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.