Nunc profana tractemus. Detecting Code-Switching in a Large Corpus of 16th Century Letters

Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Schrofenegger, Raphael Schwitter, Phillip Ströbel, Benjamin Suter

The Bullinger Letter Collection

This paper introduces a corpus of 12'000 letters in Latin and Early New High (ENH) German for studies in history, linguistics, and theology. The corpus also contains a few letters in French, Greek and Italian. We focused on the detection of code-switching between Latin and ENH German.

Latin to German

German to Latin

Crastino comitia erunt Domus tantum Dei propter dissidium Zuziensium et Samadensium

Dann Galli nostri treuwend unnd erschreckend mengen das hertz, das er hinschlichen last, ne privetur stipendio.

Examples from our corpus with intra-sentential code-switching.

Motivations for Code-Switching Detection

1. Linguistic Studies
2. Language-specific Search
3. Training Language-specific Hand-written Text Recognition (HTR)
4. Machine Translation from Latin to Modern German
5. Normalisation of ENH-German

Sentence-based Language Identification

We trained our language identifier FurL on
• 150 sentences of (16th century) Latin and
• 150 sentences of ENH-German

We tested FurL on Caesar’s “Bello Gallico” (314 sentences cut down to 20 characters): 100% correctly classified as Latin.

In our corpus, FurL classified around
• 165,500 sentences as Latin (with 2.7 million tokens)
• 39,600 sentences as ENH-German (with 0.8 million tokens)

If the number of characters for a letter exceeded 3% for either language, OR if it has at least two sentences with at least 30 characters in the other language, then we counted it as code-switching letter.

<table>
<thead>
<tr>
<th>Code-sw ENHG</th>
<th>Code-sw Latin</th>
<th>ENHG</th>
<th>Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td>688</td>
<td>1330</td>
<td>920</td>
<td>5309</td>
</tr>
</tbody>
</table>

2018 of 8247 letters (24%) contain code-switching on the sentence level.

Word-based Language Identification

Examples of the overlapping vocabularies:

<table>
<thead>
<tr>
<th>token</th>
<th>freq(DE)</th>
<th>freq(LA)</th>
<th>vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albrecht</td>
<td>41</td>
<td>1</td>
<td>German</td>
</tr>
<tr>
<td>Alexander</td>
<td>9</td>
<td>18</td>
<td>undec</td>
</tr>
<tr>
<td>Africa</td>
<td>2</td>
<td>10</td>
<td>undec</td>
</tr>
<tr>
<td>Augustinus</td>
<td>5</td>
<td>147</td>
<td>Latin</td>
</tr>
<tr>
<td>in</td>
<td>9298</td>
<td>50,340</td>
<td>undec</td>
</tr>
<tr>
<td>bis</td>
<td>259</td>
<td>145</td>
<td>twice</td>
</tr>
<tr>
<td>breve</td>
<td>9</td>
<td>6/ short</td>
<td>undec</td>
</tr>
<tr>
<td>briefen</td>
<td>22</td>
<td>1</td>
<td>German</td>
</tr>
<tr>
<td>dies</td>
<td>17</td>
<td>1236</td>
<td>Latin</td>
</tr>
</tbody>
</table>

2018 of 8247 letters (24%) contain code-switching on the sentence level.

Conclusion

• The pretrained model of the language identifier langid did not reliably distinguish between Latin and German.
• Our special-purpose language identifier FurL which we trained on only 150 sentences worked well for binary German vs. Latin sentence-level language labelling.
• Based on this sentence classification, we bootstrapped a word-based language identifier which works with high accuracy and reliably identifies sentence-internal code-switches.

Our method is easily applicable and guarantees high lexical coverage which is important for languages like ENH German with many spelling variants. We will make both the corpus and the digital edition available online. The corpus will size up to roughly 3.5 million tokens in Latin and 1.2 million in ENH German.

LREC, Marseille, June 2022

We gratefully acknowledge project funding provided by various sponsors through the UZH Foundation (see www.bullinger-digital.ch/about).

Contact

Department of Computational Linguistics

http://www.cl.uzh.ch

Martin Volk
Andreasstrasse 15, CH-8050 Zurich
volk@cl.uzh.ch