Why is lexical diversity important?

Use of typology in various NLP applications generated consistent improvements; hence, regarded as a promising new direction to tackle the issue of data scarcity in multilingual NLP [1]. However, most typology-informed NLP studies are limited to morphosyntactic features and have so far ignored lexical diversity. Ignoring diversity in lexicons can lead to hard-to-detect meaning-level mistakes:

For concepts with speaker gender or age undefined:

if neither a concept nor its parents have a gender or age in the lexicalization, then all concepts with these attributes are lexical gaps.

What are lexical gaps?

A concept is considered a lexical gap if it can only be expressed through free combinations of words [2].

Methods: How to use lexical typology to infer lexical gaps?

Typological knowledge:
Murdock's lexicalization patterns [3]
566 languages
4 subdomains of kinship

Existing resources:
Wiktionary
166 languages, 1681 words
6 subdomains of kinship

Native speakers:
10 languages, 230 words
6 subdomains of kinship

Published resource: Availability and visibility

Freely accessible as a stand-alone resource at:
http://github.com/kbatsuren/KinDiv
http://ukc.disi.unitn.it/index.php/kinship/

Can be browsed and visualized at:
http://ucd.datascientia.eu
http://www.livelanguage.eu

<table>
<thead>
<tr>
<th>Domain</th>
<th>Concepts</th>
<th>Languages</th>
<th>Words</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>grandparents</td>
<td>19</td>
<td>539</td>
<td>391</td>
<td>7,171</td>
</tr>
<tr>
<td>grandchildren</td>
<td>27</td>
<td>247</td>
<td>202</td>
<td>5,049</td>
</tr>
<tr>
<td>siblings</td>
<td>21</td>
<td>304</td>
<td>498</td>
<td>3,851</td>
</tr>
<tr>
<td>uncles&amp;aunts</td>
<td>31</td>
<td>625</td>
<td>312</td>
<td>16,503</td>
</tr>
<tr>
<td>nephews&amp;nieces</td>
<td>33</td>
<td>65</td>
<td>214</td>
<td>1,606</td>
</tr>
<tr>
<td>cousins</td>
<td>67</td>
<td>60</td>
<td>294</td>
<td>3,190</td>
</tr>
</tbody>
</table>

Total       | 198      | 699       | 1,911 | 37,370 |

References