	6	7	#
alp	.58	.83	.90	.91	.92	.92	**.93**	229
aoz	.70	.87	.95	**.97**	.97	.97	.97	153
bhp	.62	.82	.90	.93	**.96**	.96	.96	203
hvn	.33	.57	.70	.77	.84	**.85**	.85	393
jmd	.58	.89	.91	**.95**	.95	.95	.95	239
kei	.62	.87	.93	**.94**	.94	.94	.94	177
kje	.83	.98	.98	.98	**.99**	.99	.99	171
ksx	.78	**.95**	.95	.95	.95	.95	.95	132
lti	.87	.96	.96	.96	**.97**	.97	.97	148
mhs	.67	.80	.91	.93	.94	**.95**	.95	188
mqy	.59	.85	.89	.89	**.91**	.91	.91	278
nxg	.66	.86	.91	**.95**	.95	.95	.95	229
plh	.68	.87	.96	.97	**.98**	.98	.98	201
ski	.56	.84	.89	.95	.95	**.97**	.97	262
slp	.45	.81	.87	.90	**.92**	.92	.92	300
slu	.80	**.93**	.93	.93	.93	.93	.93	153
tet	.56	.81	.90	**.92**	.92	.92	.92	256
xbr	.26	.47	.65	.73	.80	.81	**.82**	614
total	.56	.78	.86	.89	.91	**.92**	.92	4326

Table 6: Top-n test accuracy for cognate prediction. # is number of test examples for each language.

Lang	AED2G	Lang	AED2G
alp	1.51	mhs	1.61
aoz	1.41	mqy	1.39
bhp	1.40	nxg	1.19
hvn	1.89	plh	1.52
jmd	1.48	ski	1.32
kei	1.39	slp	1.62
kje	1.17	slu	1.87
ksx	1.21	tet	1.65
lti	1.53	xbr	1.93

Table 7: Average edit distance between a language's 1-best output and the gold.

pounding model of Wu and Yarowsky (2018c). They analyze words by splitting the word into two component parts. By accumulating counts of the these components across all languages, they derive "recipes" for a concept, e.g. the concept of *hospital* is often realized as a compound of *sick* and *house* in many languages, even those unrelated to each other. We use this compounding model to analyze translations of our core vocabulary across languages. We find 278 concepts whose translations are often compounds. As presented in Table 8, the most commonly compounded concepts are numbers words, with a recipe of e.g. *twelve* = *ten* + *two* in their respective language.

We also attempted the dictionary induction task by *generating* compound words using the compound recipes. A small number of words in certain languages were recoverable using compound generation, but overall, compound generation was not successful. The fact that most of these words core words are non-compositional is actually a strong indicator that affirms their designation as a core word.

Word	# Langs
fifteen	31
chinese	29
fourteen	27
seventeen	27
twelve	25
eleven	24
daily	22
russian	22
football	22
bedroom	21

Table 8: Core vocabulary concepts commonly compounded across languages.

7. Conclusion

This paper introduced a novel criterion for selecting a core vocabulary set: high coverage in dictionaries across the world's languages. We use this simple but effective criterion to produce a core vocabulary list suited for establishing linguistic relationships and both first- and second-language learning due to its high overlap with existing manually-created lists constructed for such purposes. Words in our core vocabulary exhibit features indicative of coreness, including being cognates in related languages and often not being compositional. In addition, the core words span multiple domains and cover high frequency concepts which ought to be translatable in any language. We employed a cognate prediction model to translate the core vocabulary words with promising results. Based on the consensus of thousands of lexicographers across the world's languages, in constructing dictionaries for low-resource languages, translations of these core words can be elicited by field linguists or computationally via a cognate methods or inflectional generation methods such as Nicolai et al. (2020). Code and data used in this paper, including the full list of core vocabulary words, is available at `https://github.com/wswu/corevoc`.

8. References

Baldwin, T., Pool, J., and Colowick, S. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. *Coling 2010: Demonstrations*, pages 37–40.

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, 3(2):1–10.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Leech, G., Rayson, P., et al. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

List, J.-M., Cysouw, M., and Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, May 23-28, 2016, Portorož, Slovenia*, pages 2393–2400.

Macleod, C., Ide, N., and Grishman, R. (2000). The American national corpus: A standardized resource for Amer-

ican English. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

McCarthy, A. D., Wu, W., Mueller, A., Watson, W., and Yarowsky, D. (2019). Modeling color terminology across thousands of languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2241–2250, Hong Kong, China, November. Association for Computational Linguistics.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Nicolai, G., Lewis, D., McCarthy, A. D., Mueller, A., Wu, W., and Yarowsky, D. (2020). Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Ogden, C. K. (1932). *The ABC of basic English (in Basic)*. K. Paul, Trench, Trubner & Company Limited.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A comparative handbook*, pages 55–75.

Trask, R. L. (2000). *The dictionary of historical and comparative linguistics*. Psychology Press.

West, M. P. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green.

Wu, W. and Yarowsky, D. (2018a). A comparative study of extremely low-resource transliteration of the world's languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wu, W. and Yarowsky, D. (2018b). Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wu, W. and Yarowsky, D. (2018c). Massively translingual compound analysis and translation discovery. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.