

# Poisson-Based Inference for Perturbation Models in Adaptive Spelling Training

**Gian-Marco Baschera, Markus Gross**

*Department of Computer Science, ETH Zürich, Switzerland*

**Abstract.** We present an inference algorithm for perturbation models based on Poisson regression. The algorithm is designed to handle unclassified input with multiple errors described by independent mal-rules. This knowledge representation provides an intelligent tutoring system with local and global information about a student, such as error classification (local) and prediction of further performance (global). The inference algorithm has been employed in a student model for spelling with a detailed set of letter and phoneme based mal-rules. The local and global information about the student allows for appropriate remediation actions to adapt to their needs. The error classification, student model prediction and the efficacy of the adapted remediation actions have been validated on the data of two large-scale user studies. The enhancement of the spelling training based on the novel student model resulted a significant increase in the student learning performance.

**Keywords.** Inference, Poisson regression, spelling, error classification

## INTRODUCTION

The ability to adapt to student needs is a central feature of an intelligent tutoring system (ITS). It is based on an abstract representation of the student, which is called the student model. These adaptive systems have produced impressive gains in user studies (Shute & Psotha, 1996). A student model is mainly characterized by its form of representing the student knowledge in the domain. The most prominent representatives are overlay models (BIP; Barr et al., 1976), perturbation models (DEBUGGY; Burton, 1982) and cognitive models (model tracing; Anderson et al., 1990). A second important element of the model is an inference algorithm for the estimation of the student mastery of the domain, and the subsequent update of the student model, while the student is training. In recent years various methods have been developed for a broad range of educational applications, specialized on the respective application-specific student input data. In this paper we present a novel approach for estimating the student's proficiency in a perturbation model with independent error production rules, which we refer to as *mal-rules*. In contrast to previous work, our method can handle student input with multiple, unclassified errors, i.e., errors which cannot be unambiguously related to one of the independent mal-rules. The method was implemented in a student model for spelling training, where errors can be described by several independent mal-rules, such as visual confusion of letters, auditory confusion of phonemes, or typing errors. A detailed description of the spelling error taxonomy and the corresponding mal-rules are given in the Section entitled 'Error Taxonomy and Mal-Rules'.

Most inference algorithms rely on the constraint that every input can be assigned to one single rule or mal-rule. A well-known example is Corbet and Anderson's knowledge tracing approach (1995). This constraint requires a decomposition of tasks into pieces of single skills as provided in model tracing approaches. However, these decompositions are not desired or possible in many applications. For example, the commonly cited mixed-number fraction subtraction (Tatsuoka, 1987)

requires multiple skills for one calculation. Similar issues arise in spelling tasks, where the input of a single letter cannot be broken down further. A unique association with a mal-rule is not possible. To overcome the ambiguity of the error source in mixed-number fraction subtraction, Mislevy employs a Bayesian inference network for skill estimation (1996). Other approaches dealing with the ambiguity are multiple classification latent class models (Maris, 1999) or linear logistic test models based on item response theory (Fischer, 1973). However, these methods estimate the probability of success or failure on one given item. They do not make allowance for multiple errors in one single input. Therefore, this multiplicity requires a different viewpoint on student errors.

We regard errors as randomly occurring events, which are best described by a Poisson distribution. Unlike knowledge tracing or higher-order latent trait models (de la Torre & Douglas, 2004), we do not assume the student attributes to be either in a learned or unlearned state. Mal-rules describe the difficulties in spelling and divide them into different categories. However, these mal-rules do not represent a concept of spelling which can simply be acquired and once mastery is reached, only slipping or other rules would cause subsequent errors of the same type. For example, visual and auditory confusions of letters in dyslexic children cannot simply be comprehended and removed. Similarly, the irregularities in the phoneme-grapheme mapping inhibit a sudden mastery of the phoneme-grapheme matching process. Therefore, we represent the strengths and weaknesses in spelling by the error rates on mal-rules for every individual student. For the estimation of these error rates we propose a Poisson regression (McCullagh & Nelder, 1989). The employment of a linear link function in the Poisson regression assures the independence of the factors. The inference algorithm is described in detail in Section ‘Student Model’.

The proposed method emerged from insights gained during a first user study (Kast et al., 2007) of Dybuster, a multi-modal German spelling training for dyslexic children (Gross & Vögeli, 2007). A brief summary has been presented at AIED09 (Baschera & Gross, 2009). Dyslexia and the training software Dybuster are discussed in the Section entitled ‘Difficulty Domain: Spelling’. Our research aimed for a student model, representing the student knowledge in spelling, which allows for an adaptation of the spelling training to an individual student’s needs. Student modeling in spelling has been mostly neglected so far. Commercial spelling learning environments focus on sustaining the children’s attention and motivation by utilizing the multi-media abilities of recent computer systems (iSpellWell, SuperSpell 2, Ultimate Spelling). However, their adaptation to the student is restricted to the repetition of erroneously entered words. In collaboration with elementary school teachers and psychologist, we identified the demand for information on two levels to adapt spelling training appropriately to students’ needs:

- *Local information:* error localization and classification to enable adequate remediation on erroneous inputs.
- *Global information:* student knowledge representation to allow for optimized word selection based on further spelling performance prediction on the entire word database, and for feedback to human tutors on students’ strengths and weaknesses.

A core challenge in building such a model is the identification of patterns and similarities in spelling errors across the entire word database, and to represent them using as few mal-rules as possible. Bodén et al. propose a language-specific set of 68 letter patterns to describe spelling difficulties. In their evolutionary approach of adapting spelling exercises, they respond to erroneous inputs by selecting similar words (Bodén & Bodén, 2007). Error localization and classification with

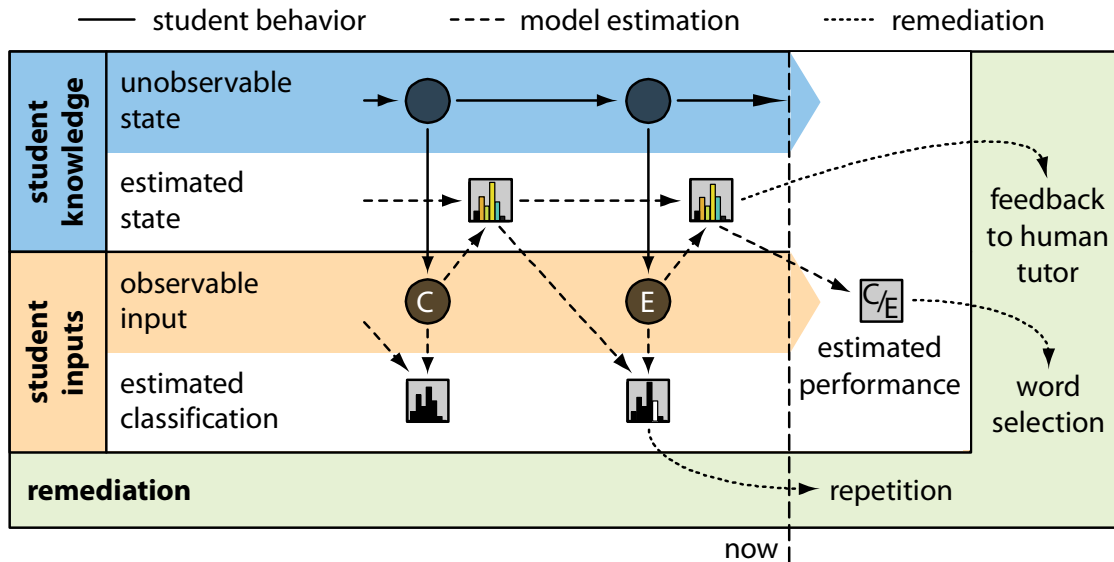


Figure 1: Workflow of the student model: The dark circles at the top represent the unobservable knowledge state of a student progressing over time. The model can only observe the student inputs, which can either be correct (C) or erroneous (E). Each input allows for an update of the student knowledge estimation, indicated by the colored bars in the rectangle. This enables a classification of subsequently committed errors, and a prediction of the further spelling performance. Appropriate remediation actions can be conducted on a local level (repetition), or a global level (word selection and feedback to human tutors).

respect to specific difficulties are not considered. This leads to an inappropriate adaptation for many error categories, e.g., capitalization: *Spiel* – *spiel* (engl. game). This capitalization error results in a selection of a similar word, such as *viel* (engl. plenty), which does not contain an error possibility for the previously committed capitalization error. Spencer predicts spelling difficulties from measures of orthographic transparency, phonemic and graphemic length, and word frequency based on English language corpora (Spencer, 2007). The resulting difficulty measure takes only a limited set of error types into account and neglects for example visual and auditory confusions or typing difficulties. It stays constant for all children and does not allow for adaptation to individual students. Additionally, language corpora are usually based on extensive collections of print samples, of which the word frequencies may be quite different from those of spoken, heard or hand-written language of children (Balota et al., 2001). Similar drawbacks appear in Bader-Natal and Pollack’s SpellBEE peer-tutoring system. Their difficulty estimation on the word database, computed from the input data of 17,000 students using the tutoring system, provides a constant relative spelling difficulty between words (Bader-Natal & Pollack, 2007).

The student model described in this paper is designed to provide both local and global information about an individual student. To render the model adaptable to different student strengths and weaknesses we developed an error taxonomy for isolated word spelling and defined corresponding mal-rules. During the spelling training, the student is represented by the error rates of each mal-rule. After each observed input of the student, the estimation of the unobservable knowledge state is updated using our Poisson-based inference algorithm (see Figure 1). Based on these estimated error

rates, the model enables a classification of subsequent erroneous inputs (local information), and a prediction of further spelling performance (global information). The information provided by the student model allows for appropriate remediation actions of the learning environment and human tutors. The 'Results' Section provides verifications of the presented modeling approach and results of a second evaluation study.

## **DIFFICULTY DOMAIN: SPELLING**

Writing is a fundamental skill in life. Although the message to be communicated is more important than the mechanics of writing, e.g., correct spelling, the competence in spelling has a high influence on the quality of written work. Spelling errors influence judgments that others make about overall quality of writing, distract readers from the message, and in extreme cases render the message incomprehensible. Perhaps even more important: problems with mechanics interfere with higher writing processes and affect the quality of writing (MacArthur, 1999). The orthographic depth of Western languages, i.e., the non-bijective correspondence between letters and phonemes, raises difficulties for children learning to spell. The process of spelling includes several additional challenges, from precise hearing to the final writing of the word. An efficient spelling training software needs to be able to detect the individual strengths and weaknesses of a child and adapt the training accordingly.

### **Dyslexia**

Correct spelling is especially hard to learn for dyslexic children. Developmental dyslexia is characterized through low reading and writing skills in spite of an (above) average IQ, adequate education and inconspicuous social background (World Health Organization, 1993). Dyslexia occurs predominantly in Western world languages including English, French, German, and Spanish. It is estimated that about 5-7% of the Western world population suffers from minor or major forms of dyslexia (Reitsma, 1989). The definition of dyslexia is purely symptom-oriented, and does not describe the causes of the disorder. These are strongly debated and a good overview of the most prominent hypotheses is given by Habib (2000). The diversity of hypotheses is partly based on the different characteristics of spelling difficulties of dyslexic children. They range from visual ('d'-b') and auditory confusions ('n'-m') to difficulties in the phoneme-grapheme matching process (/f/: 'f'-v').

### **Dybuster**

Gross and Vögeli developed a multi-modal spelling training for dyslexic children, called Dybuster. The entire framework is based on the concepts of information theory and multi-modal learning (Gross & Vögeli, 2007). The central idea of the training software is to recode a sequential textual input string into a multi-modal representation using a set of codes. These codes reroute textual information through multiple undistorted perceptual cues, including topological, color, shape, and auditory representations (see Figure 2 (left)).

The Dybuster software is structured into different games. In the first game – the color game – the students have to learn the association between a letter and a color. Based on the information theoretical model of Dybuster eight different colors are used. The mapping of letters to colors is the result of a multi-objective optimization. For example, letters easily confused by dyslexics, e.g., 't' and 'd', map to different colors. The idea is to associate colors with letters to eliminate mistakes.

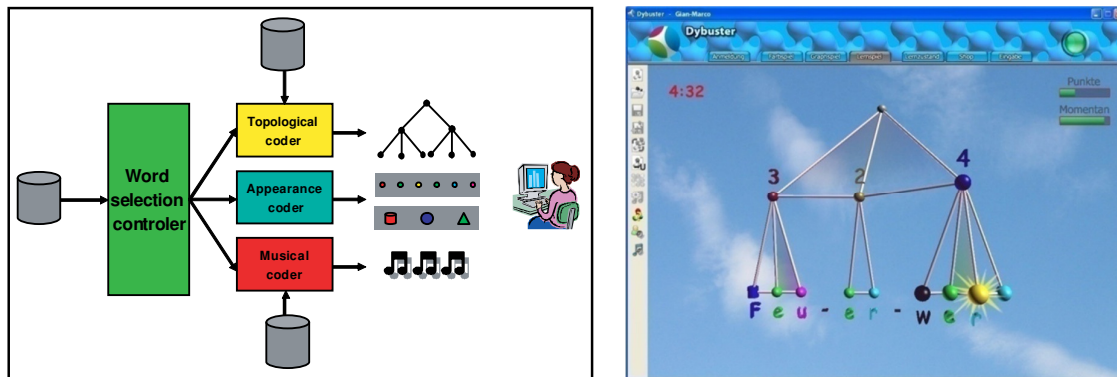


Figure 2: Left: Conceptual components of the Dybuster method and framework. Right: The learning game of Dybuster. Display of all visual cues and instantaneous feedback on an error of omitting the silent letter 'h' (correct: Feuerwehr (engl. firefighters)).

In the second game – the graph game – the students have to segment a word into its syllables and letters graphically. The structure of the word is visualized in a so-called syllable graph. In the third game – the actual learning game – Dybuster presents the alternative representations of a word. Before the students enter the word themselves using the keyboard, the graph appears on screen and the colors and shapes (spheres for small letters, cylinders for capital letters, and pyramids for umlauts) are displayed for all letters (Figure 2, right). A voice dictates a word and the students hear a melody computed from the involved letters and the lengths of the syllables.

A very close relationship between correct spelling and visual memory exists. The results of a Spanish study (López, 1987) state that 83% of spelling is learned visually. To avoid the display of entire misspelled words, Dybuster provides immediate visual and auditory feedback to erroneous inputs. Similar approaches are followed in other spelling training software (e.g., García et al., 2008). This immediate correction is paramount to effective training (Brown, 1990). However, it restricts the error analysis of the input string to the actual error symbol, making unambiguous error classification more difficult. We illustrate this with the example in Figure 3. This example shows that some error categories exhibit identical symptoms and thus cannot be classified unambiguously, not even by a human. Due to the immediate correction, a student can also commit several differing errors at one letter position, e.g., the student commits a typing error while correcting a confusion of 'n' and 'm'.

The implementation of the initial Dybuster version contained a limited representation of the student error behavior. During the training a symbol confusion matrix is maintained. This represents the student's probabilities of confusing a correct letter with a given error letter. Additionally, a local error history of every word is stored to model its individual difficulty for each student. To achieve an adaptation to the student, the word selection from a module is based on the symbol confusion matrix and the local error history. However, the allocation of all 1500 words into the 15 modules is fixed.

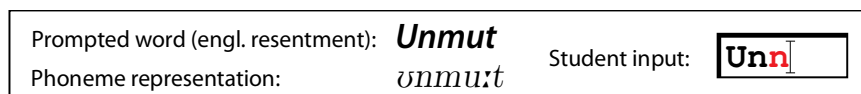


Figure 3: The confusion of the letter 'm' and 'n' could be due to a doubling of the 'n', due to a confusion of similar phonemes /m/ and /n/, or due to the small key distance of 'm' and 'n', (typing error).

The generation of the modules is based on a static word difficulty measure and word frequencies (Gross & Vögeli, 2007), which inhibits an effective adaptation to the student's needs. Additionally, the symbol confusion matrix represents only letter level errors, which consist of a limited set from all error categories of isolated word spelling (see Section "Error Taxonomy and Mal-Rules").

### User Studies

To evaluate the multi-modal spelling training approach, an extensive user study has been conducted at ETH Zurich over a period of six months in collaboration with the Institute for Neuropsychology of the University of Zurich (Kast et al., 2007). The user groups, aged 9-11, involved 43 German-speaking children with developmental dyslexia, and 37 matched children with normal reading and writing skills. They were divided into four different groups. A group of children with developmental dyslexia (DW) and a control group (CW) both practiced with the training software during a first period. The first period was for three months, four times a week, 15-20 minute sessions. This amounted to a total of about 950 minutes of interactive training. The second dyslexic group (DO) and control group (CO) received no training. In a second cross-over period the conditions were swapped: the groups DW and CW had to suspend training and the groups DO and CO started their three month training period.

The children's writing amelioration was measured by a dictation containing 100 words. A random half of the words were used during the training session and the second half served for testing the children's ability to generalize to novel words. All the children had to pass the writing test before training, after three months and at the end of the study. The results of the spelling tests showed a significant improvement. The writing skill of the children with dyslexia DW improved by 27% on average. Whereas the counterparts DO without training improved only 4%. There was no improvement at all on 1/3 of the DO group; thus proving the effectiveness of the multi-modal training method. Furthermore, the DW group improved by 32% on words from the learnt subset and 23% on the generalization dataset. This result leads to the conclusion that the recoding can effectively generalize to new and unknown words – a highly desirable property. Finally, compared to non-dyslexic children, the groups CW and CO improved by 27% and 17% respectively. Figure 4 gives a graphical summary of the results.

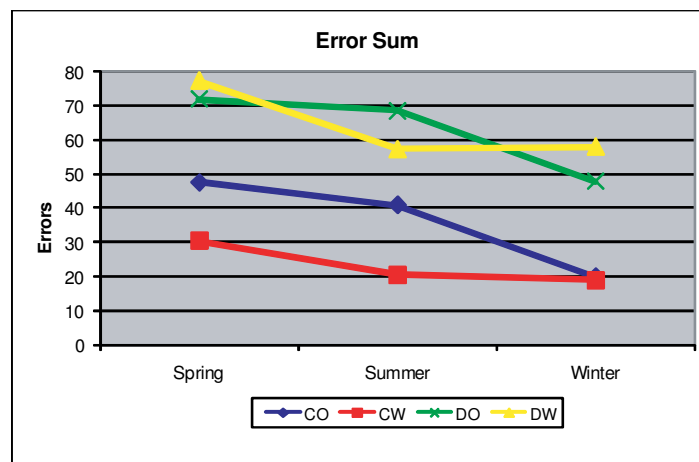


Figure 4: Improvements in dictation error for all participating groups (Kast et al., 2007).

Table 1: Training frequencies for first and second user study

	<b>Mean number of training sessions</b>	<b>Mean minutes per session</b>	<b>Mean total training minutes</b>	<b>Mean inputs per session</b>
First study	51.0	18.9	965.3	58.0
Second study	57.3	16.3	937.5	60.2

The pre- and post-spelling tests provide information about the progress of spelling skills over the entire three months of training. To design appropriate mal-rules and a student model for spelling, we have to investigate the spelling process of individual students in more detail. During the spelling training, every keystroke was timestamped and stored in a logfile. Due to a technical problem some logfiles were corrupted during the training. Despite this loss of data, we obtained 54 usable logfiles (28 dyslexic/26 control) from the first study. The recorded student input data enables an exact reconstruction of the training process. The student model described in this publication emerged from an extensive analysis of the available user data. In addition, the data serves as input for the verification process in the results section.

A second user study has been conducted to evaluate the student model presented in this publication. The novel student representation and corresponding remediation actions have been implemented into the Dybuster framework, and tested under real-world conditions. 40 dyslexic and 27 matched control children were participating in the second user study. Children were recruited via responses to letters distributed in elementary schools, or presentations in school classes where the program was demonstrated. Additionally, the recruitment of dyslexic children was done via therapists or school psychological services. All children were native German speakers. Subjects were classified as dyslexic if their scores in the standardized writing and reading tests were below the 10<sup>th</sup> percentile. Non-dyslexic children showed standard reading and spelling skills. As in the first study, the test battery included classical German spelling (“Salzburger-Lese und Rechtschreibtest SLRT”; Landerl et al., 1997 or “Diagnostischer Rechtschreibtest für fünfte Klassen DRT5”; Grund et al., 1995) and reading tests („Zürcher Lesetest ZLT“; Linder et al., 2000) in order to quantify writing and reading performances. Also, a standard German intelligence test (HAWIK III; Tewes & Rossmann, 1999) was administered to assure (above) average general cognitive skills in subjects (IQ>85). Children with an IQ below 85 were excluded from the study. In both studies the test results show (see Appendix A), as expected, that dyslexic and non-dyslexic children differ significantly in reading and spelling performance. However, there are no significant differences in age, school grades and IQ.

Like in the previous study, half of the children were training for three months and suspended their training for the other three months, whereas the other half participated in the cross-over setting. Again, the children were training for 15–20 minutes four times a week amounting to a total of about 950 minutes of interactive training (see Table 1 for details). Generally, the training took place at their home computers. Also, once a week the participants had the option of supervised training at our lab. During training, children had to study the same 1500 words as in the first study, with a level of difficulty corresponding to their elementary grade. The results of the student model evaluation study are shown in the results section.

## ERROR TAXONOMY AND MAL-RULES

As stated in the previous section, dyslexic children suffer from independent difficulties in many aspects of spelling. The possible error sources of the example shown in Figure 3 include phoneme-grapheme matching, auditory letter confusion, and typing difficulties. Although these difficulties lead to the same error, they are independent of each other. To allow for an adequate adaptation to the student's needs, the model has to represent the student knowledge based on mal-rules describing different error sources. This section examines the error sources relevant to isolated word spelling. Additionally, it presents our corresponding, independent mal-rules. A crucial factor for the design of mal-rules is the requested and available level of information to detect an error category. Spell checkers, e.g., Aspell (Atkinson, 2006), rely on the entire input of a word to build the suggestion set of correct spelling. However, due to the immediate feedback on errors in recent spelling training software, our analysis of errors is restricted to the input up to the error letter. In addition to the error letter, the student input allows for an estimation of the error phoneme. The mal-rules have to be constructed according to this special setting and work only on correct word, input up to the error letter and an estimation of the error phoneme. Neither supplementary information about the surrounding phonemes, nor the entire input is available.

James presents a spelling error taxonomy in 'Errors in Language Learning and Use' (1998). The error categories are structured according to the sources of errors. For example, James' category

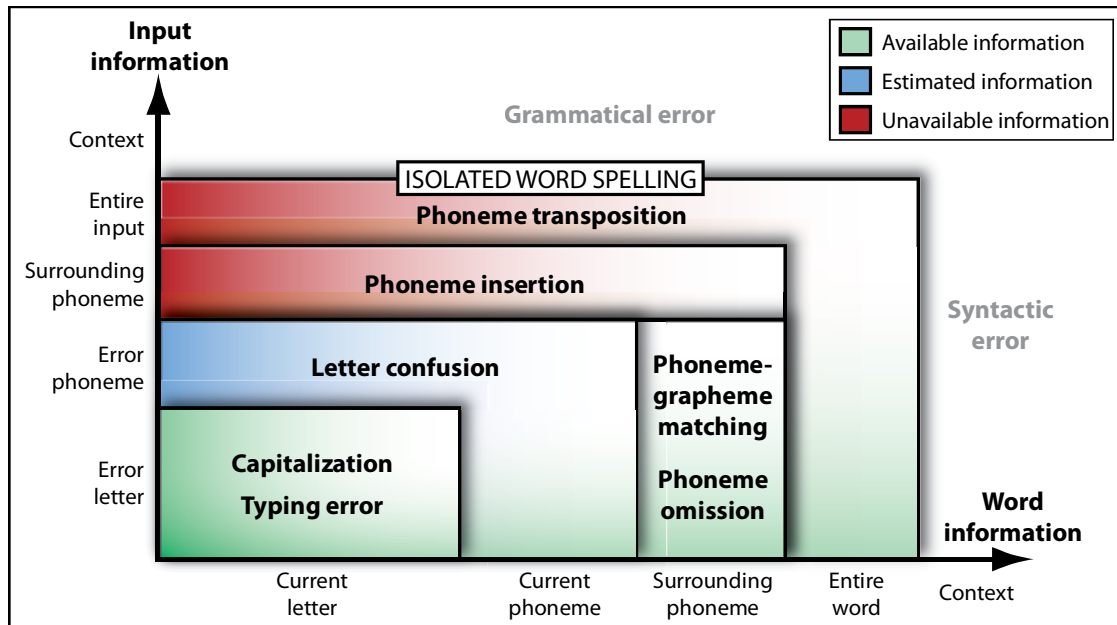


Figure 5: Error taxonomy for isolated word spelling, structured according to the information required about correct word and student input to detect the error category. Information about correct word and its phoneme representation are fully available. Reliable information about the erroneous input is only available with respect to the error letter. The error phoneme can be estimated, but the surrounding phoneme and the entire input are unknown.



*Dyslexic Errors* includes some letter confusions ('d'-'b'), as well as parts of the phoneme-grapheme matching errors, which have different requirements about available input information to be detectable. In the following we describe an error taxonomy for isolated word spelling, structured according to the required information about the correct word and the student input to detect an error category (see Figure 5). Additionally, we present our set of mal-rules for each category, if detection is possible based on the available input data.

### **Capitalization (Cap)**

Capitalization errors are upper- and lower-case confusions, which occur more in the German language versus other Western languages. Over 15% of the spelling errors made in scholarly essays are capitalization errors (Augst, 1985). The German-specific difficulty is that letters are not only written upper-case if they are at the beginning of a sentence or a name, but also every noun is capitalized. However, detecting capitalization errors is simple and unambiguous, and requires only local information about correct and error letters.

*Mal-rules:*

- *ToLowerCase (binary):* Typing a capital letter lower-case.
- *ToUpperCase (binary):* Typing a lower case letter upper-case.

### **Typing Error (Typo)**

Typos are errors committed due to typing difficulties. James states that typing errors can be divided into spatial and temporal errors. In spelling software utilization we mostly face users of the second or third grade of school, which do not master the touch typing system yet. Due to their slow average typing speed, permutations of the letter sequence (temporal errors) like *sehr* (engl. very) to *sher*, are not good indicators for typing errors. The investigation of our input data of the first user study has shown that the probability of typos strongly depends on the distance between correct and error letters on the keyboard (spatial errors), e.g., *sehr* to *aeher*. Therefore, a detection of the typo error category requires only correct and error letters, and is dependent on the input device used for the training.

*Mal-rules:*

- *KeyDistance (categorical):* The spatial distance between correct and error key. The analysis of the student data showed that the typo error probability for all keys more distant than the surrounding keys of the correct letter does not differ significantly from zero. Therefore, we discarded attempts to model the error probability - key distance relation by a Gaussian decay function and introduced three categories of key distances to avoid a non-linear optimization problem. As shown in Figure 6, we have the most error-prone category *Left/Right*, the keys to the top left, top right and on the bottom are combined to *Top/Bottom* and all the other keys belong to the category *Distant*.
- *Technical (binary):* German umlauts cause problems to type on the keyboard. From handwriting, the children are used to write the vowel first and subsequently put additional dots on top of it. This causes a high confusion rate between umlauts and their corresponding unmutated vowels. To model this input device-specific difficulty, we introduce a binary mal-rule *Technical*.

	T	Z	U	I
F	G	H	J	K
	V	B	N	M

Figure 6: Key distance from the correct (green) letter 'H'. The two keys 'G' and 'H' belong to the category *Left/Right* (red). The keys 'Z', 'B' and 'N' build the category *Top/Bottom* (orange). All the other keys are *Distant* (yellow).

### Letter Confusion (LetC)

Letter confusion denotes all permutations of letters. These can occur due to visual similarity of letters, e.g., 'd'-'b', or due to an auditory similarity of corresponding phoneme, e.g., /n/-/m/. The visual confusion can be detected on a letter level, e.g., the horizontal mirror image of 'd' equals 'b'. Confusions caused by an auditory similarity of the correct and the erroneous input require information about the current correct and error phoneme. The underlying difficulty of confusing *Niete* [nirte] (engl. rivet) with *Miete* [mi:te] (engl. rent) is only revealed on a phonological level.

*Mal-rules:*

- *VisualSimilarity (continuous):* We introduce a visual similarity mal-rule based on the cross-correlation between the images of letters. This is computed using the *CH3 Steinschrift* font, which corresponds to the hand writing taught in Swiss schools (see Figure 7). Since most letters in the written language are lower-case, the confusion should be calculated on their lower-case representation. However, due to the capitalized letters on the keyboard, we compute three visual distances: one between lower-case letters *VS(LowerCase)*, one between capitals *VS(UpperCase)*, and a distance between lower- and upper-case letters *VS(LUCase)*. Especially for dyslexic children, confusions of letters appear more frequently if the letter pairs have a high visual similarity when mirrored horizontally, for example the common 'd'-'b' confusion. To evaluate the visual similarity of a letter pair, we therefore take the maximum of the two cross-correlation values of actual and horizontally mirrored image.



Figure 7: The CH3 Steinschrift font used to compute the visual similarity of letters.

- *AuditorySimilarity (categorical):* Our auditory similarity measure between a correct and a false phoneme is based on the hierarchical phoneme structure proposed by Dekel et al. (Dekel et al., 2005). We modified the structuring of vowels to better address our findings regarding vowel confusion probabilities in the user data (see Figure 8). Confusions between nearby phonemes along the edges /i/-/a/ and /a/-/u/ of the so-called vowel triangle (Hall, 2000) are more likely to happen, and are thus labeled as similar. We define auditory similarity (AS) as a

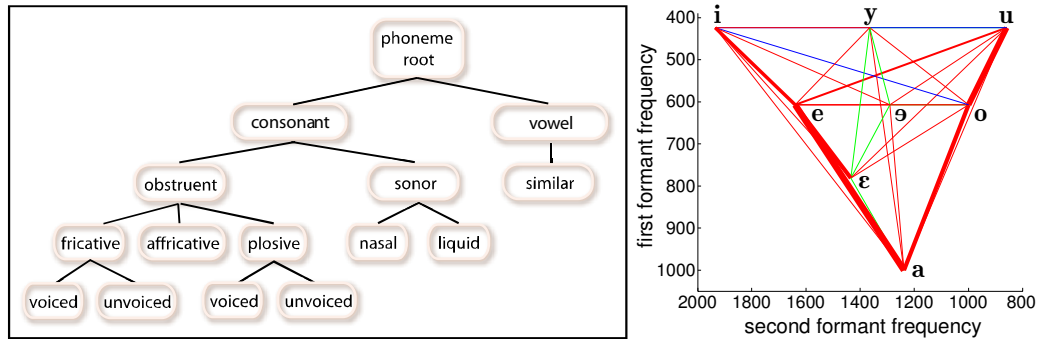


Figure 8: Hierarchical phoneme structure (*left*): Each phoneme is assigned to one of the leaf nodes, each representing a phonetic attribute. Phoneme triangle (*right*): Vowels are positioned with respect to their first and second formant frequencies on the Mel-scale (Fant, 1960). Empirical confusion probabilities are represented by the thickness of the connecting red lines. Blue and green lines indicate a significant influence of other features (key distance and technical).

categorical feature representing the nearest common ancestor node of the correct and the closest possible error phoneme.

### Phoneme-Grapheme Matching (PGM)

Phoneme-grapheme matching is one of the most frequent error categories. PGM errors are caused by the non-bijectionality of the phoneme-grapheme correspondence, i.e., many phonemes can be represented by multiple graphemes. In German, this non-bijectionality can be divided into three subcategories:

- *Elongation*: Several vowel phonemes exist in a short and an elongated version. For example, the phoneme /a:/ can be represented by the grapheme 'a', as well as by the graphemes 'aa' and 'ah'. This doubling of vowels and the additional silent 'h' are called elongation.
- *Sharpening*: Doubling of consonants, like 'ss', 'nn' or 'pp', are called sharpening.
- *Phoneme Matching*: The remaining phoneme-grapheme correspondences are simply named phoneme matching. Common error sources are the phonemes /ɔy/ ('äu', 'eu') and /f/ ('v', 'f').

All of the three subcategories have different grapheme representations that are pronounced the same way. The correct one cannot be determined by careful listening alone. The phoneme-grapheme correspondence must be deduced from language rules, or learned by heart. To detect these errors the information about the surrounding phonemes of the correct word and about the current phoneme of the erroneous input needs to be known, as will be shown in the following.

#### Mal-rules:

To detect phoneme-grapheme matching (PGM) errors, we align the user input and the phonological structure of the correct word. Kondrak (2003) presents an algorithm to phonologically align words based on string alignment. However, such methods request the entire string of input and correct word. To allow for an identification of PGM errors based only on the input up to the error letter, we use a novel, local alignment of the phonological structure of input and correct word. The algorithm compares the error phoneme with the current, following, and previous phoneme:

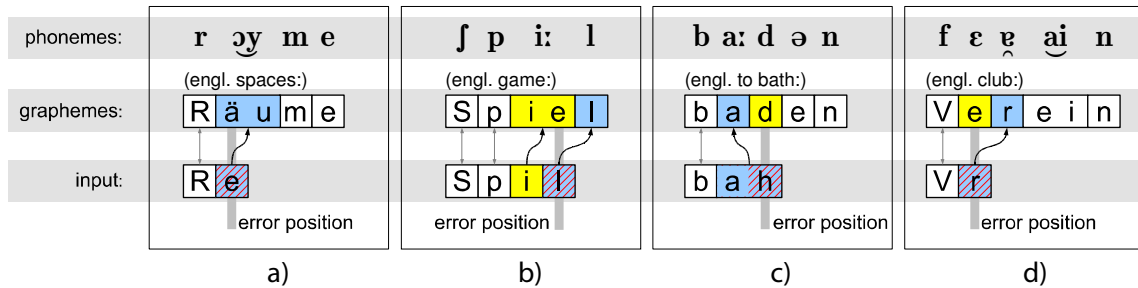


Figure 9: Alignment of correct and input phonemes and resulting mal-rules: a) Phoneme matching b) Letter omission c) Letter addition d) Phoneme omission

- *PhonemeMatching*: A false letter can be part of a wrong representation of the correct phoneme. For example, in Figure 9.a) the false letter 'e' is the beginning of the grapheme 'eu', which is a representative of the correct phoneme /əy/.
- *LetterOmission*: If a false letter marks the beginning of the next phoneme and the current phoneme is falsely represented by the previous input grapheme, we face a *LetterOmission*, such as in Figure 9.b): the error letter 'l' matches the following phoneme, and the current phoneme /i:/ is incorrectly represented by the grapheme 'i'.
- *LetterAddition*: The previous input grapheme concatenated with the false letter can match the previous phoneme. In Figure 9.c) the false letter 'h' appended to the previous input grapheme 'a' results in the grapheme 'ah' - which is a representative of the previous phoneme /a:/.

To discriminate the errors in greater detail, we further subdivide the mal-rules presented above. In *PhonemeMatching*, we distinguish between *Vowel* and *Consonant* phonemes, as well as between *Main* and *Special* graphemes. The attributes *Main* and *Special* are manually attached to every grapheme. They indicate whether a grapheme is the most likely (main) representative of the phoneme or an unusual (special) one. *LetterOmission* and *LetterAddition* are both subdivided into *Elongation* and *Sharpening* based on the type of phoneme the error occurred in (*Vowel/Consonant*). These features are language specific and the phoneme-grapheme correspondence has to be adapted if the features are applied in other languages.

### Phoneme Omission (PhoO)

The error of omitting a complete phoneme while entering a word is called phoneme omission. It is a common error among dyslexic children. A typical cause of omitting an entire phoneme is, e.g., the phoneme /ɛ/ in the word *Verein* [fɛrain] (engl. club). To detect a phoneme omission, the information about current and surrounding phonemes of the correct word is required. From the erroneous input we only need the current phoneme, which has to match the next phoneme in the correct word to indicate a phoneme omission.

*Mal-rules*:

- *PhonemeOmission*: If a false letter marks the beginning of the next phoneme and the current phoneme is completely omitted, we detect a *PhonemeOmission*. As displayed in Figure 9.d), the incorrectly entered grapheme 'r' matches the following phoneme /ɛ/. The current phoneme /ɛ/ has been omitted.

## Phoneme Insertion & Phoneme Transposition

The insertion of an additional error phoneme and the transposition of two phonemes in a word are called phoneme insertion and phoneme transposition respectively. Their detection requires information about the surrounding phonemes, or even the entire input of the word, which are not available to the error analysis. Therefore, these error categories are not detectable from the available student inputs.

## Feature Vector

The presented mal-rules allow for a detailed description of errors. Each provides error information in one of the three following forms:

- *Binary*: A mal-rule has been applied or not. For example, writing a letter in false case (Capitalization).
- *Categorical*: The described mal-rule combines multiple states. For example, key distance (Typing error) contains confusions with letter to the left or right (*Left/Right*), with letters to the top or bottom (*Top/Bottom*) or with distant keys (*Distant*).
- *Continuous*: Mal-rules, such as *VisualSimilarity*, provide a continuous value for a given error. This indicates to what extent a mal-rule is involved in an error.

For a given error  $e$ , each mal-rule returns one or several action values. In the binary case this is simply one or zero. Categorical mal-rules are partitioned into multiple, binary dummy variables. In the continuous case we obtain a value between zero and one, indicating the action level of the mal-rule. All mal-rules  $f_i(e)$  together describe a single error and are represented in a feature vector  $\mathbf{f}(e)$ .

## STUDENT MODEL

In this section we present the student knowledge representation for spelling, which provides the requested information about the student's strengths and weaknesses, the prediction of further spelling performance (global information), and a classification of committed errors (local information). We decided to design a perturbation model for spelling to avoid a specification of the strongly differing spelling processes of children and rather rely on certain mal-rules, which describe errors. In a perturbation model the student is typically represented by error probabilities for each of those mal-rules. Due to the possibility of entering multiple errors at one single letter position, we take a different viewpoint on errors. We regard errors as randomly occurring events and try to estimate their rate of occurrence dependent on the mal-rules involved in an error. Randomly occurring events are best described by a Poisson distribution, which is characterized by the expected number of events that occur during a given time interval, or a given number of exposures to risk. In the case of spelling, every possible error in a prompted word is considered as an exposure to risk and we are interested in the corresponding error expectation values. A Poisson regression allows for an estimation of the error rate of mal-rules for every student based on its currently available input data. Based on these error rates we can predict the further spelling performance on the entire word database and provide a probabilistic classification of committed errors. The prediction and classification of errors will be described in detail at the end of this section. This local and global information about a student can be used to adapt the training to the student needs and to choose appropriate remediation actions.

## Data Collection

The student model, representing the student's difficulties with individual mal-rules, has to be estimated out of the available input data of a student. We analyze every input of a student and distinguish between *possible errors*, i.e., error inputs which could potentially be entered at the given word, and *committed errors*, i.e., error inputs which the student actually entered during the training. Each letter of a word contains 29 potential errors, namely one capitalization and 28 confusions with all other letters. The collection of possible and committed errors for the *Unmut* error example (see Figure 3) is illustrated in Figure 10. Every possible error  $e$  of all the prompted words could be considered as an item on which, when presented to the student,  $r$  errors occurred. There is an average of 6.7 letters per word and a database of 1500 words, which yields approximately 300,000 items that the student can be tested on. The presented mal-rules allow for a detailed description of these errors. The feature vector  $\mathbf{f}(e)$  indicates the extent to which each mal-rule is activated in an error  $e$ . Tatsuoka (1985) proposed a **Q**-matrix, in which all the feature vectors for every item are assembled. In the case of spelling with its large amount of items, this would result in an approximately  $300,000 \times$  number-of-mal-rules **Q**-matrix. Therefore, we do not store the feature vector for every item, but continuously compute and assemble only the feature vectors  $\mathbf{x}_i = \mathbf{f}(e)$ , which were actually presented to the student. For each  $\mathbf{x}_i$  we count the number of times an error possibility described by the feature vector  $\mathbf{x}_i$  was encountered ( $N(\mathbf{x}_i)$ ), and how often such an error was actually committed by the student ( $Y(\mathbf{x}_i)$ ). This reduces the number of different feature vectors down to 363 for the first user study.

## Inference Algorithm

The inference algorithm has to estimate the student's difficulties with each individual mal-rule, based on the observed error behavior described by  $Y(\mathbf{x}_i)$  and  $N(\mathbf{x}_i)$  for all  $\mathbf{x}_i$ . Due to the possibility of multiple errors, we consider errors as randomly occurring events. These are best described by a Poisson distribution, and the probability distribution for every variable  $Y(\mathbf{x}_i)$  is defined as:

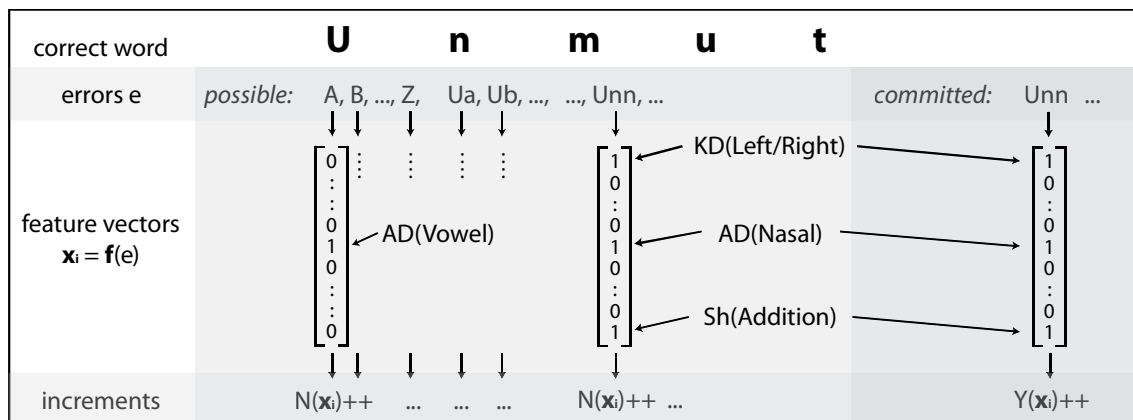


Figure 10: Data collection for the prompted word *Unmut*: Every letter of the word is permuted with each letter from the alphabet, the feature vector  $\mathbf{x}_i = \mathbf{f}(e)$  describing this possible error  $e$  is computed and the corresponding occurrence counter  $N(\mathbf{x}_i)$  is incremented. For every error  $e$  the student actually committed in this word, we compute the feature vector  $\mathbf{f}(e)$  and increment the corresponding error counter  $Y(\mathbf{x}_i)$ .

$$P(Y(\mathbf{x}_i)) = \frac{e^{-\mu(\mathbf{x}_i)N(\mathbf{x}_i)} (\mu(\mathbf{x}_i)N(\mathbf{x}_i))^{Y(\mathbf{x}_i)}}{Y(\mathbf{x}_i)!} \quad (1)$$

where  $\mu(\mathbf{x}_i) > 0$  denotes the rate parameter and  $N(\mathbf{x}_i)$  the number of exposure to risk. The expectation value of  $Y(\mathbf{x}_i)$  in a Poisson distribution is given by  $E[Y(\mathbf{x}_i)] = \mu(\mathbf{x}_i)N(\mathbf{x}_i)$ . The error rate  $\mu(\mathbf{x}_i)$  has to be related to a linear combination of the unknown student parameters  $\boldsymbol{\beta}$  and the feature vector  $\mathbf{x}_i$  by a link function  $g(\cdot)$ :

$$\boldsymbol{\beta}\mathbf{x}_i = g(\mu(\mathbf{x}_i)) \quad (2)$$

The Poisson regression to estimate the student parameters  $\boldsymbol{\beta}$  is part of the family of generalized linear models (McCullagh & Nelder, 1989). McCullagh and Nelder propose the canonical log link function for the Poisson regression, which has the beneficial property of matching the domain of the link function with the range of the non-negative rate parameter. However, they state that, since the log link function induces a multiplicative effect on the mal-rules, the appropriateness of this selection has to be checked. To make allowance for the independence of the presented mal-rules, we request an additive behavior of the error rates of individual mal-rules. For example, if a certain error activates a typo and a PGM mal-rule, we expect the error rate on such an error possibility to be the sum of the error rates of the typo and the PGM mal-rules, since these two error types are independent and do not influence each other. Therefore, we propose the usage of a linear link in the presence of independent mal-rules.

To compare the effect of the two link functions we employ the Tukey-Anscombe plot (Anscombe & Tukey, 1963) for residual analysis. The plot shows the inter-relation of fitted values of the method

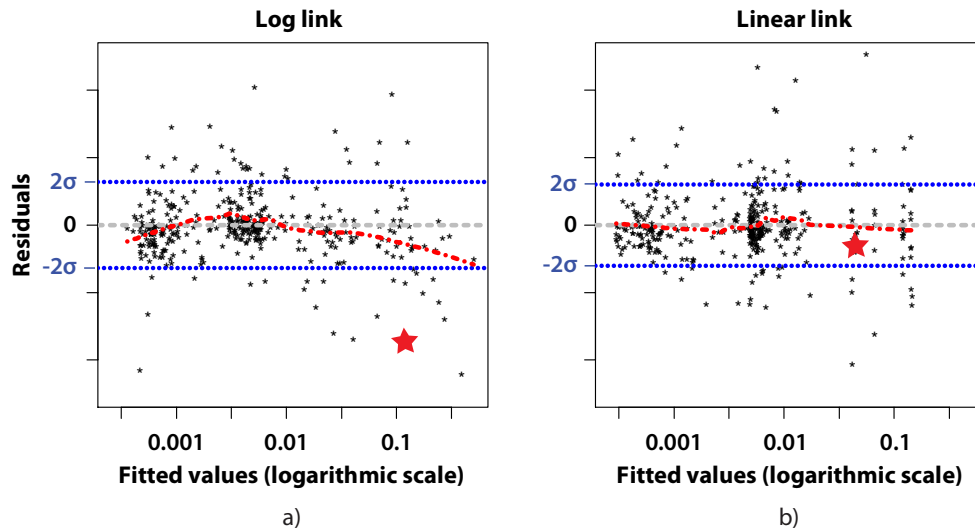


Figure 11: Tukey-Anscombe plot for log (a) and linear (b) link function on a logarithmic scale.  $\sigma$  indicates the median absolute deviation of the residuals and the red line shows the LOWESS smoothed expectation value (Cleveland, 1981) of the residuals. The log link method suffers from a non-zero error expectation (red line), due to the multiplicative interdependence of the mal-rules. This is depicted (red star) by the *Netz - Nez* error example (see Figure 11).

Prompted word (engl. net):	<b>Netz</b>	Student input:	<b>Nez</b>
Phoneme representation:	<i>nɛts</i>		

Figure 12: The omission of the letter 't' could be due to a phoneme-grapheme matching error ('tz' and 'z' are both representatives of the phoneme /ts/) or due to the small key distance of 't' and 'z' (QWERTZ keyboard).

used and its residuals. The residuals should be normally distributed with expectation zero, and a constant variance on the entire scale of fitted values. The analysis is performed on the input data of all 80 children of the first user study. As can be seen in Figure 11.a), the non-zero error expectation value (red line) of the log link clearly depicts the violation of the independence assumption of individual mal-rules. Low error expectation values with single mal-rules activated are rather underestimated (positive residuals), and the combination of mal-rules results in overestimated error expectation values (negative residuals). As Figure 11.b) shows the linear link provides a zero expectation value across the entire fitted value scale. One can see how reasonable estimates are assured even for the higher fitted values. This effect is additionally illustrated by the error example in Figure 12. The error has the two mal-rules *Left/Right* (typing error) and *ConsonantSpecial* (PGM error) activated. Table 2 shows the estimated error probabilities for an error with only *Left/Right* or *ConsonantSpecial* activated and the combination of both, as it is the case in the presented example. The multiplicative interrelation of mal-rules of the log link leads to an overly pessimistic estimation of the error probability (0.116 compared to the measured 0.036). The residual of the estimation is visualized by a red star in Figure 11. In contrast, the error rates of the linear link function show an additive behavior. The estimated error expectation (0.045) for errors having both features activated corresponds more closely with the empirical expectation value (0.036) as shown in Table 2.

This error example indicates the importance of the additive interrelation of mal-rules. The remediation actions will focus on the most severe weaknesses of the student. Therefore, if the log link function leads to strongly overestimated error rates of errors with multiple mal-rules activated, then those errors will experience an undesired bias in remediation.

The appropriateness of the linear link function can additionally be evaluated by comparing the two regression models based on Akaike's information criterion (Akaike, 1974). The AIC score is a measure of the goodness of fit of an estimated statistical model, where a lower score is better. The log link and the linear link model yield an AIC score of -10538 and -11138 respectively, which shows the superiority of the linear link in the context of independent mal-rules.

### Significance of Mal-rules

To evaluate the significance of individual mal-rules, we employed the likelihood ratio (LR) test (Cameron & Trivedi 1998). This test returns the log-ratio of the likelihood of the full model to the

Table 2: AIC score and estimated error expectation values for log and linear link. As a comparison, the empirically measured error expectation value for the Netz-Nez error example are given.

Method	AIC score	<i>KD(Left/Right)</i> activated	<i>PGM(ConsSpec)</i> activated	Both activated (e.g. Netz - Nez)
Log link	-10538	0.002	0.028	0.119
Linear link	-11138	0.004	0.041	0.045
Empirical error expectation value				0.036



likelihood of a model with the given mal-rule left out. These values asymptotically follow a  $\chi^2$  distribution with one degree of freedom. The computation of the likelihood of a Poisson regression depends on the dispersion parameter  $\Phi$  (McCullagh & Nelder, 1989). The models fitted into the data of individual students showed a mean over-dispersion of 3.1 and the likelihood has been corrected with the respective dispersion  $\Phi$ .

Not every student struggles with all difficulties represented by the set of mal-rules. To justify the appropriateness of a mal-rule, we investigate the highest LR score across all N=54 students to determine whether the mal-rule is significant at the  $\alpha=5\%$  level for at least one of the students. Consequently, we have to apply a false discovery rate correction. The Šidák correction ( $\alpha_c=1-(1-\alpha)^{1/N}$ ; Abdi, 2007), well suited for the independent tests of individual students, yields a corrected significance level  $\alpha_c$  of 0.095%.

As can be seen in Appendix B, most mal-rules are highly significant. This is especially true for all capitalization, phoneme-grapheme matching and phoneme omission features. In the letter confusion error category, 4 out of 14 auditory confusion nodes from the hierarchical phoneme structure are not significant. This is mainly due to the fact that most confusions of fricative and fluid sounds are very sparsely sampled. We estimate the closest possible phoneme for the auditory confusion and many fricative and fluid phonemes are represented by the same graphemes. This often results in a detection of phoneme-grapheme matching errors rather than auditory confusion.

An interesting finding is the non-significance of the visual similarity mal-rules for both lower and upper case letters. This indicates that dyslexics suffer from auditory and phonological processing deficits, but not from an impaired visual processing. This corresponds well with recent findings in dyslexia research (Ramus, 2003).

### Error Classification and Prediction

The student representation provides an estimate of the individual mal-rules difficulty. During the training, the representation of the student's mastery of the domain is continuously updated after each entered word. Based on these estimates we can compute a prediction of further spelling performance and a classification of committed errors for each individual student. This information is expressed by the following two values:

- $P_C(k|\mathbf{e})$ : The probability that the  $k^{\text{th}}$  mal-rule is the source a committed error  $e$ .
- $E[E|w]$ : The expected number of errors a student will make on the spelling of the word  $w$ .

To provide the probability of the  $k^{\text{th}}$  mal-rule being the cause of an error described by  $\mathbf{f}$ , we employ Bayes' theorem.  $P(\mathbf{f}|k)$ , the probability that an error occurs if the  $k^{\text{th}}$  mal-rule is causing an error, is always equal to 1.

$$P_C(k|\mathbf{f}) = \frac{P(\mathbf{f}|k)P(k)}{P(\mathbf{f})} = \frac{P(k)}{P(\mathbf{f})} = \frac{\beta_k f_k}{\beta \mathbf{f}} \quad (3)$$

The estimated error rates  $\beta$  are used as a prior probability  $P(k)$  of the  $k^{\text{th}}$  mal-rule causing an error. The probability  $P_C(k|\mathbf{f})$  corresponds to the part of the  $k^{\text{th}}$  mal-rule on the total error expectation of a given error described by  $\mathbf{f}$ .

The error expectation estimate  $E[E|w]$  for the word  $w$  can be obtained by summing over the error expectation values of the errors contained in the set  $S_c(w)$  of all possible confusions in the word  $w$ .

$$E[E|w] = \sum_{e \in S_c(w)} \mu(\mathbf{f}(e)) = \sum_{e \in S_c(w)} \boldsymbol{\beta} \mathbf{f}(e) \quad (4)$$

This allows for a prediction of the spelling performance on every word in the entire word database, even if the word has never been prompted so far.

Formulae 4 and 5 indicate that the classification and prediction of errors is dependent on the individual student parameters  $\boldsymbol{\beta}$ . Varying student characteristics influence the determination of the most likely cause of an error and change the error expectation values, as illustrated in the ‘Results’ Section.

## REMEDICATION

The main contribution of this publication is the presented student model with its novel inference algorithm and its detailed set of mal-rules for spelling. To evaluate the validity of the student representation, the intelligent learning environment had to be extended with appropriate remediation actions and tested in a second user study. The global and local information about the student, provided by the student model, allows for various forms of remediation. These range from an optimized word selection and specialized lessons for individual mal-rules to student adjusted repetition of erroneously spelled words. Additionally, the information about the student can be preprocessed and presented to human tutors to allow for additional remediation actions.

For the second user study we designed simple remediation actions, suitable for a three month training period. The study version of the spelling software contains 1500 words. On average the students worked on 800 words during their training. Therefore, we focus on a selection of words from the database, which optimizes the training gains during the three months of training. Gross and Vögeli (2007) proposed word selection in an error entropy minimizing way. Words are selected based on the average letter entropy of the corresponding student input, i.e., based on the uncertainty about the correctness of an input, expressed by the estimated error probability. In a similar fashion our novel word selection controller tries to minimize the error expectation of the student inputs by selecting words with the highest error expectation per letter ratio from the database. This results in an individual selection of words for each student, with the highest progress potential with respect to training time. To avoid frustration due to excessive demands, the word selection algorithm intersperses easy words, if three erroneous inputs occur in a row. Certainly, this word selection scheme is not optimal for long term training, where larger word sets and longer training periods allow and demand a more structured training setup.

The local repetition behavior on erroneous inputs is composed according to findings on the error repetition in the collected student data. As will be shown in the results section, the optimal time for repetition of errors due to missing spelling knowledge, such as PGM errors, is between 3 and 6 minutes after the erroneous input (see Figure 15.b)). In contrast, the repetition of randomly occurring errors, such as typos, is not time sensitive. The error repetition algorithm is designed with respect to those findings. The spelling software was extended with the presented remediation actions to evaluate the correctness of the information provided by the student model.

## RESULTS

The presented mal-rules and the associated inference algorithm provide global and local information about student input. In a first step, we show an example of error classification and prediction for three selected students to illustrate the influence of different student characteristics. Next, prediction and classification are validated based on the student data of the first user study. We then present our analysis of forgetting, which was used to optimize the remediation actions. Finally, the results of the second study are presented and we investigate the learning progress gain induced by the student model and the corresponding remediation actions.

### Example of Use

The student model provides information about the probability of an error belonging to an error category and a prediction of further spelling performance, dependent on the student characteristics. In this section, we present the student model and its application for three selected students of the first user study, to show the influence of varying student parameters. Subject 1 is dyslexic and has strong difficulties with capitalization and letter confusion. Subject 2 belongs to the control group and has the highest error rate for *Left/Right* typing errors. The main difficulties of subject 3 (dyslexic) are the phoneme-grapheme correspondences. Especially elongation and sharpening are the cause of many errors committed by subject 3. The student parameters of the three selected subjects are illustrated in Figure 13. As the logarithmic scale indicates, the error rates of mal-rules vary by orders of magnitudes. For example, the *KD(Left/Right)* and *ToLowerCase* mal-rules have an error rate around 0.005 and 0.1 respectively. Nevertheless, the *KD(Left/Right)* mal-rule is still relevant, since it is activated for two confusions for every letter in a word. A word, such as *Männer* (engl. men), contains 12 *Left/Right* confusion possibilities, but only one for the capitalization mal-rule *ToLowerCase*.

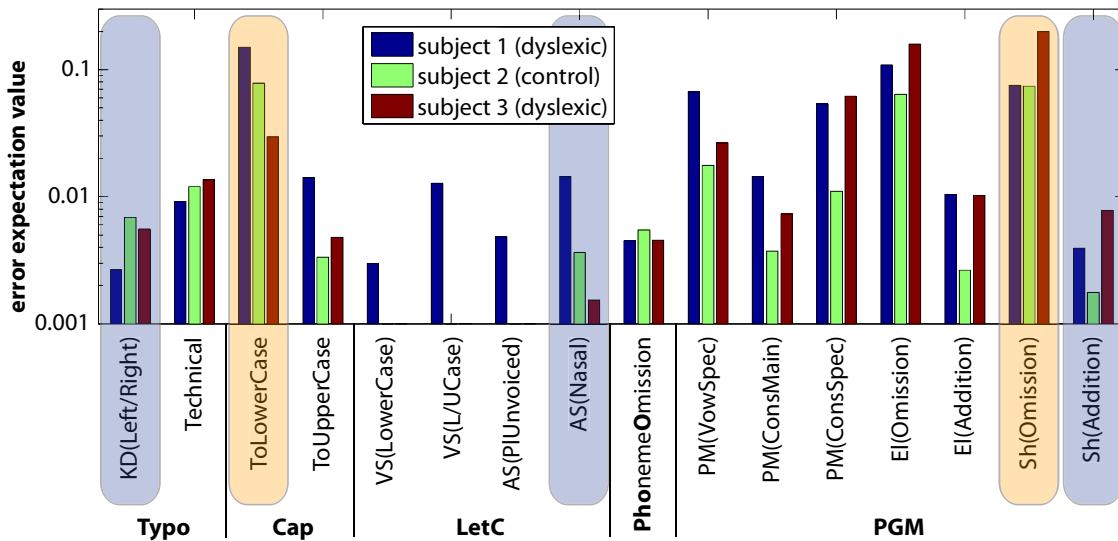


Figure 13: Student characteristics for three subjects of the first user study. All estimated parameters  $\beta$  with a value above 0.001 for at least one of the three students are displayed on a logarithmic scale. Mal-rules relevant for the error classification and prediction examples are highlighted.

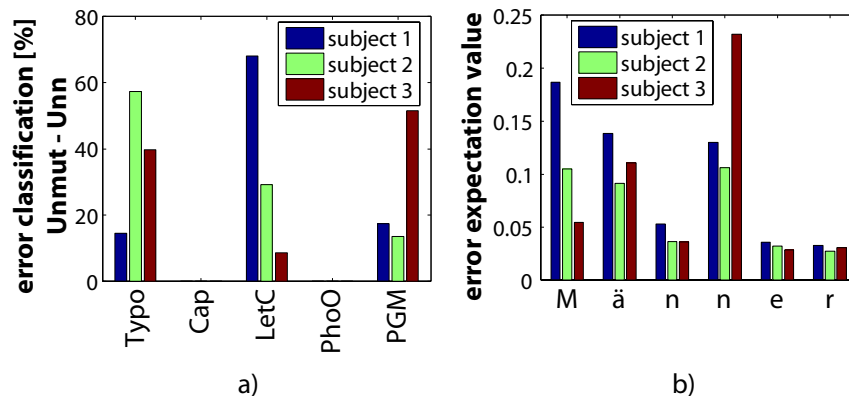


Figure 14: Application of the student model for three students: a) Probabilistic error classification of *Unmut - Unn*. b) Estimated error expectation values for each letter of the word *Männer*.

Figure 14.a) shows the probabilistic classification of the error example *Unmut - Unn* (see Figure 3). Due to the high confusion rate of nasal phonemes (*AS(Nasal)*), the error is classified as letter confusion for subject 1. Subject 2 shows little difficulties with auditory similarities and sharpening errors. Therefore, the error is classified as typing error (*KD(Left/Right)*) for subject 2. The high error rate at *Sharpening(Addition)* makes a phoneme-grapheme matching error most likely for subject 3.

Similarly, we obtain varying error expectations for the three students. Figure 14.b) shows the error expectation for each letter of the word *Männer* (engl. men). The first letter contains a capitalization possibility (*ToLowerCase*). The strong difficulty of subject 1 on capitalization results in an error expectation of almost 0.2, in contrast to 0.1 and 0.05 for subject 2 and 3 respectively. However, the error expectation at the second 'n' is estimated twice as high for subject 3 compared to subject 1 and 2, due to the high error rate on sharpening (*Sh(Omission)*).

These examples show how different student characteristics influence the error classification and prediction. The remediation actions, such as global word selection and erroneous input repetition, are adapted according to this information.

## Validation

As illustrated above, the student model provides a classification and prediction of errors for individual students. To verify the determined cause and expected number of errors, we need the true underlying source and expected number of errors. However, as shown in the error example in Figure 3, some errors are not unambiguously classifiable, even by a human. Due to the lack of a ground truth of the error classification, we investigate the error repetition behavior for the individual error categories on the data of the first study. For the verification of the spelling performance prediction, the error expectation numbers are compared with the empirical measures during the training. Further, we present the analysis of the error repetition behavior, identifying the optimal point in time for repetition, which is incorporated in the remediation of the second study.

### *Error Classification*

The error repetition behavior provides information about how much learning has taken place. The rate of forgetting over time depends on the presence of an unknown concept of spelling involved in an

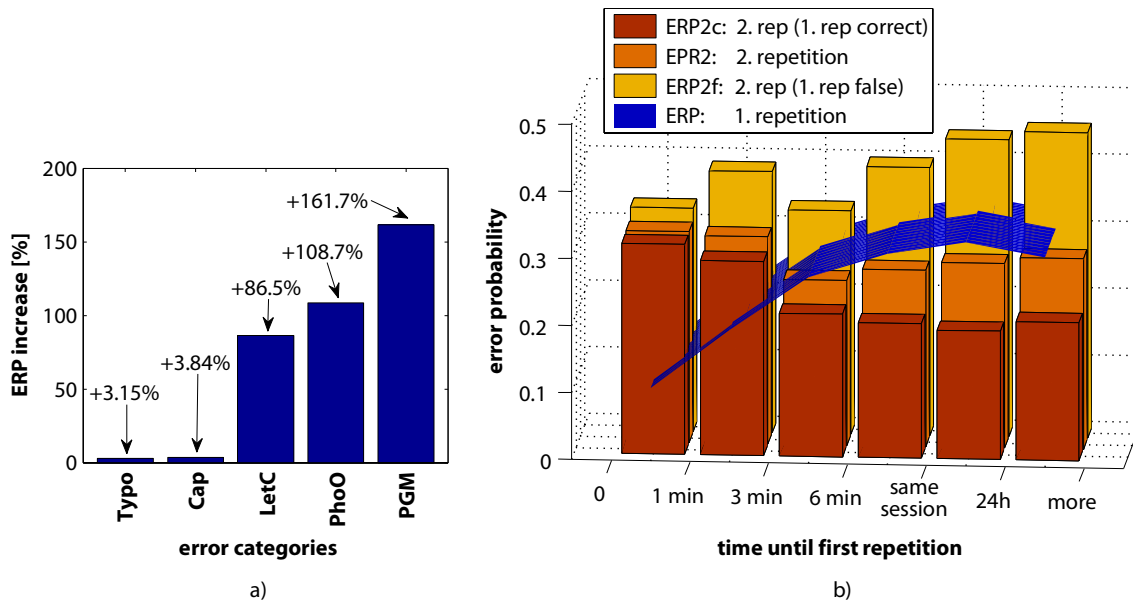


Figure 15: Error repetition behavior: a) ERP increase from less than 60s to more than 60s between error and repetition for all categories. b) Error probabilities at the first and second repetition of a PGM error dependent on the time between error and first repetition.

error. Input device dependent errors, which are randomly distributed over all inputs, are not expected to show an increase in error repetition probability (ERP:  $P(R_I = f)$ ) over time. On the contrary, if students learn and remember word spelling concepts, e.g., correct phoneme-grapheme matching, then the chance of forgetting the just learned spelling will grow over time. This results in a time-dependent increase of the ERP at the first repetition ( $R_I$ ) of the word after the erroneous input.

We investigated the error repetition probability for all five error categories on the inputs of the 54 children of the first user study. The analysis was performed on the 39,600 classified errors on which a repetition has been recorded. Figure 15.a) shows the time dependence of the error repetition probability at the first repetition. The blue bars indicate the increase of the ERP from repetitions with less than 1 minute between erroneous input and first repetition to repetitions with more than 1 minute. The ERP increase of typo ( $p=0.85$ ) and capitalization ( $p=0.91$ ) was not significant. In contrast, the increase for letter confusion, phoneme omission and phoneme-grapheme matching was highly significant (all  $p<0.001$ ). The significance was evaluated in a chi-squared congruency table test (Everitt, 1992). Typing errors are randomly distributed across all inputs and do not depend on spelling knowledge, which can be taught. The non-significant ERP increase of only 3.15% verifies the classification of typos. Letter confusion, phoneme omission, and phoneme-grapheme matching show a highly significant ERP increase, which indicates that errors classified in one of these categories are based on missing spelling knowledge and can be remediated. This finding is in accordance with the common assumption about those error categories. The low, non-significant ERP increase for capitalization shows that the students actually know when to use upper or lower case letters and the difficulty of capitalization is rather the unusual way of entering an upper case letter by using the shift key.

### *Optimal point in time for repetition*

For the verification of our remediation actions, we investigate the impact of the point in time for repetition onto the long-term learning of spelling. As Figure 15.a) indicates, the earlier an erroneous input is repeated, the lower is the corresponding ERP. At first appearance, that demands a repetition of erroneously entered words as early as possible, to avoid subsequent errors on the same word. However, the goal of repetition of errors is the long-term learning of the correct spelling. This is measured by means of the error repetition probability (ERP2:  $P(R_2 = f)$ ) for the second repetition ( $R_2$ ) of the word after the erroneous input. We consider only inputs with more than 12 hours between first and second repetition to exclude correct spelling in the second repetition by retrieval from the short-term memory. This ERP2 value provides a measure for learning efficacy, and is used to determine the optimal point in time for repetition. We expect the long-term learning to be influenced by two opposing effects:

- A short time span between error and first repetition enables a correct repetition due to correct spelling knowledge retrieved from the short-term memory. If the first repetition was correct after a longer time span, the probability is higher that the correct spelling was actually stored in long-term memory. Therefore, a large time span between erroneous input and first repetition will decrease the error repetition probability at the second repetition, if the first repetition was correct (ERP2c:  $P(R_2 = f | R_1 = c)$ ).
- A second erroneous input after a short time span between error and first repetition does not evidence as much a false representation in long-term memory, as is the case if the word is repeated erroneously after a longer period of time. Therefore, a large time span between error and first repetition will increase the error repetition probability at the second repetition, if the first repetition was false (ERP2f:  $P(R_2 = f | R_1 = f)$ ).

For PGM errors, these two effects are illustrated in Figure 15.b). The front and back row bars indicate the dependence of ERP2c and ERP2f on the time span between error and first repetition. The requested ERP2 value can now be computed by the Bayesian rule as a weighted sum of ERP2c and ERP2f, weighted by the probability of an error at the first repetition (ERP; plane in Figure 15.b)):

$$P(R_2 = f) = P(R_2 = f | R_1 = c) \cdot P(R_1 = c) + P(R_2 = f | R_1 = f) \cdot P(R_1 = f) \quad (5)$$

As can be seen in Figure 15.b), the ERP2 (middle row bars) for PGM errors is high for short time span, due to the high ERP2c value. The ERP2 decreases with the ERP2c value, reaches a minimum at 3 to 6 minutes and rises again on account of the increasing influence of the ERP2f. This identifies a repetition between 3 and 6 minutes after a PGM error as most effective. Similar effects are shown by the phoneme omission and letter confusion error categories. However, due to the random distribution of typo and capitalization errors, a distinct point in time optimal for repetition cannot be found.

### *Error expectation*

As the last part of the student model validation, we analyze the estimated error expectation by comparing it to the empirical error expectation values. As a benchmark, we consider the word difficulty measure based on the symbol confusion matrix (SCM) of the first Dybuster study (Gross & Vögeli, 2007), and a difficulty measure for spelling proposed by Spencer (2007). The latter one is

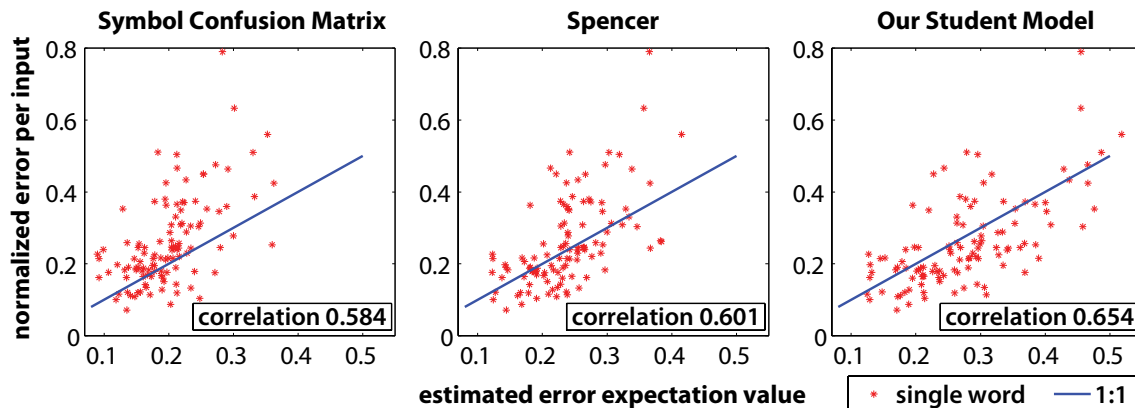


Figure 16: Empirical error expectation values plotted against three word difficulty measures.

based on the phonetic difference (difference between number of phonemes and letters), the phoneme transparency (probability of a phoneme being represented by a specific grapheme), and the frequency of a word, extracted from language corpora. We reconstructed Spencer’s difficulty measure based on the CELEX2 word database (Baayen et al., 1993). The training of the three measures was performed on all inputs committed by the students during the first user study. Subsequently, we investigated their predictive power on words entirely learned by every student, i.e., on words which have been entered twice correctly in a row. In the first user study this yields a set of 111 words, of which the estimated error expectation values and the average error per input is displayed in Figure 16. The blue line represents the actual correct prediction of errors. As can be seen, all measures provide reasonable estimates for simple words (left). However, our model outperforms on very difficult words (right) and allows for an error estimation of more than 0.5. This is compared to a maximum of approximately 0.35 and 0.4 for the SCM and Spencer approach respectively. The two related methods fail to represent specific difficulties of complicated words, which is indicated by the high density of data points above the correct prediction line for words with high empirical error expectation value. This is reflected in the correlation between expected word difficulty and empirical error per input. The presented student model exceeds the SCM and Spencer’s measure by more than 0.05.

In addition, Spencer’s measure relies on an extensive analysis of language corpora and represents general spelling difficulties, constant for all students. This performs best on a comparison with error rates averaged over all students, as presented in Figure 16. However, our student-model-based measure will adapt to the student characteristics and will even allow for an improved error expectation, if it is computed for individual students.

### Evaluation Study

To evaluate the presented student model, we implemented the student representation and corresponding remediation actions (as shown in Section ‘Remediation’) into the Dybuster framework. A second user study was conducted with the enhanced software version to obtain training data, which allows for a comparison with the old version. We compare the learning progress of both studies by means of learning curves. The concept of describing practice effects by simple nonlinear functions in a broad range of tasks was first introduced by Newell and Rosenbloom’s “Mechanisms of Skill

Acquisition and the Law of Practice” (1981). Based on Heathcote et al.’s findings (2000), we rely on an exponential law of practice:

$$E[E(t)] = a'e^{-bt} + c \quad (6)$$

where  $E[E(t)]$  is the error expectation value at time  $t$ . For the comparison of the two studies we are interested in the initial error expectation ( $a = a' + c$  : Error expectation value at time  $t = 0$ ), the learning progress ( $b$ ) and the asymptotic error expectation ( $c$  : Error expectation value for time  $t \rightarrow \infty$ ). Therefore we perform the variable transformation  $a = a' + c$  and obtain the exponential decay function

$$E[E(t)] = (a - c)e^{-bt} + c \quad (7)$$

The error expectation values  $E[E(t)]$  are investigated for individual error categories and collected for the first 30 training days, i.e., we count only the days the children were really working with the training software. To exclude repetition effects from the analysis, we only consider the first prompt of each word. The error expectation values  $E[E(t)]$  for both studies at day  $t$  are computed by dividing the number of committed errors  $Y(t)$  by the number of error possibilities  $N(t)$  of all students of the respective study. A weighted nonlinear least squares method is employed to estimate the parameters for the exponential fit to both datasets. The number of error possibilities ( $N(t)$ ) are used as weights for the estimation. To evaluate the significance of the difference between the two regressions we run a combined estimation. Every parameter  $p$  is replaced by a term  $p(1+d_pg)$ , consisting of an absolute parameter  $p$  for the group  $g=0$  and a relative parameter  $d_p$ , denoting the relative difference of the

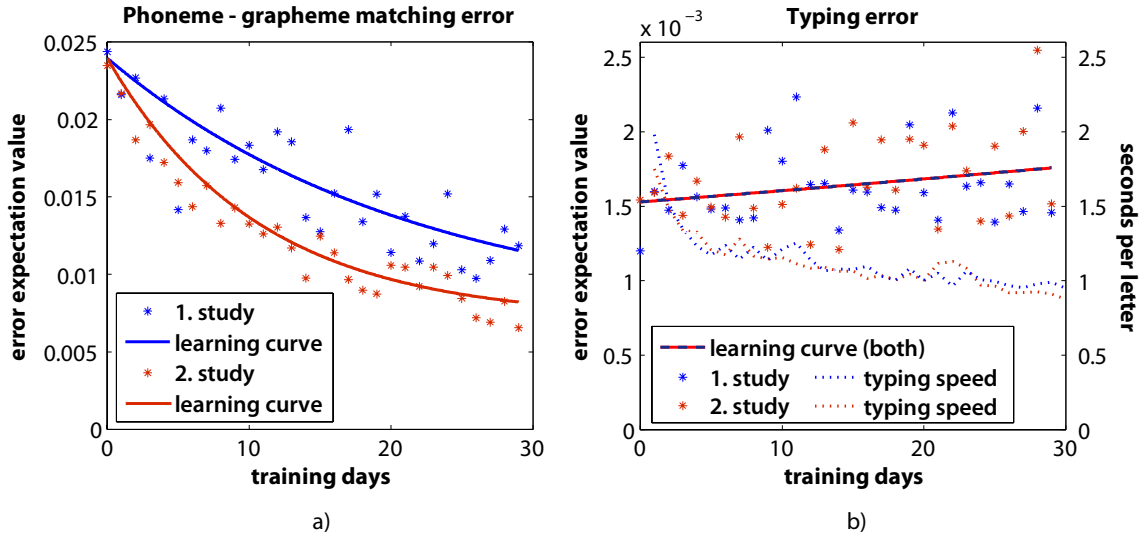


Figure 17: Learning curves comparison between all children of first and second user study for PGM (a) and Typo (b). a) Students of second user study show an increased learning progress (steeper curve). b) No significant difference between typing error expectation values of the two studies was found. The slight increase of typing errors correlates with the decrease of seconds per letter (increasing typing speed).



parameter  $p$  between the first ( $g=0$ ) and the second ( $g=1$ ) group. This results in an estimation of the following form:

$$E[E(t, g)] = (a(1 + d_a g) - c(1 + d_c g))e^{-b(1+d_b g)t} + c(1 + d_c g) \quad (8)$$

where  $g$  equals zero for the first study and one for the second;  $d_a$ ,  $d_b$  and  $d_c$  indicate the percentage difference between the corresponding parameters of the two studies and their t-tests return a measure for the significance of the difference. The introduction of the additional three parameters to the regression model can lead to overfitting. This is especially the case when the data contains no differences in initial error probability, learning progress or asymptotic error probability between the studies. To avoid overfitting and to reduce the model for estimation, we run a backward model selection based on the AIC score (Akaike, 1974). Removed features will be marked with an “RP”, to indicate that no influence of this parameter was found.

To evaluate the influence of the presented student model, we restrict our analysis to the error categories phoneme-grapheme matching and typing error. These categories comprise a major part of all errors (30% and 40% respectively) and represent, on one hand, a category where learning takes place and an improvement should be observable and, on the other hand, a category of randomly occurring errors, which should not show any variation.

Figure 17 shows the error expectation values on each day, and the corresponding learning curves for phoneme-grapheme matching and typing errors. The estimated parameters and corresponding significance ( $p$ ), evaluated with a t-test, are listed in Table 3. As can be seen, the initial error expectation and the asymptotic error expectation do not differ between the two studies ( $d_a$  and  $d_c$  were removed by the AIC criterion). This implies that the students of the first and second study possess the same average spelling skills at the beginning of the study and show equal spelling skill estimations for an infinite training time, which is essential to compare their learning progress over time. The learning progress for PGM however, is significantly higher during the second user study (+104%), which manifests in a steeper learning curve in Figure 17.a).

Table 3: Estimated learning curve parameters and corresponding significance (t-test) for both studies on phoneme-grapheme matching and typing errors. Parameters of first study ( $a$ ,  $b$ ,  $c$ ) are given in absolute values and the difference to the second study ( $d_a$ ,  $d_b$ ,  $d_c$ ) is given in relative values. RP’s indicate removed parameters based on AIC.

	Init. error prob.		Learning progress		Asym. error prob.	
	( $a$ , $d_a$ )	p	( $b$ , $d_b$ )	p	( $c$ , $d_c$ )	p
<b>PGM (all children)</b>						
First study	0.024	2e-16	0.046	4e-07	0.0071	2e-08
Second study	RP		+104%	2e-16	RP	
<b>PGM (dyslexic children)</b>						
First study	0.031	2e-16	0.049	2e-08	0.0092	1e-12
Second study	RP		+168%	3e-07	RP	
<b>Typo</b>						
First study	0.0015	2e-16	-0.0048	0.065	RP	
Second study	RP		RP		RP	

Table 3 shows the parameters of the comparison based on the data of dyslexic children only. The spelling progress on PGM of dyslexic children improved by +168%. This shows that dyslexic children can benefit even more from the detailed knowledge representation on dyslexia-typical difficulties in spelling and the corresponding remediation. On the contrary, no significant variation in typing error expectation between the two studies was found ( $d_a$ ,  $d_b$  and  $d_c$  removed). The expectation values of typing errors in both studies do not decrease at all but trend higher ( $p=0.065$ ) during the spelling training. This effect correlates with the typing speed, specified as seconds per letter in Figure 17.b). These results prove that the student-adapted remediation actions based on the proposed student model yield a doubling of the learning progress in spelling, but do not influence the rate of randomly occurring errors, such as typos.

## CONCLUSION

The presented work addresses student knowledge representations in perturbation models on input data with multiple errors and independent mal-rules. We identified an inference algorithm based on a Poisson regression with a linear link function as most suitable to allow for student model estimations on unclassified student inputs. The appropriateness of the chosen approach has been demonstrated by residual analyses of different link functions and manifests in more reliable estimations. The estimated student characteristics enable an intelligent tutoring system to compute local (error classification) and global (error prediction) information about the student behavior, and to adapt the training accordingly.

The method has been implemented in a student model for spelling based on a detailed set of mal-rules describing spelling errors. We validated the estimated error classification and prediction on the input data of a first user study. Both classification and prediction showed the expected behavior and clearly outperformed related measures. A second user study conducted with the spelling software, enhanced by the presented student representation, showed more than a doubling in the learning progress, induced by the student adapted remediation actions.

The presented approach is not based on the assumption of state transitions on individual mal-rules, but rather estimates the student specific difficulties on spelling error categories. The Poisson regression however assumes constant error rates on the mal-rules and is limited in modeling student progress consequently. For long term application with strongly changing student characteristics, the selection of student inputs for parameter estimation has to be investigated. Also the remediation actions were chosen with respect to the particular setting of the user study, and are not optimal for long-term training schedules. Investigations of improved remediation actions to provide an optimal training progress over more than three months are subject to further research.

## REFERENCES

- Abdi, H. (2007): Bonferroni and Sidak corrections for multiple comparisons. Salkind, N.J. (eds.) *Encyclopedia of Measurement and Statistics* (pp. 103-107). Thousand Oaks, CA.
- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.
- Anderson, J.R., Boyle, A.T., Corbett, A. & Lewis, M. (1990). Cognitive Modeling and Intelligent Tutoring. *Artificial Intelligence*, 42, 7-51.
- Anscombe, F.J., & Tukey, J.W. (1963). The Examination and Analysis of Residuals. *Technometrics*, 5(2), 141-160.
- Atkinson, K. (2006), GNU Aspell 0.60.4', <http://aspell.sourceforge.net/>.

- Augst, G. (1985). Kommentar zum Internationalen Vorschlag der Gross- und Kleinschreibung. *Kommission für Rechtschreibfragen: Die Rechtschreibung des Deutschen und ihre Neuregelung*, 114–142.
- Baayen, R.H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database*. Philadelphia, PA, University of Pennsylvania, Linguistic Data Consortium.
- Bader-Natal, A., & Pollack, J. (2007). Evaluating Problem Difficulty Rankings Using Sparse Student Response Data. *Supplementary Proceedings of the 13th International Conference on Artificial Intelligence in Education*.
- Balota, D.A., Pilotti, M., & Cortese, M.J. (2001). Subjective Frequency Estimates for 2'938 Monosyllabic Words. *Memory & Cognition*, 29(4), 639-647.
- Barr, A., Beard, M., & Atkinson, R.C. (1976). The Computer as a Tutorial Laboratory: The Stanford BIP Project. *International Journal of Man-Machine Studies*, 8, 567-596.
- Baschera, G.-M., & Gross, M. (2009). A Phoneme-based Student Model for Adaptive Spelling Training. In *Proceedings of Artificial Intelligence in Education 2009*, 614-616, IOS Press.
- Bodén, M., & Bodén, M. (2007). Evolving Spelling Exercises to Suit Individual Student Needs. *Applied Soft Computing*, 7(1), 126-135.
- Brown, A.S. (1990). A Review of Recent Research on Spelling. *Educational Psychology Review*, 2(4), 365-397.
- Burton, R.R. (1982). Diagnosing Bugs in a Simple Procedural Skill. Sleeman, D.H., & Brown, J.S. (eds.) *Intelligent Tutoring Systems*, 157-184, London, UK, Academic Press.
- Cameron, A.C., & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.
- Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35(54).
- Corbet, A.T., & Anderson, J.R. (1995). Knowledge Tracing: Modelling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.
- de la Torre, J., & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.
- Dekel, O., Keshet, J., & Singer, Y. (2005). An Online Algorithm for Hierarchical Phoneme Classification. *Lecture Notes in Computer Science*, 3361, 146-158.
- Everitt, B. (1992). *The analysis of contingency tables*. Chapman & Hall/CRC.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co, The Hague, Netherlands.
- Fischer, G.H. (1973). The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, 37, 359-374.
- García, R.M.C., Kloos, C.D., & Gil, M.C. (2008). Game Based Spelling Learning. *Frontiers in Education Conference*, S3B-11-S3B-15.
- Gross, M., & Vögeli, C. (2007). A Multimedia Framework for Effective Language Training. *Computer & Graphics*, 31, 761-777.
- Grund, M., Haug, G., & Naumann, C. L. B.-V., Weinheim. (1995). *Diagnostischer Rechtschreibtest 5. Klasse*. Weinheim: Beltz-Verlag.
- Habib, M. (2000). The Neurological basis of developmental dyslexia: An Overview and Working Hypothesis. *Brain*, 123(12), 2372-2399.
- Hall, T.A. (2000). *Phonologie: Eine Einführung*. Gruyter, Berlin.
- Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The Power Law Repealed: the Case for an Exponential Law of Practice. *Psychonomic Bulletin & Review*, 7(2), 185-207.
- iSpellWell. Voelker Software, <http://www.ispellwell.com/>.
- James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. Harlow, Pearson Education.
- Kast, M., Meyer, M., Vögeli, C., Gross, M., & Jäncke, L. (2007). Computer-based Multisensory Learning in Children with Developmental Dyslexia. *Restorative Neurology and Neuroscience*, 25(3-4), 355-369.
- Kondrak, G. (2003). Phonetic Alignment and Similarity. *Computers and the Humanities*, 37, 273-291.
- Landerl, K., Wimmer, H., & Moser, E. (1997). *Der Salzburger Lese- und Rechtschreibtest (SLRT)*. Bern: Verlag Hans Huber.
- Linder, M., & Grisseemann, H. (2000). *Zürcher Lesetest*. Bern-Göttingen: Hogrefe-Verlag.

- López, J.M. (1987). *Didáctica Actualizada de la Ortografía*. Santillana, D.L. (eds), Aula XXI, Madrid.
- MacArthur, C.A. (1999). Overcoming Barriers to Writing: Computer Support for Basic Writing Skills. *Reading & Writing Quarterly*, 15, 169-192.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.
- McCullagh, P., & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.
- Mislevy, R.J. (1996). Test Theory Reconciled. *Journal of Educational Measurement*, 33(4), 379-416.
- Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of Skill Acquisition and the Law of Practice. Anderson, J. R. (eds), *Cognitive Skills and their Acquisition*. 1-55.
- Ramus, F. (2003). Developmental dyslexia: specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, 13(2), 212-218.
- Reitsma, P. (1989). Orthographic memory and learning to read. Aaron, P.G., & Joshi, R.M. (eds) *Reading and Writing Disorders in Different Orthographic Systems*, 52, 51-74.
- Shute, V.J., & Psotka, J. (1996). Intelligent Tutoring Systems: Past, Present, and Future. D.H. Jonassen (ed.) *Handbook of Research for Educational Communications and Technology*, Scholastic Publications.
- Spencer, K. (2007). Predicting Children's Word-spelling Difficulty for Common English Words from Measures of Orthographic Transparency, Phonemic and Graphemic Length and Word Frequency. *The British Psychological Society*, 98, 305-338.
- SuperSpell 2. 4Mation Educational Resources Ltd. <http://www.4mation.co.uk/cat/superspell.html>.
- Tatsuoka, K.K. (1983). Rule Space: An Approach for Dealing with Misconceptions based on Item Response Theory. *Journal of Educational Measurement*, 20(4), 345-354.
- Tatsuoka, K.K. (1985). A Probabilistic Model for Diagnosing Misconceptions by the Pattern Classification Approach. *Journal of Educational Statistics*, 10(1), 55-73.
- Tewes, U., & Rossmann, P., Schallberger U. (1999). *Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK-III)*. Bern: Verlag HansHuber.
- Ultimate Spelling. eReflect Learning Solutions. <http://www.ultimatespelling.com/>.
- World Health Organization (1993). ICD-10. *The international classification of diseases, Vol. 10: Classification of mental and behavioural disorders*, Geneva.

## APPENDIX A

Table 4: Test results for subjects of first study.

Measures	Dyslexic (n=28)				Non-dyslexic (n=27)				Mann-Whitney
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	p
Age (years)	10.36	0.87	8.83	11.75	10.36	0.81	9.25	11.92	0.922
School grade	3.96	0.84	3.00	5.00	3.88	0.89	3.00	5.00	0.726
IQ	106.04	12.26	87.00	135.00	112.94	10.22	92.00	129.00	0.062
verbal IQ	108.04	12.32	85.00	142.00	115.13	9.71	97.00	132.00	0.034
performance IQ	102.93	12.52	87.00	125.00	107.38	12.48	79.00	128.00	0.164
wordlist reading error	-2.68	2.83	-14.00	0.65	-0.38	1.11	-3.00	1.00	1e-04
wordlist reading time	-4.23	5.43	-27.90	2.20	-0.34	1.09	-1.70	1.70	1e-05
text reading time	-2.93	3.48	-17.50	4.25	-0.44	1.58	-4.50	1.66	2e-04
text reading error	-4.42	5.13	-25.00	0.14	-0.18	0.67	-1.91	0.82	1e-06
writing performance	-1.20	0.67	-2.40	0.10	0.19	1.05	-1.30	1.90	4e-05
	<b>Freq.</b>				<b>Freq.</b>				
Gender (m/f)	10/18				14/13				
Handedness (r/l)	24/4				20/7				

Table 5: Test results for subject of second study.

Measures	Dyslexic (n=40)				Non-dyslexic (n=27)				Mann-Whitney
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	p
Age (years)	10.38	0.94	7.83	12.08	9.92	0.92	7.83	11.92	0.535
School grade	4.68	0.85	3.00	6.00	4.32	0.90	2.00	6.00	0.168
IQ	113.03	10.99	88.00	128.00	117.92	12.31	95.00	140.00	0.236
verbal IQ	114.34	16.07	52.00	136.00	122.80	12.59	98.00	148.00	0.046
performance IQ	105.89	17.81	39.00	144.00	109.64	14.17	88.00	135.00	0.453
wordlist reading error	-1.59	1.27	-2.85	0.99	0.11	1.45	-2.85	2.05	3e-05
wordlist reading time	-1.96	0.99	-2.85	0.30	0.06	1.31	-2.35	2.50	3e-07
text reading time	-1.81	0.99	-2.85	1.17	0.06	0.83	-1.48	1.75	1e-08
text reading error	-1.88	0.99	-2.85	-0.31	-0.17	0.84	-2.85	1.17	1e-07
writing performance	-1.48	0.56	-2.55	0.40	-0.16	0.68	-1.00	1.30	1e-09
	<b>Freq.</b>				<b>Freq.</b>				
Gender (m/f)	27/13				15/12				
Handedness (r/l)	33/7				24/3				

## APPENDIX B

Table 6: Significance of individual mal-rules evaluated by means of a likelihood ratio test. Significance code according to false discovery rate corrected  $\alpha = 5\%$  (\*),  $1\%$  (\*\*) and  $0.1\%$  (\*\*\*)).

Error category	Mal-rule	p	Sig.
<i>Typo</i>	<i>KD(Left/Right)</i>	2e-16	***
	<i>KD(Top/Bottom)</i>	0.012	
	<i>Technical</i>	2e-16	***
<i>Capitalization</i>	<i>ToLowerCase</i>	2e-16	***
	<i>ToUpperCase</i>	2e-16	***
<i>Letter confusion</i>	<i>VD(LowerCase)</i>	0.013	
	<i>VD(UpperCase)</i>	0.018	
	<i>VD(L/UCASE)</i>	8e-08	***
	<i>AD(Vowel)</i>	0.018	
	<i>AD(Similar)</i>	5e-12	***
	<i>AD(Consonant)</i>	2e-04	**
	<i>AD(Obstruent)</i>	4e-06	***
	<i>AD(Fricative)</i>	0.005	
	<i>AD(FrVoiced)</i>	0.002	
	<i>AD(FrUnvoiced)</i>	0.002	
	<i>AD(Affricative)</i>	1e-06	***
	<i>AD(Plosive)</i>	4e-08	***
	<i>AD(PIVoiced)</i>	4e-04	*
	<i>AD(PIUnvoiced)</i>	3e-04	*
	<i>AD(Sonor)</i>	5e-07	***
	<i>AD(Nasal)</i>	7e-08	***
<i>AD(Fluid)</i>	0.024		
<i>Phoneme omission</i>	<i>PhonemeOmission</i>	2e-16	***
<i>Phoneme-grapheme matching</i>	<i>PM(VowMain)</i>	3e-06	***
	<i>PM(VowSpec)</i>	2e-16	***
	<i>PM(ConsMain)</i>	2e-16	***
	<i>PM(ConsSpec)</i>	2e-16	***
	<i>El(Omission)</i>	2e-16	***
	<i>El(Addition)</i>	4e-14	***
	<i>Sh(Omission)</i>	2e-16	***
	<i>Sh(Addition)</i>	2e-16	***