### Overview of text corpora publicly available in Sketch Engine

**Category legend:**
- **main**: corpora available only for regular (paying) users
- **trial**: corpora available for both trial and regular users
- **open**: corpora available without prior registration

<table>
<thead>
<tr>
<th>Name</th>
<th>Language</th>
<th>Category</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILDES Afrikaans Corpus</td>
<td>Afrikaans</td>
<td>main</td>
<td>33,134</td>
</tr>
<tr>
<td>OPUS2 Afrikaans</td>
<td>Afrikaans</td>
<td>trial</td>
<td>743,954</td>
</tr>
<tr>
<td>OPUS2 Albanian</td>
<td>Albanian</td>
<td>trial</td>
<td>55,099,328</td>
</tr>
<tr>
<td>Arabic web corpus</td>
<td>Arabic</td>
<td>trial</td>
<td>174,239,600</td>
</tr>
<tr>
<td>arTenTen12 [sample 115M]</td>
<td>Arabic</td>
<td>main</td>
<td>131,159,130</td>
</tr>
<tr>
<td>arTenTen12 [sample 115M, lemma voc]</td>
<td>Arabic</td>
<td>main</td>
<td>131,990,130</td>
</tr>
<tr>
<td>KSUCCA</td>
<td>Arabic</td>
<td>trial</td>
<td>59,693,146</td>
</tr>
<tr>
<td>arTenTen12</td>
<td>Arabic</td>
<td>trial</td>
<td>6,637,387,738</td>
</tr>
<tr>
<td>arTenTen12 [sample 50M]</td>
<td>Arabic</td>
<td>trial</td>
<td>54,692,024</td>
</tr>
<tr>
<td>arTenTen12 sample (Stanford tagger)</td>
<td>Arabic</td>
<td>trial</td>
<td>33,092,029</td>
</tr>
<tr>
<td>OPUS2 Arabic</td>
<td>Arabic</td>
<td>trial</td>
<td>406,527,277</td>
</tr>
<tr>
<td>Quran annotated corpus [unvowelled Arabic]</td>
<td>Arabic</td>
<td>trial</td>
<td>128,243</td>
</tr>
<tr>
<td>Quran annotated corpus [unvowelled Latin]</td>
<td>Arabic</td>
<td>trial</td>
<td>128,243</td>
</tr>
<tr>
<td>Quran annotated corpus [vowelled Arabic]</td>
<td>Arabic</td>
<td>trial</td>
<td>128,243</td>
</tr>
<tr>
<td>Quran annotated corpus [vowelled Latin]</td>
<td>Arabic</td>
<td>trial</td>
<td>128,243</td>
</tr>
<tr>
<td>Turkic web -- Azerbaijani</td>
<td>Azerbaijani</td>
<td>trial</td>
<td>115,280,755</td>
</tr>
<tr>
<td>BengaliWaC</td>
<td>Bengali</td>
<td>trial</td>
<td>13,719,158</td>
</tr>
<tr>
<td>bsWaC14</td>
<td>Bosnian</td>
<td>trial</td>
<td>290,176,507</td>
</tr>
<tr>
<td>OPUS2 Bosnian</td>
<td>Bosnian</td>
<td>trial</td>
<td>55,224,138</td>
</tr>
<tr>
<td>BulgarianNC2 nonweb</td>
<td>Bulgarian</td>
<td>main</td>
<td>27,721,533</td>
</tr>
<tr>
<td>BulgarianNC2 web</td>
<td>Bulgarian</td>
<td>main</td>
<td>545,637,740</td>
</tr>
<tr>
<td>BulgarianNC2 web 10M</td>
<td>Bulgarian</td>
<td>main</td>
<td>9,478,549</td>
</tr>
<tr>
<td>bgTenTen12</td>
<td>Bulgarian</td>
<td>trial</td>
<td>849,071,579</td>
</tr>
<tr>
<td>BulgarianNC</td>
<td>Bulgarian</td>
<td>trial</td>
<td>26,518,884</td>
</tr>
<tr>
<td>EUROPARL7, bg</td>
<td>Bulgarian</td>
<td>trial</td>
<td>10,602,635</td>
</tr>
<tr>
<td>OPUS2 Bulgarian</td>
<td>Bulgarian</td>
<td>trial</td>
<td>238,945,836</td>
</tr>
<tr>
<td>CHILDES Catalan Corpus</td>
<td>Catalan</td>
<td>main</td>
<td>277,816</td>
</tr>
<tr>
<td>CHILDES Croatian Corpus</td>
<td>Croatian</td>
<td>main</td>
<td>389,674</td>
</tr>
<tr>
<td>hrWaC14</td>
<td>Croatian</td>
<td>trial</td>
<td>1,404,262,704</td>
</tr>
<tr>
<td>OPUS2 Croatian</td>
<td>Croatian</td>
<td>trial</td>
<td>156,942,211</td>
</tr>
<tr>
<td>czes</td>
<td>Czech</td>
<td>main</td>
<td>465,102,710</td>
</tr>
<tr>
<td>Corpus Name</td>
<td>Language</td>
<td>Type</td>
<td>Size</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>czTenTen12 [v. 5]</td>
<td>Czech</td>
<td>trial</td>
<td>5,197,717,396</td>
</tr>
<tr>
<td>EUROPARL7, cs</td>
<td>Czech</td>
<td>trial</td>
<td>15,176,606</td>
</tr>
<tr>
<td>OPUS2 Czech</td>
<td>Czech</td>
<td>trial</td>
<td>275,519,334</td>
</tr>
<tr>
<td>daTenTen14 sample</td>
<td>Danish</td>
<td>main</td>
<td>10,949,488</td>
</tr>
<tr>
<td>CHILDES Danish Corpus</td>
<td>Danish</td>
<td>main</td>
<td>372,811</td>
</tr>
<tr>
<td>DanishWaC</td>
<td>Danish</td>
<td>trial</td>
<td>353,703,002</td>
</tr>
<tr>
<td>daTenTen14</td>
<td>Danish</td>
<td>trial</td>
<td>2,464,690,105</td>
</tr>
<tr>
<td>EUROPARL7, da</td>
<td>Danish</td>
<td>trial</td>
<td>55,794,038</td>
</tr>
<tr>
<td>OPUS2 Danish</td>
<td>Danish</td>
<td>trial</td>
<td>153,261,335</td>
</tr>
<tr>
<td>CHILDES Dutch Corpus</td>
<td>Dutch</td>
<td>main</td>
<td>7,592,039</td>
</tr>
<tr>
<td>EUROPARL7, nl</td>
<td>Dutch</td>
<td>trial</td>
<td>59,756,704</td>
</tr>
<tr>
<td>nlTenTen [2014]</td>
<td>Dutch</td>
<td>trial</td>
<td>3,013,056,738</td>
</tr>
<tr>
<td>nlTenTen14 [sample]</td>
<td>Dutch</td>
<td>trial</td>
<td>44,607,603</td>
</tr>
<tr>
<td>nlTenTen14 [sample, terms]</td>
<td>Dutch</td>
<td>trial</td>
<td>44,607,603</td>
</tr>
<tr>
<td>OPUS2 Dutch</td>
<td>Dutch</td>
<td>trial</td>
<td>446,240,037</td>
</tr>
<tr>
<td>British National Corpus (10M)</td>
<td>English</td>
<td>main</td>
<td>9,069,227</td>
</tr>
<tr>
<td>British National Corpus (TreeTagger)</td>
<td>English</td>
<td>main</td>
<td>112,985,133</td>
</tr>
<tr>
<td>Brown Family</td>
<td>English</td>
<td>main</td>
<td>8,073,482</td>
</tr>
<tr>
<td>English Dialog Corpus</td>
<td>English</td>
<td>main</td>
<td>1,359,434</td>
</tr>
<tr>
<td>enTenTen12 [sample 40M]</td>
<td>English</td>
<td>main</td>
<td>40,920,950</td>
</tr>
<tr>
<td>London English Corpus</td>
<td>English</td>
<td>main</td>
<td>2,959,320</td>
</tr>
<tr>
<td>LEXMCI</td>
<td>English</td>
<td>main</td>
<td>1,720,056,987</td>
</tr>
<tr>
<td>NCI English</td>
<td>English</td>
<td>main</td>
<td>257,900,777</td>
</tr>
<tr>
<td>ukWaC (10M)</td>
<td>English</td>
<td>main</td>
<td>9,866,055</td>
</tr>
<tr>
<td>ukWaC</td>
<td>English</td>
<td>main</td>
<td>1,559,716,979</td>
</tr>
<tr>
<td>ACL Anthology Reference Corpus (ARC)</td>
<td>English</td>
<td>open</td>
<td>49,348,397</td>
</tr>
<tr>
<td>British Academic Spoken English Corpus (BASE)</td>
<td>English</td>
<td>open</td>
<td>1,252,256</td>
</tr>
<tr>
<td>British Academic Written English Corpus (BAWE)</td>
<td>English</td>
<td>open</td>
<td>8,336,262</td>
</tr>
<tr>
<td>Brown</td>
<td>English</td>
<td>open</td>
<td>1,175,675</td>
</tr>
<tr>
<td>British National Corpus</td>
<td>English</td>
<td>trial</td>
<td>112,181,015</td>
</tr>
<tr>
<td>e-flux</td>
<td>English</td>
<td>trial</td>
<td>6,238,592</td>
</tr>
<tr>
<td>enTenTen08</td>
<td>English</td>
<td>trial</td>
<td>3,268,798,627</td>
</tr>
<tr>
<td>enTenTen12 [sample 40M] with term grammar</td>
<td>English</td>
<td>trial</td>
<td>40,920,950</td>
</tr>
<tr>
<td>enTenTen12</td>
<td>English</td>
<td>trial</td>
<td>12,968,375,937</td>
</tr>
<tr>
<td>EUROPARL7, en</td>
<td>English</td>
<td>trial</td>
<td>60,741,877</td>
</tr>
<tr>
<td>CHILDES English Corpus</td>
<td>English</td>
<td>trial</td>
<td>29,908,984</td>
</tr>
<tr>
<td>New Model Corpus</td>
<td>English</td>
<td>trial</td>
<td>114,627,650</td>
</tr>
<tr>
<td>OPUS2 English</td>
<td>English</td>
<td>trial</td>
<td>1,441,844,046</td>
</tr>
<tr>
<td>pukWaC</td>
<td>English</td>
<td>trial</td>
<td>46,256,586</td>
</tr>
<tr>
<td>SiBol/Port</td>
<td>English</td>
<td>trial</td>
<td>387,585,716</td>
</tr>
</tbody>
</table>

Overview of text corpora publicly available in Sketch Engine
## Overview of text corpora publicly available in Sketch Engine

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susanne</td>
<td>English</td>
<td>trial</td>
<td>150,426</td>
</tr>
<tr>
<td>EstonianNC</td>
<td>Estonian</td>
<td>main</td>
<td>563,220,548</td>
</tr>
<tr>
<td>etTenTen</td>
<td>Estonian</td>
<td>main</td>
<td>330,045,196</td>
</tr>
<tr>
<td>CHILDES Estonian Corpus</td>
<td>Estonian</td>
<td>main</td>
<td>399,547</td>
</tr>
<tr>
<td>EstonianRC</td>
<td>Estonian</td>
<td>trial</td>
<td>249,923,332</td>
</tr>
<tr>
<td>EUROPARL7, et</td>
<td>Estonian</td>
<td>trial</td>
<td>13,162,640</td>
</tr>
<tr>
<td>OPUS2 Estonian</td>
<td>Estonian</td>
<td>trial</td>
<td>88,432,596</td>
</tr>
<tr>
<td>FilipinoWaC</td>
<td>Filipino</td>
<td>trial</td>
<td>31,845,404</td>
</tr>
<tr>
<td>EUROPARL7, fi</td>
<td>Finnish</td>
<td>trial</td>
<td>40,930,202</td>
</tr>
<tr>
<td>fiTenTen [2014]</td>
<td>Finnish</td>
<td>trial</td>
<td>1,706,310,900</td>
</tr>
<tr>
<td>OPUS2 Finnish</td>
<td>Finnish</td>
<td>trial</td>
<td>180,134,681</td>
</tr>
<tr>
<td>frWaC 1.1</td>
<td>French</td>
<td>trial</td>
<td>1,628,667,738</td>
</tr>
<tr>
<td>CHILDES French Corpus</td>
<td>French</td>
<td>main</td>
<td>3,287,017</td>
</tr>
<tr>
<td>French web corpus</td>
<td>French</td>
<td>main</td>
<td>126,850,281</td>
</tr>
<tr>
<td>EUROPARL7, fr</td>
<td>French</td>
<td>trial</td>
<td>66,040,174</td>
</tr>
<tr>
<td>Araneum Francogallicum Minus (Fr Web 13.10)</td>
<td>French</td>
<td>trial</td>
<td>108,202,791</td>
</tr>
<tr>
<td>frTenTen12</td>
<td>French</td>
<td>trial</td>
<td>12,369,868,562</td>
</tr>
<tr>
<td>frTenTen12 [sample] with term grammar</td>
<td>French</td>
<td>trial</td>
<td>39,472,639</td>
</tr>
<tr>
<td>OPUS2 French</td>
<td>French</td>
<td>trial</td>
<td>956,614,852</td>
</tr>
<tr>
<td>deTenTen10 [sample 50M]</td>
<td>German</td>
<td>main</td>
<td>53,766,293</td>
</tr>
<tr>
<td>deWaC</td>
<td>German</td>
<td>main</td>
<td>1,627,169,557</td>
</tr>
<tr>
<td>Araneum Germanicum Minus (De Web 13.10)</td>
<td>German</td>
<td>main</td>
<td>110,879,482</td>
</tr>
<tr>
<td>deTenTen10</td>
<td>German</td>
<td>main</td>
<td>53,766,293</td>
</tr>
<tr>
<td>deTenTen10 [sample 50M] with term grammar</td>
<td>German</td>
<td>main</td>
<td>54,899,037</td>
</tr>
<tr>
<td>EUROPARL7, de</td>
<td>German</td>
<td>trial</td>
<td>800,783</td>
</tr>
<tr>
<td>GerManC</td>
<td>German</td>
<td>trial</td>
<td>157,849,124</td>
</tr>
<tr>
<td>OPUS2 German</td>
<td>German</td>
<td>trial</td>
<td>44,097,921</td>
</tr>
<tr>
<td>EUROPARL7, el</td>
<td>Greek</td>
<td>trial</td>
<td>149,067,023</td>
</tr>
<tr>
<td>GkWaC</td>
<td>Greek</td>
<td>trial</td>
<td>305,404,357</td>
</tr>
<tr>
<td>OPUS2 Greek</td>
<td>Greek</td>
<td>trial</td>
<td>22,201,247</td>
</tr>
<tr>
<td>GujarathiWaC</td>
<td>Gujarati</td>
<td>trial</td>
<td>192,119,449</td>
</tr>
<tr>
<td>HebrewGC</td>
<td>Hebrew</td>
<td>main</td>
<td>38,426,174</td>
</tr>
<tr>
<td>HebrewGC [sample]</td>
<td>Hebrew</td>
<td>main</td>
<td>1,034,238</td>
</tr>
<tr>
<td>CHILDES Hebrew Corpus</td>
<td>Hebrew</td>
<td>main</td>
<td>60,351,738</td>
</tr>
<tr>
<td>HebWaC</td>
<td>Hebrew</td>
<td>trial</td>
<td>252,278,074</td>
</tr>
<tr>
<td>OPUS2 Hebrew</td>
<td>Hebrew</td>
<td>trial</td>
<td>65,772,188</td>
</tr>
<tr>
<td>HindiWaC</td>
<td>Hindi</td>
<td>trial</td>
<td>1,642,973</td>
</tr>
<tr>
<td>OPUS2 Hindi</td>
<td>Hindi</td>
<td>trial</td>
<td>14,655,015</td>
</tr>
<tr>
<td>EUROPARL7, hu</td>
<td>Hungarian</td>
<td>trial</td>
<td>3,184,161,466</td>
</tr>
<tr>
<td>Corpus Name</td>
<td>Language</td>
<td>Type</td>
<td>Size</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>huTenTen12</td>
<td>Hungarian</td>
<td>trial</td>
<td>32,034,152</td>
</tr>
<tr>
<td>CHILDES Hungarian Corpus</td>
<td>Hungarian</td>
<td>main</td>
<td>311,543</td>
</tr>
<tr>
<td>OPUS2 Hungarian</td>
<td>Hungarian</td>
<td>trial</td>
<td>218,409,426</td>
</tr>
<tr>
<td>Chinese GigaWord 2 Corpus: Mainland, simplified</td>
<td>Chinese, Simplified</td>
<td>main</td>
<td>250,124,230</td>
</tr>
<tr>
<td>Internet-ZH</td>
<td>Chinese, Simplified</td>
<td>main</td>
<td>277,931,664</td>
</tr>
<tr>
<td>zhTenTen11 [sample 10M]</td>
<td>Chinese, Simplified</td>
<td>main</td>
<td>11,028,308</td>
</tr>
<tr>
<td>OPUS2 Chinese Simplified</td>
<td>Chinese, Simplified</td>
<td>trial</td>
<td>299,338,099</td>
</tr>
<tr>
<td>zhTenTen11</td>
<td>Chinese, Simplified</td>
<td>trial</td>
<td>2,106,661,021</td>
</tr>
<tr>
<td>zhTenTen11 [sample 10M] with term grammar</td>
<td>Chinese, Simplified</td>
<td>trial</td>
<td>11,028,308</td>
</tr>
<tr>
<td>Chinese GigaWord 2 Corpus: Taiwan, traditional</td>
<td>Chinese, Traditional</td>
<td>main</td>
<td>455,526,209</td>
</tr>
<tr>
<td>ChineseTaiwanWaC</td>
<td>Chinese, Traditional</td>
<td>main</td>
<td>349,198,060</td>
</tr>
<tr>
<td>ChineseTaiwanWaC (Universal Sketch Grammar)</td>
<td>Chinese, Traditional</td>
<td>trial</td>
<td>349,198,060</td>
</tr>
<tr>
<td>OPUS2 Chinese Traditional</td>
<td>Chinese, Traditional</td>
<td>trial</td>
<td>622,382</td>
</tr>
<tr>
<td>IgboWaC</td>
<td>Igbo</td>
<td>trial</td>
<td>162,024</td>
</tr>
<tr>
<td>IndonesianWaC</td>
<td>Indonesian</td>
<td>trial</td>
<td>109,281,359</td>
</tr>
<tr>
<td>NCI Irish</td>
<td>Irish</td>
<td>trial</td>
<td>34,358,267</td>
</tr>
<tr>
<td>CHILDES Italian Corpus</td>
<td>Italian</td>
<td>main</td>
<td>572,217</td>
</tr>
<tr>
<td>itWaC</td>
<td>Italian</td>
<td>main</td>
<td>1,909,535,703</td>
</tr>
<tr>
<td>EUROPARL7, it</td>
<td>Italian</td>
<td>trial</td>
<td>59,177,399</td>
</tr>
<tr>
<td>iTenTen10</td>
<td>Italian</td>
<td>trial</td>
<td>3,076,908,415</td>
</tr>
<tr>
<td>OPUS2 Italian</td>
<td>Italian</td>
<td>trial</td>
<td>231,143,960</td>
</tr>
<tr>
<td>CHILDES Japanese Corpus</td>
<td>Japanese</td>
<td>main</td>
<td>2,187,308</td>
</tr>
<tr>
<td>jpTenTen11 [SUW, sample]</td>
<td>Japanese</td>
<td>main</td>
<td>279,657,781</td>
</tr>
<tr>
<td>jpTenTen11 [LUW, sample]</td>
<td>Japanese</td>
<td>trial</td>
<td>203,674,569</td>
</tr>
<tr>
<td>jpTenTen11 [LUW, sample] with term grammar</td>
<td>Japanese</td>
<td>trial</td>
<td>203,674,569</td>
</tr>
<tr>
<td>jpTenTen11 [SUW]</td>
<td>Japanese</td>
<td>trial</td>
<td>10,321,875,664</td>
</tr>
<tr>
<td>JpWaC</td>
<td>Japanese</td>
<td>trial</td>
<td>409,384,405</td>
</tr>
<tr>
<td>OPUS2 Japanese</td>
<td>Japanese</td>
<td>trial</td>
<td>6,596,733</td>
</tr>
<tr>
<td>Turkic web -- Kazakh</td>
<td>Kazakh</td>
<td>trial</td>
<td>175,445,327</td>
</tr>
<tr>
<td>CHILDES Korean Corpus</td>
<td>Korean</td>
<td>main</td>
<td>53,339</td>
</tr>
<tr>
<td>koTenTen12</td>
<td>Korean</td>
<td>trial</td>
<td>560,945,022</td>
</tr>
</tbody>
</table>

Overview of text corpora publicly available in Sketch Engine
<table>
<thead>
<tr>
<th>Corpus Name</th>
<th>Language</th>
<th>Type</th>
<th>Size (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>koTenTen12 [sample] with term grammar</td>
<td>Korean</td>
<td>trial</td>
<td>43,113,814</td>
</tr>
<tr>
<td>OPUS2 Korean</td>
<td>Korean</td>
<td>trial</td>
<td>500,152</td>
</tr>
<tr>
<td>Turkic web -- Kyrgyz</td>
<td>Kyrgyz</td>
<td>trial</td>
<td>24,084,100</td>
</tr>
<tr>
<td>LatinISE</td>
<td>Latin</td>
<td>trial</td>
<td>13,586,032</td>
</tr>
<tr>
<td>EUROPARL7, lv</td>
<td>Latvian</td>
<td>trial</td>
<td>14,253,247</td>
</tr>
<tr>
<td>LatvianWaC</td>
<td>Latvian</td>
<td>trial</td>
<td>74,447,302</td>
</tr>
<tr>
<td>OPUS2 Latvian</td>
<td>Latvian</td>
<td>trial</td>
<td>34,012,690</td>
</tr>
<tr>
<td>EUROPARL7, lt</td>
<td>Lithuanian</td>
<td>trial</td>
<td>13,733,247</td>
</tr>
<tr>
<td>LithuanianWaC</td>
<td>Lithuanian</td>
<td>trial</td>
<td>78,630,701</td>
</tr>
<tr>
<td>OPUS2 Lithuanian</td>
<td>Lithuanian</td>
<td>trial</td>
<td>40,933,573</td>
</tr>
<tr>
<td>OPUS2 Macedonian</td>
<td>Macedonian</td>
<td>trial</td>
<td>49,066,513</td>
</tr>
<tr>
<td>MalayalamWaC</td>
<td>Malayalam</td>
<td>trial</td>
<td>21,193,984</td>
</tr>
<tr>
<td>MalayWaC</td>
<td>Malay</td>
<td>trial</td>
<td>230,509,568</td>
</tr>
<tr>
<td>Maltese MLRS Corpus - sample</td>
<td>Maltese</td>
<td>main</td>
<td>12,413,213</td>
</tr>
<tr>
<td>Maltese MLRS Corpus</td>
<td>Maltese</td>
<td>trial</td>
<td>125,267,653</td>
</tr>
<tr>
<td>MaoriWaC</td>
<td>Maori</td>
<td>trial</td>
<td>8,351,983</td>
</tr>
<tr>
<td>CHILDES Norwegian Corpus</td>
<td>Norwegian</td>
<td>main</td>
<td>61,075</td>
</tr>
<tr>
<td>noTenTen12</td>
<td>Norwegian</td>
<td>trial</td>
<td>769,573,845</td>
</tr>
<tr>
<td>OPUS2 Norwegian</td>
<td>Norwegian</td>
<td>trial</td>
<td>26,467,755</td>
</tr>
<tr>
<td>TalkBank Persian [sample]</td>
<td>Persian</td>
<td>main</td>
<td>13,707,798</td>
</tr>
<tr>
<td>OPUS2 Persian</td>
<td>Persian</td>
<td>trial</td>
<td>5,367,401</td>
</tr>
<tr>
<td>TalkBank Persian</td>
<td>Persian</td>
<td>trial</td>
<td>549,165,952</td>
</tr>
<tr>
<td>BIBLE Polish, swahili-Polish</td>
<td>Polish</td>
<td>main</td>
<td>169,934</td>
</tr>
<tr>
<td>CHILDES Polish Corpus</td>
<td>Polish</td>
<td>main</td>
<td>1,247,919</td>
</tr>
<tr>
<td>Polish Web Corpus (10M)</td>
<td>Polish</td>
<td>main</td>
<td>9,541,098</td>
</tr>
<tr>
<td>Polish Web Corpus</td>
<td>Polish</td>
<td>main</td>
<td>128,185,119</td>
</tr>
<tr>
<td>EUROPARL7, pl</td>
<td>Polish</td>
<td>trial</td>
<td>15,171,493</td>
</tr>
<tr>
<td>OPUS2 Polish</td>
<td>Polish</td>
<td>trial</td>
<td>285,188,755</td>
</tr>
<tr>
<td>plTenTen12</td>
<td>Polish</td>
<td>trial</td>
<td>9,677,787,906</td>
</tr>
<tr>
<td>plTenTen12 [sample]</td>
<td>Polish</td>
<td>trial</td>
<td>55,381,476</td>
</tr>
<tr>
<td>plTenTen12 [sample, terms]</td>
<td>Polish</td>
<td>trial</td>
<td>55,381,476</td>
</tr>
<tr>
<td>CHILDES Portuguese Corpus</td>
<td>Portuguese</td>
<td>main</td>
<td>245,805</td>
</tr>
<tr>
<td>ptTenTen11 [Freeling]</td>
<td>Portuguese</td>
<td>main</td>
<td>3,268,714,354</td>
</tr>
<tr>
<td>brasileiro</td>
<td>Portuguese</td>
<td>trial</td>
<td>1,133,416,757</td>
</tr>
<tr>
<td>EUROPARL7, pt</td>
<td>Portuguese</td>
<td>trial</td>
<td>61,414,188</td>
</tr>
<tr>
<td>OPUS2 Brazilian Portuguese</td>
<td>Portuguese</td>
<td>trial</td>
<td>355,049,778</td>
</tr>
<tr>
<td>OPUS2 Portuguese</td>
<td>Portuguese</td>
<td>trial</td>
<td>377,677,225</td>
</tr>
<tr>
<td>Cetenfolha, Cetempublico</td>
<td>Portuguese</td>
<td>trial</td>
<td>66,319,147</td>
</tr>
<tr>
<td>ptTenTen11 [Freeling, sample]</td>
<td>Portuguese</td>
<td>trial</td>
<td>44,446,042</td>
</tr>
<tr>
<td>ptTenTen11 [Freeling, sample 50M] with term grammar</td>
<td>Portuguese</td>
<td>trial</td>
<td>44,446,042</td>
</tr>
<tr>
<td>Text Corpora</td>
<td>Language</td>
<td>Type</td>
<td>Size</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>pfTenTen11</td>
<td>Portuguese</td>
<td>trial</td>
<td>3,245,834,337</td>
</tr>
<tr>
<td>EUROPAL7, ro</td>
<td>Romanian</td>
<td>trial</td>
<td>10,795,858</td>
</tr>
<tr>
<td>OPUS2 Romanian</td>
<td>Romanian</td>
<td>trial</td>
<td>360,212,949</td>
</tr>
<tr>
<td>Romanian web corpus</td>
<td>Romanian</td>
<td>trial</td>
<td>53,457,522</td>
</tr>
<tr>
<td>CHILDES Russian Corpus</td>
<td>Russian</td>
<td>main</td>
<td>59,759</td>
</tr>
<tr>
<td>Russian web corpus</td>
<td>Russian</td>
<td>main</td>
<td>187,965,822</td>
</tr>
<tr>
<td>ruTenTen11 [2011, RFTagger, sample 2]</td>
<td>Russian</td>
<td>trial</td>
<td>1,572,896,814</td>
</tr>
<tr>
<td>OPUS2 Russian</td>
<td>Russian</td>
<td>trial</td>
<td>381,468,257</td>
</tr>
<tr>
<td>Araneum Russicum Minus (Ru Web 13.10) 111 M</td>
<td>Russian</td>
<td>trial</td>
<td>110,740,561</td>
</tr>
<tr>
<td>ruTenTen11 [RFTagger, sample 50M]</td>
<td>Russian</td>
<td>trial</td>
<td>59,115,079</td>
</tr>
<tr>
<td>ruTenTen11 [RFTagger, sample 50M] with term grammar</td>
<td>Russian</td>
<td>trial</td>
<td>59,115,079</td>
</tr>
<tr>
<td>OPUS2 Serbian</td>
<td>Serbian</td>
<td>trial</td>
<td>198,141,613</td>
</tr>
<tr>
<td>srWaC14</td>
<td>Serbian</td>
<td>trial</td>
<td>561,529,963</td>
</tr>
<tr>
<td>Setswana2</td>
<td>Setswana</td>
<td>trial</td>
<td>13,511,692</td>
</tr>
<tr>
<td>skTenTen11</td>
<td>Slovak</td>
<td>main</td>
<td>876,003,720</td>
</tr>
<tr>
<td>EUROPAL7, sk</td>
<td>Slovak</td>
<td>trial</td>
<td>15,042,066</td>
</tr>
<tr>
<td>OPUS2 Slovak</td>
<td>Slovak</td>
<td>trial</td>
<td>82,952,296</td>
</tr>
<tr>
<td>EUROPAL7, sl</td>
<td>Slovenian</td>
<td>trial</td>
<td>14,616,666</td>
</tr>
<tr>
<td>Fida PLUS 620m</td>
<td>Slovenian</td>
<td>trial</td>
<td>738,503,145</td>
</tr>
<tr>
<td>Kres</td>
<td>Slovenian</td>
<td>trial</td>
<td>120,447,573</td>
</tr>
<tr>
<td>OPUS2 Slovenian</td>
<td>Slovenian</td>
<td>trial</td>
<td>163,160,520</td>
</tr>
<tr>
<td>esTenTen11 (European, Freeling, Lempos)</td>
<td>Spanish</td>
<td>main</td>
<td>2,459,314,898</td>
</tr>
<tr>
<td>esTenTen11 (European, Freeling) sample</td>
<td>Spanish</td>
<td>main</td>
<td>21,408,945</td>
</tr>
<tr>
<td>esTenTen11 (European, TreeTagger)</td>
<td>Spanish</td>
<td>main</td>
<td>2,459,314,898</td>
</tr>
<tr>
<td>esTenTen11 (Eu + Am, Freeling, Lempos)</td>
<td>Spanish</td>
<td>main</td>
<td>11,173,609,480</td>
</tr>
<tr>
<td>esTenTen11 (Eu, TreeTagger) sample</td>
<td>Spanish</td>
<td>main</td>
<td>49,485,462</td>
</tr>
<tr>
<td>CHILDES Spanish Corpus</td>
<td>Spanish</td>
<td>main</td>
<td>1,358,475</td>
</tr>
<tr>
<td>Spanish web corpus</td>
<td>Spanish</td>
<td>main</td>
<td>116,900,060</td>
</tr>
<tr>
<td>esTenTen11 (American, Freeling, Lempos)</td>
<td>Spanish</td>
<td>trial</td>
<td>8,714,294,582</td>
</tr>
<tr>
<td>esTenTen11 (American, TreeTagger)</td>
<td>Spanish</td>
<td>trial</td>
<td>8,719,845,099</td>
</tr>
<tr>
<td>esTenTen11 (Eu + Am, Freeling, Lempos, sample)</td>
<td>Spanish</td>
<td>trial</td>
<td>55,582,256</td>
</tr>
<tr>
<td>esTenTen11 (Eu + Am, Freeling, Lempos, sample) with term grammar</td>
<td>Spanish</td>
<td>trial</td>
<td>55,582,256</td>
</tr>
<tr>
<td>EUROPAL7, es</td>
<td>Spanish</td>
<td>trial</td>
<td>60,862,330</td>
</tr>
<tr>
<td>OPUS2 Spanish</td>
<td>Spanish</td>
<td>trial</td>
<td>870,615,999</td>
</tr>
<tr>
<td>BIBLE Swahili, Swahili-Polish</td>
<td>Swahili</td>
<td>main</td>
<td>169,612</td>
</tr>
<tr>
<td>SwahiliWaC</td>
<td>Swahili</td>
<td>trial</td>
<td>24,968,506</td>
</tr>
<tr>
<td>CHILDES Swedish Corpus</td>
<td>Swedish</td>
<td>main</td>
<td>665,889</td>
</tr>
<tr>
<td>SwedishParole</td>
<td>Swedish</td>
<td>main</td>
<td>25,731,328</td>
</tr>
<tr>
<td>EUROPAL7, sv</td>
<td>Swedish</td>
<td>trial</td>
<td>51,759,122</td>
</tr>
<tr>
<td>Corpus Type</td>
<td>Language</td>
<td>Type</td>
<td>Size</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------</td>
<td>------</td>
<td>--------------------</td>
</tr>
<tr>
<td>OPUS2 Swedish</td>
<td>Swedish</td>
<td>trial</td>
<td>128,245,911</td>
</tr>
<tr>
<td>svTenTen [2014]</td>
<td>Swedish</td>
<td>trial</td>
<td>3,899,190,951</td>
</tr>
<tr>
<td>SwedishWaC</td>
<td>Swedish</td>
<td>trial</td>
<td>114,022,801</td>
</tr>
<tr>
<td>TamilWaC</td>
<td>Tamil</td>
<td>trial</td>
<td>32,861,569</td>
</tr>
<tr>
<td>TeluguWaC</td>
<td>Telugu</td>
<td>trial</td>
<td>4,697,932</td>
</tr>
<tr>
<td>CHILDES Thai Corpus</td>
<td>Thai</td>
<td>main</td>
<td>299,962</td>
</tr>
<tr>
<td>ThaiWaC</td>
<td>Thai</td>
<td>trial</td>
<td>108,013,897</td>
</tr>
<tr>
<td>CHILDES Turkish Corpus</td>
<td>Turkish</td>
<td>main</td>
<td>207,223,730</td>
</tr>
<tr>
<td>Turkic web -- Turkish</td>
<td>Turkish</td>
<td>trial</td>
<td>4,124,558,200</td>
</tr>
<tr>
<td>TurkishWaC</td>
<td>Turkish</td>
<td>trial</td>
<td>42,202,822</td>
</tr>
<tr>
<td>Turkic web -- Turkmen</td>
<td>Turkmen</td>
<td>trial</td>
<td>2,536,935</td>
</tr>
<tr>
<td>OPUS2 Ukrainian</td>
<td>Ukrainian</td>
<td>trial</td>
<td>3,374,552</td>
</tr>
<tr>
<td>Turkic web -- Uzbek</td>
<td>Uzbek</td>
<td>trial</td>
<td>24,570,516</td>
</tr>
<tr>
<td>VietnameseWaC</td>
<td>Vietnamese</td>
<td>trial</td>
<td>129,781,089</td>
</tr>
<tr>
<td>Welsh corpus</td>
<td>Welsh</td>
<td>main</td>
<td>62,753,279</td>
</tr>
<tr>
<td>WelshWaC</td>
<td>Welsh</td>
<td>trial</td>
<td>14,786,791</td>
</tr>
</tbody>
</table>

Overview of text corpora publicly available in Sketch Engine