Broad Linguistic Complexity Analysis for Greek Readability Classification

Savvas Chatzipanagiotidis
Universität Tübingen, Germany
sv.chtz@gmail.com

Maria Giagkou
Institute for Language & Speech Processing, Greece
mgiagkou@ilsp.gr

Detmar Meurers
Universität Tübingen, Germany
detmar.meurers@uni-tuebingen.de

Abstract

This paper explores the linguistic complexity of Greek textbooks as a readability classification task. We analyze textbook corpora for different school subjects and textbooks for Greek as a Second Language, covering a very wide spectrum of school age groups and proficiency levels. A broad range of quantifiable linguistic complexity features (lexical, morphological and syntactic) are extracted and calculated.

Conducting experiments with different feature subsets, we show that the different linguistic dimensions contribute orthogonal information, each contributing towards the highest result achieved using all linguistic feature subsets. A readability classifier trained on this basis reaches a classification accuracy of 88.16% for the Greek as a Second Language corpus.

To investigate the generalizability of the classification models, we also perform cross-corpus evaluations. We show that the model trained on the most varied text collection (for Greek as a school subject) generalizes best.

In addition to advancing the state of the art for Greek readability analysis, the paper also contributes insights on the role of different feature sets and training setups for generalizable readability classification.

1 Introduction

Automatic readability classification systems are intended to assess whether a text is appropriate for a given group of readers. The groups of readers differ in terms of their level of education in their first language (L1), their second language proficiency (L2), or in terms of their special needs (e.g., due to cognitive disabilities). Textbooks play a major role in fostering language development in school, and educational policy documents such as the US Common Core State Standards (CCSSO, 2010) make explicit that the texts should offer a "staircase of increasing complexity so that all students are ready for the demands of college- and career-level reading no later than the end of high school."[1] Correspondingly, in the US context there are a number of research and commercial approaches to measuring text difficulty for English (Nelson et al., 2012). Research continues in this domain, e.g., investigating features from second language acquisition (Vajjala and Meurers, 2012) or psycholinguistic research (Howcroft and Demberg, 2017), training neural networks (Azpiazu and Pera, 2019), broadening the corpus basis (Vajjala and Lučić, 2018), and externally grounding the ratings (Redmiles et al., 2019).

For second language learners, texts at an appropriate level of proficiency are of particular importance, with a substantial strand of computational linguistic research focusing on this target group, e.g., (François and Fairon, 2012; Pilán et al., 2014; Xia et al., 2016), and parallel automated proficiency evaluation approaches analyzing L2 writing, e.g., (Lu, 2010; Giagkou et al., 2015; Weiss and Meurers, 2019b). While much of the initial computational linguistic research on readability focused on English, research targeting other languages is also increasingly emerging, e.g., (Hancke et al., 2012; Falkenjack et al., 2013; Dell'Orletta et al., 2014; Giagkou et al., 2017). Some recent approaches employ neural networks and deep learning for classification-related tasks (Martinc et al., 2019; Conneau et al., 2017). For instance, Martinc et al. (2019) employed neural classifiers on English and Slovenian corpora using unsupervised and supervised techniques to assess readability, outperforming some previous state-of-the-art approaches.

In this paper, we focus on the Greek language and investigate readability classification using a

[1] http://purl.org/ccss/ela-keyshifts

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 48–58
April 20, 2021 ©2021 Association for Computational Linguistics

broad range of linguistic complexity features on a comprehensive collection of five textbook corpora covering various subjects, and Greek as a second language. We first explore which feature groups and combinations thereof best predict the school grade or proficiency level assigned by the publisher. Going beyond the typical within-corpus setup, we then conduct cross-corpus tests to study how well the models trained on different feature sets and corpora generalize across data sets.

The paper is organized as follows: Section 2 presents the textbook corpora and section 3 the linguistic complexity feature sets used in the readability classification experiments. Section 4 then spells out the experimental setup and the results, before concluding in section 5.

2 Data

Our data consists of five corpora of Greek texts used as educational material for different school subjects and educational contexts. We use four corpora from previous work on Greek (Giagkou, 2009; Georgatou, 2016), in addition to a Greek as a Second Language (GSL) textbooks corpus compiled for the needs of the research at hand. Table 1 provides an overview of the corpora, their number of texts and average text lengths.

Corpus	Texts	Total Tokens	Avg. Tokens
Greek	520	181.267	348
History	835	344.202	412
Science	782	199.471	255
E.DIA.M.ME.	432	106.990	248
GSL	600	164.367	273
Total Number	**3.169**	**996.297**	

Table 1: Descriptive statistics of the data set used

The *Greek* corpus (Giagkou, 2009) comprises authentic texts from the official coursebooks used in Greek schools for teaching the subject Greek from the second grade of primary school to the second grade of upper secondary school. The texts cover a wide range of domains and text types, from descriptive to argumentative texts and literature. The *History* corpus (Georgatou, 2016) consists of texts from the official history textbooks used in the Greek educational system, and the *Science* corpus (Georgatou, 2016) combines the textbooks for the *Geology, Biology and Chemistry* subjects.

For each of the three curriculum subjects, the corpora delineate textbooks at three different levels: *primary school*, *lower secondary school*, and *upper secondary school*, which correspond to ISCED 1, 2, and 3 in the International Standard Classification of Education (ISCED).

The *E.DIA.M.ME.* corpus (Georgatou, 2016) includes coursebooks for Greek as a heritage language. The texts are divided into five levels, which, in turn, correspond to grades of the greek educational system. The *E.DIA.M.ME.* project[2] aims to promote the Greek language to Greek migrants living abroad and non-native Greeks who want to learn the Greek language. As mentioned by Damanakis (2011), the learning material for *E.DIA.M.ME.* project has been designed and composed according to the CERF guidelines, allowing a partial matching between these five levels and the six CERF-proficiency levels. In particular, the fourth E.DIA.M.ME level corresponds both to the B2 and C1 CERF level. For our three-level CERF classification, we assigned the fourth E.DIA.M.ME level to the higher level (C). Finally, we compiled a Greek as a Second Language (GSL) corpus by collecting and digitizing reading material from seven coursebooks of Greek L2 and from past certification tests in Greek, for which the respective Common European Framework of Reference for Languages (CEFR) level (Council of Europe, 2001) was indicated by the publisher.[3]

For the *E.DIA.M.ME.* Greek as a heritage language corpus and the Greek as a Second Language (GSL) corpus, the CEFR framework provide a three level classification into Basic (A), Independent (B), and Proficient (C) users of Greek.

The distribution of texts per level is noted in the confusion matrices shown in Tables 3, 4, 5, 7, and 8 in the following sections.

3 Linguistic complexity analysis

We used a state-of-the-art Greek dependency parser (Prokopidis and Papageorgiou, 2017) of the Institute for Language and Speech Processing (ILSP) to annotate the corpora. The CoNLL-format files produced by the parser were used as input to our

[2] *E.DIA.M.ME.* project: http://www.ediamme.edc.uoc.gr/diasporanew/index.php?lang=en

[3] Since the C2 level was underrepresented, we supplemented it with texts crawled from native Greek websites, with minimal external validation of the level through http://www.greek-language.gr/certification/readability

code performing the complexity features extraction, returning a vector of feature values for each text.

Greek is an inflected language with rich morphology, allowing relative flexible word order in sentence construction. We computed a broad set of linguistic complexity features capturing lexical, morphological, and syntactic characteristics. We combined measures from Greek readability research (Giagkou, 2009; Giagkou et al., 2017; Georgatou, 2016) with features originally introduced for German (Hancke et al., 2012) and added cognitively-motivated features from psycholinguistic research on English (Liu, 2008; Gibson, 2000; Shain et al., 2016) used in research on L2 and academic language development (Weiss and Meurers, 2019a,b). Together with some traditional readability formulas, for each text we computed an overall set of 215 features[4]:

- Lexical (67 features): Lexical variation measures, including POS-specific ones, e.g., type-token ratios, noun verb ratio, verb variation, modifier token ratio; token-length lexical sophistication
- Morphological (77 features): Aspects of inflections, derivation, adjective degrees, common noun ratio, passive verb ratio; Mean size of paradigm (MSP) and Morphological feature entropy (MFE) (Çöltekin and Rama, 2018)
- Syntactic (68 fetures): Parse-tree derived measures, e.g., avg. parse tree height, dependent clause ratio; features based on POS or grammatical functions per sentence, e.g., subject-verb ratio; 16 features are cognitively-motivated based on Dependency Locality Theory (DLT) (Gibson, 2000; Shain et al., 2016) and Mean Dependency Distance (Liu, 2008)
- Readability formulas (3): FOG, SMOG and a version of Flesch Reading Ease Score for the Greek Language (Tzimokas and Matthaioudaki, 2014)

4 Classification experiments

The classification experiments were conducted using *WEKA* (Hall et al., 2009). We initially compared several machine learning algorithms, including Logistic Regression, Multilayer Perceptron and Sequential Minimal Optimization (SMO). The SMO classifier systematically outperformed other options, so we here focus on the results discussion of the SMO classifier[5]. For all our experiments, we report accuracy using 10-fold cross validation and the F-score.

To study the impact of the groups of complexity features (lexical, morphological, syntactic), we ran machine learning experiments with each individual linguistic feature subset, then with binary combinations thereof (lexical & morphological, lexical & syntactic, morphological & syntactic), with all 212 linguistic features, and finally with all 215 features.

To investigate the informativity of individual features, we ranked features using Weka's *Info Gain (IG) Attribute Evaluation* (Gnanambal et al., 2018) and ran experiments with the Top 30 features as well as with all features with an IG > 0.1.

4.1 Results for ISCED classification

Figure 1[6] sums up the performance of the three-level ISCED classifiers for the different feature groups for the three school subjects Greek, Science, and History.

The results indicate a strikingly similar pattern in History and Science corpora. For each of the three linguistic feature groups (lexical, morphological, syntactic), the classifier reaches around 75% accuracy (74.97%–76.88%), which, compared to the 33% random baseline, clearly demonstrates that the History and Science texts in different school levels systematically differ in each of these linguistic dimensions.

Interestingly, the accuracy increases about 5% when combining any two linguistic feature groups (78.68%–81.58%). This means that the features from the different linguistic domains do not encode the same information, but contribute distinct complexity differences. This is confirmed by the additional rise in accuracy to around 84%, when all three linguistic domains are combined (83.83%/84.78%).

The performance of the classifier on the Greek corpus is substantially poorer and more varied. This corpus comprises texts from the Greek Language subject, that is designed to provide both

[4] The full list of features and their computation is available at `https://osf.io/2qdzw/?view_only=1dfd5710735a449db14d2d84022c4adb`

[5] The comparison of different algorithms was conducted as part of Chatzipanagiotidis (2020)

[6] We here use line graphs in order to visually represent the results for the different feature groups of the different corpora in a compact way, also supporting transparent within- and cross-corpus comparison.

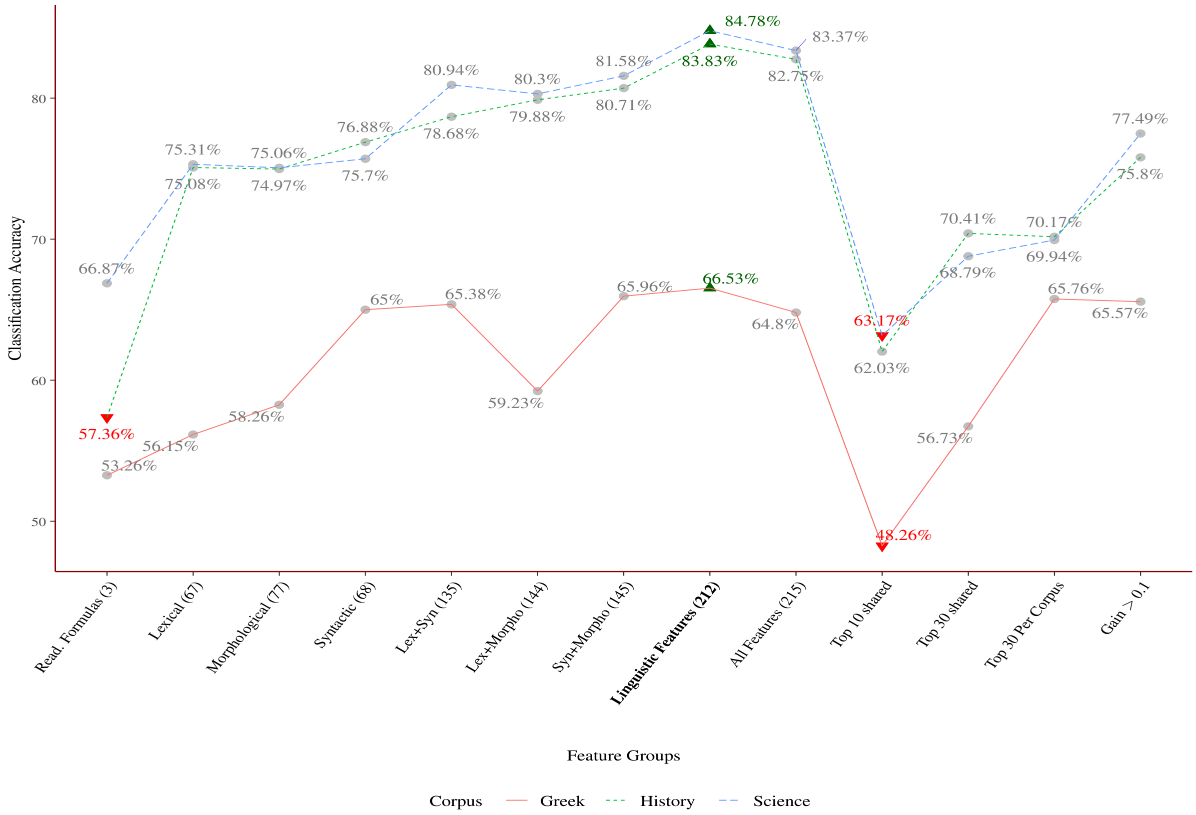

Figure 1: Ten-fold cross validation results for three-level ISCED classification

an understanding of grammar and critical reading skills, using different text types by a range of important authors. The heterogeneity of the texts could explain the low classification accuracy for the lexical features (56.15%), and the grammar component of the curriculum the slightly higher results for morphology (58.26%) and syntax (65%). While the History and Science texts seem to be systematically and coherently written in a way reflecting the incremental complexification of language in all its dimensions, that goal seems to have played much less of a role in the selection of texts for the Greek Language subject.

The classifier trained on the most informative features per corpus (*Info Gain > 0.1*) supports classification with good accuracy, though it is outperformed by any combination of two or more linguistic feature groups. When training classifiers on smaller feature groups, such as the 10 or 30 features ranked highest for the three corpora combined or the top 30 features for each corpus, we can see that the performance is systematically reduced. High classification accuracy thus is not driven by a few highly informative features, but linked to informing the classifier broadly about different dimensions of linguistic modeling. Simply adding more features is not what matters here; this is illustrated by the traditional readability formulae, which actually lower the overall results when included in the set.

Table 2 summarises the best results from the experiments discussed above. For all three corpora, the best performing classifiers are those trained on all 212 linguistic features.

Corpus	Accuracy	Weighted Avg. F-score
Greek	**66.53%**	0.660
History	**83.83%**	0.838
Science	**84.78%**	0.842

Table 2: Best results for Greek school subject corpora

Tables 3, 4, and 5 provide the confusion matrices for the three educational levels. We see that erroneous classifications mostly occur in adjacent levels, with the ISCED-2 level being the most challenging in the Greek and Science corpora.

ISCED	1	2	3	All	F-score
	predicted				
1	**313**	11	21	345	0.892
2	26	**68**	32	126	0.630
3	18	11	**282**	311	0.873

Table 5: Confusion matrix for Science corpus

ISCED	1	2	3	All	F-score
		predicted			
1	**147**	33	9	189	0.762
2	41	**82**	50	173	0.512
3	9	32	**117**	158	0.701

Table 3: Confusion matrix for Greek corpus

ISCED	1	2	3	All	F-score
		predicted			
1	**199**	10	11	220	0.907
2	15	**214**	45	274	0.782
3	5	49	**287**	341	0.839

Table 4: Confusion matrix for History corpus

4.2 Results for CEFR classification

Figure 2 shows the results of the three-level CEFR classification for the Greek as a Second Language corpus (GSL) and the Greek as a heritage language corpus (E.DIA.M.ME.) for each feature group.

The classifiers trained on the GSL corpus, with lexical, syntactic, and morphological features, as individual sets or combined, exhibit higher accuracy when compared to the ISCED classification results. Again, the combination of the 212 linguistic features resulted in the highest score (88.16%).

For the E.DIA.M.ME. corpus, we also obtain high accuracies, but a different pattern. Here the combination of lexical and syntactic features resulted in the best performing classifier (81.25%). Morphological features alone resulted in less reliable classification, and they additionally did not contribute to the classifier's performance when added to the feature set. These results indicate that for native Greeks living abroad, the focus of the material is on vocabulary and syntax, whereas morphology is less systematically introduced and complexified across levels.

The models based on a selection of the most informative features on the one hand confirm the pattern seen before, with larger feature sets performing better. But again the performance of the set of features with Info Gain > 0.1 is higher than in the previous experiments, and it reaches a noteworthy accuracy score in the E.DIA.M.ME. corpus. This finding confirms that not all dimensions of linguistic modeling are systematically complexified across CEFR classes in this corpus.

Table 6 summarises the best results for the heritage language and the second language corpora.

The confusion matrix in Table 7 illustrates where the few texts are incorrectly classified for the GSL corpus, with all but one placed in an adjacent class.

Corpus	Accuracy	Weighted Avg. F-score	Best Performing Feature Group
GSL	**88.16%**	0.861	All Linguistic Features
E.DIA.M.ME.	**81.25%**	0.803	Lexical & Syntactic Features

Table 6: Summary of highest accuracy results for Greek heritage and second language corpora

CEFR	A	B	C	All	F-score
		predicted			
A	**181**	19	0	200	0.905
B	18	**157**	25	200	0.793
C	1	20	**179**	200	0.886

Table 7: Confusion matrix for Greek as a Second Language (GSL) corpus

For the E.DIA.M.ME corpus, Table 8 shows that texts from the intermediate level B were frequently misassigned, most often to the higher level (C).

CEFR	A	B	C	All	F-score
		predicted			
A	**158**	3	6	167	0.898
B	16	**39**	28	83	0.549
C	11	17	**154**	182	0.832

Table 8: Confusion matrix for Greek as a heritage language (E.DIA.M.ME) corpus

This result may be due to the fact that the E.DIA.M.ME level alignment to CEFR B2 and C1 levels is known to be unclear (Damanakis, 2011), but also due to the smaller size of the E.DIA.M.ME level B training set in our data. The finding is also in line with previous research on the linguistic features discriminating CEFR levels in the E.DIA.M.ME corpus (Giagkou et al., 2017), according to which the linguistic features investigated capture a significant shift in Greek reading skills during the transition from C1 to C2 CEFR level, but the transition from level A to B was not as clearly reflected.

Putting the results into the context of previous research on Greek readability classification, the classifiers substantially outperform previous work. While Giagkou (2009) reported an overall accuracy of 80.59%, this was for a two-level classification task, making the results not directly comparable. A more direct comparison is possible with Georga-

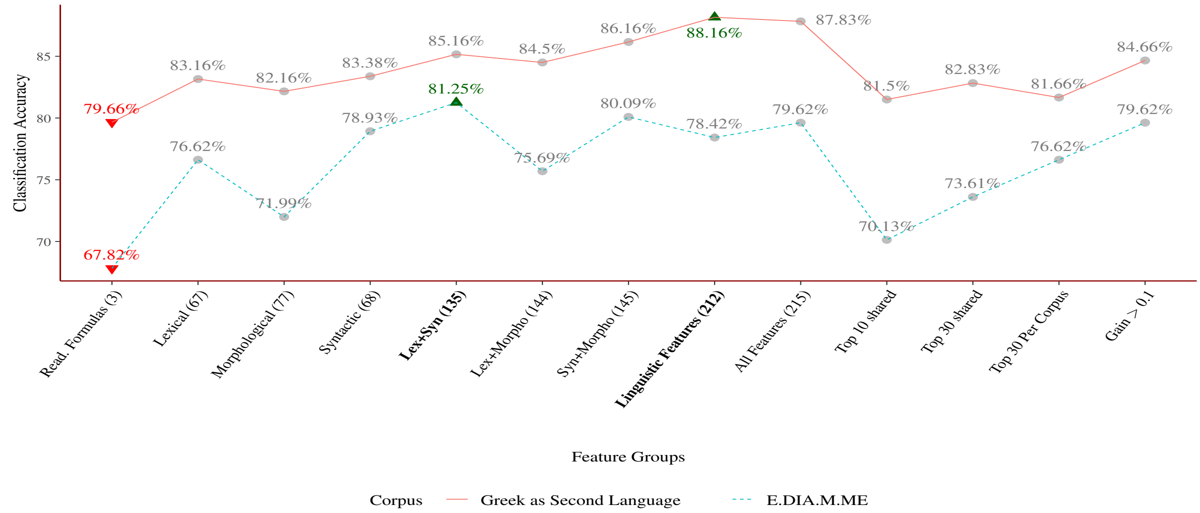

Figure 2: Ten-fold cross validation results for three-level CEFR classification

tou (2016) three-level ISCED classification results. Table 9 summarizes the accuracy of her classifiers compared to our best results in the respective corpora. For her experiments, Georgatou (2016) created a set of 153 complexity features which served as input variables to train the SMO classifier.

Corpus	Georgatou (2016)	This paper	Improv.
Greek	59.61%	66.53%	6.92
History	78.92%	83.83%	4.91
Science	76.47%	84.78%	8.31

Table 9: Comparison of results to Georgatou (2016)

Complementing this quantitative improvement of 5-8%, the empirical breadth provided by our combined and extended corpus base arguably contributes as much to advancing the state of the art – making it possible to conduct detailed cross-corpus analyses, as discussed in Section 4.4.

4.3 Analyzing the contributions of the different feature groups across corpora

It is already apparent from the findings reported so far that different feature groups and their combinations result in different classification results per corpus. To further investigate the influence of the feature groups on the classification models, we investigated the distribution of the most informative features per corpus. Figure 3 provides an overview of the number of features from the different feature groups with an Info Gain $>$ 0.1 for the three level classification for the different corpora.

The first apparent aspect in Figure 3 is the total number of informative features (Info Gain>0.1) in the five corpora. For the E.DIA.M.ME. and GSL corpora, a substantially higher number of features is informative than for the school subjects corpora (Greek, History and Science). For instance, in the GSL corpus, more than half of all features (152 out of 215) are informative. This means that the linguistic features capture and model the complexity of the Greek heritage and second language material more effectively than that of the texts written for the different subjects taught in Greek schools. This may be due to the fact that the adoption of the CEFR guidelines in the development of Greek heritage and second language learning materials highlights the need to incrementally complexify the linguistic properties of the materials as the language proficiency develops; teaching the language here is the undisputed focus. In the development of the textbooks for the different school subjects in Greece, the subject to be taught will be the main concern so that the increased complexification of language to foster mastery of complex, academic language receives less of a central role. Also for the school subject Greek, the textbooks comprise authentic texts (writings of well-known writers among them) rather than being selected or written to foster academic language development. In the absence of a respective framework and descriptors for the development of academic language, the developers of school textbooks in Greece seem to focus more on the coverage of the curriculum topics, without equally catering for the linguistic properties of the texts that would correspond to the students' academic language development. Nevertheless, remember that the classifiers for the History and Sci-

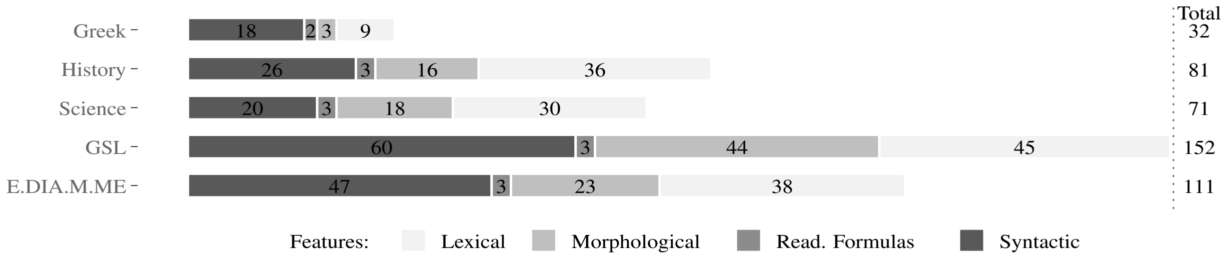

Figure 3: Feature evaluation: Features with Info Gain > 0.1 per corpus

ence corpora achieved accuracies of around 84% when considering a broad set of linguistic features. So incremental complexification of the language does already receive attention in the state-organized school books in Greece – which contrasts positively with the general lack of systematic language complexification in Geography textbooks published by commercial school book publishers in Germany as reported by Berendes et al. (2018).

A closer investigation of the History and Science corpora reveals a comparable distribution of the top informative features; lexical features proved to be more predictive with the syntactic ones following in the second place. On the other hand, only 32 features in total proved to be informative for the Greek corpus. More than half of the features are syntactic ones. These results, in addition to the classification accuracy scores in the three school subject corpora, highlight another characteristic: the domain-specific corpora (History and Science) make use of multiple dimensions of linguistic complexification. In contrast, the classification for the Greek corpus is more challenging, since only a small number of features, mainly the syntactic ones, show a significant distinctive variation throughout the school stages.

Moving to Greek as heritage or second language, the top features have a similar distribution in both corpora. A significant number of lexical, morphological and syntactic features are strong predictors, with the latter feature group being the most prominent. The prominence of syntactic features is in line with Giagkou et al. (2017), where the width and height of syntactic trees, the use of the genitive case and of adjectives, among others, were identified as successful discriminators of CEFR levels in the E.DIA.M.ME. corpus.

While the performance of readability formulae as a feature group was poor, they are among the informative features across corpora. This confirms that formulas based on simple measures of sentence and word length are an easy way to approximate complexity, though our experiments showed that they become irrelevant or even counterproductive when deeper linguistic complexity analyses are available.

4.4 Cross-corpus evaluation

To investigate the generalizability of the trained models, we conducted cross-corpus analyses, by systematically testing the models trained on one corpus on the other corpora. Table 10 summarises the results for classifiers trained on the different feature subsets. To facilitate comparison, the Table also includes the 10-fold cross-validation within-corpus results discussed in the previous sections.

Compared to the within-corpus analyses, the accuracies for cross-corpus evaluation are substantially reduced, though all still perform considerably above the random baseline of 33%. The classifiers trained on the heritage (E.DIA.ME.E.) or second language corpora (GSL) still exhibit acceptable accuracies of over 60%, when either is tested against the other. For the three school subject corpora (Greek, History, Science), the cross-corpus results drop below 60% (with the History classifier trained with Lexical and Syntactic features faring slightly better when evaluated on the Science corpus). Accuracy is higher when training and testing happens across the domain-specific curriculum subjects (History and Science). For interpreting the results, note that these two corpora share the same text genre, with both consisting of informational texts. Together with the lower performance for the Greek corpus, which belongs to a different genre, consisting mostly of literary texts, our investigation confirms the importance of genre effects in readability estimation (Sheehan et al., 2008). The classifiers generalize well across corpora of the same genre (informational texts on History and Science), but the Greek corpus classifier trained mostly on literary texts fails to generalize to informational texts.

Corpus		Feature set								
Models from:	Evaluated at:	Lex	Morpho	Syn	Lex & Syn	Lex& Morpho	Syn& Morpho	Ling. Feat	All Feat	Info Gain > 0.1
Greek	10-Fold	56.15	58.26	65.00	65.38	59.23	65.96	**66.53**	64.80	65.57
	History	56.52	51.97	55.44	**58.56**	51.73	52.93	55.20	55.80	56.64
	Science	38.10	**51.40**	40.15	36.18	47.44	40.28	40.28	40.66	36.82
History	10-Fold	75.08	74.97	76.88	78.68	79.88	80.71	**83.83**	82.75	75.80
	Greek	45.96	48.46	**54.03**	49.42	49.42	51.15	51.53	50.96	50.96
	Science	60.99	43.35	**61.89**	58.05	49.48	43.73	41.04	41.30	49.61
Science	10-Fold	75.31	75.06	75.70	80.94	80.30	81.58	**84.78**	83.37	77.49
	Greek	45.57	46.92	42.88	43.84	44.42	46.73	46.69	**48.07**	45.76
	History	54.85	42.63	**57.96**	55.32	54.37	46.22	51.61	51.73	50.17
GSL	10-Fold	83.16	82.16	83.38	85.16	84.5	86.16	**88.16**	87.83	84.66
	E.DIA.M.ME.	**64.12**	59.49	59.02	61.11	58.10	59.25	61.11	59.72	59.02
E.DIA.M.ME.	10-Fold	76.62	71.99	78.93	**81.25**	75.69	80.09	78.42	79.62	79.62
	GSL	**63.50**	56.66	60.16	61.00	57.33	59.83	59.99	59.00	60.83

Table 10: Cross-corpus analysis

The lower performance of the models based on the Greek corpus may be due to the smaller corpus size compared to the History and Science corpora.

Interestingly, while the full set of 212 linguistic features outperformed all other feature combinations in almost all of the within-corpus experiments, this is not the case in the cross-corpus experiments. In most cross-corpus tests the classifiers trained with single feature groups achieved the best results. This is consistent with our exploration of the most informative features per corpus (Section 4.3), where we found that different numbers of features from different linguistic sets were found to be informative in a given corpus.

Comparing the best accuracy scores achieved in within-corpus testing with the respective scores from the cross-corpus evaluation, an interesting pattern emerges: the Greek classifier, which was the lowest performer in within-corpus evaluation (66.53%) – presumably because it is trained on the most heterogeneous collection of texts, not originally written as teaching material – shows the lowest performance drop when applied to other corpora. So the classifier in cross-corpus testing seems to benefit from the wider domain and range of text types in its training material that make the within-corpus evaluation a harder task. There is thus a trade-off: A less specialized classifier, trained on a general corpus instead of a domain-specific one, is preferable in exhibiting a more stable behaviour when generalizing to other domains. At the same time, a generic readability modeling solution, pertinent to various text types and domains, then cannot be as finely attuned to domain- and text type-specific readability assessment.

5 Conclusions

In this paper, we addressed the issue of Greek readability classification based on a broad range of textbooks, from History and Science via Greek as a school subject to Greek as a heritage language and as a second language. We analyzed these textbook corpora covering all levels of school or language proficiency in terms of a set of complexity features covering three domains of linguistic modeling. Given that Greek is a highly inflected language with relatively free word order, particular emphasis was placed on the coverage of morphological and syntactic characteristics. While we designed the feature set to advance the state of the art for Greek readability research, we hope that it also encourages the development of such rich linguistic complexity feature sets for the broadening set of languages being investigated under related research perspectives.

In a series of three-class classification experiments, our feature set supported accuracy scores of up to 88%, outperforming the previous results for Greek readability classification. The lower accuracy obtained for the Greek as school subject corpus can be explained by its more heterogeneous nature as a collection of original texts and its smaller size. It probably also indicates that complexification of Greek as a school and academic language so far lacks a systematic framework and descriptors. In comparison, the adoption of the CEFR guidelines in the development of Greek as a Second Language learning materials arguably contributed to the very high accuracies achieved by our approach for the GSL corpus.

When comparing the classifiers trained on single feature groups to classifiers trained on combinations of feature sets, we observed that combinations of different linguistic feature domains systematically improved the results. In almost all within-corpus experiments, the results indicate that the richer the linguistic information, the better the classification performance.

The cross-corpus evaluation testing the generalizability of the classification models reduced classification accuracy. While the linguistic complexity features are capable of capturing complexification of language in general, the weighting of the evidence for readability classification in a strict sense thus is domain specific. Interestingly, the most successful model in terms of generalizability was the one with the lowest performance in the within-corpus validation results, namely the classifier trained on the heterogeneous Greek as a school subject corpus. While models trained on data covering a wide range of text types and domains generalize better, their top performance is lower than that of domain-specific models.

To foster further research on Greek readability classification, the four non-commercial textbook corpora we used for the reported research will become freely accessible through CLARIN.EL.

6 Acknowledgments

The authors would like to thank the Institute for Language and Speech Processing (ILSP) for its support and for providing access to the ILSP NLP pipeline for Greek. Special thanks to Spyridoula Georgatou for sharing three corpora for the purpose of this study. We thank the anonymous reviewers for their insightful comments and suggestions.

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. 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*, 110(4):518–543.

CCSSO. 2010. Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C. `http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf`.

Savvas Chatzipanagiotidis. 2020. Combining machine learning techniques with broad linguistic analysis of readability in Greek. Master's thesis, international studies in computational linguistics, Eberhard Karls Universität Tübingen.

Çağrı Çöltekin and Taraka Rama. 2018. Exploiting universal dependencies treebanks for measuring morphosyntactic complexity. In *Proceedings of First Workshop on Measuring Language Complexity*, pages 1–7.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.

Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press. Council for Cultural Co-operation. Education Committee. Modern Languages Division.

Michail Damanakis. 2011. Assessment of Greek-language education abroad by 2010 and its prospects [in Greek]. `http://ediamme.edc.uoc.gr/diaspora/index.php?id=248:apotimisi-tis-mexri-to-2010`.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of the International Journal of Applied Linguistics*, 165(2):163–193.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

Spyridoula Georgatou. 2016. Approaching readability features in Greek school books. Master's thesis, international studies in computational linguistics, Eberhard Karls Universität Tübingen.

Maria Giagkou. 2009. *Corpora and language education: exploitation potentials in teaching Greek and construction of pedagogically relevant corpora* [in Greek]. Ph.d thesis, National and Kapodistrian University of Athens, Department of Philosophy, Athens.

Maria Giagkou, Giorgos Fragkakis, Dimitrios Pappas, and Harris Papageorgiou. 2017. Feature extraction and analysis in Greek L2 texts in view of automatic labeling for proficiency levels. In *Proceedings of the ICGL12*, volume 1, pages 357–368. Romiosini/CeMoG, Freie Universitat Berlin.

Maria Giagkou, Vicky Kantzou, Spyridoula Stamouli, and Maria Tzevelekou. 2015. Discriminating CEFR levels in Greek L2: a corpus-based study of young learners' written narratives. *Bergen Language and Linguistics Studies*, 6.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

S Gnanambal, M Thangaraj, VT Meenatchi, and V Gayathri. 2018. Classification algorithms with attribute selection: an evaluation study using weka. *International Journal of Advanced Networking and Applications*, 9(6):3640–3644.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

J. Nelson, C. Perfetti, D. Liben, and M. Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, et al. 2019. Comparing and developing tools to measure the readability of domain-specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4833–4844.

Cory Shain, Marten Van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 49–58.

Kathleen M Sheehan, Irene Kostin, and Yoko Futagi. 2008. When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington DC*. Citeseer.

Dimitrios Tzimokas and Marina Matthaioudaki. 2014. Deiktes anagnosimotitas: zitimata efarmogis kai axiopistias [in Greek]. In *Major Trends in Theoretical and Applied Linguistics 3*, pages 367–384. De Gruyter Open Poland.

Sowmya Vajjala and Ivana 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*, pages 297–304. Association for Computational Linguistics.

Sowmya Vajjala and Detmar 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*, pages 163–173. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019a. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393, Florence, Italy. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019b. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.