Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Machine learning for spelling acquisition: How accurate is the prediction of specific spelling errors in German primary school students?

Richard Boehme [a,*,1], Stefan Coors [a,b,1], Patrick Oster [a], Meike Munser-Kiefer [a], Sven Hilbert [a]

[a] University of Regensburg, Department of Human Sciences, Sedanstrasse 1, 93055, Regensburg, Germany
[b] Ludwig Maximilians University, Department of Statistics, Ludwigstrasse 33, 80539, Munich, Germany

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Multi-label classification
Spelling
Literacy acquisition
Primary school
Elementary school

ABSTRACT

In Germany (similar to other countries), 30 % of students demonstrate insufficient spelling skills at the end of primary school – partly owing to the challenge for teachers to manage a variety of students' learning needs. Digital tools using Machine Learning can enable teachers to individualise students' learning. However, there are still no suitable approaches for demographics of students who are not yet proficient in spelling.

With an aim to adapt Machine Learning for students of all proficiencies, we investigate how accurately specific spelling errors can be predicted across different skill levels, and what the content-related reasons for incorrect predictions are.

To that end, we developed a web application to record the spelling efforts of $N = 685$ first- and second-graders in Bavaria, Germany. A total of 18,133 different misspellings were recorded. Using this dataset, we trained six Machine Learning models and compared their performances to predict misspellings.

Comparing all Machine Learning models employed in this work, the Random Forest performed best on average as a predictor of spelling errors. Errors at the syllable- and morpheme-levels were predicted best, and errors at the basic phoneme-grapheme-level were predicted slightly less accurately. Confusions often concerned cases that are considered linguistically ambiguous or occurred in complex error entanglements. The implications of these results are discussed.

1. Introduction

Approximately 130 million people speak German as their first or second language, making it the most-spoken native language in the European Union and one of the most-spoken languages worldwide (German Federal Statistical Office, 2022). The acquisition of skills in reading and writing in this language (just as in most other languages) is critical for educational success and successful participation in modern society (Göpferich & Neumann, 2016). In particular, the acquisition of spelling is a very demanding task, where a large proportion of children lag behind the skill level that would be expected for a given age. For example, a 2016 study from the Institute for Educational Quality Improvement (IQB) examining trends in student achievement shows that on average in Germany, more than 22 % of children (including 6 % of children with officially stated special needs) achieve neither the normal standard nor the minimum standard by the end of fourth grade (Stanat, Schipolowski, Rjosk, Weirich, & Haag, 2017). Results from the more recent 2021 IQB study (Stanat et al., 2022) show that this proportion increased to over 30 %. According to the authors, it is likely that the worsening is not exclusively the result of the COVID-19 pandemic. One reason for the poor performance in Germany is seen in the great diversity of the students' prerequisites for learning, which can result in a variety of different learning needs (Stanat, Schipolowski, Rjosk, Weirich, & Haag, 2017).

Primary school teachers are challenged to identify the learning needs of their students and to provide adaptive learning opportunities for all of them in terms of formative assessment (for the construct, see Black & Wiliam, 2009; for empirical evidence, see Hebbecker & Souvignier, 2018). This not only poses an issue of time, but also requires strong diagnostic skills (Black & Wiliam, 2009; for an overview, see Kärner, Warwas, & Schumann, 2021). International findings show that teachers often lack knowledge of basic language constructs for this purpose (for Canada, England, New Zealand and the USA, see Washburn, Binks-Cantrell, Joshi, Martin-Chang, & Arrow, 2016; for Finland, see

* Corresponding author.
E-mail address: richard.boehme@ur.de (R. Boehme).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.caeai.2024.100233
Received 21 November 2023; Received in revised form 29 April 2024; Accepted 1 May 2024
Available online 9 May 2024

Aro & Björn, 2016; for Germany, see Corvacho del Toro, 2013) and that the accuracy of their diagnostic judgments is limited (for an overview, see Kärner et al., 2021).

A successful start in literacy acquisition has a positive long-term effect on later performance, and early deficits are difficult to make up for later (see longitudinal study by Sparks, Patton, & Murdoch, 2014). Moreover, students who show early signs of lagging behind can benefit from individualised exercises to improve their spelling. To that end, procedures to diagnose students' individual spelling abilities may help identify individual spelling deficits and, thus, facilitate appropriate application of such exercises (c.f., Lee, Chung, Zhang, Abedi, & Warschauer, 2020).

In this context, approaches from the field of Artificial Intelligence in Education (AIED) aiming to improve learning processes have been discussed and tested since the early 1980s (Holmes, Bialik, & Fadel, 2019). One such approach is the development of Intelligent Tutoring Systems (ITSs), which provide students with tailored tasks, support, and feedback, as well as provide teachers with diagnostic information. These are key components of formative assessment (Black & Wiliam, 2009). However, despite the potential of these approaches to enhance the learning process, their influence in the field of education and educational research has, to date, been rather minimal (cf., Hilbert et al., 2021).

2. Machine learning in the educational sciences

Parts of the computer-assisted learning processes described above are directly related to Machine Learning (ML). Likewise, the statistical approach in our study is inseparably connected to this long-existing field. However, the approach taken here utilises a combination of computer-supported learning, inclusion of knowledge and theories from educational psychology, and ML techniques for the analysis of learning processes. Thus, this work can be placed within the realm of learning analytics (even though there are some points of contact with AIED, as mentioned above, as well as educational data mining). For an overview of the similarities and distinctions between these fields, see the review of Rienties, Køhler Simonsen, and Herodotou (2020). For simplicity, we will refer to the analytical approach taken here as ML and discuss its role in the educational sciences, particularly with regard to quantitative data analysis in research.

Even though ML as an analysis tool in research has been embraced by many fields of behavioural sciences – such as psychology (Stachl et al., 2020) or the health sciences (Chen, Liu, & Peng, 2019) – the use of ML models in the educational sciences has to date been much sparser. This is currently changing, as we observe an immense increase in available digital data on all levels of the educational system (Jarke & Breiter, 2019), which in turn has made the use of data-intensive models much more feasible within the last decade. Innovative research on reading and writing acquisition has also resulted in a variety of ML-based models and tools, such as for the prediction of reading comprehension through lexical and syntactic features (Sinclair, 2020) or readability formulas (François & Miltsakaki, 2012).

Moreover, growing fields of research concerning the use of ML techniques for retention prediction (e.g., Delen, 2010; Jimenez, Paoletti, Sanchez, & Sciavicco, 2019) and automated essay scoring (AES) have evolved substantially. In particular, AES has received increased attention and resulted in a wide range of useful applications (Chai & Gibson, 2015; Ke & Ng, 2019), which is partially attributable to the close relationship to the field of Natural Language Processing (NLP) and its role in education (see Alhawiti, 2014).

Recent ML approaches have used deep learning algorithms, due to their superior capability to handle unstructured data, such as texts. Prominent examples are translation engines like *DeepL* or *Google Translate*, which are based on neural networks and use deep learning models to solve their tasks. Additionally, neural networks can learn a similarity mapping of, e.g., words into the real number space – or a so-called embedding. One example is the *Word-to-Vector* embedding (Word2Vec; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which consists of a pre-trained two-layer neural network to detect not only similarities between words but also connections between them – for example, if they frequently appear close together within a sentence. This approach transforms each word into a single numeric vector, where the numerical distances of related word vectors are close together. Another popular algorithm is *Global Vectors for Word Representations* (GloVe; Pennington, Socher, & Manning, 2014), which uses co-occurrences of words for training. Both approaches are context-free, meaning that a single word, such as 'fall', will have only a single representation. Newer approaches, such as the popular *Generative Pre-trained Transformer* (GPT; Brown et al., 2020) or *Bidirectional Encoder Representations from Transformers* (BERT; Devlin, Chang, Lee, & Toutanova, 2018) provide contextual embeddings, so that 'it was a hard fall' will have a different embedding than 'the leaves turn orange in fall'. Transformer models have also been adapted to solve more specialised tasks, such as predicting typos (e.g., CharacterBERT; see Boukkouri et al., 2020).

Depending on the task and algorithm, different transformations must be applied to the raw data. For example, automatic translation between languages often requires a more nuanced approach than a simple word-for-word translation, largely due to different language grammar structures. Rather, content and context are key elements to achieve a natural-sounding result. Hence, not only must the learning algorithm be capable of recognising complex patterns and their dependency structures, but the text data (converted to numeric values) must not lose these structures through conversion. One way to achieve this is by mapping both single words as well as combinations of several words as sentence components. These sentence components are so-called *n*-grams, where *n* equals an arbitrary natural number. While 1-grams reflect a one-word-one-number vocabulary, 2-grams decompose a sentence into all word pairs contained in it. The sentence "I like writing a lot." results in the following possible *n*-grams, up to 3-grams.

1-grams: I – like – writing – a – lot.
2-grams: I like – like writing – writing a – a lot.
3-grams: I like writing – like writing a – writing a lot.

Despite strong connections to this field, the use of ML for the analysis of spelling acquisition presents unique challenges, as we show in the following section with regard to modelling requirements.

3. German orthography and acquisition processes that must be modelled

Given a sufficiently large database, NLP can function almost completely unsupervised and bottom-up. In contrast, theoretical knowledge and a comprehensive set of rules are almost inevitably necessary for ML-based prediction of spelling errors in children. These requirements arise from the nature of the data source: Beginners both (1) produce less data and they, quite naturally, (2) produce a large number of incorrect spellings, with different frequencies of errors, from different areas with different realisations, and for different developmental reasons.

Spelling acquisition requires that children learn to associate written symbols with spoken language – neither of which carries inherent meaning. Depending on the language, the symbols represent different units, such as consonants (as in Arabic or Hebrew), syllables (as in Japanese), or so-called phonemes, i.e., speech sounds (as in German, English or Spanish). Written languages belonging to the latter group are alphabetic scripts.

A characteristic of alphabetic scripts is that a so-called grapheme – which is a single letter (e.g., ⟨h⟩ for /h/ as in *Hut*, /huːt/, Eng. *hat*) or a set of letters (in German the maximum is three, e.g., ⟨sch⟩ for /ʃ/ as in *Schule*, /ˈʃuːlə/, Eng. *school*) – corresponds to a phoneme (alphabetic principle). However, languages differ in the consistency of this mapping: For

Table 1
Stage model of basal spelling development (based on Scheerer-Neumann, 2015).

Spelling strategy		Description of the strategy	Approx. time of occurrence
1	logographic	Writing “as if”, doodling, drawing individual letters, but no insight into the alphabetic principle yet – e.g., * for *Eis* (/aI̯s/, Eng. *ice*)	pre-school
2a	alphabetic	First insight into the alphabetic principle, spelling of easily distinguishable sounds (skeleton-like spelling) – e.g., **LP* for *Lampe* (/ˈlampə/, Eng. *lamp*)	first 3 months in 1st grade
2b		Advanced insight into the alphabetic principle, for almost every sound a (set of) letter(s) is assigned – e.g., **BUME* for *Blume* (/’blu:mə/, Eng. *flower*)	until the end of 1st grade
2c		Fully developed insight into the alphabetic principle, for each sound a (set of) letter(s) is assigned – e.g., **ROLA* for *Roller* (/’rɔlɐ/, Eng. *scooter*)	from the 2nd half of 1st grade onwards
3a	orthographic	First insight into the orthographic and morphemic structure of words, (over-generalised) use of syllabic and morphemic strategies – e.g., **belld* for *bellt* (/bɛlt/, Eng. *[the dog] barks*)	from 2nd grade onwards
3b		Increasingly comprehensive insight into the orthographic and morphemic structure of words – e.g., *Roller* for *Roller* (/’rɔlɐ/, Eng. *scooter*)	from 3rd grade onwards

example, while the mapping of graphemes to phonemes (i.e., the writing direction – hearing a phoneme and associating it with a grapheme) and phonemes to graphemes (i.e., the reading direction – recognising a grapheme and associating it with a phoneme) is relatively consistent in Spanish, English has several inconsistencies in both reading and writing directions.[2] In turn, German is relatively consistent in the reading direction, but has some inconsistencies in the writing direction (cf., Kargl & Landerl, 2018). These inconsistencies in German result largely from syllable-based reading aids (syllabic principle) and a tendency towards greater morpheme-based form consistency in the written language than in the spoken language (morphemic principle), among other causes. A large fraction of spellings – based on the syllabic principle, the morphemic principle, or other principles – can be systematically derived with the help of corresponding strategies. For example, the plural word *Hunde* (/ˈhʊndə/, Eng. *dogs*) can be used to infer that the singular word is not **Hunt* but *Hund*, although the audible final phoneme is /t/ (/hʊnt/) (i.e., the morphemic principle).

At the beginning of their spelling acquisition in school, children mainly use the alphabetic strategy, where words are written as they are heard. As children progress, they use a growing number of strategies and can apply these strategies with increasing flexibility. This manifests not only in the quantity but also in the quality of spelling errors. For example, the following spellings show a continuous improvement, even if they all deviate orthographically from the target word *Fahrrad* (/ˈfa:ɐ̯ ra:t/, Eng. *bicycle*): *FT* – *Fart* – *Farat* – *Farad* – *Fahrad* (example taken from May, 2013, p. 18).

Such development-related errors and the underlying strategies are described for many alphabetic scripts in developmental models (for English, see e.g., Frith, 1985; for German, see e.g., Günther, 1986, or Scheerer-Neumann, 2015). The basic stages are shown in Table 1.

Spelling development begins in early childhood when children are first confronted with script and continues beyond primary school age. However, most strategies should be mastered by the end of primary school (for Germany, by the end of grade 4 at the age of 10, see Standing Conference of the Ministers of Education and Cultural Affairs of the States in the Federal Republic of Germany, 2005).

Since spelling errors can be related to the mastery of particular spelling rules, spelling errors are an excellent setup for supervised ML techniques. Nevertheless, the approaches described in Section 2 are of limited use for spelling acquisition, as the following unique challenges must be addressed:

1. Words are produced by individuals not yet proficient in writing (or, in the case of second-language acquisition, even speaking) the language. Consequently, the models must learn not only the errors, but also the structures and processes involved in the acquisition. So far, related research has only been carried out on language acquisition in early childhood (e.g., Stella, 2019) and second-language acquisition (e.g., Crossley, 2013; Garcia & Pena, 2011). However, because of the specific processes involved, this research is not directly transferable to spelling acquisition.
2. The training must be performed at the character/grapheme-level. However, common models are typically constructed for the word- or sentence-level. Several deterministic approaches have been previously developed – for example, algorithms for sequence comparisons, such as those proposed by MacKenzie and Soukoreff (2002) or Wobbrock and Myers (2006). Typically, these methods attempt to categorise the errors into a manageable number of predefined error categories. Yet manual derivation of classification rules quickly becomes infeasible when considering more than a dozen different error categories (as is required here).
3. The present application is not a typical multi-class classification task (with a single possible outcome for each observation). Instead, it concerns a multi-label problem, where several errors can be made (or avoided) simultaneously without a fixed upper limit. There are only a few algorithms that naturally support multi-label classification tasks.

This paper addresses these challenges in the existing approaches and develops a novel ML-based approach for analysing spelling acquisition. To obtain a systematic insight into the predictive accuracy of specific errors that first- and second-graders produce in their spelling acquisition process, we train several ML models – including Logistic Regression (LR), a Convolutional Neural Network (CNN), a Random Forest (RF), and others – and compare their performances in a benchmark experiment. It thus provides a potential vantage point for future investigations.

3.1. Summary and research gap

In the sense of the “Matthew Effect” (cf., Stanovich, 1986; Walberg & Tsai, 1983), a fast and successful start in spelling acquisition is crucial, as it has a positive long-term effect on later achievement (Sparks et al., 2014). However, it is challenging for teachers to provide an adaptive schedule for all students based on their very different learning needs (Hebbecker & Souvignier, 2018). As a result, a significant proportion of students have insufficient spelling skills at the end of primary school (Stanat et al., 2017, 2022).

Some ML approaches for reading and writing have already proven to be useful in supporting teachers to individualise students’ learning (e.g., Sinclair, 2020). However, these approaches are mostly limited to

[2] For example, in English, the *phoneme* /k/ can be represented by the graphemes ⟨k⟩ (kind), ⟨c⟩ (cat), ⟨ck⟩ (back), ⟨ch⟩ (chord) and ⟨cc⟩ (account) (Fry, 2004, pp. 91–92). The *grapheme* ⟨ea⟩ can be assigned to the phonemes /e/ (bear), /a/ (heart), /eɪ/ (break), /ɛ/ (head), /i:/ (eat), /ɜ/ (earth) (Fry, 2004, pp. 88–90).

addressing sentence-level deficiencies in student demographics who are already proficiently literate. So far, no applications have been developed and tested specifically for spelling acquisition at the (initial) primary school level. In the context of a strong variation in spelling errors and their realisations (which depend on the children's development), it is necessary to investigate how accurately error matching succeeds across different developmental stages and to what extent the error categories can be mapped as such.

To that end, the work presented here focuses on the comparison between the normative regulation of the target words and the discriminatory power of the corresponding misspellings. This is of practical importance for ML diagnostics and the individualisation of learning opportunities. With the present contribution, we aim to close this research gap.

4. Research questions

The study addresses the following research questions.

Question 1. How accurate is the prediction of error categories with different ML algorithms?

This question arises because there is no research to date on how well ML algorithms can predict spelling errors within words of primary school students. As it is vital that digital tools (such as ITSs) utilise proven methods, only very reliable algorithms should be integrated to offer students more individualised spelling support (e.g., by assessments, feedback, and task selection).

Question 2. Which error categories can be predicted most accurately (and which least accurately) with the best-performing model?

Just as different models have different prediction accuracies, it can also be assumed that the prediction accuracies of one model are different for the individual error categories. As they are linked to the stages of spelling development, this information is essential to know at which stage a particular algorithm can be used effectively.

Question 3. In terms of content, what reasons can be identified for the varying predictive accuracy?

On the one hand, it can be assumed that the linguistic foundations for the different prediction accuracies for the individual error categories are starting points for further improvements of the classifications in the context of supervised learning. On the other hand, the linguistic foundations provide information on the implications for the pedagogical practice when it comes to possible misclassifications.

5. Methods

5.1. Materials

5.1.1. Web application for spelling

To investigate the research questions defined above, we developed a web application (app) which allows students to write automatically-given words trough dictation and logs the input online. The app was developed using Apple's Swift programming language and administered via GitHub. When using the app, all student inputs were logged online.

Fig. 1 gives an overview of the functional components and the app's workflow, which we describe in more detail below.

Before starting the spelling task, students completed a registration process. In the first step of the registration, students set a selected picture-password combination so that the next time they logged in, they could continue working on the task where they left off. In the second step, students provided information about their demographic data. This included their gender, age, and grade attended, as well as their first learned language and the language spoken at home. Subsequently, students completed a keyboarding and spelling test in the form of a dictation and transcription. The keyboarding test was designed to provide information about typing speed and typing precision. For this purpose, we adapted the graphomotor test developed by Abbott and Berninger (1993), which was later used for both handwriting and typewriting (Berninger, Abbott, Augsburger, & Garcia, 2009; Berninger et al., 2006). In the test, students were asked to type dictated target letters as quickly as possible. The spelling test served as a baseline and contained 12 words. The input was logged with timestamps. After the registration process, the spelling task started automatically.

When designing the app, we aimed to consider the diverse learning needs of students in primary schools under the maxim of simple handling. For example, we developed a custom virtual keyboard featuring only the letters needed to spell the given words. To minimise the risk of stereotype threats (see e.g., Cadinu, Maass, Rosabianca, & Kiesner, 2005), the user interface was designed to be diversity-sensitive (see Fig. 2).

In the app, the respective target word was displayed as a picture (Fig. 2, A) and could be played aurally as often as necessary via headset – as a single word (Fig. 2, A), as well as in a contextualising sentence (Fig. 2, B). Students were tasked with spelling the target word correctly (Fig. 2, C). If the input was incorrect in at least one position, the entire incorrect input was displayed (Fig. 2, D), and, for scaffolding, the incorrect position was marked with a blank space (Fig. 2, E). Students were given a maximum of three attempts to spell each target word

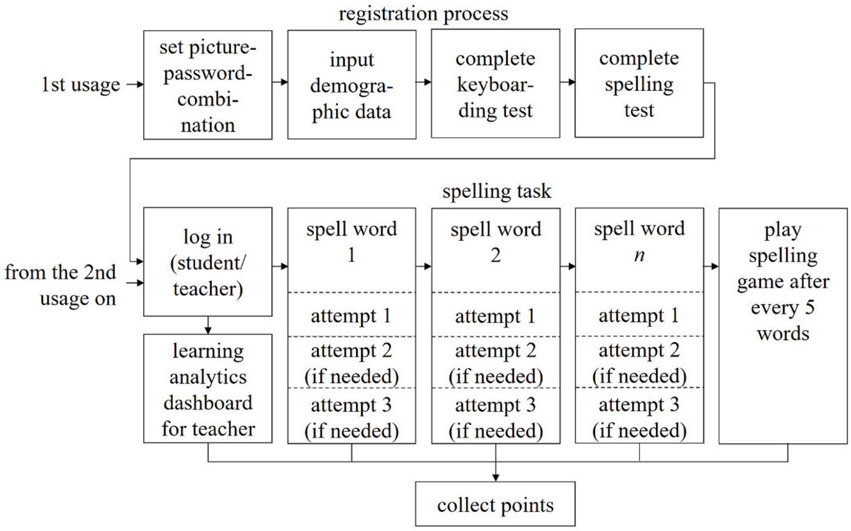
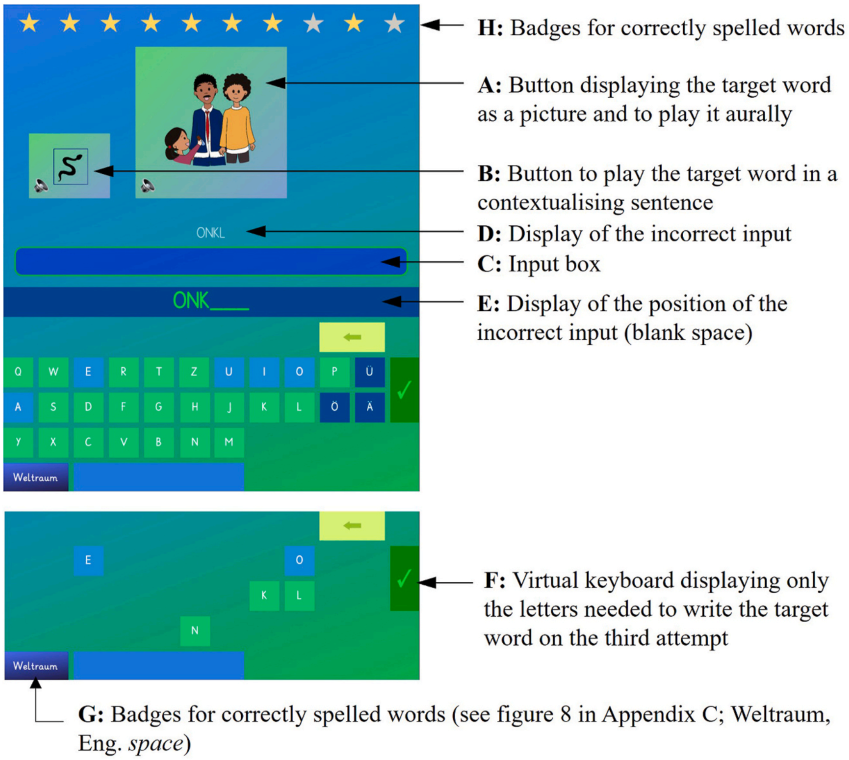


Fig. 1. Model of the functional components and the workflow of the developed app.



H: Badges for correctly spelled words

A: Button displaying the target word as a picture and to play it aurally

B: Button to play the target word in a contextualising sentence

D: Display of the incorrect input

C: Input box

E: Display of the position of the incorrect input (blank space)

F: Virtual keyboard displaying only the letters needed to write the target word on the third attempt

G: Badges for correctly spelled words (see figure 8 in Appendix C; Weltraum, Eng. *space*)

Fig. 2. Spelling task in the app using the example word Onkel (/’ɔŋkl̩/, Eng. uncle).

correctly. Before the third attempt, for didactic reasons, the correct spelling was shown for 3 sec. In addition, on the third attempt, the virtual keyboard displayed only the letters needed to write the target word (Fig. 2, F). This way, we intended to achieve a high hit rate in order not to overwhelm the students and, thereby, demotivate them.

Students received feedback and badges in the form of stars and planets, corresponding to the number of words they had already spelt correctly (Fig. 2, G and H). Additionally, a spelling game was presented at regular intervals (see Appendix C, Fig. 7). Both elements served to increase student engagement, which in turn has been shown to increase learning activity (cf., Clark, Tanner-Smith, & Killingsworth, 2016).

A vocabulary generally recognised as relevant for primary school students does not exist (Blumenthal, Sikora, & Mahlau, 2021). Instead, the scientific community has defined criteria that such a vocabulary should fulfil. These criteria include that the vocabulary should contain words with a characteristic structure concerning the orthographic principles. In addition, a relevant vocabulary should cover high-frequency words that are individually meaningful for children (Hoffmann-Erz, 2019).

Working with such a vocabulary is mandatory in some German federal states, such as Bavaria. For this reason, we chose the spelling vocabulary compiled by the Bavarian school authorities (State Institute for School Quality and Educational Research [ISB], 2017, pp. 272 et seq.), and selected 461 words therefrom. The spelling of 90 % of the words in the vocabulary can be explained using orthographic principles. It also contains 59 of primary school students’ 100 most frequently used words (see corpus analysis by Siekmann, 2023).

5.1.2. Categories of spelling errors

In order to classify various spelling errors, we developed an instrument for recording spelling errors on a single-grapheme basis. Furthermore, linguistics and primary education experts validated the instrument’s content according to the procedure proposed by Koller, Levenson, and Glück (2017). We distinguish 15 error categories that refer to strategies to tackle characteristic spelling phenomena. As depicted in Table 2, these categories, in turn, can be clustered to the principles described in Section 3. When categorising errors, we distinguish between systematic and unsystematic errors. Systematic errors are misspelt graphemes that indicate that a student has not used the required strategy or has applied a spelling rule in a position not covered by that specific rule (a so-called overgeneralisation). These are error categories 1 to 14. Unsystematic errors (category 15) are those not due to overgeneralisation or the phonological structure of a particular word. These include spellings that deviate considerably from the target word. Possible causes for error category 15 spellings can be, for example, (un-) intentional misspellings, unintentional pressing of the Enter key, or barely comprehensible concatenations of errors.

A more detailed explanation of the error categories as well as their underlying orthographic rules can be found in Appendix A.

5.2. Procedure

We contacted the principals of both urban and rural schools in Bavaria by email, who in turn informed the relevant teachers of the opportunity for voluntary participation in the study. We provided all interested schools that did not have tablets and headsets themselves with these tools for the survey period, which lasted about five weeks. We made the app available to the schools via Testflight – an Apple app for the iOS operating system that allowed us to install the spelling app on tablets without publishing it in the App Store. On the first day of the survey, trained test leaders presented the app to students following a standardised presentation scheme. The students had the opportunity to familiarise themselves with the app’s functions. Subsequently, they undertook the registration process, which included the completion of the keyboarding and spelling tests. The presentation and the registration could typically be completed in 45 min. After this period, the teachers were given a detailed explanation of the app’s functions. They were then free to decide how and how often they wanted to integrate the app into their lessons. From this point on, we supported the use of the app remotely. We collected data both from the registration on the first day of the study and throughout the phase when teachers used the app independently in class. All data from both the registration and regular

Table 2
Error categories, abbreviations of the error categories, and typical errors clustered according to orthographic principles.

No.	Error category	Abbreviation of the error category	Typical error
Alphabetic principle			
1	Basic phoneme-grapheme correspondence (_without_ additional rule)	PGC	*_Hs_ for _Haus_
2	Basic phoneme-grapheme correspondence (_with_ additional rule)	PGCandRule	*_Schport_ for _Sport_
Syllabic principle			
3	Unstressed syllable	SylUnst	*_Spiegl_ for _Spiegel_
4	Doubling of consonants (gemination)	Gemin	*_Mate_ for _Matte_
5	Silent _h_ in intervocalic position	Hvocal	*_Rue_ for _Ruhe_
6	Vowel length marker with silent _h_	Hlong	*_Zal_ for _Zahl_
7	Vowel length marker with German grapheme _ß_	ßlong	*_Strase_ for _Straße_
Morphemic principle			
8	German Umlaut graphemes _ä_ and _äu_	UmlÄ	*_Beume_ for _Bäume_
9	Final-obstruent desonorisation	Desonor	*_Hunt_ for _Hund_
10	Morpheme connector	MorphC	*_entarnen_ for _enttarnen_
Lexical principle			
11	Compound spelling	Comp	*_Apfel Saft_ for _Apfelsaft_
12	Capitalisation	Capital	*_haus_ for _Haus_
13	Irregularities	Irreg	*_Fater_ for _Vater_
Other			
14	Overgeneralisation	Overgen	*_Däcke_ for _Decke_
15	Unsystematic errors	Unsys	*_solllllllll_ for _soll_

Note. For the sake of simplicity, only the addressed error is modelled in each of the examples.

implementation by teachers was factored into the ML training and analysis.

5.3. Participants

The app was used by $N = 685$ students (50.5 % female; 49.5 % male; age in years: $M = 8.0$; $SD = 0.7$; min $= 6.3$; max $= 10.0$) in 40 classes at 11 schools in Bavaria at the end of the school year. Of these, $n_1 = 280$ students were first-graders (divided into 15 classes), $n_2 = 245$ students were second-graders (divided into 14 classes), and $n_{1/2} = 160$ students attended a mixed first- and second-grade class (divided into 11 classes). 75 % of the students spoke German as their first language, and 84 % mainly spoke German at home. The remaining 16 % spoke more than 16 other languages at home, with Russian (3 %) and Turkish (2 %) being the largest groups.

According to the teachers at 6 of the 11 participating schools, tablets were already used regularly in class. Regardless, all students had basic experience using tablets in their private lives.

5.4. Analysis

We carried out the analysis following the standard ML procedure in the context of educational sciences (see Hilbert et al., 2021). The standard procedure includes data _preprocessing_ for the spelling errors as targets and as the input data, _modelling_ of different ML models, and _evaluation_ of their performances. We conducted the evaluation by 10-fold cross-validation. For simplicity with regard to computational power and time, we did not optimise the hyperparameters of each learning algorithm, with this matter instead remaining an open point for further research. We describe each step in more detail below.

5.4.1. Preprocessing

5.4.1.1. Preprocessing of the target variables. Since we are presented with a multi-label problem when analysing misspelt words, a linguistics expert prepared the misspelt words in the first step by classifying errors according to the error categories described in Section 5.1.2, with a binary code indicating if an error was present (1) or absent (0). Thus, if several graphemes are misspelt in a word, several error categories were assigned to the word, as exemplified in Table 3 and described thereafter.

The exemplary spelling attempt *_schpikl_ contains the following five errors, two of which belong to the same error category:

- The rule that nouns are capitalised was not considered (error category 12: capitalisation).
- The rule on the phoneme-grapheme correspondence in the connection of /ʃp/ was not considered (error category 2: basic phoneme-grapheme correspondence with additional rule).
- The rule that the long vowel /iː/ usually corresponds with the grapheme ⟨ie⟩ was not considered (error category 2: basic phoneme-grapheme correspondence with additional rule).
- The phoneme /g/ was not recognised correctly (error category 1: basic phoneme-grapheme correspondence without additional rule).
- The rule on the obligatory vowel in each syllable was not considered (error category 3: unstressed syllables).

5.4.1.2. Preprocessing of the input data. To be processed by an ML model, the raw text input data must be transformed into a numeric space. To model spelling errors _within_ words, we transferred the principle of _n_-grams from the sentence-level to the character-level. In order to reproduce graphemes, which in German can consist of 1, 2 or 3 letters, we have used 1-, 2-, and 3-grams. Table 4 shows such a decomposition using the aforementioned example word _Spiegel_ and the error word _schpikl_.

As outlined above, the errors refer to the following graphemes or grapheme combinations:

- Capital letter ⟨S⟩ instead of lowercase letter ⟨s⟩ when the word is a noun (error category 12)
- Grapheme ⟨S⟩ instead of grapheme ⟨sch⟩ when followed by grapheme ⟨p⟩ (error category 2)
- Grapheme ⟨ie⟩ instead of grapheme ⟨i⟩ when the phoneme is a long vowel (error category 2)

Table 3
Scheme of error categorisation using the example word Spiegel (/ˈʃpiːgl̩/; Eng. mirror).

target word	graphemes	student's attempt	error categories (15 elements)
Spiegel	S,p,ie,g,el	schpikl	1,1,1,0,0,0,0,0,0,0,0,1,0,0,0

Note. Each digit in the binary code corresponds sequentially to the presence or absence of errors 1 through 15.

Table 4
Scheme of n-gram integer mapping (up to 3-grams) using the example word Spiegel (/ˈʃpiːɡl̩/; Eng. mirror).

word	n	n-grams text representation	n-grams integer mapping
Spiegel	1	S – p – i – e – g – e – l	1 – 2 – 3 – 4 – 5 – 4 – 6
	2	Sp – pi – ie – eg – ge – el	7 – 8 – 9 – 10 – 11 – 12
	3	Spi – pie – ieg – ege – gel	13 – 14 – 15 – 16 – 17
schpikl	1	s – c – h – p – i – k – l	18 – 19 – 20 – 2 – 3 – 21 – 6
	2	sc – ch – hp – pi – ik – kl	22 – 23 – 24 – 8 – 25 – 26
	3	sch – chp – hpi – pik – ikl	27 – 28 – 29 – 30 – 31

Vocabulary: 1:S, 2:p, 3:i, 4:e, 5:g, 6:l, 7:Sp, 8:pi, 9:ie, 10:eg, 11:ge, 12:el, 13:Spi, 14:pie, 15:ieg, 16:ege, 17:gel, 18:s, 19:c, 20:h, 21:k, 22:sc, 23:ch, 24:hp, 25:ik, 26:kl, 27:sch, 28:chp, 29:hpi, 30:pik, 31:ikl

- Grapheme ⟨g⟩ instead of grapheme ⟨k⟩ when the phoneme is a /g/ (error category 1)
- Grapheme ⟨el⟩ instead of grapheme ⟨l⟩ when the syllable is unstressed (error category 3)

All variants are included in the vocabulary listed in Table 4. However, this is only because we employed 1-, 2- and 3-grams. Thus, the example shows that all three *n*-grams are necessary to represent orthographic rules and errors in German.

5.4.2. Modelling

As only a few algorithms naturally support multi-label classification tasks, we have decomposed the whole error composition problem into individual binary classification tasks for each error category. In effect, this allowed us to apply any ML algorithm capable of handling binary classification.

Due to the word-to-*n*-grams transformation of the input space and the binary classification task decomposition, several standard ML models could be trained to perform a benchmark experiment investigating their different performances. As a baseline, we included LR. Furthermore, we trained models for the *k* Nearest Neighbours (KNN), a Support Vector Machine (SVM), a glmnet, and a RF. Detailed introductions on most of the included model classes can be found in Hilbert et al. (2021).

Following Zhang, Zhao, and LeCun (2015), we also included a CNN in our benchmark, consisting of an embedding layer and three 1D-convolutional layers following three further hidden layers. Using cross-entropy loss, the CNN naturally handles the multi-label task and delivers independent probabilities for each error category.

We performed the benchmark using the mlr3 ecosystem by Lang et al. (2019), which is based on the programming language **R**. We have taken the ML models (except the CNN) from the mlr3learners package using the default hyperparameter setting. One exception is the KNN algorithm, whose number of considered neighbours have been set to 10 (default is 7). We used the package mlr3multioutput as an extension to solve multi-label tasks with the mlr3 framework. Moreover, we exclusively included the CNN within this framework using keras and TensorFlow.

5.4.3. Evaluation

We evaluated the performances of each algorithm included in the benchmark experiment by 10-fold cross-validation: Here, the dataset is split into ten disjunct folds (i.e., parts), while each fold serves as a test set in one iteration, the other nine folds combined are used for model training. To assure comparability, folds are the same for all models. Each test set is necessary to compare their performances on new data yet unseen during training. The error made on the test set is an estimate of the model's generalisability.

Due to the naturally varying frequency of spelling rules and errors, the classes in the present dataset are highly imbalanced. To adequately evaluate the performance of predicting the error categories, we calculate F_1 scores and the Area Under the Receiver Operating Characteristic Curve ([ROC] AUC), as suggested for such cases and described in detail by Hilbert et al. (2021). The F_1 measure is based on the actual class labels, reflecting the harmonic mean of the precision and recall. The best possible value of 1 indicates perfect precision and recall, while a score of 0 indicates either precision or recall must be null. Meanwhile, the AUC measure is based on probabilities and shows a binary classifier's discrimination ability described by its resulting ROC curve. A score of 0.5 reflects the score of random labelling using prior class probabilities, while 1 indicates a perfect classifier.

For a more detailed analysis of the best-performing model and to reflect the nature of the underlying multi-label task, we additionally describe the joint confusion matrix for all 15 error categories considered together. This becomes helpful for multi-label tasks, where (in contrast to multi-class classification problems) a correctly predicted class does not necessarily mean that all other classes cannot be present. Consequently, following Heydarian, Doyle, and Samavi (2022), we show a multi-label confusion matrix and describe the confusions in terms of content.

6. Results

We begin with reporting the descriptive statistics (see Section 6.1). We then discuss the results of the research questions. First, we compare the performance of all six ML models included in the analysis (research question 1, see Section 6.2). Then, we briefly describe which error categories are best and worst predicted by the best model (research question 2, see Section 6.3). Finally, we show the above-mentioned multi-label confusion matrix based on the best model and describe the confusions from a linguistic perspective using representative sets of examples (research question 3, see Section 6.4).

6.1. Number of errors and correlations

As depicted in Fig. 3, test words were spelt correctly on the first attempt in 67.6 % of cases. 10.9 % were spelt correctly on the second attempt, 10.7 % were spelt correctly on the third attempt, and another 10.8 % were spelt *in*correctly on the third attempt. Across all spelling attempts, this resulted in 112,480 errors in 79,576 misspelt words. We identified 18,133 *different* misspellings.

All students had basic experience of using tablets from their private lives, and tablets were already being used regularly in 6 of the 11 schools. We found a statistically significant correlation between age and typing speed ($\rho = 0.294, p < 0.001$) as well as between typing speed and spelling performance ($\rho = 0.280$, $p < 0.001$). According to Cohen (1992), these correlations can be classified as small to medium. In contrast, age and typing accuracy ($\rho = 0.002$, $p = 0.952$) as well as typing accuracy and spelling performance ($\rho = 0.019, p = 0.621$) are not significantly correlated.

6.2. Prediction accuracy of the 15 error categories with different ML models

With respect to research question 1, and as illustrated in Fig. 4, the AUC scores revealed a noteworthy pattern across all 15 error categories regarding the performances of each model.

With an average AUC of 0.9849 on the 10 cross-validation test folds, the RF performed best overall, demonstrating a slight advantage over the CNN with an average score of 0.9810. In contrast, the LR (serving as a baseline) performed worst with an average score of 0.6100. The SVM, KNN ($k = 10$), and glmnet performed significantly worse than the CNN and the RF (but better than the LR), with respective average scores of 0.9089, 0.9454, and 0.8233.

The general pattern of these results remains the same when considering the F_1 score for performance evaluation. There is only a small significant difference in overall ranking, with the CNN performing best with an average F_1 score of 0.9770, compared to an average score of

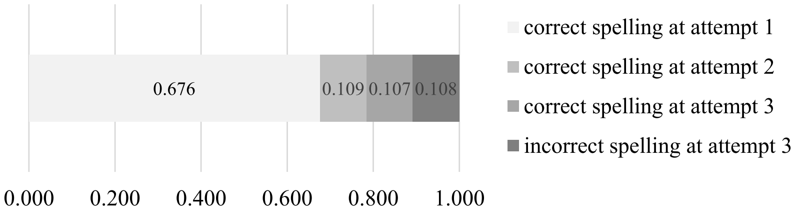


Fig. 3. Proportion of (in)correct children's spellings, broken down by attempt 1, 2 and 3.

0.9732 of the RF. Fig. 5 (Appendix B) shows the boxplots for the cross-validated results for all models. As shown, the CNN and RF models tend to outperform all other learning algorithms – frequently by a wide margin in many error categories.

6.3. Best- and worst-predicted error categories by the best-performing ML model

As noted in Section 6.1, the RF is the overall best-performing ML model. For this reason, and according to research question 2, we now highlight the performance of this model in a detailed analysis. Table 5 shows that error categories 5, 7, and 10 have AUC scores of $>$ 0.9990, and are thus best predicted. Additionally, error categories 2, 3, 4, 6, 8, 9, 11, and 13 also show satisfactory performances with AUC scores of $>$ 0.9900. Notably, this includes all error categories that can be assigned to the syllabic or morphemic strategy (see Table 2). In contrast, errors 1, 12, 14, and 15 have AUC scores of $<$ 0.9900 and, thus, are predicted slightly less accurately.

6.4. Description of the errors confounded by the best ML model

Table 6 shows the multi-label confusion matrix. The grey boxes in the table correspond to the number of correctly predicted errors per category. The white boxes represent the error categories that were incorrectly predicted. It is noteworthy that there are very few misclassifications for most categories, which is also reflected in the high AUC scores described above. Most of the confusions are between error 1 and other categories, likely owing in part to the high number of errors made in this category.

Following research question 3, we now provide deeper insights into the nature of these confusions. We focus on errors 1, 12, 14 and 15, which were the least accurately predicted errors by the algorithm (see Section 6.3). We describe two representative sets of examples, each for error 1 as a true label, and error 1 as a predicted label. In addition to error 1, we describe two sets of examples with a higher number of confusions. We refer to the errors highlighted in the confusion matrix.

In the confusion of error 1 as the true label with error 15 as the predicted label, it is notable that in the majority of the 309 cases, there are only a few graphemes assigned to phonemes of the respective target word (as demonstrated in the examples in Table 7). Several phoneme-grapheme mappings are still missing, resulting in skeleton-like spellings. However, the confusion does not include any cases with error words that evidently have intentional misspellings, although error category 15 does contain those as well. The complexity of the distinction between these two categories becomes apparent with, for example, the error word ⟨gst⟩ for ⟨Gespenst⟩ (error 1) in opposition to an actual error of category 15 such as ⟨gffs⟩ for ⟨scheinen⟩: Here, the transition from a meaningful categorisation of the individual errors (in the sense of error category 1) to a word structure that is no longer clearly recognisable (in the sense of error category 15) is fluid.

In the case of error 1 as the true label being confused with error 2 as the predicted label, almost all of the 27 error words show transpositions of letters or alphabetic errors (as in the examples shown in Table 8) such that they belong to error category 1. However, in all cases, these errors occur at positions where an additional rule (in the sense of error category 2) must be considered. For example, in the word ⟨Spaβ⟩, the student wrote the grapheme ⟨b⟩, which represents the voiced plosive /b/, whereas an unvoiced plosive /p/ (⟨p⟩) was required (error category 1). However, in the case of the phoneme compound /ʃp/, an additional rule must be observed: this compound is always written with ⟨sp⟩ instead of ⟨schp⟩ at the beginning of a syllable (error category 2). Thus, it is a complex combination of errors.

A total of 561 confusions occurred between category 14 as a true label and category 1 as a predicted label. These primarily concerned cases that, from a linguistic perspective, cannot be clearly assigned to one or the other category. On the one hand, this is because different causes can have the same symptoms and are thus linguistically ambiguous. On the other hand, it is because spelling errors sometimes occur in complex error entanglements, as is the case in the previous example (see Table 8). For instance, with the target word ⟨backen⟩ in Table 9 – and more precisely, with the graphemes ⟨a⟩ and ⟨ck⟩ – the starting question is whether the student, firstly, can recognise the short vowel and, secondly, can distinguish the voiceless plosive /k/ from the voiced plosive /g/. If the problem can be found here (e.g., due to dialectal influences), it is indeed an error that would have to be assigned to error category 1. If the error is instead in the use of the correct orthographic strategy (which we assume here), then the error would necessarily be assigned to error category 14. Thus, this kind of error can justifiably be assigned to both categories and can only be clearly classified with consideration for the context, as well as errors in other written words.

Furthermore, a large proportion of the words actually contain category 1 errors (such as in all examples of the target words ⟨Gespenst⟩ and ⟨läuft⟩ [exception: **läufd*]), so that it is less a case of confusion than of cases in which only parts of the overall error composition were recognised.

In addition, as exemplified in Table 9, the overgeneralisation of capitalisation was often assigned to error category 1. Even if such cases are not ambiguous from a linguistic perspective, it must be taken into account that upper-case letters represent fundamentally different symbols for the algorithm than lower-case letters. In this context, there is a similarity between error categories 14 and 1: error category 1 also concerns confusion of letters (or sounds), but for phonological reasons rather than lexical reasons.

The 248 confusions between error 12 as the true label and error 1 as the predicted label (see examples in Table 10) are similar to the last case mentioned above. In 70 % of the spellings with this type of confusion, only lower-case letters instead of an upper-case first letter were used (corresponding to category 12). The remaining 30 % of the spellings

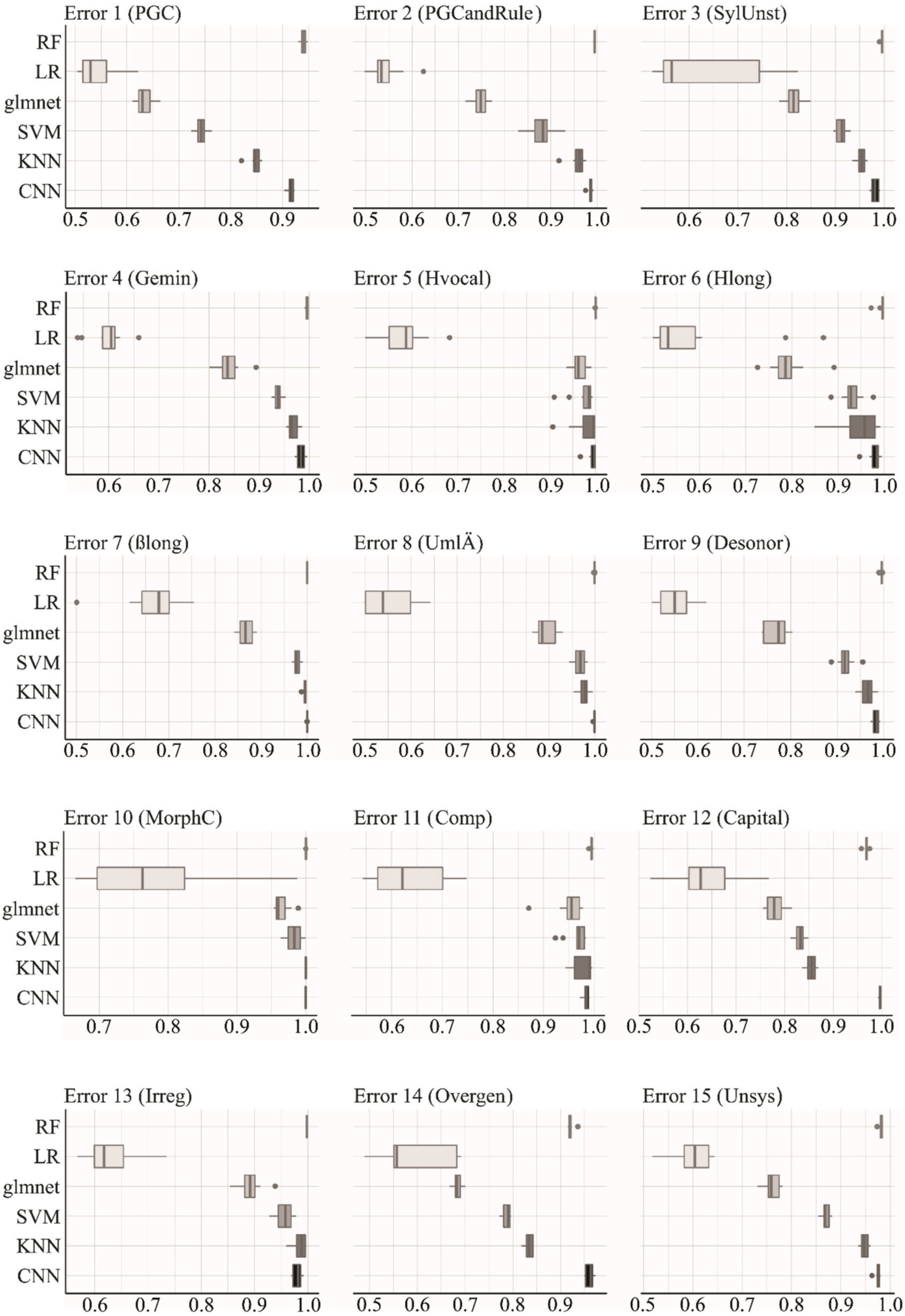


Fig. 4. Boxplots of AUC model scores of six ML models (RF, LR, glmnet, SVM, KNN, CNN) for each error category based on 10-fold cross-validation.

Table 5
Mean scores of the RF model evaluated by 10-fold cross-validation for each error category and mean score over all error categories for the AUC measure.

Error 1 *PGC*	Error 2 PGC andRule	Error 3 SylUnst	Error 4 Gemin	Error 5 Hvocal	Error 6 Hlong	Error 7 ßlong	Error 8 UmlÄ
0.9400	0.9944	0.9954	0.9949	0.9995	0.9940	0.9994	0.9985

Error 9 *Desonor*	Error 10 MorphC	Error 11 Comp	Error 12 Capital	Error 13 Irreg	Error 14 Overgen	Error 15 Unsys	Mean Score
0.9951	1.000	0.9944	0.9677	0.9978	0.9206	0.9816	0.9849

Note. The grey boxes correspond to the best-predicted errors (AUC scores >0.9900), and the framed boxes correspond to the worst-predicted errors (AUC scores <0.9900); descriptions of the abbreviations are provided in Table 2.

Table 6
Multi-label confusion matrix with the number of true and false predictions for the 15 error categories based on the RF model.

		Predicted labels															
		Error 1	Error 2	Error 3	Error 4	Error 5	Error 6	Error 7	Error 8	Error 9	Error 10	Error 11	Error 12	Error 13	Error 14	Error 15	NPL
True labels	Error 1: PGC	7329	27	5	45	5	5	4	2	12	0	8	131	12	591	309	1047
	Error 2: PGCandRule	160	1045	1	0	0	0	2	0	0	0	7	36	2	77	24	452
	Error 3: SylUnst	147	4	747	14	2	2	0	0	3	0	13	14	0	55	25	409
	Error 4: Gemin	89	0	1	822	0	2	0	0	4	2	10	21	0	53	30	192
	Error 5: Hvocal	9	2	0	0	140	0	0	0	0	0	4	3	0	4	6	38
	Error 6: Hlong	40	0	0	0	0	112	0	0	2	0	0	4	0	29	23	192
	Error 7: ßlong	74	3	0	3	0	0	1085	0	2	0	33	6	0	97	0	12
	Error 8: UmlÄ	36	0	2	4	0	3	0	393	2	0	3	10	0	7	14	109
	Error 9: Desonor	60	1	2	2	0	0	0	0	475	0	4	5	0	30	14	294
	Error 10: MorphC	6	4	4	2	0	2	2	0	20	447	18	18	2	46	2	0
	Error 11: Comp	9	0	0	0	0	0	0	0	4	0	52	5	0	1	0	110
	Error 12: Capital	248	16	2	18	2	2	13	2	13	0	24	2060	9	47	78	1047
	Error 13: Irreg	26	0	0	0	0	0	0	0	0	0	4	8	239	27	2	141
	Error 14: Overgen	561	6	2	24	1	1	71	1	10	1	16	45	4	3841	106	1834
	Error 15: Unsys	379	9	4	11	2	2	0	6	5	0	0	71	3	170	4272	1258
	NTL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note. NTL = situations with no true label; NPL = situations where no label was predicted; fields with a grey background correspond to the true positives; confusions in the framed boxes are examples for a closer examination; descriptions of the abbreviations are provided in Table 2.

contain additional errors. This indicates that the use of the lower-case letter was incorrectly recognised as a letter transposition (corresponding to category 1), which is presumably because the algorithm treats capital letters as fundamentally different symbols from their corresponding lower-case letters.

The confusion of error 14 as the true label and error 7 as the predicted label can be described similarly to the above-mentioned confusion of error 14 and error 1. All 71 cases of this confusion follow the same logic as the example error word ⟨flie<u>ss</u>en⟩ shown in Table 11. This spelling error is in a complex entanglement with the other error category, legitimising both categories. On the one hand, the grapheme ⟨ss⟩, which according to the orthographic rule may only be written when preceded by a short vowel, was written after a long vowel. This represents an overgeneralisation of the rule (corresponding to error category 14). On the other hand, the grapheme ⟨ß⟩ was not used (corresponding to error category 7). This indicates that the student is not yet using the strategy correctly.

In the confusion of error 7 as the true label with error 11 as the predicted label, it is striking that all 33 cases are composites (which is a core element of error category 11), where one of the constituents should have ended with the grapheme ⟨ß⟩ but was written with ⟨s⟩ (error category 7). If the predicted label of error 11 were correct, the error words in Table 12 would be ⟨Fuß ball⟩, ⟨Gieß kanne⟩, Reiß verschluss⟩, ⟨Fuß boden⟩ and ⟨Fuß ballmannschaft⟩. Thus, as the two errors occur in the same

Table 7
Confusion of error 1 (true label) and error 15 (predicted label).

Example of target word	Example of incorrect spelling
Gespenst (/ɡəˈʃpɛnst/, Eng. *ghost*)	gst, gest, gstp
Zähne (/ˈt͡sɛːnə/, Eng. *teeth*)	zne, cne, zen
Spiegel (/ˈʃpiːɡl̩/, Eng. *mirror*)	spe, schpi, stge
Sätze (/ˈzɛt͡sə/, Eng. *sentences*)	ses
Turnschuhe (/ˈtʊʁnˌʃuːə/, Eng. *trainers*)	tusue

Table 8
Confusion of error 1 (true label) and error 2 (predicted label).

Example of target word	Example of incorrect spelling
Spaβ (/ʃpaːs/, Eng. *fun*)	schbahs
Stoβ (/ʃtoːs/, Eng. *knock*)	schdohs, schdos
Quelle (/ˈkvɛlə/, Eng. *source*)	gwele, kwejee
Quatsch (/kvaʧ/, Eng. *nonsense*)	gwatsch
aussteigen (/ˈaʊ̯sˌʃtaɪ̯ɡn̩/, Eng. *to exit*)	auschdaigen, auschdaign

Table 9
Confusion of error 14 (true label) and error 1 (predicted label).

Example of target word	Example of incorrect spelling
backen (/ˈbakn̩/, Eng. *to bake*)	Bagn, Baggen, Bacen
Ente (/ˈɛntə/, Eng. *duck*)	Änte, änte
läuft (/lɔɪ̯ft/, Eng. *he/she/it goes*)	lövt, lüfd, Fäuft, läufd
Kaiser (/ˈkaɪ̯zɐ/, Eng. *emperor*)	keihser, Kheiser
Gespenst (/ɡəˈʃpɛnst/, Eng. *ghost*)	Geschpänz, geschbenc, gäspänt

Table 10
Confusion of error 12 (true label) and error 1 (predicted label).

Example of target word	Example of incorrect spelling
Taxi (/ˈtaksi/, Eng. *taxi*)	taxi
Buch (/buːx/, Eng. *book*)	buch
Quatsch (/kvaʧ/, Eng. *nonsense*)	quatsch, kwhatsch
Onkel (/ˈɔŋkl̩/, Eng. *uncle*)	ongkel
Brot (/bʁoːt/, Eng. *bread*)	brot, brood

Table 11
Confusion of error 14 (true label) and error 7 (predicted label).

Example of target word	Example of incorrect spelling
flieβen (/ˈfliːsn̩/, Eng. *to flow*)	fliessen
Spieβ (/ʃpiːs/, Eng. *skewer*)	Spiess
fleiβig (/ˈflaɪ̯sɪç/, Eng. *diligent*)	fleissig
Spaβ (/ʃpaːs/, Eng. *fun*)	Spass, spass
Grüße (/ˈɡʁyːsə/, Eng. *greetings*)	Grüsse

Table 12
Confusion of error 7 (true label) and error 11 (predicted label).

Example of target word	Example of incorrect spelling
Fuβball (/ˈfuːsˌbal/, Eng. *football*)	Fusball, vusball
Gieβkanne (/ˈɡiːsˌkanə/, Eng. *watering can*)	Giskane, giskane, Geiskane, gieskannä
Reiβverschluss (/ˈʁaɪ̯sfɛɐ̯ˌʃlʊs/, Eng. *zip*)	Raisverschluss, Raisferschluss, Reisverschluss
Fuβboden (/ˈfuːsˌboːdn̩/, Eng. *floor*)	Vusboden, vusboden, Fusbodn, fusbodn
Fuβballmannschaft (/ˈfuːsbalˌmanʃaft/, Eng. *football team*)	fusbalmanschaft

position, it can be inferred that the algorithm has recognised central elements of both error categories. Nevertheless, the error (which in this case is not ambiguous) is incorrectly predicted. However, when considering the total number of errors made across all categories, confusions of this nature are very rare.

7. Discussion

When comparing six ML models, we found that two of them – the CNN and the RF – achieved satisfying results with AUC scores close to 0.99, with the RF performing best on average. Further analysis of the RF showed that the prediction accuracies differed notably for the various error categories. While errors related to basic phoneme-grapheme correspondence were confused in some cases with the categories related to capitalisation, overgeneralisation, or unsystematic errors, errors related to syllable- and morpheme-levels of spelling were predicted with notable reliability. The two latter levels represent the core area of spelling acquisition at primary school age (see Kargl & Landerl, 2018; Scheerer-Neumann, 2015). Thus, the RF algorithm seems particularly suitable for detecting spelling deficiencies in children who have already mastered the alphabetic strategy (i.e., when they write one [set of] letter [s] for each sound).

It is striking that most errors recorded in this study related to basic phoneme-grapheme correspondence. Due to the typical spelling skills of the chosen demographic of first- and second-graders, the emergence of some of these errors was expected (see Table 1). However, errors of this type should predominantly occur in the first three months of the first school year (see Scheerer-Neumann, 2015). Consequently, the students' overall spelling performance seems to lag behind expected spelling acquisition for this age group. These results are consistent with the IQB trend studies from 2017 and 2022, which showed that students (still) struggle with spelling at the end of primary school. Stanat and colleagues attribute the low performance in these studies to the fact that students have diverse prerequisites and learning needs, making it challenging for teachers to manage. We also found various prerequisites in the present study – at least with regard to the languages in use. For example, 25 % of the sample spoke German as a second language and spoke more than 16 other first languages combined – an example of "linguistic super-diversity" (Gogolin, 2010; see also Gogolin & Duarte, 2017).

Since early deficits are difficult to make up for later (Sparks et al., 2014), the spelling performances shown here and in the IQB trend studies is a cause for concern. There is a need to focus more on the students' learning needs. As a possible means to better identify the learning needs of individual students, the algorithm can serve as a key enabler for more individualised support for primary school children's acquisition of spelling (as suggested by Hilbert et al., 2021). This can be done in various practical ways. For example, the algorithm could be implemented in ITSs to provide students with continuous elaborated feedback in the learning process. This seems promising, as feedback is one of the most important measures for improving students' performance (Hattie, 2008). The algorithm could also provide teachers with differentiated diagnostic information about the learning process. Both is now possible with the developed instrument consisting of 15 different categories and the precise algorithm.

It may be assumed that the algorithm is not only more accurate, but also provides more differentiated results than what some teachers are able to provide. Considering that teachers must be experts in many areas of education, it is not surprising that they can only meet these expectations to a limited extent. For example, previous findings indicate there may be knowledge gaps in the area of basic language constructs (see Aro & Björn, 2016; Corvacho del Toro, 2013; Washburn et al., 2016) as well as limited accuracy in teachers' diagnostic judgements (Kärner et al., 2021).

Aro and Björn (2016) found that teachers' knowledge of morphology was significantly lower than for phonology and phonics. However, the inverse was observed for the best-performing algorithm examined here (RF): while morphemic errors were predicted particularly well, the RF algorithm apparently predicted basic phoneme-grapheme correspondence (without an additional rule) slightly less accurately. One reason is that the algorithm is less capable of distinguishing whether it is *still* an unsystematic transcription without a clearly recognisable word structure or *already* a meaningful transcription.

Furthermore, the algorithm is evidently less able to distinguish between (i) upper- and lower-case letters as identical representatives of the same phoneme and (ii) graphemes representing different phonemes. Upper-case letters have a one-to-one relationship with lower-case letters, and only one error category is possible for each case of confusion between upper-case and lower-case letters (upper-case instead of lower-case: overgeneralisation of capitalisation; lower-case instead of upper-case: disregarding of capitalisation). However, the steps of data pre-processing employed do not seem sufficient for this error category, even though we included 1-grams within the data (which should reflect those one-to-one relationships).

For other error categories, confusions are often plausible because these are cases of linguistic ambiguity, error entanglements, etc. For example, this may be observed when confusing error category 1 "basic phoneme-grapheme correspondence without additional rule" with error category 12 "capitalisation", error category 14 "overgeneralisation", or error category 15 "unsystematic errors". This illustrates the specific challenges for the algorithm described in a similar vein by Stella (2019) for language acquisition in early childhood.

We note that in this study, each word was examined individually and independently of other words. However, in pedagogical settings, a greater focus is given to the students' spelling behaviour in the learning *process*, taking into account the context, analysing the number of specific errors, and relating them to other specific errors. This enables educators to identify the *key aspects* in the errors made, so that these can be further improved upon and be practised (cf., Schründer-Lenzen, 2013). Under this condition, the results can be enriched with contextual information by considering and reporting correlation structures between the different error categories. In turn, these results can then be further processed – depending on the context of the application. The performance of the RF achieved is thus overall satisfactory.

8. Limitations and implications for future research

Since we concluded that the computationally heavy multi-label learning approach is already appropriate and sufficient for its current purposes, we decided not to perform Hyperparameter Optimisation (HPO). However, if the ML algorithm were to be used for major decisions (e.g., to support diagnostic decisions, like for dyslexia), the algorithm would need further optimisation. There are several possibilities to improve our current approach. Naturally, the first would be to perform thorough HPO and subsequently compare model performances. As a next step, models that heavily rely on HPO, such as tree boosting (e.g., xgboost), could be included. Furthermore, more advanced neural networks should be able to improve the current performances even more. Those could include recurrent neural networks, as well as the use of advanced pre-trained models, such as BERT. In this case, transformer models constructed for modelling at the character/grapheme-level within words would be necessary. However, only models for the sentence-level are available so far (including CharacterBERT, despite its name). As stated above, further improvements could be achieved by more advanced preprocessing of the input text data, which would be of particular use for errors related to capitalisation.

Regardless of the model's eventual performance, if the errors are intended to be analysed even more precisely, the 15 error categories used in this study would need to be further subdivided. We have already made such a subdivision to cover a total of 79 side factors. This concerns, for example, conjugated verbs like *geht* (/geːt/, Eng. he/she/it *goes*), where the silent *h* that is inserted in the infinitive *gehen* (/ˈgeːən/) to avoid a visual vowel clash (error category 5) is also retained in the conjugated form due to morpheme consistency. Therefore, the infinitive must first be formed to derive the silent *h* in *geht*. However, in order to train an ML algorithm on this basis, a significantly larger sample of students or a larger pool of error words would be necessary.

With regard to the sample, it must also be noted that we only collected data from first- and second-graders in Bavaria. Therefore, students in higher grades may produce different entanglements of spelling errors, so the algorithm's performance may differ slightly for this demographic. This is possible even though the number of errors decreases throughout educational development, and errors made at higher skill levels are more likely to occur on the syllabic and morphemic levels (Scheerer-Neumann, 2015; see also Table 1) – errors that were predicted particularly well by the RF algorithm. Moreover, we have only referred to the standard pronunciation (cf., Krech, Stock, Hirschfeld, & Anders, 2009) and the orthography in use in Germany (Dudenredaktion, 2016). Therefore, some spelling errors from students of other German-speaking regions may not be recorded or may be classified incorrectly. For example, this is most likely with error category 7 "vowel length marker with German grapheme *ß*", which does not exist in Swiss orthography.

Nevertheless, the present study has shown that the obtained error classification algorithm is very accurate. These results support the potential for this algorithm to provide diagnostic information to teachers and feedback to students – for example, through integration into existing tools, such as ITSs. Although there is still little empirical evidence on the predictive validity of data-informed teaching, a positive effect on students' performance has been found for teachers' use of data in the field of literacy acquisition (Wayman, Shaw, & Cho, 2017).

Moreover, teachers need specific diagnostic competencies to analyse the data (Zeuch, Förster, & Souvignier, 2017). In this context, it should be noted that diagnostic competencies do not become less important through the use of ML-based digital tools. On the contrary, evidence suggests that diagnostic competencies should be reinforced in (pre-service) teacher training (cf., Böhme, Brühl, Reisemann, Munser-Kiefer, & Hilbert, 2023).

With regard to feedback presented in a machine-automated way, it should be noted that there is still little and inconclusive evidence on the specific effects. For example, Meurers, Kuthy, Nuxoll, Rudzewitz, and Ziai (2019) found improved performance using elaborated feedback in digital textbooks at the secondary school level, while Vasalou et al. (2021) showed that children with reading difficulties struggle to understand elaborate game-based feedback at the primary school level. Therefore, the question arises of how machine-automated feedback can be designed to be beneficial for children in spelling classes and to individualise learning in a practical way.

Finally, the algorithm learned to classify spelling errors according to labels assigned by human experts – an approach also known as supervised ML (see Hilbert et al., 2021). Therefore, the algorithm learned to mimic human classification, but this classification may have contained mistakes (i.e., "human error"). Considering that a key prerequisite for the algorithm's usefulness is that the classification on which it was trained is correct, it is possible that the algorithm's performance reflects such underlying "human errors".

9. Conclusion

The aim of the present study was to investigate how accurately spelling errors at different stages of development can be predicted using different ML models, which specific errors are predicted best (and which worst), and what the content-related reasons for these prediction performances may be.

Comparing six ML models, the RF performed best on average, especially at the syllable- and the morpheme-levels. Errors at the basic phoneme-grapheme level were predicted slightly less accurately. Confusions often concerned linguistically ambiguous cases or occurred in complex error entanglements.

These results indicate that ML approaches are not only useful for demographics who are already proficiently literate (as shown in previous works), but that they also may be useful tools to investigate specific spelling deficiencies (within words) in a differentiated way, support diagnosis of deficient spelling behaviour, and help individualise students' learning.

10. Statements on open data and ethics

Due to the research approval conditions of the Bavarian authorities, research data is not shared. This study was approved by the Bavarian State Ministry of Education, and all procedures were conducted in accordance with applicable laws and institutional guidelines. Informed consent was obtained from all participants, and their privacy rights were strictly observed.

Funding

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) (grant number 01JA 2010) and by the German Federal Ministry of Economic Affairs and Climate Action (BMWK) (project number 7101147 talidu/16GP303401).

CRediT authorship contribution statement

Richard Boehme: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stefan Coors:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Patrick Oster:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation. **Meike Munser-Kiefer:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sven Hilbert:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors are involved in a start-up that develops intelligent learning software which is supported by the German Federal Ministry of Economic Affairs and Climate Action (BMWK). However, the results presented are part of a university project and doctoral theses - RB, SC, MMK, SH. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank our partners for the good and constructive cooperation, and especially the students and teachers at the associated schools for their intensive participation and important contributions.

Abbreviations

AIED	Artificial Intelligence in Education
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
GloVe	Global Vectors for Word Representations
GPT	Generative Pre-trained Transformer
HPO	Hyperparameter Optimisation
IQB	Institute for Educational Quality Improvement (Institut zur Qualitätsentwicklung im Bildungswesen)
ISB	State Institute for School Quality and Educational Research (Staatsinstitut für Schulqualität und Bildungsforschung)
ITS	Intelligent Tutoring System
KMK	Standing Conference of the Ministers of Education and Cultural Affairs of the States in the Federal Republic of Germany (Kultusministerkonferenz)
KNN	*k* Nearest Neighbours
LR	Logistic Regression
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
(ROC) AUC	Area Under the Receiver Operating Characteristic Curve
SVM	Support Vector Machine
Word2Vec	Word-to-Vector

Appendix A. Categories of spelling errors

The following explanations of spelling errors are based, on the one hand, on graphemic regularities reflected in the orthographic rules of German (for an overview, see Dudenredaktion, 2016). On the other hand, they are based on common didactic considerations for teaching spelling rules (for an overview, see Schründer-Lenzen, 2013).

Alphabetic principle

The majority of German graphemes can correctly be written following the alphabetic principle. Thus, students can apply the following basic strategy: “Write one (set of) letter(s) for each sound.”

1. *Basic phoneme-grapheme correspondence (without additional rule)*

Many graphemes in German words follow the alphabetic principle without any further rules to be observed (e.g., *Auto*, /ˈaʊ̯to/, Eng. *car*; *Tomate*, /toˈmaːtə/, Eng. *tomato*; *Zelt*, /ˈt͡sɛlt/, Eng. *tent*). However, to write these words correctly, students need the awareness to recognise individual phonemes in the spoken language. At the beginning of literacy acquisition, phoneme awareness is still poorly developed (Pfost, Blatter, Artelt, Stanat, & Schneider, 2019), which at this stage frequently leads to errors such as transpositions, omissions or additions of graphemes. Example: *Haus* (/haʊ̯s/, Eng. *house*); typical error: **Hs.*[3]

2. *Basic phoneme-grapheme correspondence (with additional rule)*

Some German words contain graphemes, that deviate from the basic phoneme-grapheme correspondence shown above (error category 1). Hence, the memorisation of said grapheme compounds and the conditions of their use are a basic requirement for the correct spelling of these words. They then can be applied to almost all words because the rule is very regular. For example, the *sh* sound /ʃ/ is always represented by the grapheme ⟨s⟩ if it is the first letter of a syllable and followed by a /p/ or /t/. Example: *Sport* (/ʃpɔrt/, Eng. *sport*); typical error: **Schport*.

Syllabic principle

In German, vowels are obligatory elements of each syllable. They vary in quantity and stress. German words typically follow a Trochaic meter with most two-syllabled words carrying stress on their first syllable. Therefore, students can follow another basic strategy: "Divide the word into syllables and pay attention to the vowels."

3. *Unstressed syllables*

In German, unstressed syllables are subject to vowel reduction, depending on the precision of articulation. For example, the *schwa* sound /ə/ (as in *given*) is reduced in some cases when it is followed by a /l/, /m/ or /n/. Since vowels are obligatory elements of each syllable, students should make sure to assign a vowel to each syllable, even if they cannot hear it. Example: *Spiegel* (/ˈʃpiːɡl̩/, Eng. *mirror*); typical error: **Spiegl*.

4. *Doubling of consonants (gemination)*

Short vowels in stressed syllables trigger the closing of that syllable with a consonant. If there is only one consonant between the first and the second syllable, a gemination of said consonant is required. Example: *Matte* (/ˈmatə/, Eng. *mat*); typical error: **Mate*.

5. *Silent h in intervocalic position*

If one syllable ends with a vowel and the next syllable begins with a vowel, then a visual vowel clash is typically avoided by the insertion of a silent *h*. This serves as a reading aid. Example: *Ruhe* (/ˈruːə/, Eng. *silence*); typical error: **Rue*.

6. *Vowel length marker silent h*

Long vowels are orthographically marked in German in some phonological environments, most commonly by the insertion of silent *h*, which serves as a reading aid. This rule applies only if the long vowel is followed by an *l*, *m*, *n* or *r*. However, since numerous exceptions exist to this rule, the corresponding words are often not introduced systematically in class and have to be memorised instead. Example: *Zahl* (/ˈt͡saːl/, Eng. *number*); typical error: **Zal*.

7. *Vowel length marker German grapheme ß*

Another marker of vowel length is the German *ß*. It is used exclusively in cases where, in the basic form of a word, one syllable ends with a long vowel and the next begins with a voiceless /s/. Example: *Straße* (/ˈʃtraːsə/, Eng. *street*); typical error: **Strase*.

Morphemic principle

The German script is characterised by a rich inflectional morphology (cf., Kargl & Landerl, 2018). This means that there is a strong tendency to mark the morphological relation between words by consistent spelling, even if they differ in pronunciation. Therefore, students can derive spellings using the following strategy: "Break words down into their significant units, identify word stems and derive the spelling from these."

8. *German Umlaut graphemes ä and äu*

The German *Umlaut* graphemes *ä* and *äu* represent the *e* sound /ɛ/ and the *oy* sound /ɔɪ̯/. They are used in most cases to maintain morpheme consistency and to indicate correct phonological articulation. In particular, they are used for plural formation. Example: *Baum* – *Bäume* (/baʊ̯m/ – /ˈbɔɪ̯mə/, Eng. *tree* – *trees*); typical error: *Baum* – **Beume*.

[3] For the sake of simplicity, only the addressed error is modelled in each of the examples.

9. *Final-obstruent desonorisation*

Obstruents in syllabic offsets are always voiceless in German (such as /t/ in *helped* – /hɛlpt/). Voiced obstruents are subject to the so-called final-obstruent desonorisation, a process in which they lose their sonorant qualities. However, following the principle of morpheme consistency, the spelling is retained in the written language. Example: *Hund* – *Hunde* (/hʊnt/ – /ˈhʊndə/, Eng. *dog* – *dogs*); typical error: **Hunt* – *Hunde*.

10. *Morpheme connector*

The last sound of a morpheme can be similar (or even identical) to the initial sound of the following morpheme. While in articulation, these two single, yet identical sounds are typically reduced to one sound – depending on the precision of articulation – both corresponding graphemes remain in spelling. Example: *enttarnen* (/ɛnˈtarnən/, Eng. *to unmask*); typical error: **entarnen*.

Lexical principle

German has a few special characteristics in spelling at the word level. The semantic and morpho-syntactic dimensions (word class) influence the orthographic dimension. Consequently, students must apply the following basic strategy: "For spelling, take into account the meaning of the word and the context of content."

11. *Compound spelling*

Compared to other languages, writing compound words as one unit is a very productive word-formation pattern in German, which can lead to particularly long words. The individual compounds are synthesised into one word or connected with a hyphen for better readability. However, they are not written as separate words as in English. In some cases, this serves to reduce ambiguity on a semantic level. For example, *wiederkehren* means to *come back* and *wieder kehren* means to *sweep again*. Another example: *Apfelsaft* (/ˈa͡pf l̩ˌzaft/, Eng. *apple juice*); typical error: **Apfel Saft*.

12. *Capitalisation*

As in English, words at the beginning of sentences are capitalised. In addition, to compensate for the high flexibility of word order in German syntax, nouns are systematically capitalised so that the reader can easily identify them as nominalised elements at any position. Example: *Haus* (/haʊ̯s/ , Eng. *house*); typical error: **haus*.

13. *Irregularities*

Beyond the rules and strategies presented so far, there are some words whose spelling is irregular and must be memorised. Some of them serve to distinguish so-called homophones, i.e., words that sound identical but have different meanings (e.g., *Lied*, /liːt/, Eng. *song*, and *Lid*, /liːt/, Eng. *eyelid*). Others are completely irregular, for instance, words that contain the grapheme ⟨v⟩. Example: *Vater* (/faːtɐ/, Eng. *father*); typical error: **Fater*.

Other

All the above categories are based on German spelling principles. The following categories are subordinate to these and refer to developmental errors on the one hand and unsystematic errors on the other.

14. *Overgeneralisation*

Children's spelling develops in stages (see Table 1). It is typical for children at certain stages to use spelling rules even in positions where the respective rule does not apply. This is called overgeneralisation. It can often be observed, for example, that children who have recently gained insight into the morphemic strategy write the *e* sound /ɛ/ with the ⟨ä⟩ grapheme, although morpheme consistency does not apply (cf., error category 8). Example: *Decke* – *Decken* (/ˈdɛkə/ – /ˈdɛkn̩/, Eng. *blanket* – *blankets*); typical error: **Däcke* – **Däcken*.

15. *Unsystematic errors*

Spellings that deviate too much from the target word were excluded. Reasons can be, for example, (un-)intentional misspellings, unintentional pressing of the Enter key or hardly comprehensible concatenations of errors. Example: *soll* (/zɔl/, Eng. *should*); typical error: **solllllll* (as an intentional misspelling).

Appendix B. Results

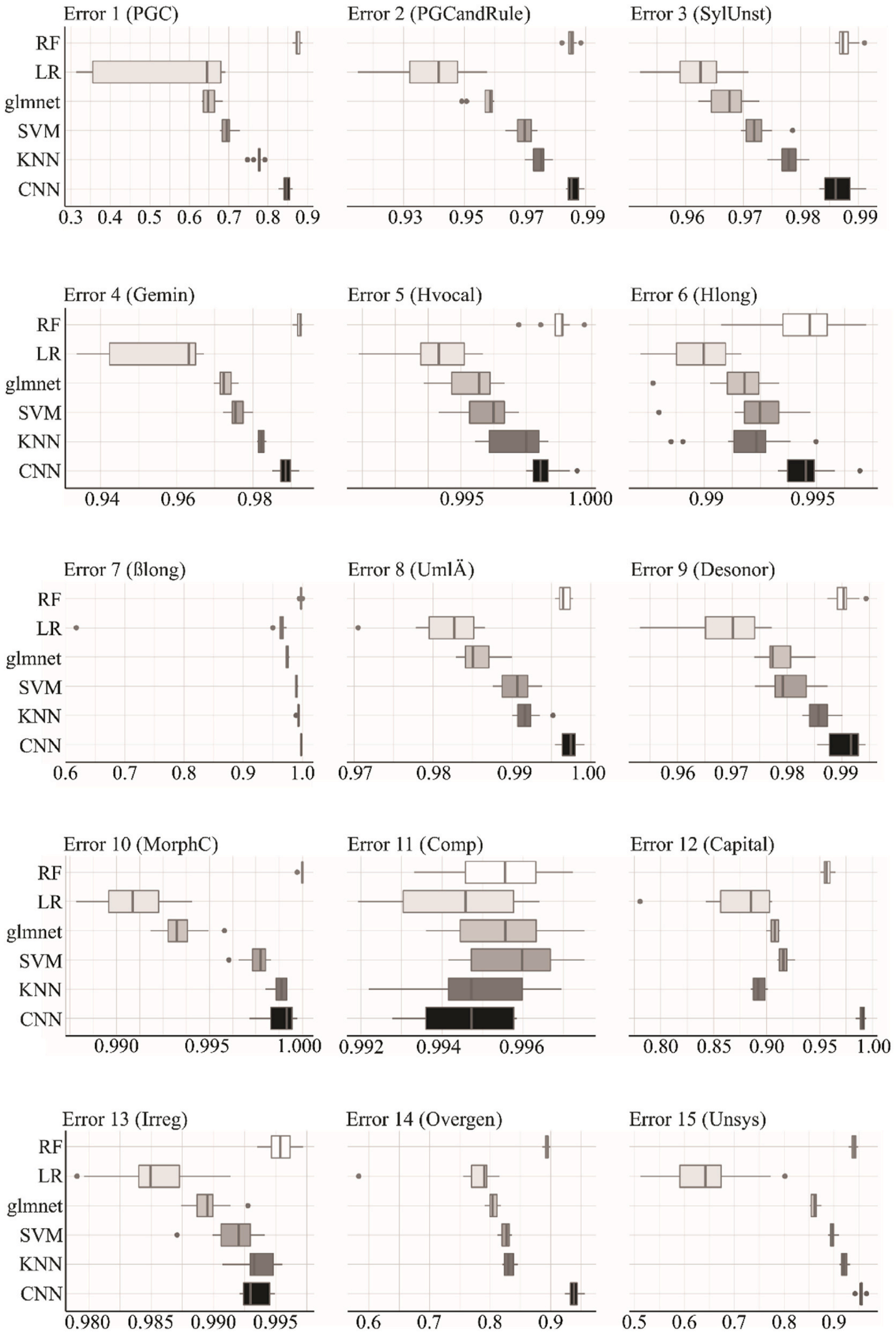


Fig. 5. Boxplots of F_1 model scores of six ML models (RF, LR, glmnet, SVM, KNN, CNN) for each error category based on 10-fold cross-validation.

Table 13
Mean scores of six ML models (RF, LR, glmnet, SVM, KNN, CNN) evaluated by 10-fold cross-validation for each error category and mean score over all error categories for the AUC and F_1 measures.

Model	Measure	Error 1	Error 2	Error 3	Error 4	Error 5	Error 6	Error 7	Error 8	Error 9	Error 10	Error 11	Error 12	Error 13	Error 14	Error 15	Mean Score
		PGC	PGC andRule	SylUnst	Gemin	Hvocal	Hlong	βlong	UmlÄ	Desonor	MorphC	Comp	Capital	Irreg	Overgen	Unsys	
RF	AUC	0.9400	0.9944	0.9954	0.9949	0.9995	0.9940	0.9994	0.9985	0.9951	1.0000	0.9944	0.9677	0.9978	0.9206	0.9816	0.9849
CNN	AUC	0.9153	0.9854	0.9815	0.9830	0.9928	0.9793	0.9997	0.9982	0.9835	0.9996	0.9843	0.9955	0.9799	0.9607	0.9752	0.9809
KNN	AUC	0.8470	0.9584	0.9521	0.9675	0.9774	0.9470	0.9943	0.9761	0.9644	0.9998	0.9773	0.8540	0.9838	0.8342	0.9480	0.9454
SVM	AUC	0.7435	0.8798	0.9129	0.9373	0.9733	0.9304	0.9776	0.9668	0.9181	0.9829	0.9662	0.8311	0.9562	0.7873	0.8690	0.9088
glmnet	AUC	0.6332	0.7463	0.8141	0.8406	0.9649	0.7907	0.8674	0.8937	0.7675	0.9649	0.9494	0.7796	0.8927	0.6841	0.7608	0.8233
LR	AUC	0.5401	0.5438	0.6337	0.5981	0.5825	0.5897	0.6650	0.5528	0.5545	0.7871	0.6363	0.6411	0.6280	0.5985	0.5983	0.6100
RF	F_1	0.8749	0.9853	0.9877	0.9921	0.9987	0.9944	0.9975	0.9966	0.9904	0.9999	0.9954	0.9572	0.9954	0.8935	0.9396	0.9732
CNN	F_1	0.8470	0.9861	0.9864	0.9885	0.9982	0.9946	0.9978	0.9973	0.9905	0.9988	0.9946	0.9898	0.9933	0.9380	0.9541	0.9770
KNN	F_1	0.7750	0.9748	0.9779	0.9822	0.9971	0.9920	0.9930	0.9919	0.9860	0.9988	0.9949	0.8929	0.9935	0.8308	0.9205	0.9534
SVM	F_1	0.6964	0.9695	0.9724	0.9759	0.9959	0.9923	0.9901	0.9905	0.9802	0.9976	0.9958	0.9162	0.9916	0.8258	0.8975	0.9459
glmnet	F_1	0.6514	0.9569	0.9673	0.9727	0.9955	0.9915	0.9747	0.9857	0.9784	0.9934	0.9955	0.9071	0.9896	0.8047	0.8609	0.9350
LR	F_1	0.5412	0.9387	0.9620	0.9550	0.9941	0.9898	0.9304	0.9815	0.9689	0.9910	0.9943	0.8727	0.9850	0.7678	0.6481	0.9014

Note. Descriptions of the abbreviations are provided in Table 2.

Appendix C. App screenshots

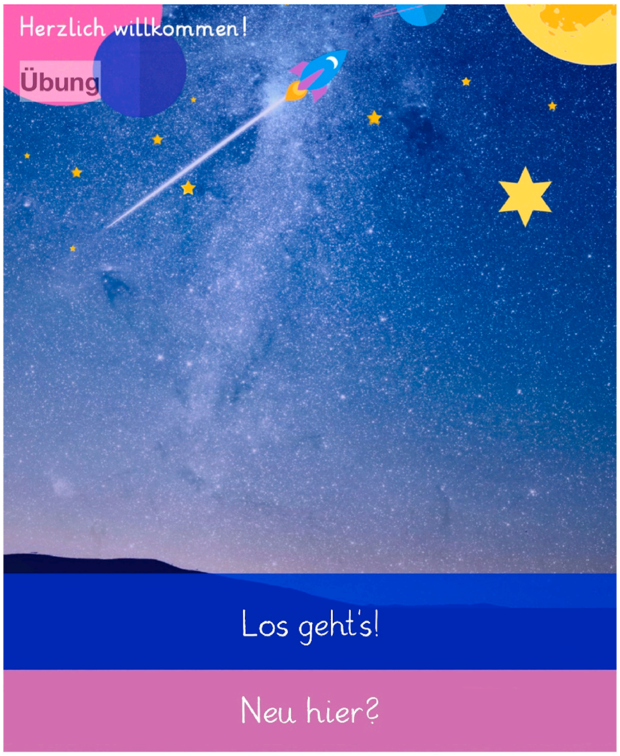


Fig. 6. Welcome page with registration, login and exercise area.

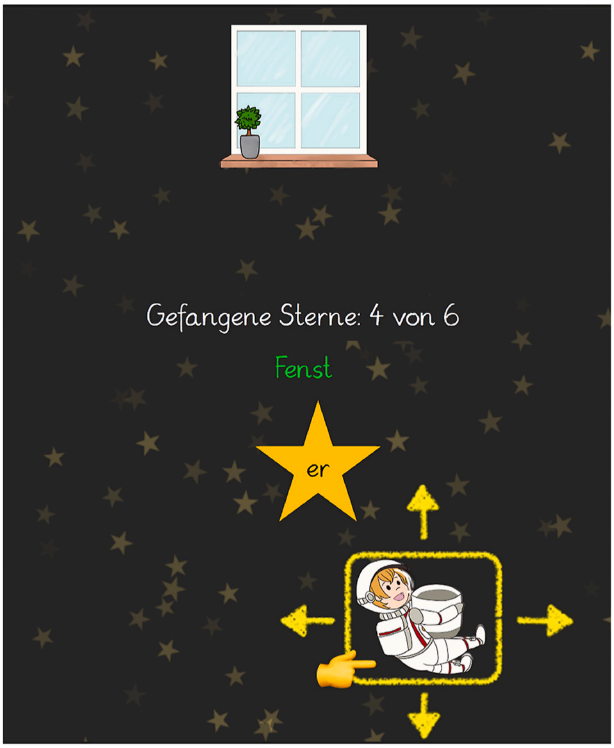


Fig. 7. Spelling game.

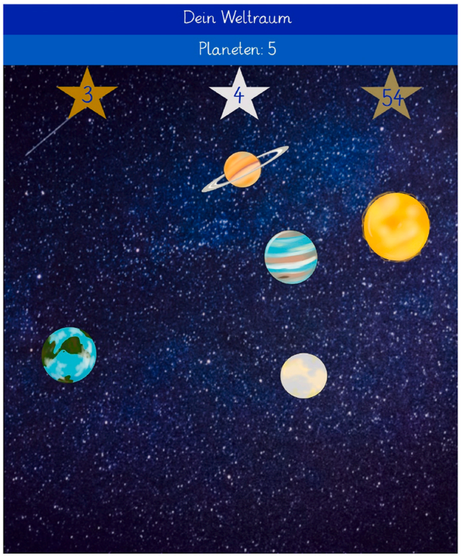


Fig. 8. Space where points can be collected.

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*(3), 478–508. https://doi.org/10.1037/0022-0663.85.3.478

Alhawiti, K. M. (2014). Natural Language processing and its use in education. *International Journal of Advanced Computer Science and Applications, 5*(12), 72–76. https://doi.org/10.14569/IJACSA.2014.051210

Aro, M., & Björn, P. M. (2016). Preservice and inservice teachers' knowledge of language constructs in Finland. *Annals of Dyslexia, 66*(1), 111–126. https://doi.org/10.1007/s11881-015-0118-7

Berninger, V. W., Abbott, R. D., Augsburger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123–141. https://doi.org/10.2307/27740364

Berninger, V. W., Abbott, R. D., Jones, J., Wolf, B. J., Gould, L., Anderson-Youngstrom, M., et al. (2006). Early development of language by hand: Composing, reading, listening, and speaking connections; three letter-writing modes; and fast mapping in spelling. *Developmental Neuropsychology, 29*(1), 61–92. https://doi.org/10.1207/s15326942dn2901_5

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Blumenthal, S., Sikora, S., & Mahlau, K. (2021). Lernverlaufsdiagnostik im Rechtschreibunterricht der Grundschule. *Diagnostica, 1–13*. https://doi.org/10.1026/0012-1924/a000261

Böhme, R., Brühl, D., Reisemann, K., Munser-Kiefer, M., & Hilbert, S. (2023). Auf- und Ausbau von Kompetenzen zur digital gestützten Diagnose und Förderung im schriftsprachlichen Anfangsunterricht. Ein phasenübergreifendes Seminarkonzept für (angehende) Grundschullehrkräfte. *Herausforderung Lehrer*innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion (HLZ), 6*(2), 31–48. https://doi.org/10.11576/hlz-6247

Boukkouri, H. E., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., & Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary Representations from characters. *arXiv*. https://doi.org/10.48550/arXiv.2010.10392

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16*(7), 572–578. https://doi.org/10.1111/j.0956-7976.2005.01577.x

Chai, K. E., & Gibson, D. (2015). Predicting the risk of attrition for undergraduate students with time based modelling. *12th international conference on cognition and exploratory learning in digital age (CELDA 2015)*.

Chen, P.-H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. *Nature Materials, 18*(5), 410–414. https://doi.org/10.1038/s41563-019-0345-0

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Corvacho del Toro, I. M. (2013). Fachwissen von Grundschullehrkräften: Effekt auf die Rechtschreibleistung von Grundschülern. In *Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg* (Vol. 13). University of Bamberg Press.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46*(2), 256–271. https://doi.org/10.1017/S0261444812000547

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-Training of deep bidirectional transformers for language understanding. arXiv https://doi.org/10.48550/arXiv.1810.04805.

Dudenredaktion. (2016). *Duden: Die Grammatik (9th ed., Vol. 4)*. Dudenverlag.

François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2012)*, 49–57.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neurological and cognitive studies of phonological reading* (pp. 301–330). Hillsdale: Erlbaum.

Fry, E. (2004). Phonics: A large phoneme-grapheme frequency count revised. *Journal of Literacy Research, 36*(1), 85–98. https://doi.org/10.1207/s15548430jlr3601_5

Garcia, I., & Pena, M. I. (2011). Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning, 24*(5), 471–487. https://doi.org/10.1080/09588221.2011.582687

German Federal Statistical Office. (2022). *Anzahl deutschsprachiger Menschen weltweit*. Retrieved from https://de.statista.com/statistik/daten/studie/1119851/umfrage/deutschsprachige-menschen-weltweit/.

Gogolin, I. (2010). Stichwort: Mehrsprachigkeit. *Zeitschrift für Erziehungswissenschaft, 13*(4), 529–547. https://doi.org/10.1007/s11618-010-0162-3

Gogolin, I., & Duarte, J. (2017). Superdiversity, multilingualism and awareness. In J. Cenoz, D. Gorter, & S. May (Eds.), *Language awareness and multilingualism: Encyclopedia of Language and education* (3rd ed., Vol. 10, pp. 375–390). Springer. https://doi.org/10.1007/978-3-319-02240-6_24.

Göpferich, S., & Neumann, I. (Eds.). (2016). *Developing and assessing academic and professional writing skills*. Frankfurt am Main: Peter Lang.

Günther, K. B. (1986). Ein Stufenmodell der Entwicklung kindlicher Lese- und Schreibstrategien. In H. Brügelmann (Ed.), *ABC und Schriftsprache: Rätsel für Kinder, Lehrer und Forscher* (pp. 32–54). Konstanz. Faude.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hebbecker, K., & Souvignier, E. (2018). Formatives Assessment im Leseunterricht der Grundschule – Implementation und Wirksamkeit eines modularen, materialgestützten Konzepts. *Zeitschrift für Erziehungswissenschaft, 21*(4), 735–765. https://doi.org/10.1007/s11618-018-0834-y

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). Mlcm: Multi-label confusion matrix. *IEEE Access, 10*, 19083–19095. https://doi.org/10.1109/ACCESS.2022.3151048

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., … Stachl, C. (2021). Machine learning for the educational sciences. *The Review of Education, 9*(3). https://doi.org/10.1002/rev3.3310

Hoffmann-Erz, R. (2019). Die Wiederentdeckung des Grundwortschatzes. *Lernen und Lernstörungen, 8*(3), 133–139. https://doi.org/10.1024/2235-0977/a000269

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning.* Boston, MA: Center for Curriculum Redesign.

Jarke, J., & Breiter, A. (2019). Editorial: The datafication of education. *Learning, Media and Technology, 44*(1), 1–6. https://doi.org/10.1080/17439884.2019.1573833

Jimenez, F., Paoletti, A., Sanchez, G., & Sciavicco, G. (2019). Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Transactions on Learning Technologies, 12*(2), 225–236. https://doi.org/10.1109/TLT.2019.2911070

Kargl, R., & Landerl, K. (2018). Beyond phonology: The role of morphological and orthographic spelling skills in German. *Topics in Language Disorders, 38*(4), 272–285. https://doi.org/10.1097/TLD.0000000000000165

Kärner, T., Warwas, J., & Schumann, S. (2021). A learning analytics approach to address heterogeneity in the classroom: The teachers' diagnostic support system. *Technology, Knowledge and Learning, 26*(1), 31–52. https://doi.org/10.1007/s10758-020-09448-4

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In T. Eiter, & S. Kraus (Eds.), *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 6300–6308). California: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/879.

Koller, I., Levenson, M. R., & Glück, J. (2017). What do you think you are measuring? A mixed-methods procedure for assessing the content validity of test items and theory-based scaling. *Frontiers in Psychology, 8*, 1–20. https://doi.org/10.3389/fpsyg.2017.00126

Krech, E.-M., Stock, E., Hirschfeld, U., & Anders, L. C. (2009). *Deutsches ausspracheworterbuch.* Berlin: De Gruyter.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., … Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software, 4*(44), 1903. https://doi.org/10.21105/joss.01903

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in us K-12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

MacKenzie, I. S., & Soukoreff, R. W. (2002). A character-level error analysis technique for evaluating text entry methods. In O. W. Bertelsen (Ed.), *Proceedings of the second nordic conference on human-computer interaction* (pp. 241–244). New York: ACM. https://doi.org/10.1145/572020.572056.

May, P. (2013). *HSP 1-10: Hamburger schreib-probe: Manual/handbuch: Diagnose orthografischer kompetenz.* Stuttgart: Klett.

Meurers, D., Kuthy, K. de, Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics, 39*, 161–188. https://doi.org/10.1017/S0267190519000126

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of words and phrases and their compositionality. Retrieved from https://arxiv.org/pdf/1310.4546 (preprint).

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In Q. C. R. I. Alessandro Moschitti, G. Bo Pang, & U.o. A. Walter Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Pfost, M., Blatter, K., Artelt, C., Stanat, P., & Schneider, W. (2019). Effects of training phonological awareness on children's reading skills. *Journal of Applied Developmental Psychology, 65*, Article 101067. https://doi.org/10.1016/j.appdev.2019.101067

Rienties, B., Køhler Simonsen, H., & Herodotou, C. (2020). Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.00128

Scheerer-Neumann, G. (2015). Lese-Rechtschreib-Schwäche und Legasthenie: Grundlagen, Diagnostik und Förderung. *Lehren und Lernen.* Stuttgart: Kohlhammer.

Schründer-Lenzen, A. (2013). Schriftspracherwerb. *Lehrbuch* (4th ed.). Wiesbaden: Springer. https://doi.org/10.1007/978-3-531-18947-5

Siekmann, K. (2023). *Grund- und Orientierungswortschatz für die Primarstufe: Häufigkeitsbasierter Wortschatz, Phonem-(Basis-)Graphem-Korrespondenzen, Fehlerverteilungen und didaktische Implikationen.* Band 1 (2nd ed.). Schönau am Königssee: Siekmann Verlag.

Sinclair, J. (2020). *Using machine learning to predict children's reading comprehension from lexical and syntactic features extracted from spoken and written language (doctoral dissertation).* Toronto: University of Toronto.

Sparks, R. L., Patton, J., & Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of '10 years later'. *Reading and Writing, 27*(1), 189–211. https://doi.org/10.1007/s11145-013-9439-2

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., … Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences of the United States of America, 117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., & Haag, N. (Eds.). (2017). *IQB trends in student achievement 2016: The second national assessment of German and mathematics proficiencies at the end of fourth grade.* Münster: Waxmann.

Stanat, P., Schipolowski, S., Schneider, R., Sachse, K. A., Weirich, S., & Henschel, S. (2022). *Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe: Erste Ergebnisse nach über einem Jahr Schulbetrieb unter Pandemiebedingungen.* Berlin: Waxmann.

Standing Conference of the Ministers of Education and Cultural Affairs of the States in the Federal Republic of Germany (KMK). (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich: Board decision on 15 October 2004*. Munich: Luchterhand. Retrieved from https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Deutsch-Primar.pdf.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–406.

State Institute for School Quality and Educational Research (ISB). (2017). In *LehrplanPLUS Grundschule in Bayern* (4th ed.) Munich.

Stella, M. (2019). Modelling early word acquisition through multiplex lexical networks and machine learning. *Big Data and Cognitive Computing, 3*(1), 10. https://doi.org/10.3390/bdcc3010010

Vasalou, A., Benton, L., Ibrahim, S., Sumner, E., Joye, N., & Herbert, E. (2021). Do children with reading difficulties benefit from instructional game supports? Exploring children's attention and understanding of feedback. *British Journal of Educational Technology, 52*(6), 2359–2373. https://doi.org/10.1111/bjet.13145

Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359–373. https://doi.org/10.2307/1162605

Washburn, E. K., Binks-Cantrell, E. S., Joshi, R. M., Martin-Chang, S., & Arrow, A. (2016). Preservice teacher knowledge of basic language constructs in Canada, England, New Zealand, and the USA. *Annals of Dyslexia, 66*(1), 7–26. https://doi.org/10.1007/s11881-015-0115-x

Wayman, J. C., Shaw, S., & Cho, V. (2017). Longitudinal effects of teacher use of a computer data system on student achievement. *AERA Open, 3*(1), Article 233285841668553. https://doi.org/10.1177/2332858416685534

Wobbrock, J. O., & Myers, B. A. (2006). Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction, 13*(4), 458–489. https://doi.org/10.1145/1188816.1188819

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice, 32*(1), 61–70. https://doi.org/10.1111/ldrp.12126

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *arXiv (preprint)*. https://doi.org/10.48550/arXiv.1509.01626