A Cross-platform Investigation of Complexity for Russian Learners of English

N. Ballier, University of Paris, France
A. Buzanov, HSE University, Moscow, Russia
T. Gaillat, University of Rennes, France
O. Vinogradova, HSE University, Moscow, Russia

Eurocall 2021, Paris, 26-27 Aug.

OUTLINE

Background

Research questions

Data

Method

Results

Discussion

Further Research

Conclusion

References

Background

L2 analysis: Complexity, Accuracy and Fluency

- Construct: Complexity > linguistic complexity (systemic and structural) (Housen *et al.* 2012)
- Operational measures: lexical diversity, syntactic complexity, cohesion …

L2 proficiency prediction tasks

- Several languages and some experiments on CEFR levels
- Rely on different types of features including complexity metrics
- Show encouraging results BUT B level remains an issue.

Need:

Explore the B level to identify discriminatory features.

Research questions

RQ1: Do we find similar features across two automatic NLP pipelines of automatic metrics : Inspector (Lyashevskaya et al., 2021) and AIDALLA (Sousa *et al*, 2020, Gaillat *et al*. 2021)

RQ2: Which linguistic features are criterial in distinguishing between the B1 and B2 levels?

Data

1,171 of REALEC essays (https://realec.org/index.xhtml#/exam/),

268,563 words (301,272 tokens) total

- written in English examination by 2nd-year Bachelor program university students (~20 yo) with Russian L1
- 2 task types - description of the graphical material & opinion essay; the former of 183 words average length, the latter, 274 words on average; each examinee writes both types of essays
- A set of 72 parameters of text complexity for each essay from Inspector

Inspector Parameters

Inspector = a set of text complexity features (Lyashevskaya et al, 2021), namely:

- 9 lexical diversity measures
- lexical density parameter
- 14 lexical sophistication metrics
- 24 syntactic complexity parameters
- 13 measures of morphological complexity
- 4 discursive complexity measured by numbers of discourse-organising nouns and linking units, as well as 4-grams and functional *n*-grams
- 7 parameters of error counts

AIDALLA parameters and tools

A set of complexity metrics (768) (Gaillat et al., 2021)

- Syntactic e.g. amount of coordination, subordination, microsystems
- Semantic e.g. ambiguity
- Lexical e.g. density, sophistication
- Pragmatic e.g. cohesion

Annotation and pattern frequency tools

- LCA (TreeTagger) - TAACO - TAALES - TAASC -TEXTSTAT - PYENCHANT
- Modified version of L2SCA New features based on paradigmatic microsystems L2SCA_MS

Automatic scoring output

First received predictions of CEFR level text at:

- DUOLINGO

- GRAMMARLY

- WRITE AND IMPROVE

Then applied an algorithm to obtain a single CEFR-level prediction for each text:

https://github.com/soimmary/REALEC

Defining classes B1 and B2 from the automatic prediction

Dataset 5 classes	
At least two out of three predictions at level A (A1 or A2)	removed
One prediction at level A, two at level B (B1 or B2)	B1-
All three predictions B1	B1-
One prediction at level A, one at level B or C, and one at level C	B1
One prediction B1, one B2, and one at level B (B1 or B2)	B1
One prediction B1, one at level B (B1 or B2), and one C1	B1+
All three predictions B2	B1+
Two predictions B1, and one C2	B2
Two predictions C1, and one at level B (B1 or B2	B2
Two predictions B2, and one C1	B2
One prediction C2, one B2, and the third at level B (B1 or B2)	B2+
One prediction B1, one at level C (C1 or C2), and one C2	B2+
Two predictions at level C (C1 or C2), and one C1 or B2	removed

Dataset 2 classes

B1-, B1, B1+	B1
B2, B2+	B2

Distribution

B1 B1- B1+ B2 B2+
361 90 368 250 78

Classification

Two datasets: Training (75%) and test (25%) sets

- Inspector: 72 Metrics (Lyashevskaya et al., 2021)
- AIDALLA: Pipeline of 768 metrics (Sousa *et al*, 2020, Gaillat *et al*. 2021)

Model: regression (Elastic Net) >> Minimise effect of non-informative variables

- penalty mechanism
- feature dimension reduction

Stage 1: Multinomial

Stage 2: Binomial

Stage 3: Regular binary logistic regression for variable importance based on the Stage 2 model's significant features

Evaluation metrics: Balanced Accuracy

Results

Stage	Method	Elastic Net: Inspector metrics	Elastic Net: AIDALLA metrics
1	Balanced accuracy (averaged over 5 classes)	0.551	0.509
2	Balanced accuracy (2 classes)	0.623	0.652

Results (Inspector) Significant features

1. **number of derivational suffixes (level 4)**
2. number of derivational suffixes (level 5)
3. number of derivational suffixes (level 6)
4. **Lexical density**
5. corrected **verb sophistication**
6. lexical sophistication
7. uber type token ratio
8. **number of shell nouns**
9. **Lexical Frequency Profile (first 1000 most frequent words)**
10. verb variation ii
11. **number of T-units**
12. number of complex T-units
13. number of coordinate phrases
14. Levenshtein distance between lemmatized sentences (between all sentences)
15. **frequency of verbs in present simple tense (plur)**
16. **frequency of verbs in past simple**

Other features

- **number of misspelled tokens**
- number of punctuation mistakes in participle phrases
- **number of punctuation mistakes connected with the conjunction *but***
- **number of punctuation mistakes connected with the conjunction *because***

Results (AIDALLA) Significant features

Stage 1

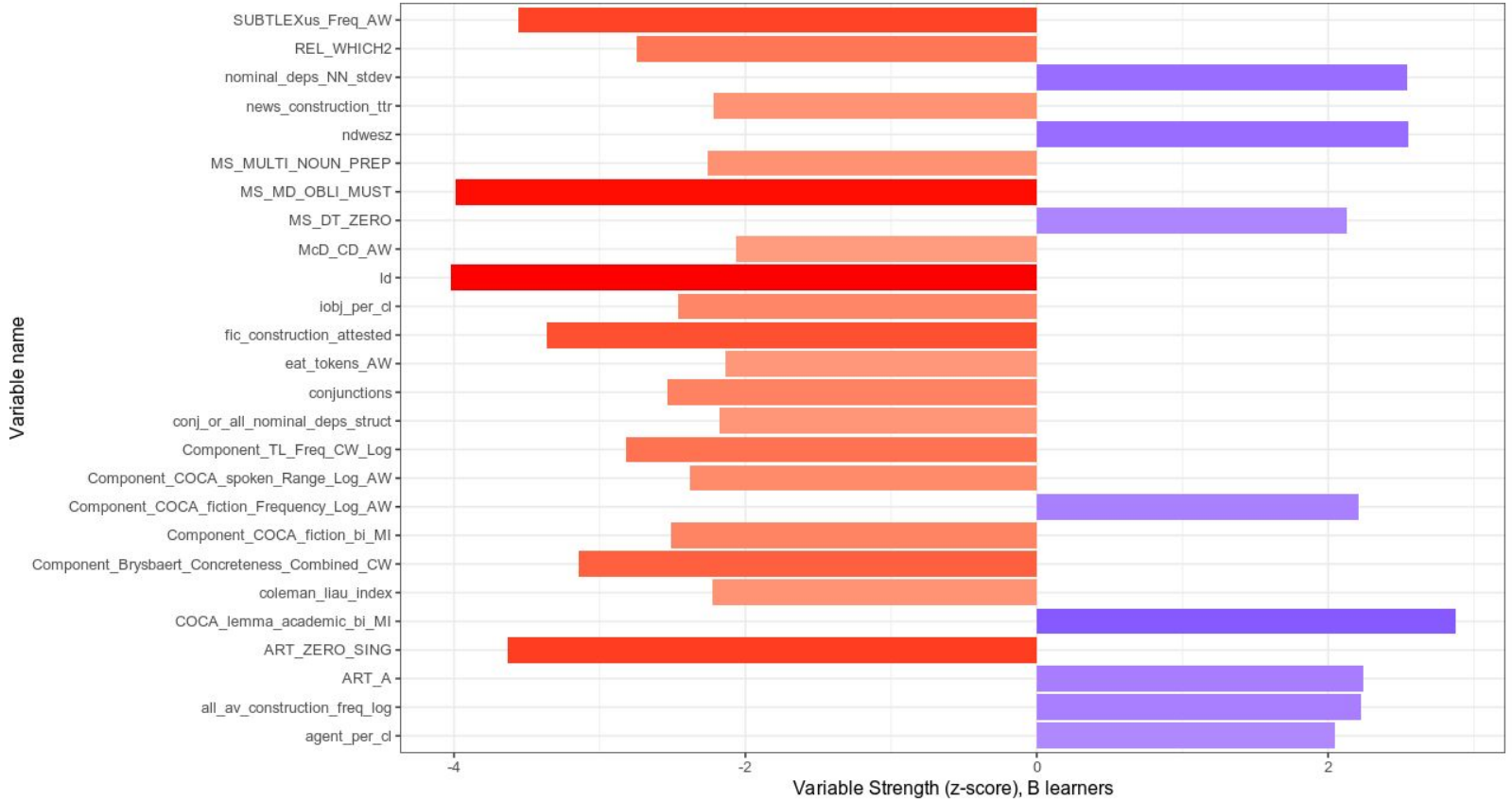
- Lexical (25) > sophistication metrics based on COCA
- Syntactic (71) > NP complexity indices and COCA-based construction freq indices
- Functional (10) > Microsystems: possibility (might/may), proforms (this/that)
- Cohesive (11) > Lexical overlap between adjacent sentences
- Readability (1) > coleman-Liau

Stage 2

- Lexical (43) > sophistication metrics based on COCA + concreteness + familiarity + meaningfulness
- Syntactic (76) > COCA-based construction freq indices + NP complexity (incl prepositions, conjunctions, relatives, possessives)
- Functional (14) > Microsystems: Prepositional constructions, determiners, obligation modality, countability
- Cohesive (4) > Lexical overlap between adjacent sentences, link words (nonetheless, therefore, although, therefore, that is why, for this reason)
- Misspellings (1)

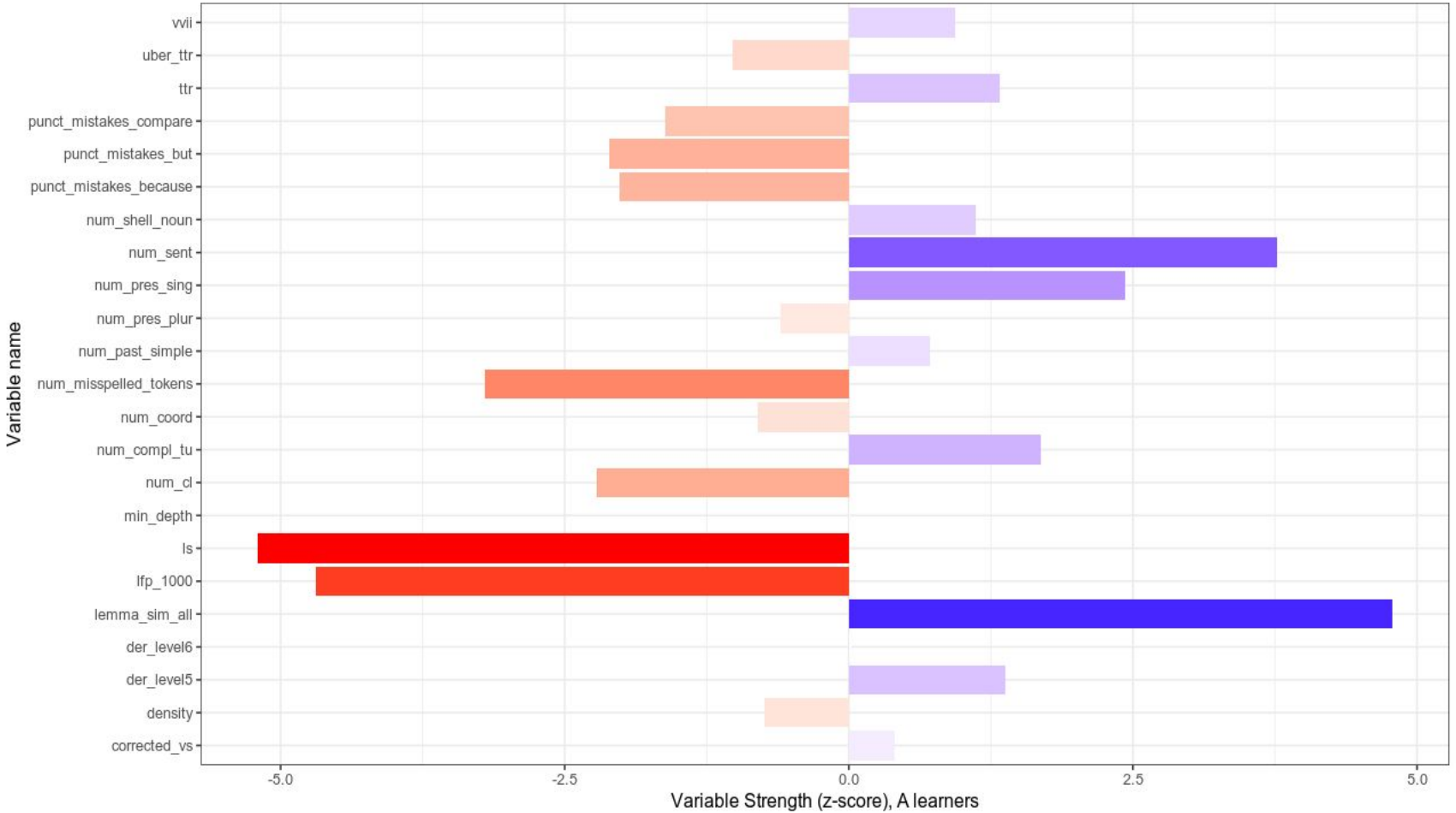
Discriminatory features - AIDALLA

Stage 3



Discriminatory features - Inspector

Stage 3



Discussion and conclusion

Features in common

- lexical features - Sophistication (Lexical Frequency Profile and reference corpora) ; lexical density;
- Cohesion
- Accuracy - number of misspelled words and number of punctuation errors

Differences

- Lexical: Inspector favours internal counts while AIDALLA favours reference-corpus indices
- Morphology: Inspector points out differences between B1 and B2 texts in the use of suffixes *-able*, *-er*, *-ish*, *-less*, *-ly*
- Syntactic: Inspector > Internal freq counts of sentence, phrasal and verbal units; AIDALLA > reference-corpus indices and nominal complexity counts

Conclusion:

Combination of internal frequency counts + and reference-corpus indices

Further research

Other statistical models with cross-validation (those in R?)

Better balanced numbers of texts at each predicted level

Datasets based on different L1s e.g. EFCAMDAT

References

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., & Zarrouk, M. (2019). A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors. In European Conference on Technology Enhanced Learning (pp. 308-320). Springer.

Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, Thomas Gaillat (2020) Machine Learning for learner English. A plea for creating learner data challenges, 30 p. *International Journal of Learner Corpus Research*, Issue 6.1, 72-103.

Stephen Bax, Fumiyo Nakatsuhara & Daniel Waller (2019). Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels. *System*, Volume 83, July 2019, pp. 79-95

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (1998). TiMBL: Tilburg Memory-Based Learner-version 1.0-Reference Guide.

Sousa, A., Ballier, N., Gaillat, T., Stearns, B., Zarrouk, M., Simpkin, A. and Bouyé, M. (2020) From Linguistic Research Projects to Language Technology Platforms : A Case Study with Learner Data . LREC2020 1st International Workshop on Language Technology Platforms. Marseilles, 16 May 2020, ACL / LREC2020 1st International Workshop on Language Technology Platforms. Marseilles, 16 May 2020, 112-120.

Vinogradova, O. The role and applications of expert error annotation in a corpus of English learner texts. Computational Linguistics and Intellectual Technologies. Proceedings of "Dialog 2016", 15, pp. 740–751

Vinogradova, O. I., Lyashevskaya, O. N., & Panteleeva, I. M. Multi-level Student Essay Feedback in a Learner Corpus. In Proceedings of the International Conference "Dialogue 2017", 2, pp. 370-382, 2017.

Olga Lyashevskaya , Irina Panteleeva, Olga Vinogradova. Automated assessment of learner text complexity // Assessing Writing - International Journal, https://www.journals.elsevier.com/assessing-writing. 2021. No. 49. Article 100529.

https://github.com/soimmary/REALEC

https://docs.google.com/spreadsheets/d/1_LhLHkcrZPGMIpygs2-tLDxEPQSaKvHLLEkxXz0uQB4/edit#gid=1111529494

Acknowledgements

Thanks to

stats:

Clément Levrard

Taylor Arnold

Andrew Simpkin

corpus data:

Learner Corpora Laboratory

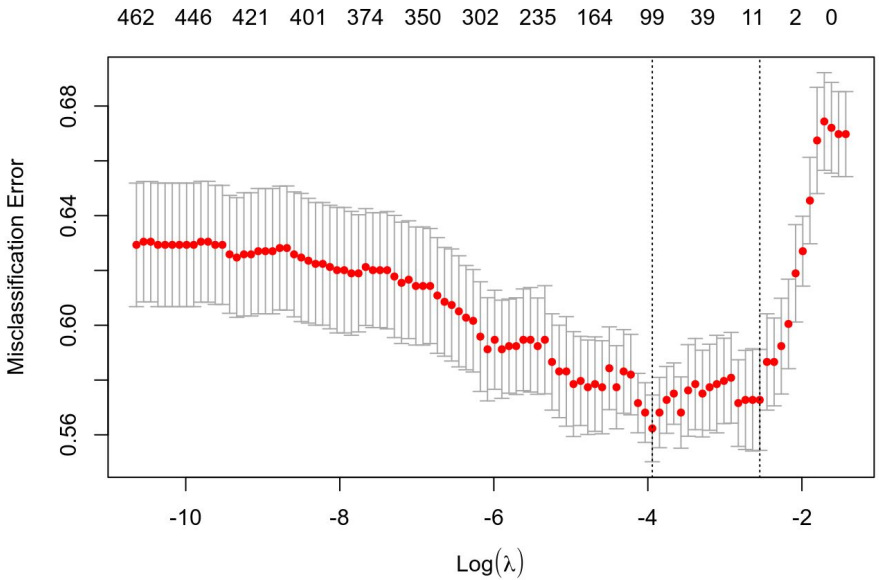
Maria Bocharova

Inspector - text complexity tool

Irina Panteleeva

THANK YOU !

Alternates


462 446 421 401 374 350 302 235 164 99 39 11 2 0
Misclassification Error
0.56
0.60
0.64
0.68
-10
-8
-6
-4
-2
Log(λ)

Take home message

Limitation of the model based on Spanish and French learners for Russian learners

REsults AIDALLA

Binary classification

Confusion Matrix and Statistics

```
          Reference
Prediction   1   2
         1 194  71
         2   6  15

               Accuracy : 0.7308
                 95% CI : (0.6754, 0.7813)
    No Information Rate : 0.6993
    P-Value [Acc > NIR] : 0.1361

                  Kappa : 0.1841

 Mcnemar's Test P-Value : 3.021e-13

            Sensitivity : 0.9700
            Specificity : 0.1744
         Pos Pred Value : 0.7321
         Neg Pred Value : 0.7143
             Prevalence : 0.6993
         Detection Rate : 0.6783
   Detection Prevalence : 0.9266
      Balanced Accuracy : 0.5722
```

Confusion Matrix and Statistics

```
          Reference
Prediction  1  2  3  4  5
         1 49 16 28 11  0
         2 36  9 36 22 13
         3 13  1 16 29  6
         4  0  0  0  0  0
         5  0  0  0  0  0
```

Overall Statistics

```
               Accuracy : 0.2596
                 95% CI : (0.2097, 0.3146)
    No Information Rate : 0.3439
    P-Value [Acc > NIR] : 0.9991

                  Kappa : 0.0427

 Mcnemar's Test P-Value : NA
```

Statistics by Class:

```
                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.5000  0.34615  0.20000   0.0000  0.00000
Specificity           0.7059  0.58687  0.76098   1.0000  1.00000
Pos Pred Value        0.4712  0.07759  0.24615      NaN      NaN
Neg Pred Value        0.7293  0.89941  0.70909   0.7825  0.93333
Prevalence            0.3439  0.09123  0.28070   0.2175  0.06667
Detection Rate        0.1719  0.03158  0.05614   0.0000  0.00000
Detection Prevalence  0.3649  0.40702  0.22807   0.0000  0.00000
Balanced Accuracy     0.6029  0.46651  0.48049   0.5000  0.50000
```

Results

ElasticNet with 5 classes based on Inspector parameters

Accuracy: 0.3993

Poor second and fifth classes

10 selected features: "der_level4", "density", "vs", "lfp_1000", "num_tu", "num_adj_noun", "num_past_simple", "num_misspelled_tokens", "punct_mistakes_pp", "punct_mistakes_but"

Results

ElasticNet with 2 classes based on Inspector parameters

Accuracy: `0.7063`

22 selected features: "der_level4", "der_level5", "der_level6", "density", "ls", "corrected_vs", "lfp_1000", "uber_ttr", "vvii", "num_shell_noun", "min_depth", "num_cl", "num_tu", "num_compl_tu", "num_coord", "lemma_sim_all", "num_pres_plur", "num_past_simple", "num_misspelled_tokens", "punct_mistakes_because", "punct_mistakes_but", "punct_mistakes_compare"

Misc to be inserted

2019 badge

University logos

Kitchen sink method for the

ReCALL: https://www.eurocall-languages.org/publications/recall-journal

Method: data processing

	Inspector	AIDALLA
Lexical complexity	Y	Y
Syntactic complexity	Y (21)	Y
Morphological complexity	Y	N
Cohesion	Y	Y
Accuracy (spelling, capitalization, punctuation)	Y (7)	Y (1)

AIDALLA detailed results

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 49 16 28 11 0

2 36 9 36 22 13

3 13 1 16 29 6

4 0 0 0 0 0

5 0 0 0 0 0

Overall Statistics

Accuracy : 0.2596

95% CI : (0.2097, 0.3146)

No Information Rate : 0.3439

P-Value [Acc > NIR] : 0.9991

Kappa : 0.0427