Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment

Zarah Weiss Xiaobin Chen Detmar Meurers

Department of Linguistics & LEAD Research Network
University of Tübingen

10th Workshop on NLP for Computer Assisted Language Learning
Online workshop, May 31st, 2021

Cross-lingual readability assessment

- Readability assessment identifies suitable texts given
  - the language proficiency of the target audience
  - the reading purpose (e.g., learning, information retrieval)
- Readability classification often uses linguistic complexity
  - Alternative: neural classification approaches
  - Often higher performance but lack of linguistic insight
  - More easily applicable across multiple languages
- Increasing interest in multi-/cross-lingual readability
  - Lack of multi-level training data for non-English languages
  - Often compensated with binary classification
    (e.g. Madrazo Azpiazu & Pera 2020a,b; Weiss & Meurers 2018)

Research questions

- Present novel resources for cross-lingual feature-based readability classification
  - Multi-lingual complexity features for English and German
  - New multi-level readability corpus for English and German
- Research questions
  1. Are our feature-based readability classifiers successful?
  2. Do they generalize in zero-shot cross-lingual learning?
  3. Do reading levels differ linguistically across languages?

Automatic complexity analysis

- Complexity, Accuracy, and Fluency used to describe language performance in SLA (Housen et al. 2019)
  - Complex language is elaborate and varied (Ellis 2003)
- Assessment of text readability for specific target group (Vajjala & Meurers 2012; Chen & Meurers 2017; Crossley et al. 2019)
  - ⇒ Exclusively language-specific (but De Clercq & Hoste 2016)
- Propose broad feature set for English and German (N=312)
  - Syntactic, lexical, and morphological complexity featues
  - Measures of discourse cohesion
  - Human language use and processing measures
  - Based on Chen (2018); Weiss & Meurers (2018)

Broad multi-lingual complexity modeling

1. Shared NLP pipeline to obtain
   - Sentences, words, POS tags, constituency/dependency parses with Stanford CoreNLP (Manning et al. 2014)
   - Lemmas with Mate tools (Bohnet & Nivre 2012)
   - Stems with OpenNLP Snowball stemmer
2. Identify linguistic constructions with extraction rules
   - Language-specific rules for syllables, POS, constituents
   - Language-independent rules for all other
   - Frequency features based on Subtlex-US and Subtlex-DE (Brysbaert et al. 2011a,b)
3. Fully language-independent calculation of feature ratios

Spotlight corpus

- Articles from two monthly language learning magazines
  - *Spotlight* for L2 English (www.spotlight-online.de)
  - *Deutsch perfekt* for L2 German (www.deutsch-perfekt.de)
- Leveled reading materials aligned with CEFR levels
  - Beginning(A2), medium (B1/B2), advanced (C1)
- Articles extracted from type-set PDFs using OCR
  - Spotlight-EN from 110 issues (01/2012-12/2019)
  - Spotlight-DE from 45 issues (01/2018-12/2019)

Spotlight corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Spotlight-EN					
Easy	1.030	206±166	137	53	877
Medium	1.528	588±555	493	23	4.497
Advanced	727	606±509	489	26	2.940
Spotlight-DE					
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- More English than German & easy than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Spotlight-EN					
Easy	1.030	206±166	137	53	877
Medium	1.528	588±555	493	23	4.497
Advanced	727	606±509	489	26	2.940
Spotlight-DE					
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- More English than German & easy than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Spotlight-EN					
Easy	1.030	206±166	137	53	877
Medium	1.528	588±555	493	23	4.497
Advanced	727	606±509	489	26	2.940
Spotlight-DE					
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- More English than German & easy than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Spotlight-EN					
Easy	1.030	206±166	137	53	877
Medium	1.528	588±555	493	23	4.497
Advanced	727	606±509	489	26	2.940
Spotlight-DE					
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- More English than German & easy than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Set-up

1. Remove invariable features ⇒ from 312 to 301 features
   - Criterion for removal: 95% identical values in all corpora
2. Calculate feature z-scores separately for Spotlight-EN/-DE
3. Compare (ordinal) random forest, SVM (poly/radial)
   ⇒ Ordinal random forest performing best on both
4. Within-language training/testing (10-fold cross-validation)
5. Cross-language testing on respective other data set
   - Previous research focus: augmenting with multi-lingual data
   - Here: form of zero-shot learning

Results within languages

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
EN	CV	74.5	[73.0, 76.0]	46.5	$< 2.2 \cdot 10^{-16}$
DE	CV	88.0	[86.1, 89.6]	52.8	$< 2.2 \cdot 10^{-16}$
EN	DE	55.5	[52.9, 58.1]	52.8	.02118
DE	EN	53.4	[51.7, 55.1]	46.5	$1.284 \cdot 10^{-15}$

- English and German classifiers in 10-CV highly successful
  - Sufficiently narrow confidence intervals
  - Clear performance above baseline

Confusion Matrix and Level-Wise Performance

Spotlight-EN 10-fold CV

Pred\Obs.	Easy	Medium	Advanced
Easy	816	171	37
Medium	208	1,210	268
Advanced	6	147	422
Precision	79.7	71.8	73.4
Recall	79.2	79.2	58.1
F1	79.5	75.3	65.0

- Level-wise performance relatively balanced
- For English much lower recall for advanced materials

Confusion Matrix and Level-Wise Performance

Spotlight-DE 10-fold CV

Pred\Obs.	Easy	Medium	Advanced
Easy	727	83	1
Medium	34	399	27
Advanced	2	27	146
Precision	89.6	86.7	83.4
Recall	95.3	78.4	83.9
F1	92.4	82.4	83.7

- Level-wise performance relatively balanced
- For German much higher precision/recall for easy materials

Results across languages

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
EN	CV	74.5	[73.0, 76.0]	46.5	$< 2.2 \cdot 10^{-16}$
DE	CV	88.0	[86.1, 89.6]	52.8	$< 2.2 \cdot 10^{-16}$
EN	DE	55.5	[52.9, 58.1]	52.8	.02118
DE	EN	53.4	[51.7, 55.1]	46.5	$1.284 \cdot 10^{-15}$

- Both models generalize to some degree
  - Especially for German classifier on English data
  - Considerable performance drop across languages

Results across languages

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
EN	CV	74.5	[73.0, 76.0]	46.5	$< 2.2 \cdot 10^{-16}$
DE	CV	88.0	[86.1, 89.6]	52.8	$< 2.2 \cdot 10^{-16}$
EN	DE	55.5	[52.9, 58.1]	52.8	.02118
DE	EN	53.4	[51.7, 55.1]	46.5	$1.284 \cdot 10^{-15}$

- Both models generalize to some degree
  - Especially for German classifier on English data
  - Considerable performance drop across languages

Results across languages

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
EN	CV	74.5	[73.0, 76.0]	46.5	$< 2.2 \cdot 10^{-16}$
DE	CV	88.0	[86.1, 89.6]	52.8	$< 2.2 \cdot 10^{-16}$
EN	DE	55.5	[52.9, 58.1]	52.8	.02118
DE	EN	53.4	[51.7, 55.1]	46.5	$1.284 \cdot 10^{-15}$

- Both models generalize to some degree
  - Especially for German classifier on English data
  - Considerable performance drop across languages

Results across languages

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
EN	CV	74.5	[73.0, 76.0]	46.5	$< 2.2 \cdot 10^{-16}$
DE	CV	88.0	[86.1, 89.6]	52.8	$< 2.2 \cdot 10^{-16}$
EN	DE	55.5	[52.9, 58.1]	52.8	.02118
DE	EN	53.4	[51.7, 55.1]	46.5	$1.284 \cdot 10^{-15}$

- Both models generalize to some degree
  - Especially for German classifier on English data
  - Considerable performance drop across languages

Confusion Matrix and Level-Wise Performance

Spotlight-EN on Spotlight-DE

Pred\Obs.	Easy	Medium	Advanced
Easy	341	73	0
Medium	408	343	56
Advanced	14	93	118
Precision	82.3	42.5	52.4
Recall	44.6	67.4	67.8
F1	57.8	52.1	59.2

- English classifier overestimates true reading difficulty
  - Especially easy reading materials labeled as medium

Confusion Matrix and Level-Wise Performance

Spotlight-DE on Spotlight-EN

Pred\Obs.	Easy	Medium	Advanced
Easy	827	635	216
Medium	193	732	315
Advanced	10	161	196
Precision	49.3	59.0	53.4
Recall	80.3	47.9	27.0
F1	61.1	52.9	35.8

- German classifier underestimates true reading difficulty
  - Medium reading materials labeled as easy
  - Advanced reading materials labeled as medium

Set-up

- Zooming in on individual features both Spotlight corpora
  - Which linguistic features distinguish reading levels?
  - How does this compare across languages?
- Use correlation-based feature subset selection (Hall 1999)
  - Maximize correlation with reading level
  - Minimize inter-correlation of selected features
- Aggregate selected features across linguistic domains
  - 16.3% of features selected for German (49/301)
  - 14.3% of features selected for English (43/301)

  $\Rightarrow$ Zoom in on individual linguistic domains

Comparison of lexical features

Group	EN (%)		DE (%)		All
Density	7	(25.9)	5	(18.5)	27
Diversity	1	(11.1)	1	(11.1)	9
Richness	4	(7.5)	5	(9.4)	53

- Lexical density and richness relevant in both languages
  - Lexical richness: POS independent type-token ratios
  - Lexical density: POS specific ratios (lexical words / word)
- Lexical diversity hardly relevant for either language
  - Lexical diversity: POS specific type-token ratios

$\Rightarrow$ Feature relevance overall very similar for English & German

Comparison of lexical features

Group	EN (%)		DE (%)		All
Density	7	(25.9)	5	(18.5)	27
Diversity	1	(11.1)	1	(11.1)	9
Richness	4	(7.5)	5	(9.4)	53

- Lexical density and richness relevant in both languages
  - Lexical richness: POS independent type-token ratios
  - Lexical density: POS specific ratios (lexical words / word)
- Lexical diversity hardly relevant for either language
  - Lexical diversity: POS specific type-token ratios

$\Rightarrow$ Feature relevance overall very similar for English & German

Comparison of lexical features

Group	EN (%)		DE (%)		All
Density	7	(25.9)	5	(18.5)	27
Diversity	1	(11.1)	1	(11.1)	9
Richness	4	(7.5)	5	(9.4)	53

- Lexical density and richness relevant in both languages
  - Lexical richness: POS independent type-token ratios
  - Lexical density: POS specific ratios (lexical words / word)
- Lexical diversity hardly relevant for either language
  - Lexical diversity: POS specific type-token ratios

$\Rightarrow$ Feature relevance overall very similar for English & German

Comparison of syntactic features

Group	EN (%)		DE (%)		All
Clausal	1	(5.0)	8	(40.0)	20
Phrasal	1	(3.6)	5	(17.9)	28
Variation	2	(16.7)	0	(0.0)	12

- Syntactic complexity measures little relevance for English
  - Syntactic variation over syntactic elaboration
- Syntactic elaboration very informative for German
  - Syntactic variation irrelevant

$\Rightarrow$ Syntactic readability differences more language-specific

Comparison of syntactic features

Group	EN (%)		DE (%)		All
Clausal	1	(5.0)	8	(40.0)	20
Phrasal	1	(3.6)	5	(17.9)	28
Variation	2	(16.7)	0	(0.0)	12

- Syntactic complexity measures little relevance for English
  - Syntactic variation over syntactic elaboration
- Syntactic elaboration very informative for German
  - Syntactic variation irrelevant

$\Rightarrow$ Syntactic readability differences more language-specific

Comparison of syntactic features

Group	EN (%)		DE (%)		All
Clausal	1	(5.0)	8	(40.0)	20
Phrasal	1	(3.6)	5	(17.9)	28
Variation	2	(16.7)	0	(0.0)	12

- Syntactic complexity measures little relevance for English
  - Syntactic variation over syntactic elaboration
- Syntactic elaboration very informative for German
  - Syntactic variation irrelevant

$\Rightarrow$ Syntactic readability differences more language-specific

Comparison of syntactic features

Group	EN (%)		DE (%)		All
Clausal	1	(5.0)	8	(40.0)	20
Phrasal	1	(3.6)	5	(17.9)	28
Variation	2	(16.7)	0	(0.0)	12

- Syntactic complexity measures little relevance for English
  - Syntactic variation over syntactic elaboration
- Syntactic elaboration very informative for German
  - Syntactic variation irrelevant

$\Rightarrow$ Syntactic readability differences more language-specific

Comparison of syntactic features

Group	EN (%)		DE (%)		All
Clausal	1	(5.0)	8	(40.0)	20
Phrasal	1	(3.6)	5	(17.9)	28
Variation	2	(16.7)	0	(0.0)	12

- Syntactic complexity measures little relevance for English
  - Syntactic variation over syntactic elaboration
- Syntactic elaboration very informative for German
  - Syntactic variation irrelevant

$\Rightarrow$ Syntactic readability differences more language-specific

Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

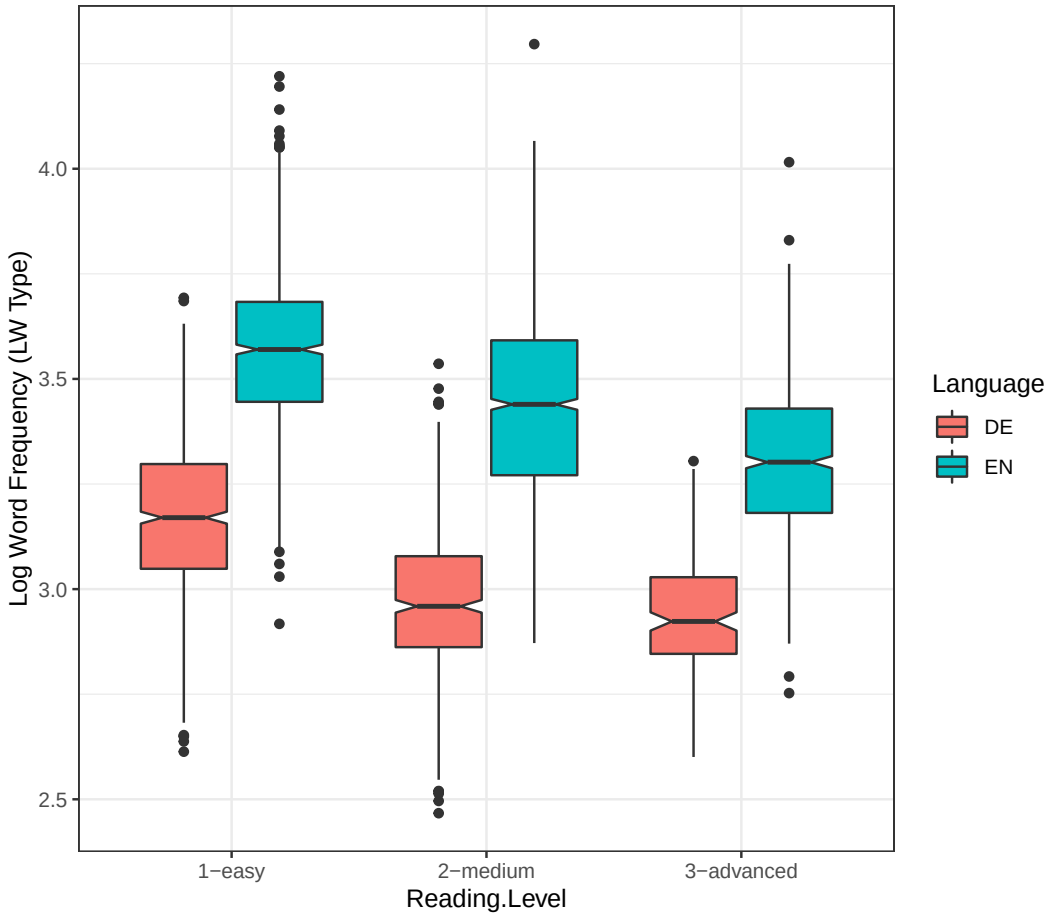
Comparison of other feature groups

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11

- Surface and language use features relevant for both
- Morphology more relevant for English than German
- Discourse and processing features hardly/not relevant

$\Rightarrow$ Linguistic differences mostly comparable across languages
  - Morphological complexity only exception

Illustrate feature differences for English/German



- More frequent vocabulary in English than German texts
- Word frequency decreases with increasing reading levels

Conclusion

- First broad multi-lingual feature set for English and German
  - Currently being extended to other languages
- Presented new multi-lingual multi-level readability corpus
  - Spotlight-DE first multi-level corpus for German
  - Extracting Spotlight data for Italian, Spanish, and French
- Show applicability of readability models across languages
  - Investigate zero-shot learning with feature-based classifiers
- Comparable reading level differences across languages
  - Esp. in lexical domain and for frequency / surface measures

  $\Rightarrow$ Feature-based approach allows detailed linguistic insights

  $\Rightarrow$ Studies including more languages in future work

References

Bengoetxea, K., I. González-Dios & A. Aguirregoitia (2020). AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural* 64, 61–68.

Bohnet, B. & J. Nivre (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1455–1465. URL https://www.aclweb.org/anthology/D12-1133.

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl (2011a). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Brysbaert, M., E. Keuleers & B. New (2011b). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology* 2(27).

Chen, X. (2018). Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research. Ph.D. thesis, Eberhard Karls Universität Tübingen Germany. URL http://hdl.handle.net/10900/85888.

Chen, X. & D. Meurers (2017). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading* 41(3), 486–510.

Crossley, S. A., S. Skalicky & M. Dascalu (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading* 42(3-4), 541–561.

De Clercq, O. & V. Hoste (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics* 42(3), 457–490.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

Housen, A., B. De Clercq, F. Kuiken & I. Vedder (2019). Multiple approaches to complexity in second language research. *Second Language Research. Special Issue on Linguistic Complexity* 35(1), 2–31.

Madrazo Azpiazu, I. & M. S. Pera (2020a). An Analysis of Transfer Learning Methods for Multilingual Readability Assessment. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* pp. 95–100.

Madrazo Azpiazu, I. & M. S. Pera (2020b). Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology* .

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard & D. McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pp. 55–60. http://aclweb.org/anthology/P/P14/P14-5010.

Martinc, M., S. Pollak & M. Robnik-Šikonja (2019). Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779* .

Vajjala, S. & I. Lučić (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*. pp. 297–304.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. pp. 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Weiss, Z. & D. Meurers (2018). Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and its Cross-Corpus Validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

OneStopEnglish corpus (OSE; Vajjala & Lučić 2018)

- Corpus of 567 Guardian articles for English L2 learners
  - Each article as elementary, intermediate, advanced version
- Established for text simplification/readability assessment
  - State-of-the art: 90.09% acc. (Bengoetxea et al. 2020)
  - Best neural approach: 78.7% acc. (Martinc et al. 2019)

Level	N. docs	μ words ($\pm SD$)	M	Min	Max
Ele.	189	556($\pm$109)	561	267	948
Int.	189	679($\pm$117)	691	315	1.083
Adv.	189	860($\pm$171)	857	357	1.465

Benchmarking feature-based classification (OSE)

- Remove invariable features ⇒ from 312 to 301 features
- Compare (ordinal) random forest, SVM (poly/radial)
  - ⇒ Best performance for SVM with polynomial kernel
- Train and test with 10-fold cross-validation
- Overall accuracy: 92.1%[89.5%; 94.2%]
  - Exceeds random baseline: 33.3%
  - Exceeds state-of-the art: 90.1% (Bengoetxea et al. 2020)

⇒ Our feature-based classification approach is competitive

Confusion Matrix and Level-Wise Performance

OneStopEnglish 10-fold CV

Pred\Obs.	Ele.	Int.	Adv.
Ele.	**179**	9	4
Int.	9	**173**	15
Adv.	1	7	**170**
Precision	93.2	87.8	95.5
Recall	94.7	91.5	90.0
F1	94.0	89.6	92.6