A Data-Driven Approach to Feedback Optimization in Language Learning Apps

Timo Landzettel

s4751914

Master Thesis

Department of Psychology, University of Groningen

Supervisor and first evaluator: Prof. Dr. Hedderik van Rijn

Second evaluator: Dr. Simone Sprenger

August 22nd, 2025

Abstract

Effective feedback helps learners understand what went wrong and how to improve. In language-learning apps, feedback is often created by comparing a learner's answer with a target answer, but this works best when we know the typical errors learners make. This enables creating automated feedback messages tailored to explain specific mistakes that were made. However, typical error categories are unknown. Here we study English vocabulary training by German Gymnasium students mostly aged 10–13, analyzing more than 3.2 million responses to identify recurring mistake patterns that could support explanatory feedback. We find twelve categories, including missing elements (for example, "to " in infinitives), spacing, capitalization, punctuation, double-letter errors, and phonological similarity, among others. After this coverage, 22% of wrong responses and 15% of near-correct responses remain unexplained. These results provide a practical foundation for automated feedback that names the error and suggests a fix, which can reduce uncertainty for learners. More broadly, they inform the design of language-learning applications for early secondary students and support more effective vocabulary acquisition in second-language learning.

Keywords: language learning app, formative feedback optimization, automated feedback, mistake analysis, second language acquisition (SLA)

A Data-Driven Approach to Feedback Optimization in Language Learning Apps

Traditional methods for studying English as a Foreign Language (EFL) often involve reading textbooks and memorizing lists of vocabulary, typically accompanied by translations into the learner's native language. While this approach has been widely used, it can be monotonous and lacks personalization. In recent years, the emergence of language learning apps has transformed this landscape. These digital tools automate and individualize the learning process, offering students access to automated and effective study methods such as spaced repetition and active recall (Abbas et al., 2022; Xu et al., 2024). Such apps are increasingly being integrated both inside and outside the classroom, offering various benefits. Fan et al. (2023) indicate that this includes intuitive and user-friendly designs, a broad selection of study materials, improved language proficiency, and enhanced student motivation. Language learning apps have also shown potential to support traditional teaching methods by complementing classroom instruction. Furthermore, the portability of mobile phones enables learning to occur anytime and anywhere, significantly enhancing accessibility.

Formative Feedback in Language Learning Apps

Another key advantage of language learning apps is their ability to provide immediate feedback. By offering real-time responses, language learning apps help facilitate the learning process. Shute (2008) claimed that feedback in this context can also be understood as a type of formative feedback, as it involves providing the learner with information aimed at influencing their thinking or behavior to enhance learning. Febriani and Abdullah (2018) argued that formative feedback can enhance the quality of learning, particularly in blended learning environments. Klimova (2019) also noted that immediate feedback increases students' confidence and classroom participation, while motivating them to continue using mobile apps for studying. Moreover, students report appreciating the immediate response provided after each input, as noted by Klimova and Polakova (2020). In their research, corrective feedback was defined as binary feedback – indicating whether an answer was right

or wrong. Research therefore agreed feedback to be beneficial for students, however, some researchers pointed out limitations in this approach. Fan et al. (2023) observed that while feedback is present in many language learning apps, it often lacks depth, such as explanations or corrective guidance. Heil et al. (2016) also reported feedback in language learning apps to lack explanation. This aligns with Andersen's (2013) findings, which highlighted that most English-learning apps offer minimal feedback beyond simple correct/incorrect notifications. In conclusion, while feedback is a beneficial and motivating feature of language learning apps, its current implementation may be overly simplistic. This conclusion leads to the following question: What constitutes good formative feedback?

Effective Formative Feedback

Effective Feedback is Explanatory, Concise, and Targeted

Moreno (2004) assessed whether simple corrective feedback or extended explanatory feedback is advantageous in the learning process. She concluded that explanatory feedback led to higher test scores and students also rated explanatory feedback to be more helpful. Those results were seen because explanatory feedback reduces students' mental load and therefore facilitates the learning process. Furthermore, effective formative feedback is specific, explanatory, and concise (Shute, 2008). According to Hattie and Timperley (2007) and Shute (2008), feedback that clearly explains what the problem is, how it occurred, and why it matters significantly enhances the learning process. Mayer and Moreno (2002) further emphasized that formative feedback should remain simple and focused to prevent cognitive overload or distraction from the core learning objective. Unnecessary information should thus be avoided. Furthermore, unclear and non-specific feedback can lead to frustrated learners (Moreno, 2004). This suggests that basic correct/incorrect responses may be insufficient. Instead, feedback should aim to guide learners by providing short, meaningful explanations. For instance, rather than merely indicating an incorrect answer in a language learning app, a more helpful message might read: *You missed important punctuation! The correct response is:*

'Come here!' Such feedback not only signals that a mistake was made but also clarifies its nature and directs the learner's attention to the specific area needing improvement. By making feedback in vocabulary training more explanatory and targeted – while keeping it concise – the learning process may become more effective and meaningful for students.

Performance Uncertainty Should Be Decreased

Formative feedback is intended to reduce a learner's uncertainty regarding their performance (Shute, 2008). Ashford et al. (2003) suggested that when learners are unsure about their performance, they are more likely to avoid feedback altogether. To address this, it may be beneficial to expand beyond simple binary feedback and provide more detailed information about the learner's level of performance. Rather than categorizing responses as merely correct or incorrect, incorporating an additional category such as 'almost correct' could help reduce performance uncertainty. For example, a more informative feedback message might state: *Your response was almost correct, but you missed important punctuation. The correct response is: 'Come here!'.* By providing more specific feedback on a learner's performance, uncertainty can be reduced, potentially enhancing the learning process.

Effective Feedback is Unbiased and Non-Evaluative

Formative feedback should be unbiased and non-evaluative (Shute, 2008). Kluger and DeNisi (1996) argue that computer-based feedback offers an advantage over person-delivered feedback by eliminating perceived biases, making it more likely to be trusted by learners. Consequently, an automated feedback mechanism that generates individualized feedback based on an analysis of the learner's response may be preferable to in-person feedback. In addition, effective formative feedback does not evaluate a learner's performance by comparing it to a standard, a norm, or the performance of others (Shute, 2008). Feedback that draws attention to the learner's self can pose a threat to self-esteem, which may negatively affect learning (Kluger & DeNisi, 1996). Instead, feedback should remain task-focused and avoid self-referential language. For example, using neutral phrasing such as: *The response*

was almost correct, but important punctuation was missing. The correct response is: 'Come here!' keeps the focus on the task rather than the individual. Automating the feedback process while maintaining a neutral, non-personal tone may contribute to reducing performance anxiety and fostering a more effective learning environment.

Including the Correct Solution in the Feedback Message

Providing the correct solution as part of the feedback message has been shown to positively influence learner performance compared to feedback that does not include the correct answer (Kluger & DeNisi, 1996). In the context of language learning apps, this implies that when a learner provides an almost correct or incorrect response, the feedback should explicitly present the complete and correct answer. For example, a feedback message such as: *The response was almost correct, but important punctuation was missing. The correct response is: 'Come here!'* is more effective than a message that omits the correct solution, such as: *The response was almost correct, but important punctuation was missing.* Therefore, feedback should always include the correct answer to enhance clarity and support the learning process.

Feedback Timing

Research has examined the role of feedback timing by distinguishing between immediate and delayed feedback to determine whether timing influences learning outcomes. In the context of vocabulary training, immediate feedback appears to have a similar effect to delayed feedback when controlling for the lag to test – the interval between the last encounter with an item and the subsequent post-test (Metcalfe et al., 2009; Nakata, 2015). In these studies, feedback was provided in a simple format, where the correct response was shown after an incorrect answer, and learners were required to produce the correct input before proceeding. Shute (2008) emphasized that the effectiveness of feedback timing depends on the perceived difficulty of the task. Learners who perceive a task as more difficult tend to benefit from immediate feedback, as it offers support and helps reduce frustration (Knoblauch

& Brannon, 1981). Conversely, for tasks perceived as relatively easy, delayed feedback may be more effective to avoid learner annoyance (Clariana, 1990). In summary, while feedback timing generally plays a role in learning, its effectiveness depends on the learner's perception of task difficulty. However, studies by Metcalfe et al. (2009) and Nakata (2015) suggest that in vocabulary training, feedback timing does not significantly impact learning outcomes when individual differences in difficulty perception are not considered. Therefore, feedback timing becomes relevant primarily when individual differences, such as task difficulty perception, are considered.

The Current Study

The aim of the present study is to propose ways to enhance the quality of feedback messages within language learning applications. Ensuring feedback is unbiased can be achieved through computer-generated feedback, which reduces the influence of human biases (Kluger & DeNisi, 1996). Similarly, feedback can be made non-evaluative by carefully phrasing messages to avoid drawing attention to the learner's self, thereby reducing potential threats to self-esteem (Kluger & DeNisi, 1996). However, making feedback explanatory and simultaneously reducing performance uncertainty presents a more complex challenge. One potential solution is to develop an algorithm capable of extending feedback messages by explicitly identifying the learner's specific mistake. Such an algorithm requires predefined knowledge of common mistake patterns to accurately analyze learner responses. Only when these mistake patterns are clearly defined can an automated system detect errors and provide meaningful, explanatory feedback. The envisioned system compares the learner's input to the correct, expected response, identifies deviations, and categorizes these according to recognized mistake patterns. Based on this analysis, the algorithm can generate targeted, explanatory feedback that informs the learner of the exact nature of their mistake, thereby improving the clarity and effectiveness of the feedback.

Using Data to Detect Mistake Patterns

To lay the foundation for such a system, this study adopts a data-driven approach to identify the most frequent mistake patterns made by German school students when learning English vocabulary using a German language learning app (Phase6). The overarching objective is to reduce the proportion of unexplained errors by systematically categorizing learner mistakes. The results of this study are intended to inform the development of an algorithm capable of detecting these specific error patterns and providing individualized, explanatory feedback. The identification of mistake patterns in this study is based exclusively on string comparisons between the correct answer and the learner's response. Several analyzes rely on the Damerau-Levenshtein distance, a metric that quantifies the difference between two strings by calculating the minimum number of operations required to transform one string into the other (Levenshtein, 1966). The Damerau-Levenshtein distance accounts for character insertions, deletions, substitutions, and transpositions, making it a suitable tool for detecting common typographical and structural errors in vocabulary learning. Based on the rules presented in Table 1, the Damerau-Levenshtein distance was used to assign student responses to one of the following categories: correct, almost correct, or wrong. In summary, this research seeks to address the following research question: What common mistake patterns can be identified in English vocabulary training for German students that can serve as the basis for generating explanatory feedback, thereby reducing the proportion of unexplained errors?

Table 1

Damerau-Levenshtein Distance Rules for Responses to Qualify as Almost Correct

Correct Answer Length (in Characters)	Allowed Damerau-Levenshtein distance Between Response and Correct Answer
3 – 5	1

6 – 10	<= 2
11 – 15	<= 3
16 – 20	<= 4
21 – 25	<= 5
> 25	<= 6

Note. Rules were created through personal assessment. They were constructed in a way that no more than a third of the characters in the correct answer are allowed to deviate for a response to be considered almost correct.

Methods

Materials

Data

Two datasets were used for the retrospective data analysis, one consisting of data from publisher Klett and one from publisher Cornelsen. This results in a total of 711,282 unique answer-response mappings, based on 3,214,723 total responses by learners to English vocabulary questions in the Phase6 language learning app. No specific demographics are available. However, the dataset comprises responses associated with textbooks used in German secondary education (Gymnasium) in two grade levels: 5th and 7th grade. As a result, it can be expected that most participants were between 10 and 13 years old and native German speakers. The data was collected over a period of more than one year.

Software

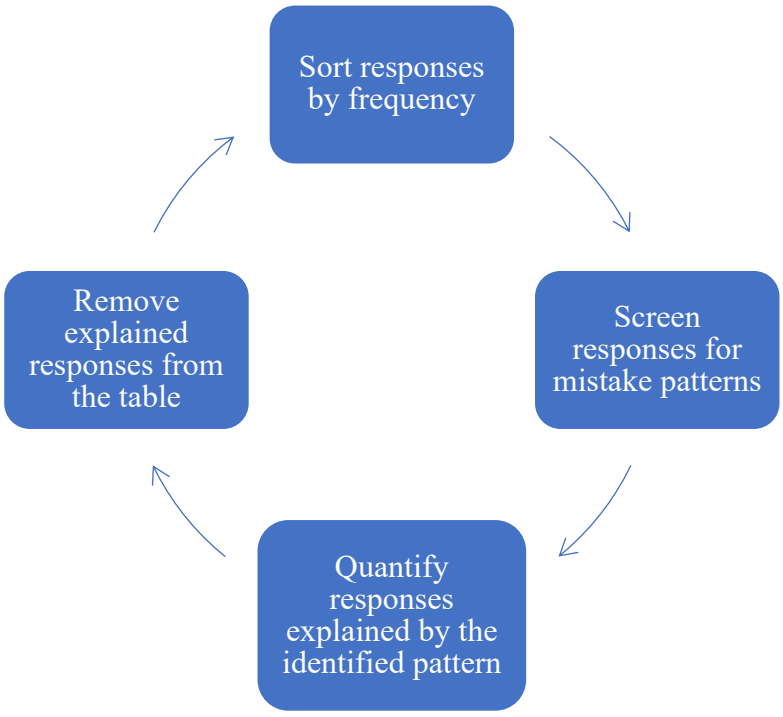
The analysis was done in Python 3.12 (Python, 2024). The packages pandas (McKinney, 2010), numpy (Harris et al., 2020), matplotlib (Hunter, 2007), weighted-levenshtein (Su, 2023), and gruut (Hansen et al., 2024) were used. All analysis code is available on the Rijksuniversiteit Groningen Psychology servers.

Procedure

After preprocessing the dataset and removing all correct responses, the remaining unexplained mistakes were systematically analyzed to identify recurring mistake patterns. To begin, these unexplained mistakes were sorted by frequency, starting with the most common. The most frequently occurring mistakes were then examined to identify potential mistake patterns. For example, one pattern observed in the dataset involved students consistently making capitalization errors. Once a specific mistake pattern was identified, the dataset was reviewed to determine how many of the currently unexplained mistakes could be attributed to this pattern. Any mistakes that could be explained by the identified pattern were subsequently removed from the table, as they were no longer considered unexplained. This process was then repeated with the updated table: the remaining unexplained mistakes were again sorted by frequency, screened for new mistake patterns, and those explained by each newly identified pattern were removed. The cycle continued iteratively until no further recurring mistake patterns could be identified. The entire procedure is summarized visually in Figure 1.

Figure 1

Mistake Pattern Analysis Process Visualization



Preprocessing of the Data

The two separate datasets provided by Phase6 and based on the learning materials of the publishers Cornelsen and Klett were first merged into a single table containing all available data. Since the analysis did not focus on differences between the publishers, responses from both sources were treated as one unified dataset. The analysis specifically targeted the German-to-English (G2E) translation condition, in which students were shown a German word or phrase and asked to translate it into English. Accordingly, the dataset was filtered to retain only rows corresponding to the G2E condition remaining 3,214,723 separate responses. In the column containing the correct answers, additional information, such as publisher identifiers and encrypted metadata, was embedded alongside the correct answer string. This extraneous information was removed, leaving only the correct answer itself in that column. Furthermore, some student responses were exceptionally long. For example, the longest recorded response exceeded 2.7 million characters, while the longest valid correct answer was only 40 characters. To ensure the quality and validity of the analysis, responses exceeding 80 characters were excluded from the dataset, based on the conservative assumption that such entries were not valid answers. This left 3,209,269 separate responses to be analyzed.

Filtering Out Correct Responses

Before analyzing the mistakes, it was necessary to first identify and remove all correct student responses from the dataset. This was achieved by comparing each student's response to the corresponding correct answer and checking for exact matches. Rows where the student response matched the correct answer were counted and subsequently removed from the table. When interpreting the table, it is important to consider how responses are recorded: each row represents a unique response attempt. If multiple students provided the same answer to the same item, this is reflected by a count in a designated column, rather than multiple identical

rows. By removing all correct responses, the table was reduced to contain only unexplained mistakes, which formed the basis for the subsequent analysis.

Mistake Categorization

Responses That Either Miss a 'to ' or a '(to) '

The first mistake pattern that was identified was a missing 'to ' or a missing '(to) ' when students were asked to translate a verb to English. Many students responded with the correct translation of the verb but missed the preceding 'to ' or '(to) '. The reason behind round brackets being present in one of the variants of the mistake and not present for the other one is that correct answer formulation was neither normed nor automated. This means that sometimes the correct answer to a verb question starts with 'to ' and sometimes with '(to) '. To quantify this mistake category, the table was filtered to count all responses by the students that would be an exact match to the correct answer if either a 'to ' or a '(to) ' was added to the response.

Double-Letter Mistakes

The next identified mistake pattern was related to double-letter mistakes. In many cases, student responses were nearly correct, but either a double letter was missing from their response or incorrectly added. For example, a student wrote *adress* instead of the correct answer *address*, or *hopefull* instead of *hopeful*. To quantify this type of mistake, the occurrence of double letters in both the correct answer and the student's response was compared. If the double-letter patterns matched, no double-letter mistake was recorded. If they differed, the response was categorized as containing a double-letter mistake. It is important to note that this analysis was limited to responses classified as "almost correct" according to the Damerau-Levenshtein distance rules described earlier. In such near-correct responses, mismatched double letters are typically the primary reason the response is incorrect. In contrast, for completely incorrect responses, double-letter mismatches may also appear but are rarely the sole cause of the error.

Spacing Mistakes

Another frequently observed mistake pattern involved incorrect spacing. Students often either omitted a necessary space or added an unnecessary space within their response. For example, a student wrote *livingroom* instead of the correct answer *living room*, or *school bag* instead of *schoolbag*. To quantify this type of mistake, two new columns were created in the table. One contained the correct answer with all spaces removed, and the other contained the student's response with spaces removed. These two columns were then compared across all rows. If the modified strings matched exactly, the response was categorized as a spacing mistake.

Capitalization Mistakes

Another common mistake pattern identified was related to capitalization. For instance, a student might have entered *Address* while the correct answer was *address*. To quantify this type of error, a new column was created in the dataset that converted all student responses to lowercase letters. These lowercase responses were then compared to the already existing column in the table that contains the correct answers in lowercase. If the two matched, the response was categorized as a capitalization mistake.

Hyphen Mistakes

Another identified mistake pattern involved incorrect use of hyphens. Students either omitted a required hyphen or added a hyphen where it was not necessary. For example, the correct answer might have been *pencil-case*, while the student response was *pencil case*. Since hyphens are sometimes replaced with spaces by students, this had to be considered during the analysis. To account for this, the columns previously created to detect spacing mistakes were reused, as those columns already controlled for additional spaces introduced by students. In the next step, all hyphens were removed from the strings in these columns. The modified student responses and correct answers were then compared across all rows. If they

matched after the removal of spaces and hyphens, the response was categorized as a hyphenation mistake.

Punctuation Mistakes

The next identified mistake pattern involved incorrect punctuation. To detect this, two new columns were created: one containing the correct answer with all punctuation removed, and the other containing the student response with punctuation removed. The following symbols were considered punctuation and deleted from both columns: ., ..., ?, ,, /, \, !, ;, :, (,), ", ', `, [,], {, }, @, #, &, *, and ^. After removing these symbols, the two columns were compared across all rows. If the modified strings matched, the response was categorized as a punctuation mistake. The results were then used to update the unexplained mistakes table in the same manner as with previous mistake patterns.

Combination of Spacing, Capitalization, Hyphen, and Punctuation Mistakes

In many cases, the observed mistakes could not be attributed to isolated issues such as incorrect spacing, capitalization, hyphenation, or punctuation alone, but rather to a combination of these factors. To account for such cases, two new columns were created that simultaneously controlled for all these mistake types. These columns were then compared across all rows. If the modified student response and correct answer matched in these columns, the response was categorized as a combination mistake and subsequently removed from the set of unexplained mistakes.

Wrong Orthographical Response but Better Phonological Response

It became apparent that some responses, although orthographically incorrect, might have sounded correct or nearly correct when spoken. For instance, a response such as *winsday* instead of *Wednesday* would be categorized as incorrect under the Damerau-Levenshtein distance orthographic assessment but is closer to the correct answer in its phonological form. The aim of this step was to detect and quantify unexplained responses that either sounded correct or sounded better than their written form suggested.

From Orthographical Representation to IPA

To analyze phonological similarity, the orthographic representations of both the correct answers and student responses were first converted into phonological form using *gruut* (Hansen et al., 2024). *gruut* is a tool that processes written input and returns an output in American English IPA (International Phonetic Alphabet) notation, based on a pre-trained pronunciation dictionary. Crucially, it can generate approximate phonetic transcriptions for non-standard or invented words, such as *winsday*, using a grapheme-to-phoneme model. In some cases, *gruut*'s IPA output includes the symbols ˈ and ˌ, which indicate stress on specific syllables. However, it was observed that these stress markers were inconsistently applied, particularly when processing made-up words. For example, *uncle* was transcribed as ˈʌŋkəl, while *unkle* was transcribed as ʌŋkəl, with the stress marker omitted. To ensure consistency in string comparisons between student responses and correct answers, both stress markers were removed from all *gruut* outputs.

From IPA to DISC

The next step involved translating the IPA representations into DISC notation. DISC offers a significant advantage for string comparisons, as it uses a one-to-one character mapping for each sound, unlike IPA, which may use multiple characters for diphthongs or long vowels (Baayen et al., 1995). A custom function was developed to convert *gruut*'s American English IPA output into DISC characters, following the IPA-to-DISC translation rules provided in the English Linguistic Guide. However, certain adjustments were necessary to align the translation with the American English pronunciations generated by *gruut*, as the guide includes examples based on a different accent. For example, *gruut* transcribes *another* as ənʌðɚ, distinguishing between the first vowel and the final sound. In contrast, the guide translates both to the same DISC symbol (@), disregarding accent-specific pronunciation differences. For the sake of consistency and alignment with American English output, the examples in the guide were treated with caution. Consequently, the IPA symbol ə was

translated to the DISC character @, while ɚ was translated to 3, preserving the American English rhotic pronunciation. The custom translation function also accounted for rhotic pronunciation throughout. In several cases, the guide's examples reflected British English pronunciations with omitted *r* sounds, such as in the word *born*, which had to be disregarded to maintain consistency with the American English output from *gruut*.

Considering Phonological Similarity Between Letters

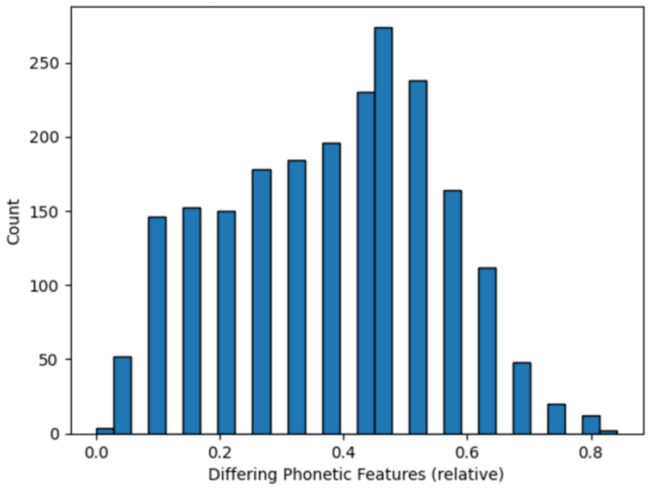
To refine phonological comparisons, it was necessary to account for the fact that some DISC letters represent sounds that are more like each other than others. For instance, the middle sound in *bean* (DISC letter *i*) is phonetically like the middle sound in *bin* (DISC letter *I*), whereas the sounds represented by *p* and *o* are markedly different. In the next stage, a weighted Damerau-Levenshtein distance function was applied to the DISC representations of both student responses and correct answers to assess whether a response sounded correct, almost correct, or incorrect. To do so, substitution costs within the Damerau-Levenshtein distance function were modified based on phonological similarity: substituting two DISC letters that sound similar incurred a lower cost, while substitutions involving dissimilar sounds incurred a higher cost. Phonological similarity was assessed based on shared phonological features. These features offer a systematic way to describe and compare speech sounds (Chomsky & Halle, 1968). The Phonological Corpus Tool (PCT) incorporates research by Mielke (2012), and Chomsky and Halle (1968) to provide a table detailing how DISC characters differ across various phonological features (Hall et al., 2022). Before applying this table, adjustments were made. Rows corresponding to DISC letters not used in the custom translation function were removed. Additionally, redundant columns representing irrelevant features were excluded. Columns containing only - values indicated features absent across all DISC letters, making them unsuitable for comparison, and were also removed.

Lowering Substitution Costs for Similar-Sounding DISC Letters

The refined table was then used to calculate custom substitution costs between all possible pairs of DISC letters. For example, if two DISC letters differed in just 1 out of 19 phonological features, such as *t* and *d*, their substitution cost was set to 1/19 = 0.05. By contrast, the two most dissimilar DISC letters, *$* (representing the *o* sound in *born*) and *J* (representing the *ch* sound in *cheap*), differed on 16 out of 19 features, resulting in a substitution cost of 16/19 = 0.84. To normalize the maximum substitution cost to 1, the default value in Damerau-Levenshtein distance, the calculated costs were all divided by 0.84. This adjustment ensured that substitution costs for highly dissimilar sounds, like *$* and *J*, equalled 1, while substitutions for similar-sounding letters were assigned proportionally lower costs. An overview of the relative phonological differences between DISC letters is provided in Figure 2.

Figure 2

Relative Phonological Differences Between DISC Letters

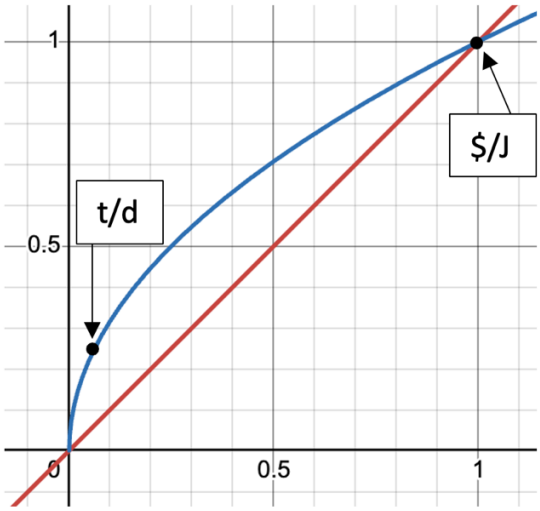


Currently, the relationship between phonological similarity and substitution cost is linear (see red line in Figure 3). However, this is suboptimal, as ideally, substitution costs should decrease more sharply as phonological similarity increases. This effect was achieved by applying a square root transformation to the substitution cost values, introducing a negative curvature to the graph (see blue line in Figure 3). The final substitution cost between two DISC letters was calculated as follows:

$$Substitution\ Cost = \sqrt{\frac{Relative\ Phonological\ Feature\ Mismatch}{0.84}}$$

Figure 3

Substitution costs per amount of mismatching phonological features



Note. The x-axis indicates the relative amount of mismatching phonological features divided by 0.84. The y-axis indicates the final substitution cost for a given pair of DISC letters.

Determining Phonological Similarity Between Words

The described procedure was implemented within a weighted Damerau-Levenshtein distance function to assess how similar two words sounded. Student responses and correct answers were compared using their DISC representations, and the modified substitution costs were applied accordingly. This process made it possible to quantify how many student responses, although orthographically incorrect, sounded correct or nearly correct. Specifically, the number of responses categorized as wrong in writing but almost correct in their DISC (phonological) representation, and wrong/almost correct in writing but correct in their DISC (phonological) representation was determined. This allowed the identification of responses that sounded better than they appeared in written form.

Responses with a Different Meaning than the Correct Answer

It was frequently observed that students provided English words as responses that were incorrect because their meaning differed from the correct answer. To identify such mistakes, it was assumed that if a student's response was not the correct answer but could be found in an English dictionary, the response was incorrect due to semantic difference. Since the focus was on meaning rather than form, capitalization mistakes in student responses were disregarded during this analysis. For example, if a student responded with *Tiger* to a prompt where the correct answer was *mouse*, the analysis would only detect this semantic error after correcting for capitalization. This step is particularly relevant, as German nouns are always capitalized, making this a common mistake among German students writing in English. To perform this analysis, the Spell Checker Oriented Word Lists (SCOWL) were used, as they provide comprehensive English dictionaries suitable for this purpose (Atkinson, 2020).

Responses in the Dictionary that Were Also Part of the Correct Answer

The analysis based on dictionary entries could not be applied indiscriminately. Upon reviewing the phrasing of correct answers within the dataset, it became apparent that many correct answers were poorly formulated. Overly complex or inconsistent phrasing occasionally led to technically correct student responses being marked as incorrect during automated string comparison. For instance, one question required the translation of the German word *Frau*, for which the correct English answer is *woman*. However, in the learning app, the correct answer was phrased as *woman, women (pl)*. Consequently, if a student correctly responded with *woman*, the system marked the answer as incorrect because it did not exactly match the stored solution. To address this, the table was filtered for student responses that were valid English dictionary entries and appeared as part of the correct answer string. Several additional precautions were implemented to avoid false positives in this process. For example, student responses to verb translation questions sometimes consisted of only *to*,

which would pass the dictionary and answer-part filter, even though such responses lacked semantic correctness. These single-word responses, such as *to*, were therefore excluded from the analysis. Similarly, isolated characters like *a* occasionally appeared in the data, which could pass the filter if they coincidentally matched part of a longer correct answer, such as *exactly*. To avoid this, only responses with an occurrence count of at least 50 were considered, as lower-frequency responses were often irregular or accidental. In summary, student responses found in the English dictionary were classified as semantically incorrect unless they also appeared as part of the correct answer string. In cases where such responses were both dictionary-valid and part of the correct answer, they were assigned to a separate mistake category indicating a somewhat correct response by students.

German Responses When English Responses Were Expected

It was further observed that some student responses were German words, even though English translations were required. To quantify these instances, a German dictionary was used to check whether a student response appeared within it (Wendt, 2017). Any response found in the German dictionary was assumed to be a German word and was therefore classified as incorrect.

Typographical Mistakes

Upon reviewing random samples of unexplained mistakes, various types of typographical errors were identified. One specific type occurred when students missed the intended key on the keyboard and instead pressed an adjacent one. For example, the correct answer *dog* was incorrectly typed as *eog*. Such mistakes are referred to here as substitution typos. To account for substitution typos, a custom weighted Damerau-Levenshtein distance function was developed. This version of the function does not penalize letter substitutions if the letters are adjacent to each other on a German QWERTZ keyboard. Since typographical errors are typically unintentional and random, they should not form systematic patterns. In theory, substitutions involving adjacent keys should occur with similar likelihood across the

keyboard. However, when substitution typos were sorted by frequency within the dataset, certain substitutions occurred far more often than others, specifically, *f* substituted for *v*, *v* for *f*, *i* for *o*, and *o* for *i*. This indicates that these substitutions are not purely random typos but rather systematic confusions. Indeed, this pattern can be explained linguistically: in German, the letters *f* and *v* are often pronounced identically, leading to confusion when writing in English. Similarly, students frequently confused the English prepositions *in* and *on*, explaining the *i/o* substitutions. Given this, the custom Damerau-Levenshtein distance function was designed to exclude *f/v* and *i/o* substitutions from being treated as random typographical errors. These specific substitutions continued to be penalized, as they represent systematic confusions rather than accidental typos. Finally, the effect of discounting legitimate substitution typos was evaluated by analyzing how many responses changed from orthographically *wrong* to *almost correct*, and from *almost correct* to *correct*, when adjacent-key typos were no longer penalized.

Additional Analyses

Quantifying Mistake Categories Across the Entire Dataset

In the previous analyses, all mistake categories, except for the first, were quantified only within the subset of unexplained mistakes remaining at each stage of the analysis. To gain a more comprehensive understanding of how each mistake category impacted the overall dataset, all mistake categories were additionally quantified across the entire dataset.

Students' Self-Assessment of Their Responses

The learning platform *Phase6* automatically marked responses as correct or incorrect based on a string comparison with the correct answer. In addition, *Phase6* offers a *gelten lassen* button (translated as "let my response count anyway"), allowing students to override the system's judgment and accept their own response, even if it deviated from the expected answer. It was examined how frequently students used this feature to accept their own incorrect responses, depending on the specific mistake category their response belonged to.

Results

Preprocessing

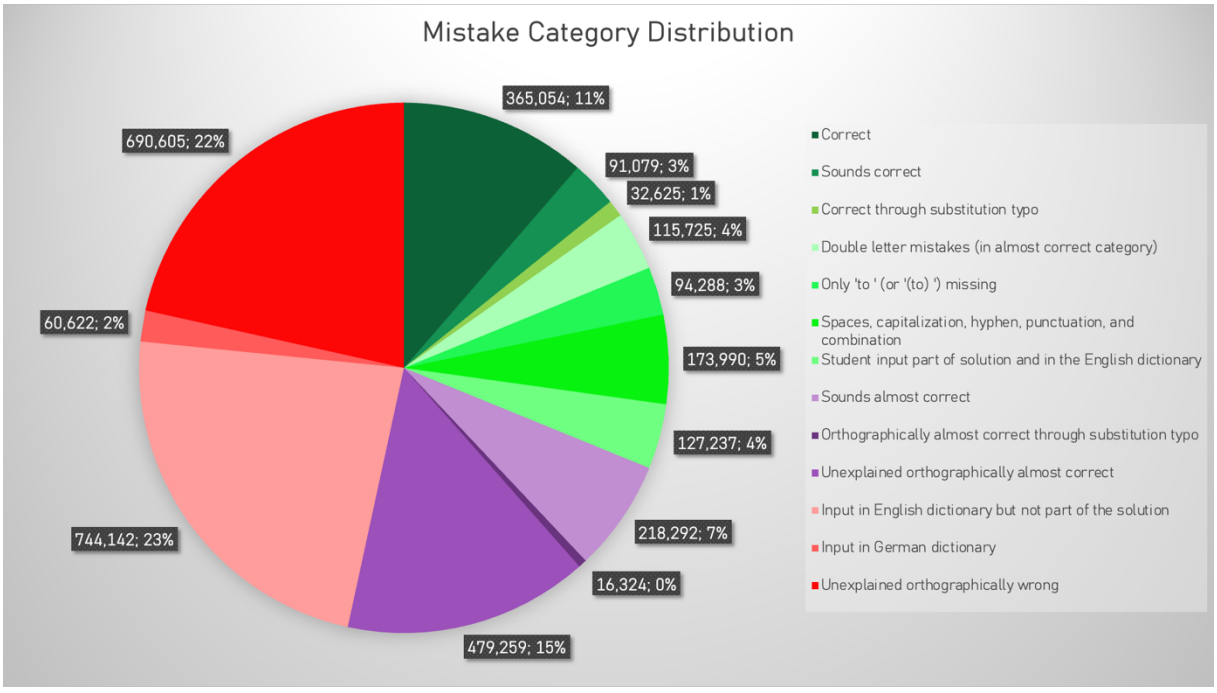
Responses exceeding 80 characters were removed from the table. This implied the deletion of 5,447 rows from the table that represented 5,454 separate responses.

Mistake Categorization

Figure 4 shows the outcome of the mistake categorization analysis. The responses shown in each category sum up to the total amount of responses in the data (3,209,269); each response cannot be part of more than one response category as they were removed from the table once assigned to one category.

Figure 4

Response Category Distribution



Note. The order of operations of quantifying mistake categories is consistent with the order presented in the procedure.

Quantifying Mistake Categories Across the Entire Dataset

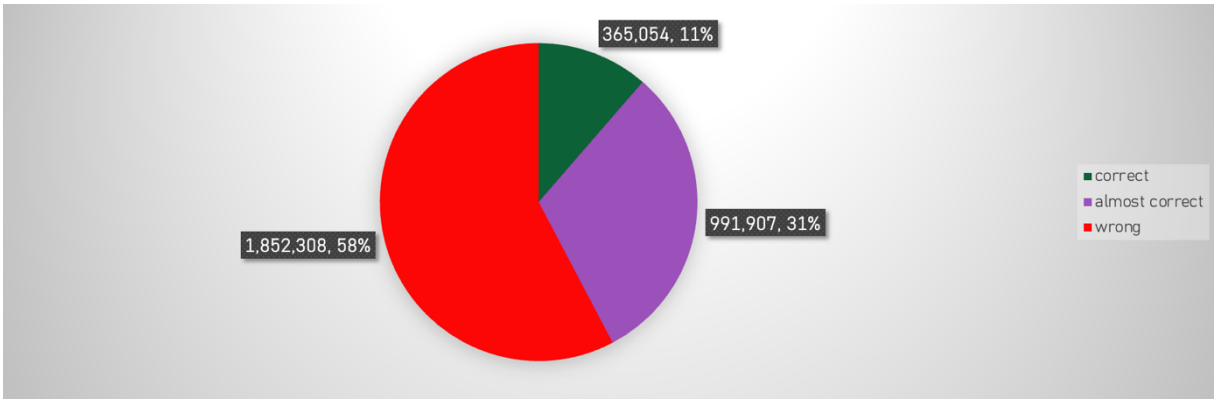
The results below show how many mistakes in the total dataset can be explained through each mistake category.

Orthographical Similarity to the Solution

Figure 5 shows the percentage of total responses that are exact matches to the correct response, almost correct, or wrong. The assignment is based on the Damerau-Levenshtein distance and its assessment rules shown in Table 1.

Figure 5

Orthographical Similarity Between Student Responses and the Correct Answer



Responses that either miss a 'to ' or a '(to) '

Out of the total responses, 83,359 (or 2.6%) only missed a 'to ' to be correct responses. 10,929 responses (or 0.34%) missed a '(to) ' instead.

Double-Letter Mistakes

Out of the 991,907 almost correct responses, 115,725 responses became correct when controlling for double-letter mistakes in both the solution and the student response. Thus, 3.61% of all responses and 11.67% of the subset of almost correct responses were double-letter mistakes.

Spacing, Capitalization, Hyphen, Punctuation, and Combination Mistakes

Table 2 shows the absolute and relative amount of spacing, capitalization, hyphen, punctuation, and combination mistakes in the total data.

Table 2

Spacing, Capitalization, Hyphen, Punctuation, and Combination Mistakes in the Total Data Set

Mistake Category	Absolute Number	Relative Number
Spacing	43,299	1.35%
Capitalization	25,773	0.8%
Hyphen	12,613	0.39%
Punctuation	33,803	1.05%
Combination	60,861	1.9%
Total	176,349	5.5%

Phonological Representation

Whether a response was correct, almost correct, or wrong in its phonological representation can be seen in Figure 6. Again, this assessment is based on rules presented in Table 1. Part of translating orthographical representations of correct answers and responses was an adjustment of substitution costs based on how similar two sounds are from each other. To get an idea about the impact of the substitution cost adjustment, the response assessment with standard substitution costs (all substitution costs 1) can be seen in Figure 7, resulting in 5% fewer “almost correct” responses when similarity is not considered.

Figure 6

Response Assessment in its Phonological Form (Adjusted Substitution Costs)

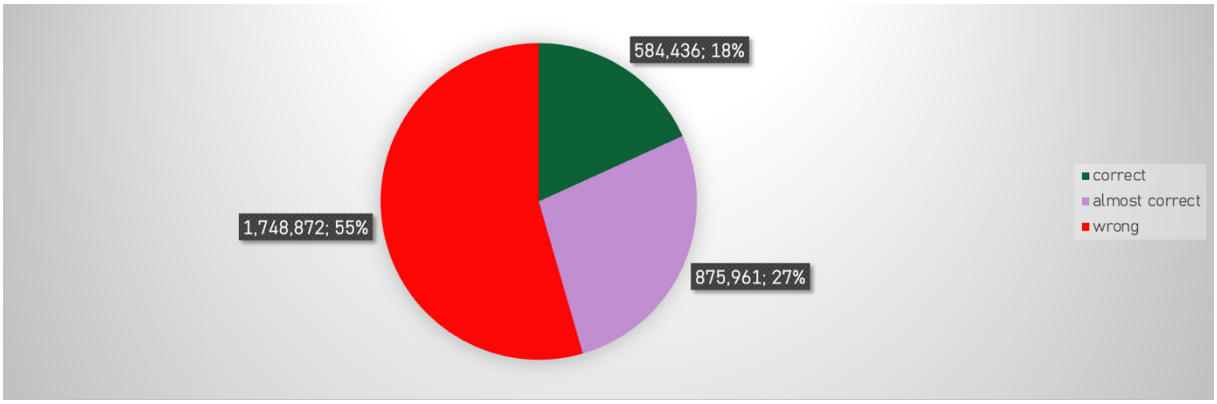
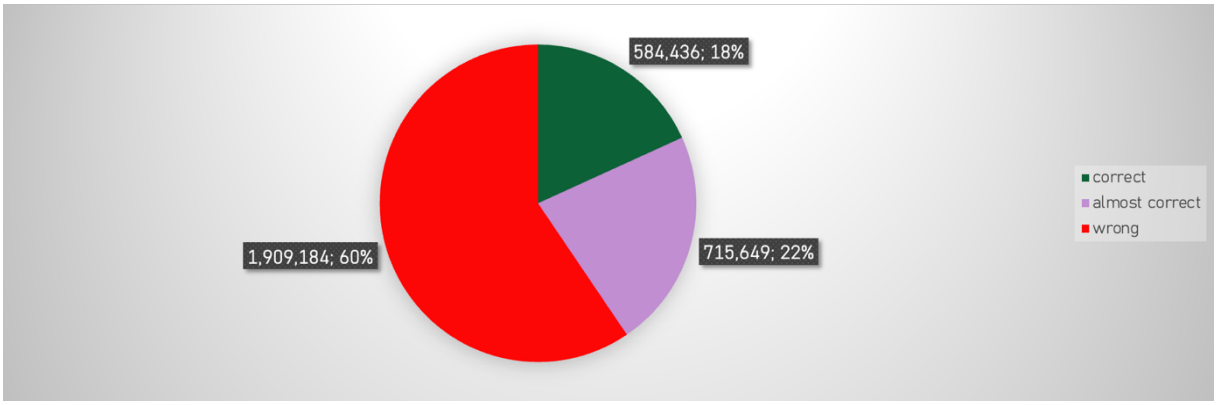


Figure 7

Response Assessment in its Phonological Form (Standard Substitution Costs)

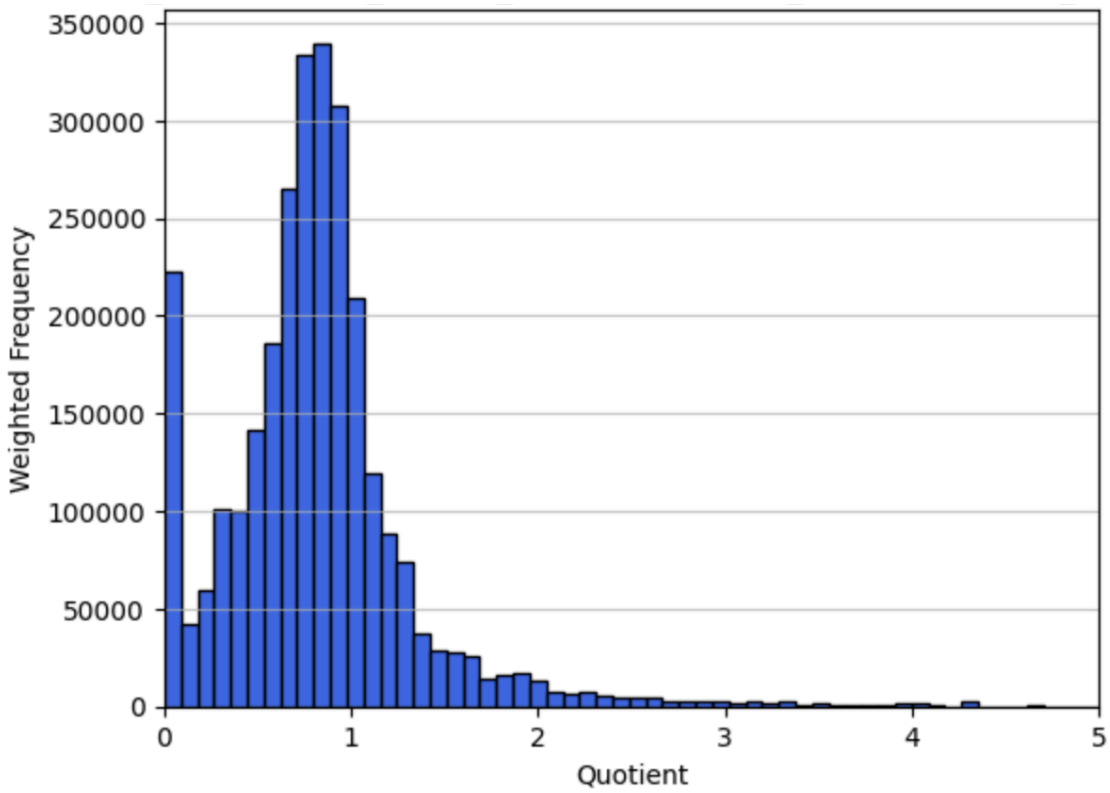


How Phonological and Orthographical Representations Compare

In addition, Figure 8 gives an idea about how many responses in the total dataset of incorrect responses had a better phonological mapping than orthographic. In numbers, 2,139,136 responses (i.e., 75.21% of all incorrect responses) sounded better than they looked, 627,142 responses (i.e., 22.05% of all incorrect responses) sounded worse than they looked, and 77,937 responses (i.e., 2.74%) sounded exactly like they looked.

Figure 8

Overview of how the Phonological and Orthographical Representation of Responses Differs in Terms of Damerau-Levenshtein Distance to the Correct Answer



Note. Only considers incorrect responses. The quotient concerns the following formula:

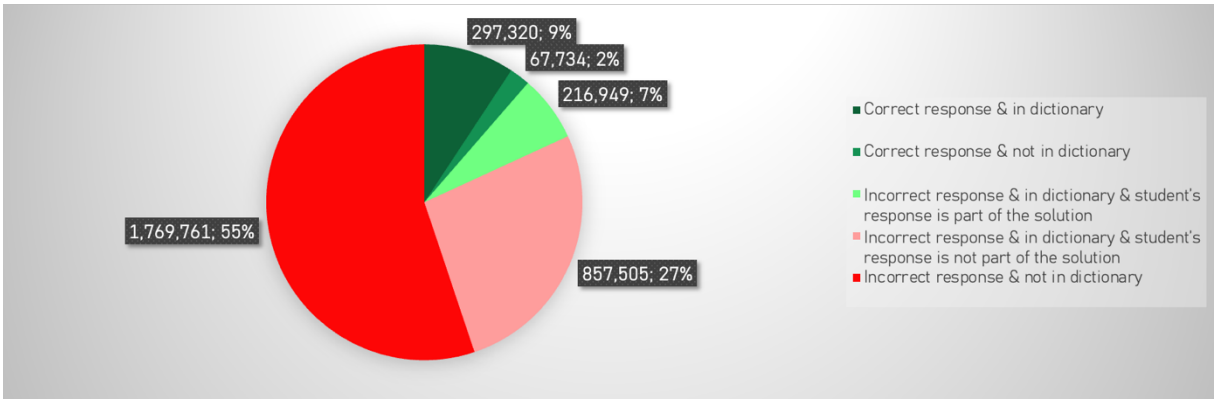
$$Quotient = \frac{DLSD\ (in\ DISC)/Correct\ Answer\ length\ (in\ DISC)}{DLSD\ (in\ orth.)/Correct\ Answer\ length\ (in\ orth.)}$$

Responses with a Different Meaning than the Correct Answer

Figure 9 shows the English dictionary analysis.

Figure 9

Checking for Responses to be in the English Dictionary



German Responses When English Responses Were Expected

Out of a total 3,209,269 responses, 483,988 (i.e., 15.05%) were in the German dictionary (Wendt, 2017). However, only 84,990 (2.65% of total responses) of those responses were not in the English dictionary.

Substitution Typos

Through not penalizing substitution typos, 45,482 out of the 991,907 (i.e., 4.59%) orthographically almost correct responses turn correct. Additionally, 58,650 out of the 1,852,308 (i.e., 3.17%) orthographically wrong responses turn almost correct. 1,250 (i.e., 0.067%) turn from orthographically wrong to correct.

Students' Self-Assessment of Their Responses

Table 3 shows an overview about how often learner responses were assessed correct by themselves based on the mistake category of their input.

Table 3

Overview on Self-Assessed Correct Responses per Mistake Category

Mistake Category	Total Responses	Self-Assessed Correct	% Self-Assessed Correct
Missing 'to '	83,359	70	0.08%
Missing a '(to) '	10,929	10,518	96.24%
Double-letter	115,725	1,068	0.92%
Spacing	43,299	1,230	2.84%
Capitalization	25,773	15,995	62.06%
Hyphen	12,613	5	0.04%

Punctuation	33,803	22,853	67.61%
Orth. incorrect but phonologically correct	219,382	62,412	28.45%
Response in English dictionary and part of the correct answer	216,949	19,938	9.2%
Responses in German wordlist	389,418	15,575	4%
Responses in German wordlist and not in English dictionary	84,990	51	0.06%
Responses turning correct through forgiving substitution typo	45,482	10	0.02%
Total non-correct responses	2,844,215	119,381	4.2%

Note. Only concerns the following subset of all data: Responses that are not orthographically correct.

Discussion

Summary of Key Findings

A variety of mistake categories were identified to address the research question. These include (Figure 4):

- Responses that sounded correct despite being orthographically incorrect
- Responses that became correct when accounting for substitution typos
- Double-letter mistakes
- Responses missing a leading “to ” or “(to) ”
- Spacing, capitalization, hyphenation, and punctuation mistakes
- Combinations of the above

- Responses that were English dictionary-valid and part of the correct answer string
- Orthographically incorrect responses that sounded almost correct
- Semantically incorrect responses
- German words submitted when English input was required

Additionally, certain categories of mistakes were most frequently overruled by students using the "let my response count" feature (Table 3). These included responses missing only a "(to) " (overrule rate 96.24%), capitalization errors (62.06%), and punctuation errors (67.61%). In contrast, responses that would have been correct if substitution typos were ignored were rarely overruled by students – only 0.02% of such cases were accepted as correct.

Implications

Figure 4 suggests that considering the identified mistake categories leaves 37% unexplained mistakes. Of these, 15% are unexplained almost correct responses and 22% are unexplained wrong responses. This means that an algorithm that can detect the mistake categories identified in this thesis can improve the quality of 63% of the feedback messages, as it makes them more explanatory (Hattie & Timperley, 2007; Moreno, 2004; Shute, 2008). Table 3 shows that learners were very likely (96.24%) to let a '(to) ' mistake count anyways, which is an interesting insight, especially comparing it to the number for passing 'to ' mistakes, which is very low (0.08%). This implies that learners see round brackets in the answer to indicate optional rather than mandatory input. Instead of including round brackets in the correct answer, correct answer formulation should therefore either avoid using round brackets or adjust their response assessment procedure that allows input between round brackets to be omitted. In the end, a learner passing a response that was assessed as a wrong response by the system implies that a learner believes the response assessment was incorrect, which might, in turn, lead to frustration. Research suggests that emotional triggers might interfere with cognitive activities such as learning, thus it should be avoided (Picard et al.,

2004). Furthermore, learners were very likely to pass their capitalization mistakes (62.06%) and their punctuation mistakes (67.61%). This implies that many learners believe capitalization and punctuation mistakes not to be meaningful mistakes. Again, to avoid frustration due to overruling perceived to be wrong response assessment, the importance of punctuation and capitalization should be taught, or a more lenient approach to responses should be taken. Another interesting finding Table 3 shows is that substitution typos were almost never overruled (0.02%). This implies that learners consider typographical mistakes to be meaningful mistakes in a language learning app setting and thus should be assessed as incorrect.

Limitations

Correct Answer Formulation

A key limitation of this dataset is that all items were copied verbatim from textbooks; as a result, the designated correct answers are not always compatible with the string-matching methods typically used in language-learning applications. For example, the correct answer to the question *Luchs* (engl. singular lynx) was *lynx (pl lynx or lynxes)*. This formulation does not only include redundant information about the plural forms that were not asked in the question but also include rather complicated formulation. To get a correct response, the student needs to remember every character in the solution, which might get difficult for questions and answers phrased like this and is probably not the intended goal. A learner responding with *lynx* practically knows the correct answer but will still be punished by the system with negative assessment. A better correct answer formulation in that example would be *lynx* as it avoids redundant information and unnecessary complicated formulation. There are several examples of non-optimal correct answer formulation such as *crisp (BE)*; *It's ... /They're ...*; *mouse (sg), mice (pl); (to) stay (at/with)*; *towards (the station/Mr Bell)* and many more. It is apparent that there are norms and rules missing for optimal correct answer

formulation in language learning apps. For ensuring easier string comparison analysis in the future and avoiding unnecessary frustration in the learner, a set of rules was formulated:

- Keep answer formulation as simple and as consistent as possible. Keep in mind that the responses are checked letter by letter by the computer.
- Avoid using (pl) in the answer – use it in the question instead (e.g. ‚What is the plural form of woman („Frau“)?‘).
- Avoid using (BE) in the answer – use it in the question instead (e.g. ‚What is the British word for potato chips („Chips“)?‘).
- Avoid questions that ask for two or more things at the same time – If you want to ask for the singular and plural translation of a word, create a question for each – avoid e.g. *woman, women (pl)* as a single correct response, use first *woman* and second *women* in separate questions.
- If you want to include separate correct answers (e.g. *is not* and *isn't*) separate them with a semicolon like this: *isn't; is not*.
- If you want to indicate a verb use ‘*to* ‘ – avoid using ‘*(to)* ‘ or other variants.

Generalizability

Since the data only included responses made by German learners, this must be considered when making claims about the generalizability of the results. For example, this analysis suggests German learners make systematic f/v confusions based on their knowledge of the German language and how to pronounce German letters. This means mistakes made in second language acquisition (SLA) are dependent on the learners' mother language. Ortega (2014) supports this claim stating that, in general, a learner's mother language influences SLA. Therefore, it might be possible to find different distributions of mistake categories for non-German learners in English vocabulary training. For non-German native speakers, there might even be mistake categories, that could not be identified in the analysis of this paper.

Thus, it is important to consider the learners' origin regarding mistakes in vocabulary training and the implications this has in terms of generalizability of the results.

Analysis Did Not Consider Other Kinds of Typographical Mistakes

The current analysis considered substitution typos. However, this kind of typo is not the only one that can happen. Another typo can happen when accidentally the target key plus an adjacent key instead of only the target key was pressed. In case this kind of typo happens, it would be necessary to delete the accidently pressed character from the response string. Notably, this kind of addition typo can only happen when the device used for the language learning app includes a mechanical keyboard. On digital keyboards, such as the one on phones or tablets, it is not possible to register two keys with only one touch on the screen. Thus, this typo is much more likely on mechanical keyboards. Responses turning correct or almost correct through controlling for addition typos is a mistake category missing in this analysis.

Analysis Constraints

The German Wordlist Itself and Overlap with the English Dictionary

There are English words in the German dictionary that was used in this analysis (Wendt, 2017). Throughout the analysis, this is controlled for by first running the English dictionary check before running the German dictionary check. This ensured only German words were detected by the German dictionary. However, there is additional overlap between the English and the German dictionary that needs to be considered. When not taking capitalization into account, there are words that exist, that can be found in both dictionaries and have different meanings. Examples of those words are *fell, kindergarten, an, bank, chef,* or *gift*. This implies that the numbers reported in the English and German dictionary check in this analysis might not be 100% accurate, as the interpretation of responses by learners found in both dictionaries was unknown.

Substitution Typos Can Turn Words Semantically Different

Substitution typos might change a word in a way that makes it semantically different. For example, a learner wants to respond with *dog* to a vocabulary question. The d-key and f-key are adjacent on a keyboard. A random substitution typo might change the input from *dog* to *fog*. Even though unintended, the input word has a semantically different meaning because of a substitution typo. This mistake will then be detected through the English dictionary check, even though a random typo happened and there was no intention to write a word with the meaning of *fog*. Again, those mistake assessment errors might have influenced the validity of the quantification analysis.

Future Research

Currently, not only feedback but also mistake assessment is often binary in language learning apps (Andersen, 2013). Future research could explore how a more nuanced approach to response assessment based on the findings this paper reports influences the learning process and performance. It is of interest how predictive validity regarding vocabulary test performance changes through reinforcing almost correct responses in the learning process. It might also be interesting to explore how assessment lenience on specific mistake categories influences predictive validity. For example, some might argue that capitalization mistakes in the English language are not as severe as other mistakes. Only proper nouns require capitalization in the English language, such as London or Ferrari, making wrong capitalization of random words a minor mistake. Therefore, some might argue capitalization mistakes should be treated more leniently. This paper can work as a foundation to experiments testing how different response assessment in the learning process impacts the predictive validity regarding vocabulary test performance as it provides an overview about the most common mistake categories made by German English learners.

Shute (2008) claimed that both delayed and immediate feedback both have their general advantages. However, learners perceiving a task to be rather simple benefit from delayed feedback. Learners finding a task more difficult benefit more from immediate

feedback (Clariana, 1990; Knoblauch & Brannon, 1981). Thus, future research could explore whether feedback timing tailored to learners' characteristics about difficulty perception significantly impact the learning outcome and predictive validity in terms of vocabulary test performance.

Conclusion

This thesis investigated how feedback in language learning apps can be optimized by identifying common mistake patterns among German learners of English. Thirteen distinct response categories were identified across more than 3.2 million learner responses, enabling the classification of 63% of previously unexplained responses. These findings provide a foundation for developing intelligent feedback algorithms that offer targeted and explanatory feedback, improving the learning process. Integrating these insights into language learning apps can lead to more accurate response assessment systems, better learner support, and more personalized feedback delivery. As the dataset only includes German learners, findings may not generalize to speakers of other languages. Future work could explore how mistake patterns differ across language backgrounds or how learners respond to tailored feedback during active learning sessions.

References

Abbas, A. M., Hamid, T., Iwendi, C., Morrissey, F., & Garg, A. (2022, March). Improving learning effectiveness by leveraging spaced repetition (SR). In *International Conference on Big data and Cloud Computing* (pp. 145-160). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1051-9_10

Andersen, I. (2013). *Mobile Apps for Learning English. A Review of 7 Complete English Course Apps: Characteristics, Similarities and Differences* (Doctoral dissertation). http://hdl.handle.net/1946/14524

Ashford, S. J., Blatt, R., & VandeWalle, D. (2003). Reflections on the looking glass: A review of research on feedback-seeking behavior in organizations. *Journal of Management*, *29*(6), 773-799. https://doi.org/10.1016/S0149-2063(03)00079-5

Atkinson, K. (2020). Spell Checker Oriented Word Lists (SCOWL) v2. *GitHub*. https://github.com/en-wl/wordlist

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *English linguistic guide* (File EUG_LET). In *The CELEX lexical database* (Release 2) [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania. https://catalog.ldc.upenn.edu/docs/LDC96L14/eug_let.pdf

Chomsky, N., & Halle, M. (1968). The sound pattern of English.

Clariana, R. B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction*.

Fan, X., Liu, K., Wang, X., & Yu, J. (2023). Exploring mobile apps in English learning. *Journal of Education, Humanities and Social Sciences, 8*(1), 2367-2374. https://doi.org/10.54097/ehss.v8i.4996

Febriani, I., & Abdullah, M. I. (2018). A systematic review of formative assessment tools in

the blended learning environment. *International Journal of Engineering & Technology, 4*(11), 33-39. https://doi.org/10.14419/ijet.v7i4.11.20684

Hall, Kathleen Currie, Blake Allen, Edith Coates, Michael Fry, Serena Huang, Khia Johnson, Roger Lo, Scott Mackie, Stanley Nam, & Michael McAuliffe (2022). Phonological CorpusTools Version 1.5.1. *GitHub.* https://github.com/PhonologicalCorpusTools/CorpusTools

Hansen, M., Barnig, M., "fedecosta", Hermann, E., Palombo, L., Weber, J., "ZingBlue", H., Saad, K., Bachmann, M., "fijipants" (2024). gruut v2.4.0. *GitHub*. https://github.com/rhasspy/gruut

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A review of mobile language learning applications: Trends, challenges, and opportunities. *The EuroCALL Review*, *24*(2), 32-50. https://doi.org/10.4995/eurocall.2016.6402

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Klimova, B. (2019). Impact of mobile learning on students' achievement results. *Education Sciences, 9*(2), 90. https://doi.org/10.3390/educsci9020090

Klimova, B., & Polakova, P. (2020). Students' perceptions of an EFL vocabulary learning mobile application. *Education Sciences, 10*(2), 37. https://doi.org/10.3390/educsci10020037

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a

historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254. https://doi.org/10.1037/0033-2909.119.2.254

Knoblauch, C. H., & Brannon, L. (1981). Teacher commentary on student writing: The state of the art. *Freshman English News*, *10*(2), 1-4. http://www.jstor.org/stable/43518564

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning?. *Language Teaching Research, 19*(4), 416-434. https://doi.org/10.1177/1362168814541721

Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and instruction*, *12*(1), 107-119. https://doi.org/10.1016/S0959-4752(01)00018-4

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference (SciPy 2010)* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & cognition, 37*(8), 1077-1087. https://doi.org/10.3758/MC.37.8.1077

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua, 122*(2), 145-163. https://doi.org/10.1016/j.lingua.2011.04.006

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, *32*(1), 99-113. https://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., ... & Strohecker, C. (2004). Affective learning—a manifesto. *BT technology journal*, *22*(4), 253-269. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

Python Software Foundation. (2024). *Python* (Version 3.12) [Computer software]. https://www.python.org/

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Su, D. (2023). weighted-levenshtein (Version 0.2.2) [Computer software]. *GitHub*. https://github.com/infoscout/weighted-levenshtein

Ortega, L. (2014). *Understanding second language acquisition*. Routledge. https://doi.org/10.4324/9780203777282

Wendt, M. (2017). wordlist-german. *GitHub.* https://gist.github.com/MarvinJWendt/2f4f4154b8ae218600eb091a5706b5f4

Xu, J., Wu, A., Filip, C., Patel, Z., Bernstein, S. R., Tanveer, R., ... & Kotroczo, T. (2024). Active recall strategies associated with academic achievement in young adults: A systematic review. *Journal of Affective Disorders*. https://doi.org/10.1016/j.jad.2024.03.010