Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and Its Cross-Corpus Validation

Zarah Weiss & Detmar Meurers
University of Tübingen, ICALL Research Group (http://icall-research.de)

Introduction

- We are presenting an **empirically broad cross-corpus analysis** of German educational media language targeting adults and children.
- **Readability assessment** refers to the task of linking a text to the appropriate target audience based on its complexity.
- Potential application domains include: the design and evaluation of education materials, information retrieval, and text simplification.
- While for English reference corpora for cross-corpus testing are available (*Common Core*, *WeeklyReader*), there are no similar resources for German.
- We illustrate how empirically broad models successfully generalize across German educational media corpora.

Automatic Complexity Analysis

- We automatically extract **400 complexity measures** covering clausal, phrasal, lexical, morphological, and discourse complexity, cognitive complexity, and language use.
- The features are theoretically grounded in SLA research, where Complexity is defined in terms of **elaborateness and variation** of language (Housen et al., 2012).
- This is to our knowledge currently the most extensive feature set for German; for details, see Weiss & Meurers (accepted).
- The pipeline will be integrated into CTAP (Chen & Meurers, 2016) by the end of fall 2018.

Data

- We compiled two balanced corpora of German news media for adults and children:
  - **GEO/GEOlino**: online articles from the leading German monthly educational magazine *GEO* and its children counterpart *GEOlino* (similar to Hancke et al. (2012)).
  - **Tagesschau/Logo**: official subtitles of the German daily news broadcasts for adults (*Tagesschau*) and children (*Logo!*).
- All subtitles and all article links are **available for research**.

Data Profiles

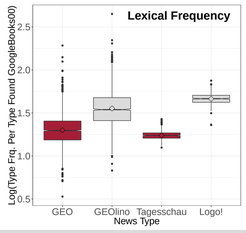
	GEO	GEOlino	Tagesschau	Logo
Documents	2,480	2,480	421	415
Median sent.	23	25	167	125
Median words	383	350	1,631	1,322
	GEO_4	$GEOlino_4$	$Tagesschau_{\frac{1}{5}}$	$Logo_{\frac{1}{5}}$
Documents	420	420	2,049	2,049
Median sent.	112.5	122.5	32	24
Median words	1,797	1,741	325	259

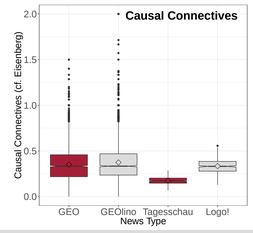
- Corpus size and document lengths differ across corpora.
- **Tagesschau/Logo$_{\frac{1}{5}}$**: We split all Tagesschau/Logo subtitles into five parts and sampled 2,049 texts per target audience.
- **GEO/GEOlino$_4$**: We appended up to four GEO/GEOlino articles from the same topic domain and sampled 420 texts each.

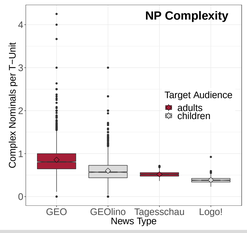
Informativeness of Complexity Measures

- We calculated the information gain (IG) of each feature for the target audience with 10-fold cross-validation (10-fold CV).
- When we rank all features by IG and compare the top 20 features with an Pearson inter-correlation $r \leq .8$, we see that
  - language use and cohesion measures cover overall 55% of the top 20 features for both data set.
  - also measures of phrasal, sentential, lexical, and morphological complexity are important.

GEO/GEOlino			Tagesschau/Logo		
Merit	Group	Feature	Merit	Group	Feature
0.33	USE	sumTypesMinAoAPerTypeInKCT	**0.98**	USE	logTypeFreqsPerTypeInGoogle00
0.33	LEX	syllablesPerToken	0.90	USE	typesNotInSubtlexPerLexicalType
0.29	USE	logTypeFreqsPerTypeInSubtlex	0.83	COH	2PPersPronounsPerNoun
0.23	USE	typesNotInSubtlexPerLexicalType	0.75	COH	probNotSubsPerTransition
0.21	COH	2PPersAndPossPronounsPerToken	0.72	COH	causalConnectivePerSentence
0.16	USE	typesNotInDlexPerLexicalType	0.69	COH	localArgOverlapsPerSentence
0.15	PHR	complexNominalsPerTUnit	0.67	SEN	sumParseTreeHeightsPerFiniteClause
0.14	USE	logLemmaFreqsPerTypeInKCT	0.66	SEN	NPsPerTUnit
0.14	SEN	syllablesInMiddleFieldPerMiddleField	0.66	COH	1PPersPronounsPerToken
0.13	COH	persPronounsPerToken	0.66	MOR	genitivesPerNoun
0.13	SEN	PPsPerTUnit	0.63	PHR	determinersPerNP
0.13	MOR	secondPersonMarkingsPerFiniteVerb	0.62	USE	lemmaFreqsPerTypeInKCT
0.12	MOR	ionTPerToken	0.62	COG	sumLongestDependenciesPerClause
0.12	LEX	synsetPerTypeInGnet	0.62	MOR	compundNounsPerNP
0.12	PHR	complexNominalsPerFiniteClause	0.61	USE	typeFreqsPerTypeInDlex
0.12	SEN	sumNonTerminalNodesPerTUnit	0.57	LEX	MTLD
0.12	USE	typeFreqsPerTypeInSubtlex	0.56	LEX	nonAuxVerbTypesPerNonAuxVerbToken
0.11	COH	pronounsPerNoun	0.55	COH	globalStemOverlapsPerSentence
0.11	USE	logAnnotatedTypeFrqBd5PerTypeInKCT	0.51	SEN	conjunctionalClausesPerSentence
0.11	COH	3PPersAndPossPronounsPerNoun	**0.50**	USE	logAnnotatedTypeFrqBd4PerTypeInDlex







Readability Models

Model	Training	Testing	Features	Acc.	SD
Baseline		GEO/GEOlino		50.0	
		Tagesschau/Logo		50.0	
10-fold CV	GEO/GEOlino	GEO/GEOlino	400	89.4	±0.09
			20	85.1	±0.09
	Tagesschau/Logo	Tagesschau/Logo	400	99.9	±0.04
			20	99.8	±0.03
Cross-Corpus	GEO/GEOlino	Tagesschau/Logo	400	**98.9**	
			20	98.8	
	Tagesschau/Logo	GEO/GEOlino	400	**52.2**	
			20	56.7	
Balanced CC	GEO/GEOlino$_4$	Tagesschau/Logo	400	99.2	
	Tagesschau/Logo$_{\frac{1}{5}}$	GEO/GEOlino	400	59.0	

- We train linear SMO classifiers using WEKA (Hall et al., 2009) and evaluate them with 10-fold CV and cross-corpus testing.
- Training on GEO/GEOlino yields high performance in within and cross-corpus testing, but training on Tagesschau/Logo does not generalize to GEO/GEOlino.
- This difference is not due to the larger corpus size of GEO/GEOlino: also GEO/GEOlino$_4$ generalizes well across corpora while Tagesschau/Logo$_{\frac{1}{5}}$ does not.

Summary & Outlook

- We designed a highly accurate model which i) is based on a very broad coverage of linguistic features, and ii) successfully generalizes across corpora and text types.
- We show that German educational media language is successfully and broadly adapted towards their target audiences, unlike, e.g., German school textbooks (Berendes et al., 2017).
- Next steps include comparing the data to *language produced by children belonging to the target group* (linguistic complexity $\hat{=}$ proficiency, see Weiss & Meurers (accepted))

References

Berendes, K., S. Vajjala, D. Meurers, D. Bryant, W. Wagner, M. Chinkina & U. Trautwein (2017). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology* .

Chen, X. & D. Meurers (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Hancke, J., D. Meurers & S. Vajjala (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080.

Housen, A., F. Kuiken & I. Vedder (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency*, John Benjamins, Language Learning & Language Teaching.

Weiss, Z. & D. Meurers (accepted). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the 4th Learner Corpus Research Conference 2017*. Presses Universitaires de Louvain.