Motivation and Benefits

Linguistic preprocessing is a fundamental step in corpus research. Its accuracy has a high impact on subsequent analysis. Especially research in the digital humanities may require de facto error-free preprocessing and thus manual correction. Thus, given a text corpus the question arises: Where to start? Find answers using *TTLab Tagging Analyzer* ([prepro.hucompute.org](http://prepro.hucompute.org)).

- Perform linguistic preprocessing of single documents or entire corpora and map them to TEI P5.
- Quickly browse through statistics and views on your data.
- Upload manually checked TEI P5 as gold standard to evaluate different preprocessors.
- Browse through evaluation results to identify typical errors and estimate effort for manual correction.
- Use it online from wherever you are with many input formats supported.

Process Model

- Multiple documents in various formats, possibly compressed, can be uploaded in one step.
- Uploaded TEI documents are automatically interpreted as reference against which an evaluation is performed.
- Supported preprocessors are automatically chosen and run, unless the user makes an explicit selection.
- The optional evaluation against a reference document includes an alignment of tokens and computation of various measures known from machine learning.
- Statistics for charts and spread sheets are computed.
- The results are returned to the client and all TEI files are ready for download.

Comparing two Documents

Example of comparing two documents to compute the F-measure, precision and recall:

```xml
<jas>
  <w lemma="leimus" type="ADV">Qualiter</w>
  <c> </c>
  <w lemma="Karolus" type="NP">Karolus</w>
  ...</jas>
```

Evaluation of Preprocessors

- Measures from machine learning like F-measure, precision and recall allow for assessing the performance of preprocessors.
- Rank-frequency charts of PoS-tagging and lemmatization errors help to quickly grasp the most frequent mistakes.
- Sortable keyword and context views enable users to browse through the quotations.

Technologies

<table>
<thead>
<tr>
<th>Name</th>
<th>Languages</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto Selection</td>
<td>English, German, Latin</td>
<td>TextTechnology Lab</td>
</tr>
<tr>
<td>Stanford NLP Group</td>
<td>English</td>
<td>Stanford NLP Group</td>
</tr>
<tr>
<td>TTLab Preprocessor</td>
<td>English, Latin</td>
<td>TextTechnology Lab</td>
</tr>
<tr>
<td>TTLab Legacy Preprocessor</td>
<td>English, German, Latin</td>
<td>TextTechnology Lab</td>
</tr>
</tbody>
</table>

References