Subjective Text Complexity Assessment for German

Laura Seiffe[1], Fares Kallel[1], Babak Naderi[2], Sebastian Möller[1,2], Roland Roller[1]

[1]German Research Center for Artificial Intelligence (DFKI),
[2]Quality and Usability Lab, TU Berlin, Berlin, Germany

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.

Corpus Creation

Data provided by project partner DATEV

Instructions, commentaries and descriptions of technical solutions or law regulations.

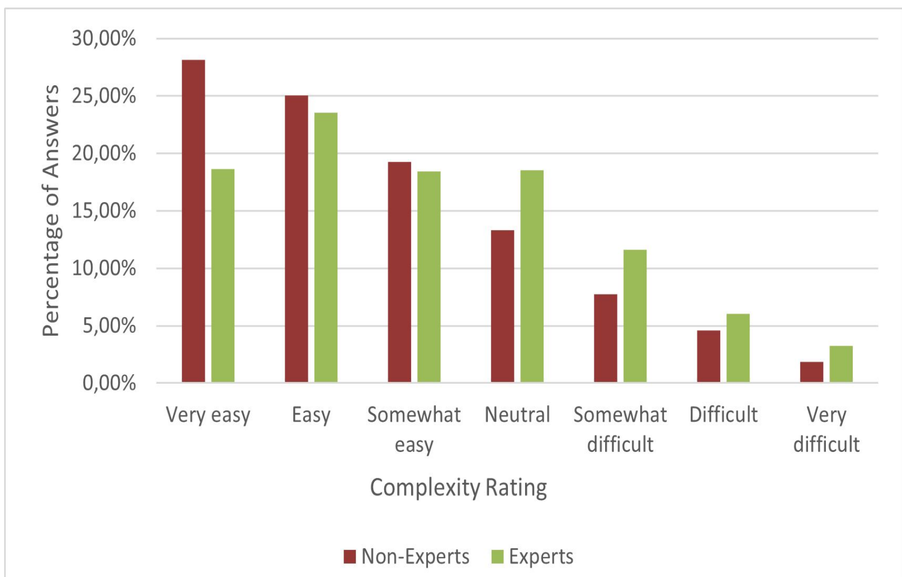
Sentence 3:

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?

Rating	Score
○ Very difficult	7
○ difficult	6
○ Somewhat difficult	5
○ Neutral	4
○ Somewhat easy	3
○ Easy	2
○ Very easy	1

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



Strong tendency towards *easy* or *very easy* rating

Non-experts rate even more *easy*

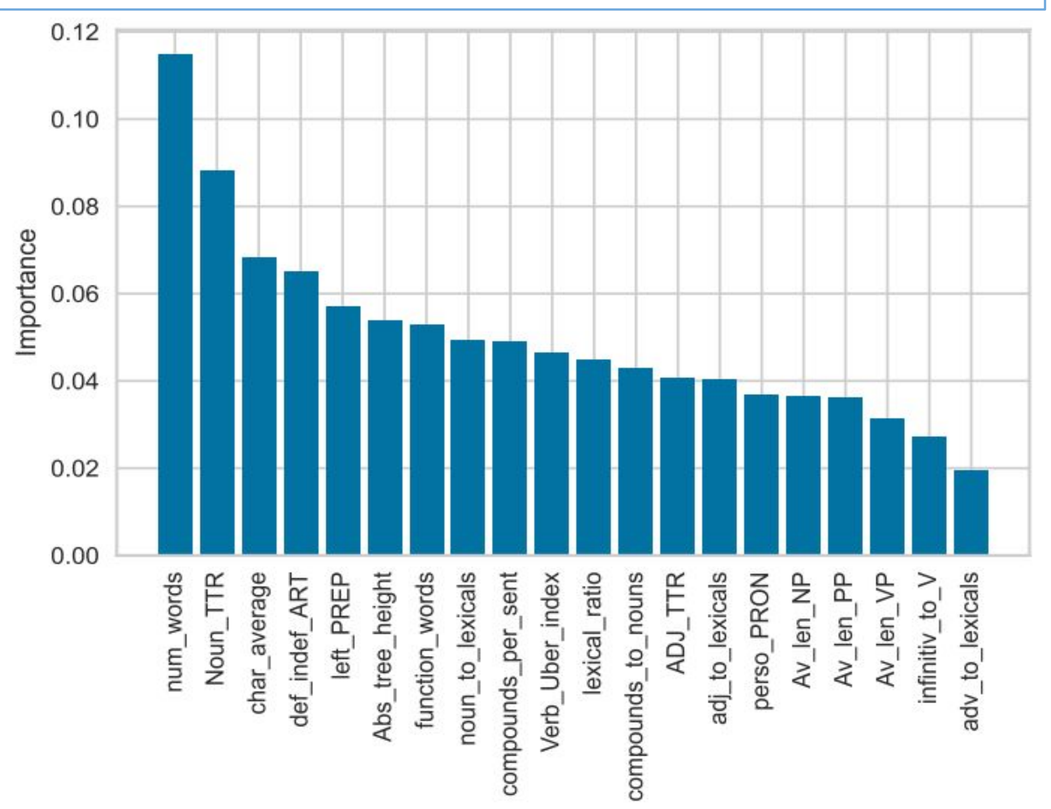
Predicting Text Complexity

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.

Feature Set	# Features	RMSE
Morpho	28	0.49
Lexicon	57	0.43
Syntax	14	0.41
Readability Formulas (RM)	5	0.35
Syntax + RM	19	0.31
Individual Set	20	0.20

Conclusion

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



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.