Overview
For different reasons, text can be difficult to read and understand for many people, especially if the text’s language is too complex. In order to provide suitable text for the target audience, it is necessary to measure its complexity.

Hypotheses
1) The target group has a significant influence on the complexity rating.
2) Non-experts perceive domain specific texts as more complex.
3) The complexity score can be predicted by linguistic features.

Predicting Text Complexity
We compiled a set of 147 linguistic features (Syntax, Morphology, Lexicon) and extracted them for each sentence. We predict the exact value of complexity using a Linear Regression model.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th># Features</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morpho</td>
<td>28</td>
<td>0.49</td>
</tr>
<tr>
<td>Lexicon</td>
<td>57</td>
<td>0.43</td>
</tr>
<tr>
<td>Syntax</td>
<td>14</td>
<td>0.41</td>
</tr>
<tr>
<td>Readability Formulas (RM)</td>
<td>5</td>
<td>0.35</td>
</tr>
<tr>
<td>Syntax + RM</td>
<td>19</td>
<td>0.31</td>
</tr>
<tr>
<td>Individual Set</td>
<td>20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Conclusion
A set of 20 features can predict a complexity score. Corpus (322 sentences) and feature set are openly available.

Corpus Creation
Data provided by project partner DATEV
Instructions, commentaries and descriptions of technical solutions or law regulations.

Sentence 3:
DATEV provides the customer with the use of the DATEV computer center / the DATEV Cloud within the scope of the service described in this service description and processes personal data on behalf of the customer within the scope of the provision of the service and for separately agreed service and support services and in the case of remote support.

Q 3.1 How do you rate the overall complexity of the sentence?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very difficult</td>
<td>7</td>
</tr>
<tr>
<td>difficult</td>
<td>6</td>
</tr>
<tr>
<td>Somewhat difficult</td>
<td>5</td>
</tr>
<tr>
<td>Neutral</td>
<td>4</td>
</tr>
<tr>
<td>Somewhat easy</td>
<td>3</td>
</tr>
<tr>
<td>Easy</td>
<td>2</td>
</tr>
<tr>
<td>Very easy</td>
<td>1</td>
</tr>
</tbody>
</table>

A) Non-experts via crowdsourcing
B) Experts recruited from DATEV staff

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the projects AuTexx (01IS17043) and vALID (01GP1903A). Moreover, we would like to thank DATEV eG and Prof. Dr. Andreas Both (Head of Research, DATEV) for providing data and helping to conduct the expert-experiments.