What neural networks know about linguistic complexity

Serge Sharoff

University of Leeds

Abstract

Linguistic complexity is a complex phenomenon, as it manifests itself on different levels (complexity of texts to sentences to words to subword units), through different features (genres to syntax to semantics), and also via different tasks (language learning, translation training, specific needs of other kinds of audiences). Finally, the results of complexity analysis will differ for different languages, because of their typological properties, the cultural traditions associated with specific genres in these languages or just because of the properties of individual datasets used for analysis. This paper investigates these aspects of linguistic complexity through using artificial neural networks for predicting complexity and explaining the predictions. Neural networks optimise millions of parameters to produce empirically efficient prediction models while operating as a black box without determining which linguistic factors lead to a specific prediction. This paper shows how to link neural predictions of text difficulty to detectable properties of linguistic data, for example, to the frequency of conjunctions, discourse particles or subordinate clauses. The specific study concerns neural difficulty prediction models which have been trained to differentiate easier and more complex texts in different genres in English and Russian and have been probed for the linguistic properties which correlate with predictions. For example, this study shows how the rate of nouns and the related complexity of noun phrases affects difficulty via statistical estimates of what the neural model predicts as easy and difficult texts. The study also analysed the interplay between difficulty and genres, as linguistic features often specialise for genres rather than for inherent difficulty, so that some associations between the features and difficulty are caused by differences in the relevant genres.
Keywords: Automatic text classification, Deep learning, Interpreting neural networks
To cite as Sharoff, Serge, 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics* 26(2), 370–389. https://doi.org/10.22363/2687-0088-30178

1 Introduction

Linguistic complexity is a complex phenomenon, as it manifests itself on different levels, through different features, and via different application tasks. In terms of levels of complexity analysis, it is natural to to analyse complexity on the level of words, as some words are naturally more difficult than others, and this allows a way of ranking them as often done in Complex Word Identification (CWI) tasks. A different set of categories is needed to analyse complexity of sentences, which primarily depends on the networks of syntactic and semantic relations between words. Yet another level of complexity analysis concerns difficulty with respect to global text properties, which is primarily about capturing the flow of argumentation: even when individual sentences are easy to understand, the links between might require greater cognitive load.

Another aspect of complexity analysis concerns the features we use in our description of complexity. For words we can refer to their frequencies or their semantic features, such as abstractness. For morphosyntactic features we can refer to the part-of-speech categories or to the dependency relations. For text-level analysis we can use rhetorical relations as well as a typology of genres. In any case, each level of analysis

(words, sentences or texts) is described computationally by a vector of such features with a fixed number of dimensions.

There is also a multitude of reasons why we are interested in the phenomenon of complexity. This determines what is considered as simple or complex in each case. A typical example of applications of complexity analysis concerns language learning, which assumes the audience of non-native speakers acquiring a foreign language either as a child or an adult. In this kind of application, we can specialise our analysis for specific language teaching tasks, as some phenomena are less likely to cause problems in understanding, but more problems in production, or we can specialise for the target audience, as different phenomena are likely to cause problems depending on the native language of the learners. Another example of applications concerns translation training, which is different from language learning, as the challenge for a trainee translator often consists in transforming various aspects of the source texts into their native language. A related case concerns analysis of complexity in the context of language acquisition for children learning their native language. Yet another example concerns specific needs of other kinds of audiences, such as production of texts for native speakers with various mental disabilities.

Finally, the results of complexity analysis will differ for different languages, because of their typological properties (such as greater complexity of syntactic relations between words vs greater morphological complexity of word forms), the cultural traditions associated with specific genres in these languages, for example, emphasis on plain language in research papers in English vs traditionally accepted forms of academic discourse in Russian. It is also important to understand the properties of individual datasets used for analysis, as occasional confounding variables for the dataset might affect the replicability of the findings, for example, a limited range of genres or authors.

This paper investigates some of these aspects by focusing on word- and sentence-level analysis while also investigating the impact of genres. In terms of the task, the focus is on studying difficulty for adult learners for two languages, English and Russian without specification of their native language and with the specific focus on the language understanding task.

In terms of the computational methodology, the study uses artificial neural networks for predicting complexity. The specific study concerns neural difficulty prediction models which have been trained to differentiate easier and complex texts in different genres in English and Russian. While neural networks produce empirically efficient prediction models by optimising millions of parameters, they operate as a black box without determining which linguistic factors lead to a specific prediction. Following the Bertology framework (Rogers et al., 2020), this paper shows how to link neural predictions of text difficulty to detectable properties of linguistic data, for example, to the frequency of conjunctions, discourse particles or subordinate clauses. More specifically, the linguistic features are primarily based on Douglas Biber's Multidimensional Analysis (Biber, 1995), such as, the rate of *that* deletion or public verbs, to explain predictions of fine-tuned transformer models, such as XLM-Roberta (Conneau et al., 2020).

2 Methodology

The study presented in this paper focuses on the fine-grained difficulty assessment, when difficulty analysis is transformed from the text level to the sentence level. The focus of this study is on a specific task, namely, prediction of complexity with respect to teaching foreign languages, more specifically, automatic assessment of reading exercises from language learning textbooks. What varies in this study is a set of properties, namely the influence of genres, syntax and lexical semantics on the predictions.

2.1 Classification methods

From the computational viewpoint, the difficulty prediction problem can be defined as a short-text classification task, which predicts a difficulty label for a short text or a segment. Since difficulty naturally operates on a scale (some texts are considered as more difficult than others), this problem can be also defined as a regression task, which predicts a numeric difficulty value for a text. This study focuses on the classification task, because many statistical operations need categorical labels, and because the original annotated corpora use a small fixed number of levels. While there is a range of methods for the short-text classification task, recent studies favoured fine-tuning pre-trained transformer models. The pre-training of neural networks aims at establishing their weights by the task of predicting missing words on large corpora, for example, Wikipedias in the case of BERT (Devlin et al., 2018) or Common Crawl in the case of XLM-Roberta transformer model (Conneau et al., 2020). In the end the pre-trained representations can be shown to reflect general linguistic phenomena, such as agreement or semantic classes (Rogers et al., 2020). Fine-tuning on a target task (difficulty prediction in this case) adapts the weights of the pre-trained representations, so that the general phenomena can be linked to the target task.

In addition to building the difficulty prediction classifiers, other text parameters can be tested. More specifically, this study applied existing neural classifiers for genres to both training and testing corpora using a well-tested automatic genre annotation model (Sharoff, 2021). This allows comparing properties of texts of the same difficulty but in different genres, as well as comparing texts in the same genre, but of different difficulty levels.

2.2 Human interpretation of neural predictions

Neural networks produce empirically efficient prediction models, especially the modern setup which is based on fine-tuning pre-trained transformer models, such as BERT. However, they act as a blackbox, as it is difficult to determine why a model with a given set of training parameters produced a specific prediction. Therefore, the NLP field recently has started developing a range of approaches under the name of Bertology to understand reasons for predictions (Rogers et al., 2020).

Bertology analysis of prediction difficulty as developed in this study extends the framework from (Sharoff, 2021), which uses Logistic Regression (LR) to detect the linguistic features associated with (more accurate) predictions of a neural model. LR is a fast and transparent Machine Learning method, which is defined as:

$$ln\frac{p}{1-p} = w_0 + w_1x_1 + ... + w_nx_n$$

It fits a linear model to predict the log-odds ratio, where p is the probability of a text having a particular label, for example, Easy or Difficult, x_i are interpretable variables, e.g., the proportion of verbs or conjunctions. Since the model is linear, the relative contribution of each feature can be determined through its weight w_i for detecting this function. To assist in comparing the weights, the variables have been standardised with respect to their values and dispersion prior to fitting the logistic regression, so that for each feature its mean is zero and its standard deviation is one. In the end, the feature weights can be directly compared. Another advantage of logistic regression over other machine learning methods is that it has been well investigated from the statistical viewpoint, thus allowing a number of tests to determine the significance of each feature. One of the approaches for testing the feature significance is based on the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under a model with one of the features removed (Hosmer Jr et al., 2013). If the behaviour of the logistic regression model changes significantly when a feature is removed, the feature can be considered as more significant for this label. The tables below show the weights of features selected under the likelihood ratio test.

The linguistic features used in this study are based on the set introduced by Douglas Biber for describing register variation via Multi-Dimensional Analysis (Biber, 1988). The features include the following categories:

Lexical features such as:

- public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny…*
- time adverbials = *afterwards, again, earlier, early, eventually, formerly, immediately,…*
- amplifiers = *absolutely, altogether, completely, enormously, entirely,…*

Part-of-speech (POS) features such as:

- nominalisations
- prepositions
- past tense verbs

Syntactic features such as:

- *be* as the main verb
- *that* deletions
- pied piping

Text-level features such as:

- average word length
- average sentence length
- type/token ratio (TTR)

This set was designed specifically for English. However, some of its features are nearly universal, for example, this concerns the text-level features, even though their exact values are language-dependent. Many lexical features are comparable across languages if they can be translated reliably, for example, public verbs. Many part-of-speech features can be used across a number of languages as well, for example, nominalisations, while many syntactic features are comparable only across a smaller set of closely related languages, for example, pied piping. Some functionally equivalent features are included into the list for Russian even when they are expressed in a different way in Russian. For example, `F18` (BYpassives according to (Biber, 1988)) is expressed via passives with the agent in the instrumental case, but for consistency this feature still keeps the same name as in English. Similarly, detecting `C12` (*do* as pro-verb in English) is based in Russian on detecting ellipsis in conditions similar to those used for detecting `C12` in English. See the list in Appendix 1 for the full description of the features. Even though the set of features was introduced to describe register variation, it is sufficiently general to provide explanations for the difficulty levels.

Table 1: CEFR-annotated datasets for English and Russian

Level	English		Russian	
	Texts	Segments	Texts	Segments
A1	0	0	178	1149
A2/KET	64	304	121	1707
B1/PET	60	516	134	2109
B2/FCE	71	1354	167	4022
C1/CAE	67	1606	120	1937
C2/CPE	69	1540	6	121

Table 2: Accuracy of XLM-Roberta for English and Russian

	English			Russian		
	Precision	Recall	F1-score	Precision	Recall	F1-score
A1				0.72	0.75	0.74
A2	0.75	0.84	0.79	0.51	0.64	0.57
B1	0.58	0.66	0.62	0.50	0.66	0.57
B2	0.53	0.74	0.62	0.71	0.59	0.65
C1	0.54	0.53	0.53	0.58	0.47	0.52
C2	0.77	0.49	0.59	0.00	0.00	0.00
macro avg	0.70	0.62	0.63	0.50	0.52	0.51
accuracy	0.60			0.60		
Binary case						
Easy	0.89	0.98	0.93	0.90	0.98	0.94
Difficult	0.99	0.97	0.98	0.92	0.65	0.76
macro avg	0.94	0.97	0.96	0.91	0.82	0.85
accuracy	0.97			0.91		

2.3 Datasets

The training datasets came from the Cambridge Readability Dataset (Xia et al., 2016) for English and from the Rufola corpus (Laposhina et al., 2018) for Russian. In both cases, the source texts have been taken from existing textbooks marked with the CEFR levels by the developers of the respective corpora, namely, the Cambridge Proficiency Tests have been mapped to the CEFR levels for English, while the levels of several textbooks have been unified into the CEFR scheme for Russian. In both cases, the corpora are annotated by the CEFR levels on the text level, which means that a text corresponds to a single reading exercise. Since the amount of data on the text level does not provide enough training samples for building reliable classifiers, each text in the respective datasets was split to smaller segments with the aim of training within a window of several sentences. The optimal window size was determined to be of three sentences (this window was expanded if the total length of three adjacent sentences was less than 15 words). The distribution of training data on the document level vs the chosen window level is given in Table 1.

Large-scale testing of the linguistic properties has been conducted with raw text corpora from the English and Russian portions of the Aranea family (Benko, 2016), which were obtained by Web crawling and post-processing of websites in the respective languages. These corpora offer a reliable snapshot of how English and Russian are used in Web pages. In addition, the Nauka-Plus portion of the Taiga corpus (Shavrina and Shapovalova, 2017) was used for testing in Russian, since it has been also annotated with difficulty levels, though the focus of its annotation was on assessing its difficulty for the native speakers of Russian. The reason for using Nauka-Plus in this study is to compare the automatic difficulty predictions aimed at the non-native speakers with the verified difficulty estimates for the native speakers.

The classifiers for difficulty were built by fine-tuning the XLM-Roberta transformer model (Conneau et al., 2020) from the HuggingFace library (Wolf et al., 2019) using the CUP and Rufola training sets respectively for English and Russian. Another set of classifiers for probing the neural predictions was built using the Multi-Dimensional Analysis features and the Logistic Regression model, see Section 2.2 below. Table 2 lists the cross-validation accuracy scores after fine-tuning on the respective training corpora. The overall accuracy of both models is 60%, but the Russian model is trailing behind with respect to the F1 score. Since C2 is a minority class for Russian (see Table 1), this class is not detected in cross-validation (its texts are all

Table 3: Confusion matrices

	A2	B1	B2	C1	C2
A2	**256**	37	10	0	1
B1	40	**343**	118	12	3
B2	18	129	**1001**	175	31
C1	4	60	505	**845**	192
C2	6	18	238	531	**747**

Table 4: Association of features with difficulty for English

DIFFICULT		EASY	
A01.pastVerbs	0.299	C07.2persProns	0.341
J43.TTR	0.229	K45.conjuncts	0.271
P67.analNegn	0.205	I39.preposn	0.206
E14.nominalizations	0.133	B04.placeAdverbials	0.160
C06.1persProns	-0.116	L54.predicModals	0.134
G19.beAsMain	-0.120	G19.beAsMain	0.120
L54.predicModals	-0.134	C06.1persProns	0.116
B04.placeAdverbials	-0.160	E14.nominalizations	-0.133
I39.preposn	-0.206	P67.analNegn	-0.205
K45.conjuncts	-0.271	J43.TTR	-0.228
C07.2persProns	-0.341	A01.pastVerbs	-0.300

classified as C1), thus bringing the macro-average F1 score down. Overall, more difficult texts (C1 and C2) are not very common in the Russian training set, which makes the task of their detection more challenging in comparison to English. Nevertheless, in the binary scenario of distinguishing between Easy (A1, A2, B1) and Difficult (C1 and C2) texts the accuracy reaches 91% for Russian and 97% for English, which is sufficient for our purposes.

3 Results

To simplify presentation of the results, this study provides the contrast of Easy vs Difficult texts, i.e., those predicted at the lowest three levels (A1, A2 and B1) vs those at the top two level (C1 and C2) with the B2 level reserved as a boundary, since the errors of the classifiers overlap over this boundary. The reason for extending the scale of Easy texts to B1 comes from the lack of data for Web pages detected as suitable for A1 and A2 levels (the total number of such pages is less than 1% for either language), so what is presented as Easy in the analysis below comes mostly from pages classified as suitable for the B1 level.

Tables 4 and 5 list associations of the positive and negative weights of the most significant features with respect to the predicted difficulty levels. Some features work in the same way in both languages. For example, the rate of the first and second person pronouns has the strongest positive association with easy texts and the strongest negative association with difficult texts. These pronouns indicate personal interaction, which is often expressed in interactive spoken-like texts, even though the classifiers were applied to written language in HTML Web pages. The rate of first and second person pronouns is likely to be higher in discourse about areas of "immediate relevance" as expected for the A-level CEFR texts (Council of Europe, 2001). Similarly, the greater rate of nominalisations and negations is consistently associated with difficult

text across both languages. This quantitative evidence supports other linguistic studies concerning the extra complexity involved in processing negations in comparison to positive sentences (Doughty and Long, 2008). Similarly, nominalisations and complex noun phrases have been linked to the conceptual difficulty of grammatical metaphors when actions, which are congruently expressed by verbs, get packed into noun phrases, for example, from *how glass cracks* into *the glass crack growth rate* (Halliday, 1992).

Some difficulty indicators are language-specific. Often they can be linked to prominent language-specific constructions. In particular, `G19.beAsMain` is associated with easy texts for English, as this construction offers a simple formulaic expression for relational predicates (*X is Y*), while other relational predicates, for example, *X involves Y*, are more likely to be found in more advanced writing. The same feature does not appear prominently in easier Russian texts, as the Russian equivalent of *to be* is not overtly expressed in the present tense and therefore it is not counted by the feature extraction mechanism.

It is interesting to note that the feature `I39.preposn` is associated with different directions of complexity in English and Russian. For English its greater rate indicates easier texts, while for Russian this is associated with more difficult ones. This can be explained from the typological differences between the two languages: what is expressed by the basic prepositions in English (*of, to* or *with*) is often rendered by the case endings in Russian (respectively, genitive, dative or instrumental). Therefore, more active use of the prepositions in Russian correlates with more complex writing styles, when sentences need to include more information than the basic Subject-Verb-Object skeleton which introduces the main participants. At the same, more accessible writing styles in English need to use prepositions at a high rate, while this rate is reduced in more complex styles because of the more active use of other features, such as negations or noun compounds.

The adverbials as a syntactic function appear in Tables 4 and 5 in three different forms, as adverbs, which are detected as a POS category, and as either time adverbials or place adverbials, which are detected via lexical lists, for example, *behind* or *South*. Therefore, the rates of adverbials of different kinds affect difficulty in different ways. General adverbs tend to occur as modifiers of adjectives and verbs, thus leading to more elaborated constructions associated with more complex styles. At the same time, time and place adverbials often occur in narratives, hence they are less likely to be associated with complex styles.

Some features do not offer an easy cross-lingual explanation, such as the greater rate of conjuncts in easier English texts or the greater rate of conditionals in more difficult Russian texts. Also, quite surprisingly, word length has a positive correlation with easier Web pages in Russian and has not been detected as a significant factor associated with difficulty in English.

Table 5: Association of features with difficulty for Russian

DIFFICULT		EASY	
A03.presVerbs	0.294	C07.2persProns	0.340
I42.ADV	0.292	J44.wordLength	0.332
E14.nominalizations	0.289	D13.whQuestions	0.024
I39.preposn	0.208	C08.3persProns	-0.077
P67.analNegn	0.207	C09.impersProns	-0.078
H37.conditional	0.098	H37.conditional	-0.132
H38.otherSubord	0.094	I39.preposn	-0.216
B05.timeAdverbials	0.094	A01.pastVerbs	-0.239
C09.impersProns	0.086	I42.ADV	-0.341
C06.1persProns	-0.205	P67.analNegn	-0.381
C07.2persProns	-0.242	A03.presVerbs	-0.390

Table 6: Association of features with difficulty for Nauka-Plus

DIFFICULT		EASY	
C10.demonstrProns	0.542	N60.thatDeletion	0.461
C08.3persProns	0.406	J43.TTR	0.431
I40.attrAdj	0.375	I39.preposn	0.184
E14.nominalizations	0.343	B05.timeAdverbials	0.162
I42.ADV	0.298	D13.whQuestions	-0.010
A03.presVerbs	0.247	H38.otherSubord	-0.041
C12.doAsProVerb	-0.137	A03.presVerbs	-0.113
P67.analNegn	-0.154	K48.amplifiers	-0.120
K45.conjuncts	-0.178	E14.nominalizations	-0.300
I39.preposn	-0.185	C08.3persProns	-0.341
B05.timeAdverbials	-0.381	I40.attrAdj	-0.348
J43.TTR	-0.397	C10.demonstrProns	-0.392

There is an apparent problem in interpreting the results of the Type-Token Ratio (TTR) score as reported in Table 6 for Nauka-Plus texts against the results reported in Table 4. The TTR rate (`J43`) in Table 4 is in line with previous studies, such as (Collins-Thompson and Callan, 2004), when the higher TTR is associated with greater lexical diversity and hence with more difficult texts. At the same time, Table 6 for Nauka-Plus associates TTR with easier texts. It seems that the answer to this discrepancy comes from differences in the corpus composition in terms of topics, genres or other text properties. In this specific case, news reporting is the most common genre category in the Nauka Plus dataset (57%) with the second most common category being academic writing (30%), Table 9. As features vary across genres, the TTR is often considerably higher in news reporting as news reporting often includes many person names and locations, thus increasing their TTR without necessarily increasing their perceived difficulty. This can be illustrated by variation of the TTR across the genre categories in this dataset. For example, the Inter-Quartile Range (IQR) of TTR on the Nauka-Plus corpus is 0.5727 to 0.6727, while texts in the top quartile of the TTR values (i.e., above 0.6727) contain a higher proportion of news reporting (72%) vs academic writing (19%) in comparison to the entire corpus (57% vs 40%). Even relatively infrequent named entities do not necessarily contribute to the greater difficulty of their texts, for example, *Британское подразделение американской компании Локхид Мартин провело испытания модернизированной боевой машины пехоты Warrior* ('The British office of Lockheed Martin tested an upgraded version of their armoured carrier Warrior'). Another indicator of easy texts for Nauka Plus happens to be the higher rate of prepositions and time adverbials, which are also more typical for news reporting. This is another evidence for the importance of genres to determining the difficulty features, as the preposition rate (`I39`) is also contrary to the observations from the general Web pages in Russian, which associate the higher rate of prepositions with more difficult texts.

Those Nauka Plus texts which are closer to academic writing contain explications, which are treated as more difficult according to the annotators. From the viewpoint of their linguistic features, they contain more verbs in the present tense and more attributive adjectives, while they tend to repeat relevant terms, thus leading to lower TTR, for example, *Burkholderia одновременно является патогенным паразитическим микроорганизмом , изменяющим геном амеб…* ('At the same time Burkholderia is a pathogenic parasitic microorganism, which alters the amoeba genome…') with words *Burkholderia, amoeba, genome, microorganism, pathogenic* repeated throughout the article.

The close link between difficulty and genres observed in the Nauka-Plus corpus calls for experiments

Table 7: Association of difficulty with communicative functions for English

Difficult	#Texts	Functions	Easy	#Texts	Functions
23.15%	945958	A12.promotion	**35.93%**	195245	**A12.promotion**
17.50%	715187	**A16.information**	**17.85%**	97005	**A7.instruction**
16.97%	693702	**A1.argumentation**	**15.80%**	85831	**A8.newswire**
12.08%	493616	A8.newswire	9.44%	51302	A16.information
9.40%	384344	A7.instruction	**7.37%**	40024	**A11.personal**
6.56%	268242	A11.personal	7.16%	38898	A1.argumentation
5.10%	208218	**A17.reviewing**	4.30%	23372	A17.reviewing
4.26%	174118	**A14.academic**	1.88%	10193	A9.legal
3.88%	158695	**A9.legal**	0.21%	1136	A4.fiction
1.09%	44571	**A4.fiction**	0.06%	349	A14.academic

comparing predictions for these categories. Tables 7 and 8 present the association between genres (expressed in terms of generic communicative functions) and difficulty levels in the Aranea corpora for English and Russian. The tables highlight the cases when the proportion of genres predicted as Difficult or Easy is **higher** than for the opposite case. For example, the proportion of texts with the predicted function of A7.instruction is higher for Easy texts in English (17.85% vs 9.4% for Difficult texts in Table 7). Overall, the classifiers predict a greater proportion of promotional, news reporting, instructional and personal reporting texts as Easy across both languages. This matches the intuition of the language teachers who tend include such texts in exercises. The Fiction category is an exception to this intuition as it is often treated as a prime example of texts useful for language learners with many exercises based on examples from novels. At the same time, this study finds that typical authentic examples of fiction (at least as found on the Web) are predicted as less suitable for the learners.

Despite the different aims of the human annotation of difficulty available in the Nauka-Plus corpus (aimed at the native Russian speakers) and the automatic difficulty predictions in terms of CEFR levels, the difficulty levels are well aligned, see Table 10. The most difficult texts according to the human annotation in Nauka-Plus receive the highest CEFR level predictions and vice versa, while the automatic classifier avoids making C2 and A-level predictions.

A7.instruction and A8.news are among the communicative functions which are common in both Easy

Table 8: Association of difficulty with communicative functions for Russian

Difficult	#Texts	Functions	Easy	#Texts	Functions
19.12%	212072	**A1.argumentation**	**29.28%**	251923	**A12.promotion**
15.37%	170401	**A7.instruction**	**19.68%**	169320	**A8.newswire**
15.34%	170121	A12.promotion	12.35%	106272	A16.information
14.64%	162356	A8.newswire	11.77%	101265	A7.instruction
13.26%	147047	**A16.information**	9.08%	78111	A1.argumentation
7.79%	86435	**A11.personal**	6.07%	52224	A11.personal
6.01%	66696	**A17.reviewing**	5.36%	46098	A17.reviewing
4.07%	45123	**A14.academic**	**3.92%**	33734	**A9.legal**
3.18%	35264	A9.legal	1.91%	16460	A14.academic
1.21%	13396	**A4.fiction**	0.56%	4843	A4.fiction

Table 9: Distribution of genres in Nauka-Plus

4463	A8.newswire
2295	A14.academic
319	A12.promotion
29	A12.promotion/A8.newswire
20	A8.newswire/A14.academic
16	A1.argumentation
16	A8.newswire/A12.promotion
13	A14.academic/A18.newswire
9	A7.instruction

Table 10: Human annotations for difficulty Nauka-Plus vs predicted CEFR levels

NP1:	Human	CEFR
1325	**L4**	**C1**
972	**L1**	**B1**
899	L3	C1
871	L2	B1
837	L2	C1

Table 11: Positive and negative features for *easy* instructional and news texts

A7.instructional		A8.news	
C07.2persProns	0.5155	K55.publicVerbs	0.2913
C06.1persProns	0.1791	H35.causative	0.2666
B04.placeAdverbials	0.1702	H38.otherSubord	0.2214
I39.preposn	0.1603	N59.contractions	0.2192
L54.predicModals	0.1371	K47.generalHedges	0.2129
N60.thatDeletion	0.1341	D13.whQuestions	0.1841
B05.timeAdverbials	0.1028	A01.pastVerbs	0.1756
L53.necessModals	0.0638	C09.impersProns	0.1525
H35.causative	-0.0784	C08.3persProns	0.0521
K56.privateVerbs	-0.0902	F18.BYpassives	-0.1857
H25.presPartClaus	-0.0984	K48.amplifiers	-0.1864
E14.nominalizations	-0.1146	K50.discoursePart	-0.2290
I42.ADV	-0.1366	L54.predicModals	-0.2427
C09.impersProns	-0.1612	E16.Nouns	-0.2705
A03.presVerbs	-0.1678	K45.conjuncts	-0.3521
E16.Nouns	-0.2482	C07.2persProns	-0.4385

and Difficult parts of Aranea. Table 11 lists the linguistic features which are specific to easy texts **within** these genres. Some features resemble what is characteristic for Easy texts in English in general, such the use of the first and second personal pronouns, as well as the prepositions and time and place adverbials for instructions. As expected, the use of nouns, nominalisations, adverbs as modifiers, as well as more complex syntactic constructions in the form of subordinate clauses of different kinds is associated with more difficult texts. At the same time, a novel feature specific to this genre concerns the use of modal verbs, either necessity or prediction modals, which can be associated with more complex writing styles in general, but in the case of instructions, the use of modals makes them clearer.

The two examples below illustrate instructional texts which are classified as respectively easy and difficult:

EASY The Executive Hire Show takes place at The Ricoh Arena , Coventry . </p> Bus Public transport from train station to the Ricoh Arena : - Number 8 bus from Coventry Train Station to Coventry Transport Museum - Then catch the number 4 or number 5 from Coventry Transport Museum to Arena Park (Tesco) - Once you arrive at Arena Park there is an underpass which takes you into Car Park B of the Ricoh Arena . Follow signs for the Ricoh Arena main entrance from here . </p> Taxi For our local taxi service please visit www.mgmtaxi.co.uk or call 02476 375550 </p> Train Please note – The last train leaving Coventry Railway Station to London Euston is 23 : 31 …[1]

DIFFICULT Introduction </p> The most important part of working with this particular linked dataset , and probably datasets in general , is understanding what the variables mean and how they are coded . This is aided by studying the codebook , where available , and by running frequency tables of categorical and ordinal variables and means / medians of continuous variables . The codebook describes (or should describe) the name of each variable , what it is supposed to measure , and the number of levels or range of the values the variable takes on in the dataset . This will tell you , for example , if sex is coded as M and F , or 0 and 1 , or 1 and 2 , or 1 , 2 and 9 etc. The codebook for the linked Census data tells you that the income variables actually refer to 1985 income , even though the Census was taken in June of 1986 . This is important to keep in mind when analyzing the data . </p> One-way or two-way frequency tables not only give information on how the variables are distributed , but also … [2]

Examples also show that the neural transformer model is able to detect the inherent difficulty of topics, for example, descriptions of a statistical procedure (Difficult) as compared to giving directions (Easy), as the latter topic is more expected in texts for learners of lower levels. However, this inherent difficulty is not reflected in the set of the Biber features, and therefore is not captured in probing experiments as reported in Tables 4 or 11.

As for distinguishing easy and difficult texts among the news reporting texts, TTR does not feature in this list, thus implying that this feature has less impact the difficulty level within news items. The strongest indicator of difficult texts in this genre is K45.conjuncts, such as *in particular, instead, otherwise, similarly*, which are linked to more complex reporting styles, also with fewer past tense verbs. The counter-intuitive link between the difficult news articles and the second person pronouns rate (which featured prominently for easy texts in Table 4) is related to incomplete cleaning of some of the Web pages, as the most frequent contexts for *you* in this collection are legalistic boilerplate privacy notes, such as *When you subscribe we will use the information you provide to send you these newsletters..*, which are not considered as simple by the classifier.

[1] `http://www.executivehireshow.co.uk/visiting/travel`

[2] `http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1244`

While the rate of nouns was not considered as a predictive feature for the full corpus, as it varies considerably across the genres, this was detected as a significant feature within the two genres of Table 11.

4 Related studies

Statistical methods for analysing text complexity can be traced to frequency studies aimed at designing systems of shorthand writing (Käding, 1897), which was followed by traditional measures of readability, such as Lorge or Flesch-Kincaid measures, initially developed in the context of American adult education (Lorge, 1944; DuBay, 2004). There has been also a long line of research in statistical frequency distribution models, which can be linked to complexity (Juilland, 1964; Orlov, 1983; Baayen, 2008).

With the rise of Machine Learning, novel methods for readability prediction appeared, initially they were based on extraction of features (Pitler and Nenkova, 2008; Collins-Thompson, 2014; Vajjala and Meurers, 2014), such as those introduced by Biber, or on various frequency measures. In particular, it has been shown that unsupervised Principal Component Analysis arrives at the two principal dimensions with groups of features resembling lexical difficulty, for example, frequencies or word length, and syntactic difficulty, such POS codes (Sharoff et al., 2008). Other studies have also experimented with expanding the models from the document to the sentence level (Vajjala and Meurers, 2014) with the specific aim of comparing sentences from the Simple English Wikipedia against aligned sentences from the standard English Wikipedia.

As in many other areas of computational linguistics, feature-less neural networks provided better efficiency in difficulty predictions (Nadeem and Ostendorf, 2018), especially with the rise of pre-trained transformer models (Khallaf and Sharoff, 2021), which outperform both the linguistic features and the traditional neural networks.

Other studies have also emphasised the influence of genres on the predictions of the classifiers. In particular, existing approaches for measuring text complexity tend to **overestimate** the complexity levels of informational texts while simultaneously **underestimating** the complexity levels of literary texts (Sheehan et al., 2013). The authors of that study had to design different difficulty models for each of the two kinds of texts.

This study uses the CUP and Rufola datasets for training the classifiers. There are also many other sources for building models to distinguish easy or difficult texts. For English a commonly used choice is the WeeBit corpus (Vajjala and Meurers, 2012), which consists of texts from the Weekly reader magazine and from the BBC Bite-Size website. The other source is the Core Standards for secondary education in the US context [3]. In all of these datasets, the aim of difficulty annotation assumes the audience of native learners aged 7-17. A related experiment investigated syntactic parameters for predicting difficulty of Russian academic texts (Solovyev et al., 2019). There are also various sources of texts with difficulty assessed for adult speakers, for example, the WikiHow corpus (Debnath and Roth, 2021), which is based on Wiki texts edited for vagueness in instructions. Yet another source comes from other training scenarios, for example, from translation training, when texts are assessed with respect to the quality of their translation by translation students, for example, for translation into Russian (Kunilovskaya and Lapshinova-Koltunski, 2019) or Chinese (Yuan and Sharoff, 2020), in which case either the drop in translation quality or time spent on translation can be an indicator of difficulty.

[3] `http://www.corestandards.org/assets/Appendix_B.pdf`

5 Conclusions and further research

This paper presents a statistical study conducted on a large corpus to determine which features contribute to difficulty of English and Russian texts. This is based on a framework which combines a transformer-based neural prediction model operating as a blackbox and well-studied linguistic features providing a statistical explanation of how these features affect difficulty. For example, this study shows how the rate of nouns and the related complexity of noun phrases affects difficulty via statistical estimates of what the neural model predicts as easy and difficult texts.

The study also analysed the interplay between difficulty and genres, as linguistic features often specialise for genres rather than for inherent difficulty, so that some associations between the features and difficulty are caused by differences in the relevant genres. In particular, the Type-Token Ratio (TTR) is a good indicator of lexical diversity and it is usually higher with more difficult texts if both texts are in the same genre. At the same time, this study shows that the TTR of easy news reporting texts is likely to be higher than that of more difficult argumentative texts which make repeated references to the same key concepts.

From the practical viewpoint, the methods of this study help in automatic assessment of texts from the Web with the aim of extending the use of authentic texts in language teaching. The methods also help in understanding what makes authentic texts difficult and what might require their manual or automatic simplification. For example, despite the popularity of Fiction in language teaching applications, this study provides statistical evidence for the higher difficulty scores associated with fiction commonly found on the Web. This should not prevent the tutors from using fiction for language teaching, as it can be beneficial for both engagement and pedagogic purposes, but this calls for more attention to choosing and simplifying such texts when necessary.

Further extensions planned for improving the neural difficulty detection models involve several lines of research. First, this study focused almost exclusively on reading exercises for language learners. We need more experiments on studying variations in the link between difficulty and linguistic features with respect to different difficulty assessment needs or the composition of the training datasets. Even within the area of studying language teaching and expressing difficulty via the CEFR levels, different datasets might have different approaches to what constitutes a B1 text, for example. Also some texts are included into a textbook for a specific level not because they are fully suitable for this level, but because they can be used in other exercises for this level. For example, an authentic interview included into a B1 textbook might include rare words or more complex grammatical constructions beyond expectations of typical B1 students, while it can be a good basis for a number of exercises for understanding how native speakers express their opinions. From the viewpoint of Machine Learning, an interview of this kind even if legitimately included in the textbook acts as noise for training neural prediction models. We need to experiment with various statistical tests to establish how annotation noise can lead to less reliable predictions and how to improve our prediction models, for example, see (Paun et al., 2018).

Second, there is a rise in research on causal models, for example, (Fytas et al., 2021), because when we have a classifier, it is important to know whether this decision has been made for the right reasons, rather than because of mere correlations in our training data. Recent causal interaction methods can explain some of the issues with interpretation of predictions reported above (Janizek et al., 2021).

Third, a related line of research involves assessment of the process of mapping CEFR levels of documents to the level of segments. The process of segmentation used in this study can lead to noise, because some 3-sentence segments coming from a textbook of a higher level can still be suitable for students on lower levels. This has been already noticed in the context of using simplified Wikipedia (Vajjala and Meurers, 2014). A

similar task exists in other areas, for example, turning models which predict the quality of sentence-level translations to models predicting word quality (Zhai et al., 2020).

Finally, we need to pay more attention to cognitive aspects of difficulty processing beyond simple scores, such as exemplified by the CEFR levels. For example, this involves adding an explicit model for processing named entities (NEs), such as people names or locations. Anecdotal experience shows that language learners can often handle NEs, even if they are very rare, either because they are similar to how they are expressed in their native languages (see the example with *Lockheed Martin* above) or because they can understand their function of a person name or a location even without knowing this particular entity. This needs to be quantified. NEs are also important in a different way, as neural models can be brittle to NE replacements. For example, replacing NEs in the co-reference task changes 85% of predictions (Balasubramanian et al., 2020).

References

Baayen, H. (2008). *Analyzing linguistic data*. Cambridge University Press, Cambridge.

Balasubramanian, S., Jain, N., Jindal, G., Awasthi, A., and Sarawagi, S. (2020). What's in a name? Are BERT named entity representations just as good for any other name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214, Online. Association for Computational Linguistics.

Benko, V. (2016). Two years of Aranea: Increasing counts and tuning the pipeline. In *Proc LREC*, Portorož, Slovenia.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Collins-Thompson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proc. of HLT/NAACL*, pages 193–200, Boston.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Council of Europe (2001). Common european framework of reference for languages: Learning, teaching, assessment (cefr). Technical report, Council of Europe, Strasbourg.

Debnath, A. and Roth, M. (2021). A computational analysis of vagueness in revisions of instructional texts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 30–35, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doughty, C. J. and Long, M. H. (2008). *The handbook of second language acquisition*, volume 27. John Wiley & Sons.

DuBay, W. H. (2004). The principles of readability. Technical report, Impact Information.

Fytas, P., Rizos, G., and Specia, L. (2021). What makes a scientific paper be accepted for publication? In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 44–60, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Halliday, M. (1992). Language as system and language as instance: The corpus as a theoretical construct. In Svartvik, J., editor, *Directions in corpus linguistics: proceedings of Nobel Symposium 82 Stockholm*, volume 65, pages 61–77. Walter de Gruyter.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Janizek, J. D., Sturmfels, P., and Lee, S.-I. (2021). Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54.

Juilland, A. (1964). *Frequency dictionary of Spanish words*. Mouton.

Käding, F. W., editor (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Selbstverlag.

Khallaf, N. and Sharoff, S. (2021). Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019). Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Laposhina, A., Veselovskaya, T., Lebedeva, M., and Kupreshchenko, O. (2018). Automated text readability assessment for russian second language learners. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*.

Lorge, I. (1944). Predicting readability. *Teachers college record*.

Nadeem, F. and Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Orlov, J. (1983). Ein modell der häufigkeitsstruktur des vokabulars. In Guiter, H. and Arapov, M., editors, *Studies on Zipf's law*, pages 154–233.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., and Poesio, M. (2018). Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proc EMNLP*, pages 186–195.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3:1–32.

Sharoff, S., Kurella, S., and Hartley, A. (2008). Seeking needles in the Web haystack: finding texts suitable for language learners. In *Proc Teaching and Language Corpora Conference, TaLC 2008*, Lisbon.

Shavrina, T. and Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. In *CORPORA, International Conference*, Saint-Petersbourg.

Sheehan, K. M., Flor, M., and Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58, Atlanta, Georgia. Association for Computational Linguistics.

Solovyev, V., Solnyshkina, M., Ivanov, V., and Batyrshin, I. (2019). Prediction of reading difficulty in russian academic texts. *Journal of Intelligent & Fuzzy Systems*, 36(5):4553–4563.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proc CoNLL 2017 Shared Task*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (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, Montréal, Canada. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xia, M., Kochmar, E., and Briscoe, T. (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.

Yuan, Y. and Sharoff, S. (2020). Sentence level human translation quality estimation with attention-based neural networks. In *Proc LREC*, Online.

Zhai, Y., Illouz, G., and Vilnat, A. (2020). Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, Barcelona, Spain (Online). International Committee on Computational Linguistics.

6 Appendix 1: Linguistic features

The order of the linguistic features and their codes are taken from (Biber, 1988). The conditions for detecting the features for English replicate the published procedures from (Biber, 1988), many of them are expressed via lists of lexical items or via POS annotations, which in this study are provided by UDPIPE (Straka and Straková, 2017). The Russian features are either based on translating the English word lists or on using identical or functionally similar constructions.

Code	Label	Condition
A01	past verbs	VERB, Tense=Past
A03	present verbs	VERB, Tense=Pres
B04	place adverbials	ADV, lex in (*aboard,above,abroad,across…*)
B05	time adverbials	ADV, lex in (*afterwards,again,earlier…*)
C06	first person pronouns	PRON, lex in (*I,we,me,us,my..*)
C07	second person pronouns	PRON, lex in (*you,your,yourself,yourselves*)
C08	third person pronouns	PRON, lex in (*she,he,they,her,him,them,his…*)
C09	impersonal pronouns	Conditions from (Biber, 1988)
C10	demonstrative pronouns	Conditions from (Biber, 1988)
C11	indefinite pronouns	PRON, lex in (*anybody,anyone,anything,everybody…*)
C12	*do* as pro-verb	Conditions from (Biber, 1988)
D13	wh-questions	Conditions from (Biber, 1988)
E14	nominalizations	lex ends with ('tion','ment','ness', 'ism')
E16	nouns	Conditions from (Biber, 1988)
F18	passives with *by*	Conditions from (Biber, 1988)
G19	*be* as main verb	Conditions from (Biber, 1988)
H23	wh-clauses	Conditions from (Biber, 1988)
H34	sentence relatives	Conditions from (Biber, 1988)
H35	causatives	CONJ, lex in (*because*)
H36	concessives	CONJ, lex in (*although,though,tho*)
H37	conditionals	CONJ, lex in (*if, unless*)
H38	other subordination	Conditions from (Biber, 1988)
I39	prepositions	ADP
I40	attributive adjectives	Conditions from (Biber, 1988)
I41	predicative adjectives	Conditions from (Biber, 1988)
I42	adverbs	ADV
J43	type-token ratio	Using 400 words as in (Biber, 1988)
J44	word length	Average length of orthographic words
K45	conjuncts	Conditions from (Biber, 1988)
K46	downtoners	lex in (*almost,barely,hardly,merely..*)
K47	general hedges	lex in (*maybe, at about, something like..*)
K48	amplifiers	lex in (*absolutely,altogether,completely,enormously…*)
K49	general emphatics	Conditions from (Biber, 1988)
K50	discourse particles	Conditions from (Biber, 1988)
K55	public verbs	VERB, lex in (*acknowledge,admit,agree…*)
K56	private verbs	VERB, lex in (*anticipate,assume,believe…*)

Continued on next page

Continued from previous page

Code	Label	Condition
K57	suasive verbs	VERB, lex in (*agree,arrange,ask…*)
K58	seem/appear	VERB, lex in (*appear, seem*)
L52	possibility modals	VERB, lex in (*can,may,might,could*)
L53	necessity modals	VERB, lex in (*ought,should,must*)
L54	prediction modals	VERB, lex in (*shall,will,would*), excluding future tense
N59	contractions	Conditions from (Biber, 1988)
N60	that deletion	Conditions from (Biber, 1988)
P66	synthetic negation	Conditions from (Biber, 1988)
P67	analytic negation	Conditions from (Biber, 1988)