PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 113 OCTOBER 2019 9–30

EVALD – a Pioneer Application for Automated Essay Scoring in Czech

Kateřina Rysová, Magdaléna Rysová, Michal Novák, Jiří Mírovský, Eva Hajičová

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
[rysova|magdalena.rysova|mnovak|mirovsky|hajicova]@ufal.mff.cuni.cz

Abstract

In the paper, we present EVALD applications (Evaluator of Discourse) for automated essay scoring. EVALD is the first tool of this type for Czech. It evaluates texts written by both native and non-native speakers of Czech. We describe first the history and the present in the automatic essay scoring, which is illustrated by examples of systems for other languages, mainly for English. Then we focus on the methodology of creating the EVALD applications and describe datasets used for testing as well as supervised training that EVALD builds on. Furthermore, we analyze in detail a sample of newly acquired language data – texts written by non-native speakers reaching the threshold level of the Czech language acquisition required e.g. for the permanent residence in the Czech Republic – and we focus on linguistic differences between the available text levels. We present the feature set used by EVALD and – based on the analysis – we extend it with new spelling features. Finally, we evaluate the overall performance of various variants of EVALD and provide the analysis of collected results.

1. Introduction

The present contribution summarizes results of a long-term research focused on automated essay scoring (AES) with emphasis on automatic evaluation of surface coherence in texts written by native as well as non-native speakers of Czech. The research resulted in three software applications (one for native, two for non-native speakers).

 Corresponding author: rysova@ufal.mff.cuni.cz

Cite as: Kateřina Rysová, Magdaléna Rysová, Michal Novák, Jiří Mírovský, Eva Hajičová. EVALD – a Pioneer Application for Automated Essay Scoring in Czech. The Prague Bulletin of Mathematical Linguistics No. 113, 2019, pp. 9–30. doi: 10.2478/pralin-2019-0004.

The applications are called Evaluator of Discourse (EVALD) and are available for the general public, especially for students, teachers, examiners or anyone involved in the language teaching process.

To create a well-structured and comprehensible piece of writing is a difficult task. A text has to form a functional complex in both content and formal aspects and, primarily, it has to faithfully fulfill the author's communicative intention. To formulate ideas into a complex piece of discourse (text) is much more difficult than to create separate sentences. The stylization skills and the ability to express thoughts and attitudes in a coherent text are thus taught in language lessons as one of the most important language competencies. At the same time, these competencies are regularly tested in official as well as unofficial language exams of various kinds.

A part of the school-leaving examination at secondary schools in the Czech Republic is writing an essay in the mother tongue. Successful passing of this exam affects student's further life (e.g. the possibility of their further study at a university). Writing an essay in Czech is a part of various official exams also for foreigners, e.g. for those who apply for permanent residence in the Czech Republic. The required language level according to the Common European Framework of Reference for Languages (CEFR) is A1. Furthermore, the compulsory CEFR level for foreigners to be granted state citizenship in the Czech Republic is B1, most of the Czech universities require B2 and companies usually B2 or C1. The motivation of foreigners to learn Czech (and to practice their writing skills in Czech) is thus clear.

Given the importance and severity of the mentioned exams (which affect the real life of the students or applicants), the assessment of the essays needs to be as objective as possible. A software application like EVALD can help in this task.[1] Similar applications already exist and successfully work abroad (for more details, see Section 2). For Czech, EVALD is so far the first tool in this category.

Section 3 introduces versions of EVALD that have been developed as well as the individual steps that led to the creation of the EVALD application. These steps almost overlap with individual levels of linguistic description that have been gradually incorporated to the application. In the paper, we thus describe the creation process of EVALD also from the linguistic point of view – concerning the individual language levels/phenomena, we present their contribution to AES.

The description of EVALD is divided into the following parts. In the methodological part (Section 3.1), we introduce the individual steps necessary for developing the software application. Both linguistic and computational parts (the process of implementation) are given in detail.

The datasets used for our task are presented in Section 3.2. One is aimed at automatic evaluation of texts written by native speakers of Czech (evaluated on the scale 1–5, i.e. excellent–fail), another one for evaluation of texts written by non-native speak-

[1] Although it is necessary to emphasize that the application is rather suitable as an assistant tool – the evaluated text should be also seen by a teacher-evaluator.

ers of Czech in the CEFR classes A1–C2, and finally, the third one for determining texts reaching at least the boundary CEFR class A1. It is necessary to mention that the first two datasets were annotated in terms of the level of surface text coherence (cohesion). The newly added third dataset, the NIE dataset, was used for a slightly different task. Instead of grades with respect to coherence (cohesion), it contains overall grades embracing all aspects of language. The dataset was created especially for the purpose of permanent residence examination requiring evaluation of the overall grade.

The result part comprises three sections. We put under scrutiny a sample of selected texts from the NIE dataset, as the clear delimitation of the A1 lower boundary appeared to be very helpful in practice, especially for the official exams compulsory for the applicants to be granted permanent residence in the Czech Republic (Section 3.3).

In Section 3.4, we present individual linguistic features across the language levels according to which it is possible to differentiate the examined texts automatically using methods of supervised machine learning. As a result of the analysis of the NIE dataset, we also design new features focused on spelling here.

Finally, we present performance of the resulting applications in Section 3.5. We evaluate not only the newly introduced spelling features, but also the EVALD system as a whole. Conducting a feature ablation analysis, we explore the key properties of the tool.

In the conclusion part (Section 4), we summarize the results achieved during the whole EVALD project and we describe the possibilities of use of EVALD applications in practice.

2. Automated Essay Scoring in Practice

The topic of automated essay scoring goes back to the 1960s. The English teacher and psychologist Ellis Batten Page, who designed the first system for the automatic assessment of student essays, became a pioneer in this field (Page, 1966, 1968). In 1973, he managed to conduct successful experiments (Ajay et al., 1973) and implemented his system under the name Project Essay Grade. However, it was difficult and costly to use the system in practice with the original technical capabilities.

Therefore, the development of the field and its connection with practice was carried out later in the 1990s with the expansion of the Internet and tools for natural language processing. It gave rise to new as well as some updated applications such as E-Rater (Burstein et al., 1998), Intelligent Essay Assessor (Foltz et al., 1999), Text Categorization Technique (Larkey, 1998), or the continuing work on Project Essay Grade (Page and Petersen, 1995).

Currently, automatic text evaluation systems are used (often together with teacher assistance, albeit sometimes on themselves) for the actual classification of student writing, mostly for English as a foreign language (L2). Automatic text evaluation is used, for example, in the Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), Graduate Management Admissions Test (GMAT), SAT, American College Testing (ACT), Test of English for International Communication (TOEIC), Analytic Writing Assessment (AWA), No Child Left Behind (NCLB), or Pearson Test of English (PTE), cf. Zupanc and Bosnić (2015).

2.1. Tools for English

For English, the choice of automatic text evaluation systems is quite varied. They differ from each other by their sophistication, the number of statistical analyses they are able to offer to the user, the number of texts they have acquired for training, the chosen computational method, or whether they are available free of charge (a large number of assessing tools are currently paid).

Project Essay Grade. For example, Project Essay Grade (PEG), the first automated text evaluation project designed in the first version in the 1960s, is now a completely commercial product, with not even a demo available freely (it commercialized in 2003). PEG analyzes the input text and calculates more than 300 features that reflect the characteristics of the writing, such as fluency, diction, grammar, and construction. According to its website,[2] it is currently used in 1000 schools and 3000 public libraries.

Text Inspector. One of the well-known models for English is the Text Inspector[3] that is designed to evaluate non-native speakers' texts. The basic functionality is offered free of charge, a subscription plan is available for those who wish to analyze longer texts or get more detailed results. The Text Inspector provides a statistical analysis of the evaluated text and finally displays also its language level according to CEFR. For example, the tool calculates number of sentences, words, syllables, their average length or relative frequency in a text, distributions for part of speech and other morphological categories, variety of vocabulary, and many other linguistic characteristics. In addition, it compares frequency of words in CEFR categories with large corpora such as the British National Corpus.[4] The system also estimates how difficult the text is for the reader's understanding, using three metrics: Flesch Reading Ease, Flesch-Kincaid Grade and Gunning Fog Index.

Readable.io. Another widely used tool for automatic evaluation of texts in English is Readable.io.[5] Again, basic functionality is available free of charge, while advanced

[2] http://www.measurementinc.com/products-services/automated-essay-scoring

[3] https://textinspector.com/

[4] http://www.natcorp.ox.ac.uk/

[5] https://app.readable.io/

features require a subscription. The application ranks the evaluated text in one of the CEFR as well as IELTS categories.[6] Similarly to Text Inspector, Readable.io quantifies the readability of the text, i.e. how difficult or easy it is to read. For this purpose, it uses eight different metrics. Some are based on the length of the words and sentences, others work with lists of "difficult" or "basic" words of the language.

Apart from detailed statistics, the application gives some specific verbal recommendations, such as "*This sentence is very long. Consider rewriting it to be shorter or splitting it into smaller sentences.*" or "*This is a hard word to read. Consider using an easier word if possible.*" The application also attempts to estimate the text style (formal – informal) and conducts sentiment analysis to display its emotional color and polarity (negative – positive).

Readable.io is also linked to other automatic language processing tools. With their help, it lists keywords from the text (separate words and especially word pairs, i.e. assumed phrases), lists the currently popular words (buzzwords), grammatical words, words pragmatically colored, words that are among the first 850 words designed to teach English to foreigners, and words that are among the thousand of most frequent words from children's books. It also estimates whether the author of the text is a man or a woman.

2.2. Tools for other languages

Some AES systems extended their scope to languages other than English. For example, IntelliMetric[7] claims it can handle 20 languages. It gives the opportunity to experience Chinese, Turkish, or Malaysian text ratings directly on the website, along with American, British and Australian English.

This tool is specific in that it evaluates essays written according to the predefined assignment and offers several possibilities of text evaluation to the user. In one task, for example, the application shows the user the beginning of a text and requires to complete it. The system then evaluates this completed part. If a user inserts a text that is not actually completing the predefined text, the ranking system refuses to evaluate it because the inserted text does not match the specified topic.

The tool also offers other possibilities of text evaluation. In another task, the user has to write a story on a given topic, namely a story that is supposed to be published in a magazine read by students throughout Australia. The specific task is e.g.: write a story about captivity or imprisonment – the narrator of the story, his friend, or his animal should get into the trap. The assignment also outlines where the story takes place, what events lead to captivity and who is involved in them, what feelings the characters have and how the issues raised are resolved.

---

[6] International English Language Testing System (IELTS) is an internationally accepted exam in English as a foreign language.

[7] `http://www.intellimetric.com/`

When the user enters the input text, the system provides grades in several areas. It assesses whether the text matches the intended audience, and provides marks on the text structure, ideas, vocabulary, cohesion, paragraph breakdown, sentence structure, punctuation or spelling in general.

This system seems to be trained also with regard to the topic of the text and its target group of readers. This can make the scorer more accurate, but it also limits the range of text types that the system can evaluate.

In general, other languages than English are less covered by AES tools but research results (sometimes accompanied by tools) have been reported also for French (Lemaire and Dessus, 2001), Japanese (Ishioka and Kameda, 2006), German (Wild et al., 2005), Spanish (Castro-Castro et al., 2008), Arabic (Al-Jouie and Azmi, 2017), Polish (Broda et al., 2014), and other languages (cf., e.g., Zupanc and Bosnić, 2015).

3. EVALD – the pioneer automated essay scoring for Czech

EVALD is a software application that serves primarily for AES of Czech written texts. Currently, EVALD exists in three versions. The first two applications were created for texts written by native speakers of Czech (with grades 1–5) and by non-native speakers (in A1–C2 of CEFR levels). These two versions were focused on evaluation of surface coherence (cohesion). Their previous development and gradual extension of the systems has been reported in Rysová et al. (2016), Novák et al. (2017), Novák et al. (2018) and Novák et al. (2019). The third, new version targets texts of non-native speakers with their language competence around the lowest CEFR level. That is, the system attempts to distinguish between the competence equal to A1 level and the competence that does not reach it (level 0). It was created especially for the purpose of permanent residence examination that requires evaluation of an overall mark. The third EVALD version (exploiting the NIE dataset from the National Institute for Education, see Section 3.2) thus does not provide evaluation of surface text coherence but a general, overall grade.

In general, EVALD processes the input text by internal procedures and then informs the user about the supposed level of surface coherence (or overall grade, respectively) in the submitted text. The online version of the EVALD for Foreigners is shown in Figure 1.

In its assessment, EVALD tries to imitate human evaluators by means of supervised machine learning. That is, using hundreds of texts evaluated by teachers (human assessors) EVALD learned how they evaluate the texts in order to be able to evaluate new texts itself. An example of the text evaluation by EVALD for Foreigners is given in Figure 2.

The software is trained to evaluate prosaic texts whose content and form (e.g. length) correspond to common school essays. The essays are usually created as a comprehensive piece of writing on a given topic, e.g. during the Czech language exams in case of non-native speakers or e.g. during the lessons of Czech at secondary

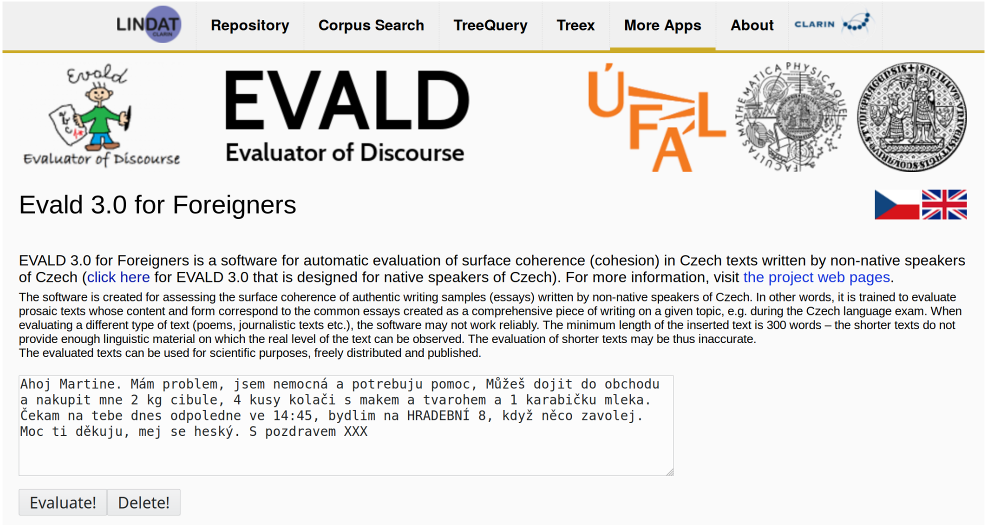

Figure 1. EVALD for Foreigners – design of the online evaluation service.

and higher grades of elementary schools in case of native speakers. When evaluating a different type of text (e.g. too short texts or poems), the software may not work reliably because it is not trained for these text genres.

3.1. Methodology – building the EVALD applications

The process of creation of EVALD applications may be divided into several steps. Firstly, suitable data has been collected. We gathered sets of texts written by native as well as non-native speakers of Czech and detailed linguistic research has been carried out on them. Texts written by learners of Czech come from the MERLIN corpus (Boyd et al., 2014), CzeSL-SGT corpus (Šebesta et al., 2014), and from exams organized by the National Institute for Education. Texts by the native speakers were taken from the corpus Skript2012/AKCES 1 (Šebesta et al., 2016). The individual datasets are described in detail in Section 3.2.

Collecting the data comprises also labeling the texts with corresponding grades. Concerning texts written by native speakers, the scale of grades was 1–5 (excellent to fail; five grades are typically used at Czech schools). As for non-native speakers of Czech, we used the scale A1–C2 (and separately 0–A1) in accordance with language levels defined by CEFR.

The texts were then examined from the perspective of the individual language areas: spelling, morphology, lexicology, syntax, semantics and discourse phenomena.

Evaluation

Evaluation class: A1
Probability of the evaluation: 0.33
The text is too short (shorter than 300 words), the evaluation may be inaccurate.

Language aspects stronger than A1:
Spelling: unrecognized words
Morphology: complexity and diversity
Syntax: complexity and diversity
Text structure: frequency of discourse connectives
Text structure: diversity of discourse connectives
Text syntax: sentence information structure

Language aspects corresponding to A1:
Vocabulary: complexity and diversity
Text structure: coreference

Evaluating scale for surface text coherence:

A1	basic user of Czech – lower level
A2	basic user of Czech – higher level
B1	independent user of Czech – lower level
B2	independent user of Czech – higher level
C1	proficient user of Czech – lower level
C2	proficient user of Czech – higher level

Figure 2. Example of text evaluation by EVALD for Foreigners – result for Text 1 (see below).

Based on the linguistic analysis of collected texts, a list of differentiating language features has been established. The features were designed to sort new texts into multiple levels labeled by the grades. Using the designed features, EVALD then learns to do the same job automatically by means of supervised machine learning. The linguistic analysis of selected texts (gained from the National Institute for Education) is presented in Section 3.3 and the differentiating features are described in detail in Section 3.4.

Based on the mentioned activities, a software application for automatic text evaluation was developed.[8] Firstly, automatic pre-processing of input texts was carried out with the help of the Treex system (Popel and Žabokrtský, 2010). The automatic text processing consists of several steps, e.g. tokenization, sentence segmentation, morphological analysis, or both surface and deep syntactic analysis. For the purpose of the EVALD project, the Treex was extended also to include detection of discourse

[8]As said in the introductory part of Section 3, the first two EVALD versions are trained to evaluate surface text coherence; the third, latest version (using the lower-level texts from the National Institute for Education) assigns an overall mark to the input text.

L1 dataset	**1**	**2**	**3**	**4**	**5**		**Total**
# documents	484	149	121	239	125		1,118
# sentences	20,986	4,449	2,913	3,382	939		32,669
# tokens	301,238	65,684	40,054	43,797	11,379		462,152
L2 dataset	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
# documents	174	176	171	157	105	162	945
# sentences	1,802	2,179	2,930	2,302	1,498	10,870	21,581
# tokens	15,555	21,750	27,223	37,717	21,959	143,845	268,049
NIE dataset	**0**	**b-line**	**A1**				**Total**
# documents	203	202	205				610
# sentences	1,166	1,112	943				3,221
# tokens	9,968	8,929	7,449				26,346
SYN dataset							**Total**
# documents							87,653
# sentences							275,349,473
# tokens							4,351,945,964

Table 1. Basic statistics on the labeled and unlabeled datasets, in total and for individual grades ('b-line' in the NIE dataset stands for the borderline cases, i.e. cases where the human annotators disagreed on the grade).

connectives and discourse (semantico-pragmatic) relations in a text, recognition of anaphoric relations, and processing of phenomena concerning sentence information structure (topic-focus articulation), see Novák et al. (2018).

Subsequently, experiments were performed on the automatically pre-processed texts. In the individual project phases, they were focused on different language layers and phenomena. The experiments were also carried out using various machine learning scenarios. A detailed evaluation of resulting applications is given in Section 3.5.

Finally, the EVALD applications themselves have been created. They are available online as a public web service of the LINDAT/CLARIN server,[9] or to download and run locally in a Docker container.[10]

3.2. Datasets

The EVALD system is based on supervised learning and as such relies heavily on labeled data. The three versions of the system have been trained using three different collections of labeled texts – collections of essays manually annotated with the overall and/or surface coherence grade: (1) L1, (2) L2, and (3) NIE dataset. In addition, EVALD takes advantage of a great number of unlabeled texts collected in the SYN dataset. Table 1 gives an overall statistics of these datasets and their short description follows.

L1 dataset. Manually labeled essays written by native speakers of Czech come from the Skript2012/AKCES 1 corpus (Šebesta et al., 2016). It comprises 1,694 students' essays written during classes of Czech language at elementary and high schools. The texts were manually labeled by coherence marks (1–5) by the authors of Novák et al. (2017).

L2 dataset. The L2 dataset was compiled from three language acquisition corpora: (1) the MERLIN corpus (Boyd et al., 2014), (2) CzeSL-SGT/AKCES 5 (Šebesta et al., 2014)[11], and (3) the Skript2012/AKCES 1 corpus (Šebesta et al., 2016). Data from these three corpora form a dataset with essays manually labeled by coherence marks on the scale of A1–C2, i.e. the full scale defined in CEFR. As the L2 sources (1) and (2) do not contain any C2 texts, the C2 class was substituted by texts taken from the L1 Skript2012/AKCES 1 corpus (see Novák et al., 2017 for details).

NIE dataset. For the task of distinguishing texts that do not even reach the lowest level A1, we newly obtained a dataset from the National Institute for Education coming from the examinations for permanent residence of foreigners in the Czech Republic.[12] In total, we deal with around 200 texts in each of the following classes: 0 (worse than A1), borderline, and A1. Documents here are in average much shorter than in the L1 and L2 datasets.

Each text was evaluated independently by three teachers – evaluators. Unlike the previous two datasets, these texts were not evaluated in terms of text coherence but they were assigned an overall mark – due to the general rules concerning the permanent residence examination. The A1 and 0 classes contain texts where all three

[9] `https://lindat.mff.cuni.cz/services/evald`
`https://lindat.mff.cuni.cz/services/evald-foreign`

[10] See `http://ufal.mff.cuni.cz/evald/documentation` for detailed installation instructions.

[11] Also for texts from CzeSL-SGT/AKCES 5, the coherence marks were manually added by the authors of Novák et al. (2017).

[12] We gratefully thank Jitka Cvejnová and Kamila Kolmašová for their kind and excellent cooperation.

evaluators agreed on the same evaluation, the borderline class includes texts where the evaluation was not unanimous.

In Section 3.5, besides experiments with classification into all three classes, we also carry out experiments that work with the A1 and 0 classes only. For these experiments, we prepared the *NIE-2* dataset, which is a limited version of NIE that does not contain borderline examples.

Example 1 demonstrates a text from the A1 class, Example 2 from the borderline class and Example 3 from the 0 class.

(1) *Ahoj Martine. Mám problem, jsem nemocná a potrebuju pomoc, Můžeš dojit do obchodu a nakupit mne 2 kg cibule, 4 kusy kolači s makem a tvarohem a 1 karabičku mleka. Čekam na tebe dnes odpoledne ve 14:45, bydlim na HRADEBNÍ 8, když něco zavolej. Moc ti děkuju, mej se heský. S pozdravem XXX*

"Hi Martin. I have a problem, I am sick and I need help, You can get to the store and buy me 2 kg of onion, 4 pieces of cake with poppy seeds and curd and 1 milk bucket. I'm waiting for you this afternoon at 14:45, staying at HRADEBNÍ street 8, if something call me. Thank you very much. Regards XXX"[13]

(2) *Dobrý den p. Holý. Oznamuji Vám, že na schůzku nemužu přijit proto ze jsem povolan na Policie ČR o 13.00 kvuli svedětstvi. Mužeme-li objednat schůzku na čtvrtek o 11.30 u Vás v ofisu? S pozdravem XXX*

"Hello Mr. Holý. I would like to inform you that I cannot come to the meeting because I am called to the Police of the Czech Republic at 13.00 because of the testimony. Can we order an appointment on Thursday at 11.30 am at your office? Regards XXX"

(3) *Ahoj Alenu. Jste nemocná. POTŘEBujes POMOC. Kup M PROSim: 4 kusy housek. MůžES PŘiJiT v 20.45 hod. DEKUJi PĚKNĚ. XXX*

"Hi, Alena. You're sick. You need help. Buy me please: 4 pieces of buns. You can come at 20.45. Thank you. XXX"

SYN dataset. As described in Novák et al. (2019), the EVALD system incorporates features that use large unlabeled data of Czech texts in two forms: (1) as a source for a language model, and (2) as a source for density estimations of other features. We use the SYN collection (version 4) of the Czech National Corpus (Hnátková et al., 2014). The language model was trained on the entire dataset, i.e. 275 million sentences, while the density estimation was counted on approx. 7% of the full SYN data.

[13] English translations under the examples are illustrative – they cannot cover the mistakes e.g. in spelling or morphology in the Czech original.

3.3. Text analysis of samples from the NIE dataset

As already mentioned, the collected texts (both by native and non-native speakers) were first subjected to a detailed linguistic analysis across the individual language layers. We dealt with phenomena from the following linguistic areas: spelling, morphology, lexicology, syntax, semantics and discourse. We conducted a complex analysis of the texts and we examined how deficiencies in certain language layer may disrupt the overall coherence of the resulting text.

Texts written by native speakers of Czech were evaluated on the scale 1–5, texts by non-native speakers were evaluated on the scale A1–C2 according to CEFR. However, it turned out to be practical to focus further on the bottom class A1 (reaching A1 is e.g. a condition for permanent residence in the Czech Republic). In practice, it is thus often necessary to state the boundary between texts already having A1 and texts that do not yet reach this level (i.e. class 0).

In this section, we focus on the analysis of these lower-level texts, specifically on the differences between texts reaching the A1 level, texts below A1 (we mark them as class 0) and borderline texts (i.e. texts between A1 and 0). It turned out that linguistically interesting is especially the difference between the A1 and the borderline level.

From the linguistic point of view, the texts are rather simple. It is, for example, very difficult to observe here linguistic phenomena from the higher language levels (such as anaphoric and coreference chains or pragmatic relations), which are practically absent in these texts. On the other hand, the texts differ from each other in lower language phenomena (concerning e.g. spelling, vowel length or writing voiced vs. voiceless consonants) that are already acquired by more advanced learners and that are thus not worth observing at higher CEFR levels.

The texts from the 0 class appeared to be rather distinct from the other two sets. These texts (as demonstrated in Example 3) contain many mistakes already in spelling, morphology or lexicology. The authors of these texts have problems already with the Czech writing system (cf. mixing of uppercase and lowercase letters within a single word, e.g. *POTŘEBujes, PROSim, MůžES, PŘiJiT, DEKUJi*). They make basic mistakes in spelling (connected with phonological issues, cf. missing vowels such as *m* instead of *mi*; mistakes in diacritics such as *potřebujes* instead of *potřebuješ*, *můžes* instead of *můžeš*, *dekuji* instead of *děkuji*, or in vowel length such as *prosim* instead of *prosím*, *prijit* instead of *přijít*) and morphology (cf. use of wrong noun cases such as *Alenu* instead of *Aleno*; mistakes in verbal grammatical categories such as mixing of person, cf. *Jste nemocná.* instead of *Jsem nemocná.*, *potřebujes* instead of *potřebuji* etc.).

The content of sentences in these texts is often hard to interpret, as they contain unrecognizable words (cf. examples from other text samples like *PANE RUŽiČKA PŠiLECE(?)*; *koupit 15 dkg sý(?), 1 kus maslo krava 3 kus rohlič(?)*, or *JSE OMLOUVAM NEMUŽU PŠIEC(?)*). Besides many formal and grammatical mistakes (and also due to them), the overall comprehensibility of these texts is disrupted – even the native speakers of Czech have problems to understand the main message of the text. The authors' communicative competence in the Czech language is thus rather low and it can be assumed that the authors could have problems with basic understanding of common (language) situations.

The other two sets (A1 vs. borderline class) appeared to be linguistically more interesting. Their overall comprehensibility was (despite the errors present) rather good and both of them contained similar error types. The problematic issues concerned especially the vowel length (cf. *problem* instead of *problém* or *dojit* instead of *dojít* in Example 1, *nemužu* instead of *nemůžu*, *přijit* instead of *přijít* or *povolan* instead of *povolán* in Example 2), confusion of voiced and voiceless consonants (*heský* instead of *hezký* in Example 1), confusion of uppercase and lowercase letters (cf. *PŘijDU, ZiTRA* or *PROSiM* from other text samples), and, rather rarely, punctuation (using a comma instead of full stop at the end of sentences, missing of a comma in subordinating constructions such as *když něco, zavolej* in Example 1 or *na schůzku nemužu přijit, proto ze jsem povolan na Policie ČR* in Example 2). Sporadic problematic issues appeared in these text sets also with higher language phenomena, namely with discourse connectives (cf. a wrong use of a connective *-li* "if" in the sentence *Mužeme-li objednat schůzku na čtvrtek o 11.30 u Vás v ofisu?* or a wrong form of the connective *protože* "because" used as *proto ze* in Example 2).

The borderline class contained also lexical errors, especially using inappropriate words and phrases (cf. the wrong phrase *objednat schuzku* "to book a meeting" instead of *sjednat schůzku* "to arrange a meeting", or stylistically inappropriate *mam bouračku* "I have a smash" used in an official letter instead of *mám nehodu* "I have an accident").

The difference between A1 and borderline sets was thus rather in the frequency of errors. Despite these errors, the main message of the texts was understandable (interpretable) and their authors proved to have a basic communicative competence in Czech.

3.4. Linguistic features

Based on a deep linguistic analysis of all datasets, we established a list of linguistic features according to which it is possible to evaluate new texts automatically, more precisely to sort them into the classes associated with individual grades.

We created such language features that EVALD is capable to track automatically in newly inserted texts and then to compare them to the already known ones (training datasets).

Currently, the EVALD application monitors approximately 200 language features from both lower and higher language levels. Additionally, the application works with the readability of a text, language model and density estimates.

As demonstrated in Section 3.3, the texts from the NIE dataset are different from the other two datasets containing texts by native speakers and non-native speakers of Czech. The low-level NIE texts are written by complete Czech language beginners, which is projected in their (non-)acquisition of various linguistic means and which should be thus reflected in the feature selection in automated scoring as well. It turned out, for example, that discourse or coreference features are not so much suitable or beneficial for this task, as they practically do not appear in the NIE texts. On the other hand, these texts differed in lower language features, especially in spelling. They contained differentiating features that were not worth observing in previous two datasets (the features concerning e.g. correct writing of letters in Czech, the use of long vs. short vowels etc. has been already adopted by the native speakers or more advanced earners of Czech). Based on the text analysis presented in Section 3.3, we thus designed new spelling features for automated scoring, not used in EVALD before. In the overview below, these features are put in **bold**.

The individual features can be arranged to multiple categories that are listed in the following overview (please note that all absolute numbers are normalized to the length of the text).

Spelling: number of typos, punctuation marks, accented characters and diphthongs, **ratio of uppercase and lowercase letters, capital letters used elsewhere than at the beginning of a word, the number of lowercase letters after the full stop, the number of *ú*, the number of *ú* used elsewhere than at the beginning of a word, occurrence of two (or more) vowels next to each other in a single word (except for diphthongs *au*, *ou*), the occurrence of soft consonants with *y*/*ý*, the occurrence of selected hard consonants with *i*/*í*, number of occurrences of the sequence *pje*, number of occurrences of two long syllables next to each other, number of characters other than letters, numbers, and punctuation** etc.

Vocabulary: richness of vocabulary expressed by several measures, average length of words, percentage of lemmas *být* [to be], *mít* [to have] and the most frequent lemma.

Morphology: percentage of individual cases, parts of speech, degrees of comparison, verb tenses, moods, etc.

Syntax: average sentence length, percentage of sentences without a predicate, number and types of dependent clauses, structural complexity of the dependency tree (number of levels, numbers of branches at various levels), distributions of functors and part-of-speech tags at the first and second positions in the sentences, etc.

Topic-focus articulation: variety of rhematizers (focalizers), number of sentences with a predicate on the first or second position, percentage of (contrastive-) contextually bound and non-bound words (more precisely: nodes in the tectogrammatical tree), percentage of subject-verb-object and object-verb-subject sentences, position of enclitics, percentage of coreference links going from a topic part of one sentence to the focus part of the previous sentence, etc.

Coreference: proportion of 21 different pronoun subtypes, variety of pronouns, percentage of null subjects and several concrete (most commonly used) pronouns, number of coreference chains (intra-sentential, inter-sentential) and distribution of their lengths, etc.

Discourse: quantity and variety of discourse connectives (intra-sentential, inter-sentential, coordinating, subordinating), percentages of four basic classes of types of discourse relations (temporal, contingency, contrast, expansion) and numbers of most frequent connectives, etc.

Readability: various readability measures combining a number of characters, syllables, polysyllables and sentences (Flesch-Kincaid Grade Level Formula, SMOG index, Coleman–Liau index, etc.)

Language model: prob. estimates of the texts with respect to an n-gram language model trained on the SYN dataset.

Density estimates: prob. estimates of all the other features with respect to SYN dataset.

3.5. Evaluation

The aim of the following experiments is to measure the overall performance of EVALD as well as contribution of individual feature sets.

We have tried multiple supervised machine learning methods[14] to train EVALD models: (1) stochastic gradient descent optimization using various loss types, (2) support vector machines with the radial basis function kernel, and (3) random forests. Apart from classification methods, we also take advantage of regression variants of the aforementioned methods. Before training, we mapped the labels of grades to integers, to which real-valued predictions of the regression models are also discretized in the end. Furthermore, we have also varied the values of hyperparameters specific to each of these methods, e.g. regularization and class balancing hyperparameters. Following the random search strategy, thousands of machine learning configurations have been tested. For the final evaluation, we pick those performing the best on the development portion of the data (see below).

The datasets are too small to be split to separate training, development and test portions. Instead, we perform a cross-validation. A standard k-fold cross-validation, however, leaves no room for development data, necessary for tuning a machine learning method and its hyperparameters. We thus take advantage of the *k*l-fold cross-validation*. It is a nested procedure, where in the outer loop the data are split into k folds and one of them is considered a test set in each iteration. In the inner loop, a data portion comprising the remaining k–1 folds is split into other l folds. The procedure then goes over them and takes one fold as a development set and the rest as a training set. The result of the inner loop is a machine learning configuration tuned on the development set. Back in a given iteration of the outer loop, this configuration is subsequently used to train a model on all k–1 folds and to test it on a corresponding test fold. Predictions on test data are thus collected in the outer loop, possibly

[14]Implemented in the Scikit-learn library (Pedregosa et al., 2011).

	L1		L2		NIE		NIE-2	
	Macro-F	**Acc**	**Macro-F**	**Acc**	**Macro-F**	**Acc**	**Macro-F**	**Acc**
Majority class	12.1	43.3	5.2	18.6	16.8	33.6	33.4	50.3
EVALD ⊖ add. spelling	53.4	63.9	65.6	68.2	52.1	53.3	78.2	78.2
EVALD	56.1	64.6	66.0	68.3	52.0	53.1	78.7	78.7

Table 2. Effect of additional spelling features to performance of EVALD on various datasets. The score in gray indicates that for this dataset the complete EVALD system is significantly better than EVALD without new spelling features.

using a different learning configuration in each of its iterations. In particular, all the following experiments are run in 10*9-fold cross-validation.[15]

Picking a single configuration as an output of the inner loop may introduce some noise for small datasets. In order to make the results of the analysis more reliable, we do the evaluation in the outer loop with an ensemble of 5 best configurations for each fold, instead of using just a single configuration. An essay is thus assigned the grade that earns the majority out of 5 possible votes.

We report performance of the tested systems using two metrics: (1) accuracy, and (2) macro-averaged F-score. Accuracy is a standard measure for classification tasks. However, it becomes less suitable for datasets with skewed distribution of classes, e.g. the L1 dataset. Therefore, we use *macro-averaged F-score* as a primary measure for the following experiments. It calculates an F-score value for every class in the dataset and then averages it over the classes.

Effect of additional spelling features. In the first experiment, we evaluate the system that takes advantage of all the features presented in Section 3.4.

We compare its performance on all labeled datasets with two baselines. The first one exploits the complete set of features except for the newly added spelling features that were designed with respect to the properties of the NIE dataset. This corresponds to the feature set as presented in (Novák et al., 2019). And the second, "majority class" baseline labels each essay with a most frequent grade.

Table 2 shows that EVALD with the new spelling features slightly outperforms the system that does not include them. The only exception is the NIE dataset, which is surprising, since these features were primarily designed to target this dataset. The

[15]Note that in (Rysová et al., 2017; Novák et al., 2017; Novák et al., 2018) we used a standard 10-fold cross-validation, since we did not tune the machine learning configuration. We introduced such tuning in (Novák et al., 2019) and performed it using a 5-fold cross-validation with non-overlapping development and test portions. In comparison to that approach, the currently used 10*9-fold cross-validation is computationally more demanding. But on the other hand, the entire dataset can be utilized as the development and test set. Furthermore, there is no doubt that the currently used nested validation is fair, without any, even an indirect, influence between the portions designated for training, tuning and testing.

	L1		L2		NIE		NIE-2	
	Macro-F	**Acc**	**Macro-F**	**Acc**	**Macro-F**	**Acc**	**Macro-F**	**Acc**
EVALD	56.1	64.6	66.0	68.2	52.0	53.1	78.7	78.7
⊖ spelling	54.4	61.9	65.1	67.3	53.4	53.6	77.9	77.9
⊖ vocabulary	53.1	63.0	65.5	67.7	51.8	52.8	76.7	76.7
⊖ morphology	53.2	62.5	63.9	66.1	50.9	51.3	78.9	78.9
⊖ syntax	57.8	66.5	66.6	69.2	53.1	53.3	77.2	77.2
⊖ readability	55.1	65.2	65.0	67.5	50.7	52.0	77.9	77.9
⊖ connectives	55.4	65.5	66.0	68.2	52.1	53.3	80.2	80.2
⊖ coreference	55.9	65.0	64.9	67.1	53.0	54.4	79.4	79.4
⊖ TFA	54.7	65.0	65.0	67.5	53.2	54.1	79.6	79.7
⊖ lang. model	54.9	65.0	65.1	67.3	53.3	54.4	77.7	77.7
⊖ dens. estim.	54.8	61.9	64.5	66.7	52.7	52.8	78.4	78.4
⊖ unlabeled	55.3	62.3	64.7	66.8	54.2	54.6	76.4	76.5
⊖ unlabeled, coherence	53.6	62.3	67.1	66.8	54.1	54.6	79.7	76.5

*Table 3. Results of the feature ablation analysis using the final EVALD system, including the new spelling features. Whereas in the upper part of the table, one category of features is removed at the time, it is a bigger group of categories in the lower part (*unlabeled*: language model and density estimation features;* coherence*: connective, coreference and TFA features). The scores in gray indicate that the performance of the complete EVALD system and its particular ablation variant is significantly different.*

improvement on NIE-2 reveals, however, that it is likely a result of the mixed nature of the borderline class. All in all, it is necessary to mention that the difference caused by removing the new spelling features from EVALD is statistically significant[16] only on L1.

The highest scores are naturally achieved on the NIE-2 dataset, which consists of solely two classes. Hence, it is important to take the majority class baseline into account. With this respect, the biggest improvement of 60 F-score points is observed on the L2 dataset.

Ablation of feature categories. In the second experiment, we investigate what is the contribution of individual feature categories to the overall quality of EVALD. We contrast the system based on the full feature set with its modifications where one feature category is left out at the time. This is repeated for each feature category. Note that each density estimation feature is derived from a single original feature. In other words, a group of density estimation features derived from a particular category encodes in some way the information captured by the category itself. Therefore, together

[16]Statistical significance was calculated by paired bootstrap resampling (Koehn, 2004) at p-level $p \leq 0.05$.

with each category we decided to remove the corresponding density estimation features, too.

Table 3 shows results of the ablation analysis for all the datasets. Although behavior of the model changes with different datasets, some common properties are evident, especially between L1 and L2. Performance gaps between the complete EVALD and its ablation variants are not large. Indeed, they are not statistically significant for most of the variants and datasets. EVALD thus seems to be robust enough, i.e. removing one category from the complete feature set does not change the results too much. Some of the categories, however, seem to be more important than the others, because we observe lowest numbers after their removal, e.g. vocabulary and morphology features. On the other hand, syntactic features appear to harm the performance on the L1 and L2 datasets, as EVALD performs even better without them.

On both NIE datasets, better scores are achieved after exclusion of discourse-related features. Due to a short average length of its documents and low Czech competence of their authors, there is most likely no room for such features to activate. In addition, recall that for the NIE dataset, we predict an overall grade, not just a grade for surface coherence.

As no obvious conclusion can be drawn when leaving out a single category, we proceeded in the ablation analysis by removing two important groups of categories. First, we left out the features that are based on the unlabeled data, i.e. language model and density estimation features. Second, apart from these, we also excluded all coherence-related features, i.e. topic-focus articulation, coreference, and discourse features. Together with the full EVALD system, these two configurations represent the three major steps in development of EVALD.

The lower part of Table 3 shows the results of these ablation experiments. Removal of SYN-based features causes a drop in accuracy, but it does not change any further after subsequent removal of coherence-related features. More important macro-averaged F-score statistics paints a different picture, though. Whereas inclusion of coherence-related and SYN-based features gradually improves the prediction quality for L1, it rather harms performance of EVALD for the other three datasets. This observation accords with the findings in (Novák et al., 2018, 2019), even if it is more emphasized most likely due to tuning of learning configuration and a more reliable cross-validation technique. Novák et al. (2018) has concluded that the effect of coherence-related features is more pronounced for essays written by native speakers of Czech, since their language competence is high enough to disclose coherence-related nuances. Similarly, Novák et al. (2019) has shown that features based on the SYN corpus are also more powerful on L1 essays. Such behavior has been justified by higher similarity of texts from the SYN corpus and L1 essays, as both were authored by native speakers of Czech.

Figure 3 illustrates the previous ablation analysis in a greater detail – on F-scores related to individual grades. Prediction naturally works best for boundary grades (i.e. 1 and 5 for L1; A1 and C2 for L2; A1 and 0 for NIE). Other grades are more difficult to

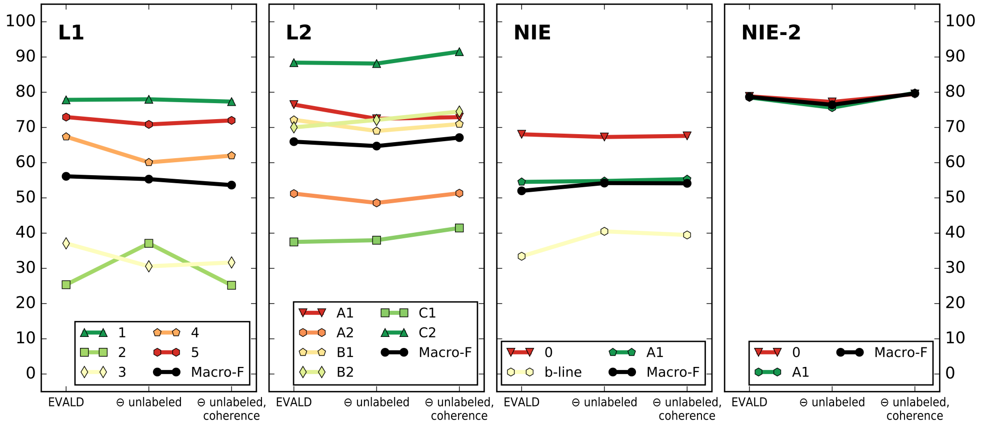

Figure 3. Changes in F-scores for individual grades achieved by EVALD when removing important feature category groups.

distinguish. Interestingly, exclusion of SYN-based features on the L1 dataset results in performance drops for rather negative grades but gains for positive grades. The most plausible explanation is that coherence-related features, which are more effective in distinguishing high-quality texts, play a stronger role in such a model. After their subsequent removal, performance for positive grades deteriorates again.

4. Conclusion

In the paper, we introduced EVALD (Evaluator of Discourse), an application for automated essay scoring in Czech. We described the process of creation of the tool and we presented its individual versions aimed at slightly different purposes.

In the linguistic part of the paper, we presented the NIE dataset – texts from the A1–0 classes gained from the National Institute for Education – and we carried out a detailed analysis of a sample of them. It turned out that these lower-level texts written by non-native speakers are linguistically rather simple. The texts from the 0 class contain many mistakes e.g. in spelling (often connected with phonological issues), morphology or lexicology. Their authors often have problems already with the Czech writing system (e.g. they mix uppercase and lowercase letters within a single word). The overall comprehensibility of the A1 and borderline texts was, on the other hand, rather good (despite the errors present). The problematic issues included especially the vowel length, confusion of voiced and voiceless consonants, and use of inappropriate words and phrases. Sporadic errors concerned higher language phenomena, namely a wrong use or wrong form of discourse connectives. The A1 and borderline texts thus differed especially in the frequency of errors.

Based on the analysis, we designed new spelling features that tried to highlight the biggest issues, so that EVALD can distinguish between the classes more easily. The experiments, however, showed that these features are surprisingly more helpful on L1 and L2, the other two datasets that EVALD operates with.

In the experiments, we also explored overall properties of the EVALD system. We observed no strong characteristic that would be common across all the datasets. Nevertheless, the present analysis confirmed (and even highlighted) the findings from previous publications on automated essay scoring in Czech: coherence-related features and the features based on great amount of unlabeled texts play an essential role in evaluation of L1 essays, i.e. the essays written by native speakers of Czech. On texts written by foreigners, these features are less important.

Acknowledgments

The research has been supported by the Ministry of Culture of the Czech Republic (project No. DG16P02B016 *Automatic Evaluation of Text Coherence in Czech*). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

Bibliography

Ajay, Helen B., P. I. Tillett, and Ellis B. Page. Analysis of Essays by Computer (AEC-II). *Final Report to the National Center for Educational Research and Development, U.S. Department of Health, Education, and Welfare (Project No. 80101)*, page 231, 1973.

Al-Jouie, Maram F. and Aqil M. Azmi. Automated Evaluation of School Children Essays in Arabic. *Procedia Computer Science: Arabic Computational Linguistics*, 117:19 – 22, 2017.

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1281–1288, Reykjavík, Iceland, 2014. European Language Resources Association.

Broda, Bartosz, Bartłomiej Nitoń, Włodzimierz Gruszczyński, and Maciej Ogrodniczuk. Measuring Readability of Polish Texts: Baseline Experiments. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 573–580, Reykjavik, Iceland, 2014. European Languages Resources Association (ELRA).

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. Computer analysis of essays. 1998.

Castro-Castro, Daniel, Rocío Lannes-Losada, Montse Maritxalar, Ianire Niebla, Celia Pérez-Marqués, Nancy C. Álamo-Suárez, and Aurora Pons-Porrata. A Multilingual Application for Automated Essay Scoring. In *Advances in Artificial Intelligence – IBERAMIA 2008*, pages 243–251, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

Foltz, Peter W., Darrell Laham, and Thomas K. Landauer. The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944, 1999.

Hnátková, Milena, Michal Křen, Pavel Procházka, and Hana Skoumalová. The SYN-series Corpora of Written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

Ishioka, Tsunenori and Masayuki Kameda. Automated Japanese Essay Scoring System Based on Articles Written by Experts. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 233–240, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Koehn, Philipp. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, 2004. Association for Computational Linguistics.

Larkey, Leah S. Automatic Essay Grading Using Text Categorization Techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 90–95, New York, NY, USA, 1998. Association for Computing Machinery.

Lemaire, Benoit and Philippe Dessus. A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, 24(3):305–320, 2001.

Novák, Michal, Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. Incorporating Coreference to Automatic Evaluation of Coherence in Essays. In *Statistical Language and Speech Processing*, number 10583 in Lecture Notes in Computer Science, pages 58–69, Cham, Switzerland, 2017. Springer International Publishing.

Novák, Michal, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. Topic–Focus Articulation: A Third Pillar of Automatic Evaluation of Text Coherence. In *Advances in Computational Intelligence*, number 11289 in Lecture Notes in Computer Science, pages 96–108, Cham, Switzerland, 2018. Springer International Publishing.

Novák, Michal, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. Exploiting Large Unlabeled Data in Automatic Evaluation of Coherence in Czech. In *Text, Speech, and Dialogue*, number 11697 in Lecture Notes in Computer Science, pages 197–210, Cham, Switzerland, 2019. Springer International Publishing.

Page, Ellis B. The Imminence of… Grading Essays by Computer. *Phi Delta Kappan*, 47(5):238–243, 1966.

Page, Ellis B. The Use of the Computer in Analyzing Student Essays. *International Review of Education*, 14(2):210–225, 1968.

Page, Ellis B. and Nancy S. Petersen. The Computer Moves into Essay Grading: Updating the Ancient Test. *Phi Delta Kappan*, 76(7):561–565, 1995.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Popel, Martin and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

Rysová, Kateřina, Magdaléna Rysová, and Jiří Mírovský. Automatic evaluation of surface coherence in L2 texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing ROCLING XXVIII (2016)*, pages 214–228, Taipei, Taiwan, 2016. National Cheng Kung University, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). ISBN 978-957-30792-9-3.

Rysová, Kateřina, Magdaléna Rysová, Jiří Mírovský, and Michal Novák. Introducing EVALD – Software Applications for Automatic Evaluation of Discourse in Czech. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 634–641, Varna, Bulgaria, 2017. INCOMA Ltd.

Šebesta, Karel, Zuzanna Bedřichová, Kateřina Šormová, et al. AKCES 5 (CzeSL-SGT), 2014. data/software, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Šebesta, Karel, Hana Goláňová, Jana Letafková, et al. AKCES 1, 2016. data/software, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Wild, Fridolin, Christina Stahl, Gerald Stermsek, Yoseba Penya, and Gustaf Neumann. Factors Influencing Effectiveness in Automated Essay Scoring with LSA. In *Proceedings of the Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, pages 947–949, Amsterdam, The Netherlands, 2005. IOS Press.

Zupanc, Kaja and Zoran Bosnić. Advances in the Field of Automated Essay Evaluation. *Informatica*, 4(39):383–396, 2015.

Address for correspondence:
Kateřina Rysová
rysova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic