'Meet me at the ribary' – Acceptability of spelling variants in free-text answers to listening comprehension prompts

Ronja Laarmann-Quante[1] and **Leska Schwarz**[2] and **Andrea Horbach**[1] and **Torsten Zesch**[1]
[1]Research Cluster D^2L^2 – Digitalization, Diversity
and Lifelong Learning. Consequences for Higher Education.
FernUniversität in Hagen, Germany
[2]g.a.s.t./TestDaF-Institut, Bochum, Germany
{ronja.laarmann-quante|andrea.horbach|torsten.zesch}@fernuni-hagen.de
leska.schwarz@testdaf.de

Abstract

When listening comprehension is tested as a free-text production task, a challenge for scoring the answers is the resulting wide range of spelling variants. When judging whether a variant is acceptable or not, human raters perform a complex holistic decision. In this paper, we present a corpus study in which we analyze human acceptability decisions in a high stakes test for German. We show that for human experts, spelling variants are harder to score consistently than other answer variants. Furthermore, we examine how the decision can be operationalized using features that could be applied by an automatic scoring system. We show that simple measures like edit distance and phonetic similarity between a given answer and the target answer can model the human acceptability decisions with the same inter-annotator agreement as humans, and discuss implications of the remaining inconsistencies.

1 Introduction

Imagine a listening comprehension task where a student listens to two people scheduling a meeting at the library. The student is then supposed to answer the question 'Where do they want to meet?' and writes 'ribary' instead of 'library'. Is this answer acceptable or not?

The answer to this question is not an easy one. Human experts perform a complex holistic decision in such a case, primarily based on whether they assume that the learner understood the right answer (see Section 2). The aim of this paper is to get a deeper understanding on which factors influence the acceptability of a spelling variant and ultimately how to model this decision automatically. Thereby, we aim at a model that is transparent and uses features which allow to explain under which conditions the system accepts a variant and under which not. To this end, we conduct a corpus study based on real learner answers and human ratings in a high stakes test of German as a foreign language and explore different operationalizations of spelling variant acceptability. We show that our classifier does not yet reach an adjudicated gold standard, but the human decisions can be approximated up to the same level as human-human agreement. Finally, we discuss possible reasons and implications of the remaining inconsistencies.

The remainder of the paper is structured as follows: In Section 2, we give some background about listening comprehension tasks and the role of orthography. In Section 3, we introduce the data set and in Section 4, we analyze the distribution of spelling variants and the human acceptability decisions. Section 5 examines different features that could be used to operationalize the holistic human acceptability decisions.

2 Background

In many high stakes language tests, listening comprehension is tested with a free-text production task (e.g. DALF[1] for French, Goethe Certificate[2] and TestDaF[3] for German, Cambridge Certificate[4] for English). This means that the test takers have to listen to an audio prompt and formulate an answer in their own words. This gives rise to variance in the answers, e.g. synonyms or different syntactic or orthographic variants (Horbach and Zesch, 2019), which makes the automatic scoring of such answers a challenging NLP task.

While variance at the level of wording or syntax is a topic extensively covered both by short-answer-scoring in general (Ziai et al., 2012) as well as computational semantic similarity (Bär et al., 2012), the implications of orthographic variance are an understudied topic in automatic scoring. In e.g. reading comprehension tasks, where test takers can often copy material from the prompt, spelling errors are

[1] https://www.france-education-international.fr/en/delf-dalf

[2] https://www.goethe.de/de/spr/kup/prf/prf.html

[3] https://www.testdaf.de/

[4] https://www.cambridgeenglish.org/exams-and-tests/

usually ignored (Horbach et al., 2017). In listening comprehension tasks, however, the assessment of orthographic variants (e.g. *ribary* or *librarie* for *library*), plays a much more central role, as we will briefly outline.

Receptive skills like listening comprehension can only be measured indirectly, i.e. by inferring comprehension from the performance in a derived task (Buck, 2001, p. 99), e.g. multiple-choice or true/false questions or free-text production tasks. All these tasks require skills that go beyond pure listening comprehension (Rost and Candlin, 2014, p. 183ff), e.g. reading comprehension for answering multiple-choice items and writing skills for free-text answers. Test designers have to carefully decide whether such a skill is considered to be relevant for the construct to be tested or not. In the context of academic listening, for example, note-taking is an important skill and therefore considered to be construct-relevant (Kecker, 2015). Orthography, in contrast, is considered a construct-irrelevant skill for the task and should thus be ignored for scoring. This means that if the test-taker had the right answer in mind without being able to express it in an orthographically correct way, the answer should be marked as correct (see e.g. Harding and Ryan (2009), Harding et al. (2011)). The crucial difficulty hereby is that the spelling of the word interferes with the assessment whether the test-taker had the right answer in mind. If the test-taker, for example, just produces some vague encoding of the relevant phonetic string, this likewise leads to a spelling variant of the correct answer but it should be marked as incorrect.

Hence, the acceptability of a spelling variant is based on a complex holistic decision that an automatic scoring system is not straightforwardly able to make in the same way. Nevertheless, an operationalization has to be found which leads to ratings that match the human ratings as closely as possible. Furthermore, in a high stakes test it is crucial that the decisions of the automatic scoring system are transparent and understandable to human experts.

3 Data Set

In this paper, we experiment with data from the digital TestDaF. It is a high stakes test designed for students planning to apply for studying at a German university. It assesses test-takers' language abilities at the TestDaF levels 3, 4 or 5, corresponding to the CEFR levels B2 to C1.

	FULL	SPELL
# prompts	17	17
# answers	3,777	310
# answer types	1,572	248
avg. # answ./prompt	222 ± 78	18 ± 15
avg. # types/prompt	92 ± 32	15 ± 9
avg. length (words)	1.6 ± 0.7	1.8 ± 0.7
avg. length (chars)	13.2 ± 6.1	16.3 ± 5.0
accepted answers	53.3 %	54.8 %
accepted answ. types	25.6 %	48.4 %

Table 1: Description of the full data sample (FULL) and the subsample consisting of spelling variants only (SPELL).

The listening comprehension section consists of seven different task types, including selected-response item formats like multiple-choice questions, as well as three constructed response tasks where test-takers are asked to write short answers, between single words and a few sentences in length. In this paper, we concentrate on the task that elicits very short answers of a maximum of five words per prompt. This task is particularly suitable to study the role of spelling variants because other sources of variation are limited compared to longer textual answers.

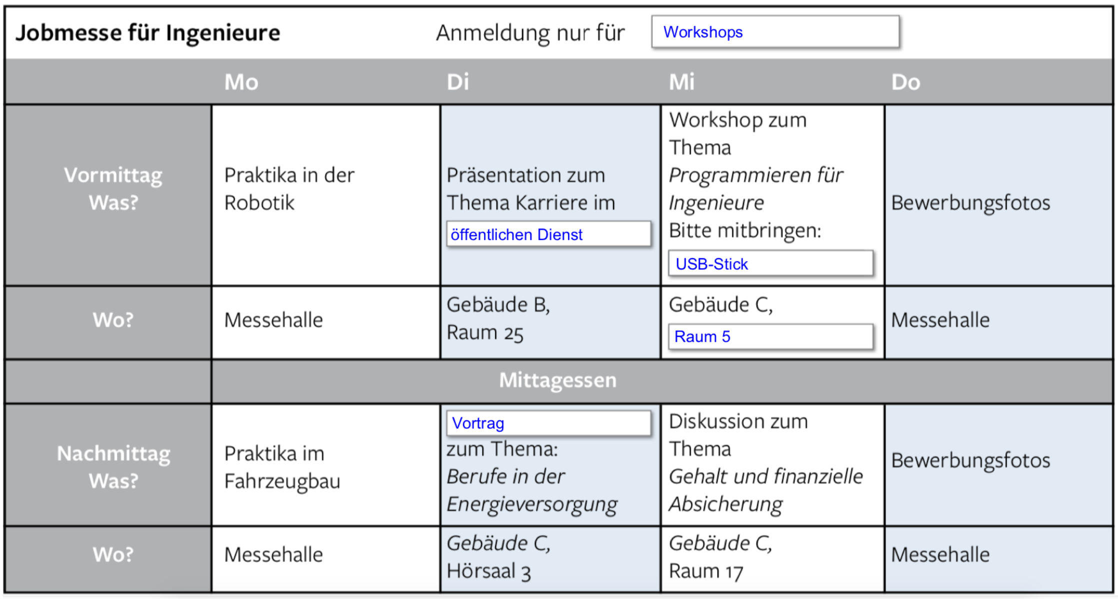
In this task, test-takers listen to a pre-recorded conversation between two or three native speakers in a situation typical for everyday student life, e.g. a conversation between a student and a professor. Test-takers are presented a table, form or chart related to the content of the listening text with five blanks that are to be filled while listening to the input text. See Figure 1 for an example. While test-takers can type in a maximum of five words per blank, all blanks can be answered correctly with one or two words.

For the analyses in this paper, we extracted all answers from 17 different prompts where each prompt corresponds to one blank in the task described above. Table 1, column FULL, shows some basic statistics of the extracted data.[5] Each answer had manually been rated by human experts for whether it was acceptable or not.

4 Human Ratings of Spelling Variants

In the following, we will focus on spelling variants in the data set.

[5] The data set cannot be made publicly available and not all target answers can be revealed in this paper. Some prompts are public, though, and the German examples used in this paper are all real answers to those prompts.



Jobmesse für Ingenieure Anmeldung nur für Workshops

	Mo	Di	Mi	Do
Vormittag Was?	Praktika in der Robotik	Präsentation zum Thema Karriere im öffentlichen Dienst	Workshop zum Thema *Programmieren für Ingenieure* Bitte mitbringen: USB-Stick	Bewerbungsfotos
Wo?	Messehalle	Gebäude B, Raum 25	Gebäude C, Raum 5	Messehalle
	Mittagessen			
Nachmittag Was?	Praktika im Fahrzeugbau	Vortrag zum Thema: *Berufe in der Energieversorgung*	Diskussion zum Thema *Gehalt und finanzielle Absicherung*	Bewerbungsfotos
Wo?	Messehalle	*Gebäude C,* Hörsaal 3	*Gebäude C,* Raum 17	Messehalle

Figure 1: Example of a listening comprehension task in the digital TestDaF that elicits short free-text answers. Target answers are given in blue. We see a timetable for a job fair with the days as columns and morning and afternoon activities (*Was?*) and the corresponding locations (*Wo?*) as rows. The upper left gap, e.g., prompts the test-taker to complete the entry *Presentation about the topic “a career in ______”* with the target answer being *public service*.

4.1 Distribution of Spelling Variants

Two annotators labeled all answer types with a category that describes in which way the answer deviates from the target answer. For a subset of about 500 answer types, we compute the agreement of our two raters on the binary decision whether the answer is a spelling variant or some other variant. Other variants include for example grammatical deviations (e.g. singular/plural), synonyms (*Speicherstick* ‘memory stick’ for *USB-Stick*), or answers that are incomplete (*Raum* for *Raum 5*), unintelligible (*OS*) or referring to something different (*Kaffee* ‘coffee’ for *Workshops*). Inter-annotator agreement is Cohen’s κ=.78, which shows that even for humans, distinguishing spelling variants from other variants, especially grammatical variants, is not trivial.

The two annotators then discussed those cases where they disagreed and decided on a final gold label. For the analyses in this paper, we extracted all answers gold-labeled as spelling variants, including real-word errors. Note that answers which differ from the target answer only with regard to capitalization, hyphenation or splitting a compound in two parts are not part of this set because they are always acceptable.

Table 1, column SPELL, shows some statistics of the spelling variant sample. In total, about 16% of the different answer variants are attributable to spelling, showing that they account for a non-negligible amount of variance in the data.

The distribution of spelling errors follows a Zipf distribution, i.e. most of the spelling variants in our data set occur only once while a few can be found several times. In other words, different test-takers make different kinds of errors, hence it is not possible to foresee all cases beforehand and include them in the rating guidelines or to hard-code them in an automatic scoring system.

The left panel of Figure 2 shows the number of different spelling variants per prompt. One can see that some prompts seem to be more prone to spelling errors than others, with some prompts triggering more than 30 different variants and others only triggering two. We found that there are more spelling variants in prompts with longer target answers than with shorter ones (Pearson correlation r =.58). As one can see in the right panel of Figure 2, the acceptance rate of spelling errors according to the human gold standard varies quite a lot. While for some prompts, most of the variants are accepted, for others, most are rejected. In total, 48% of the spelling variant types are accepted.

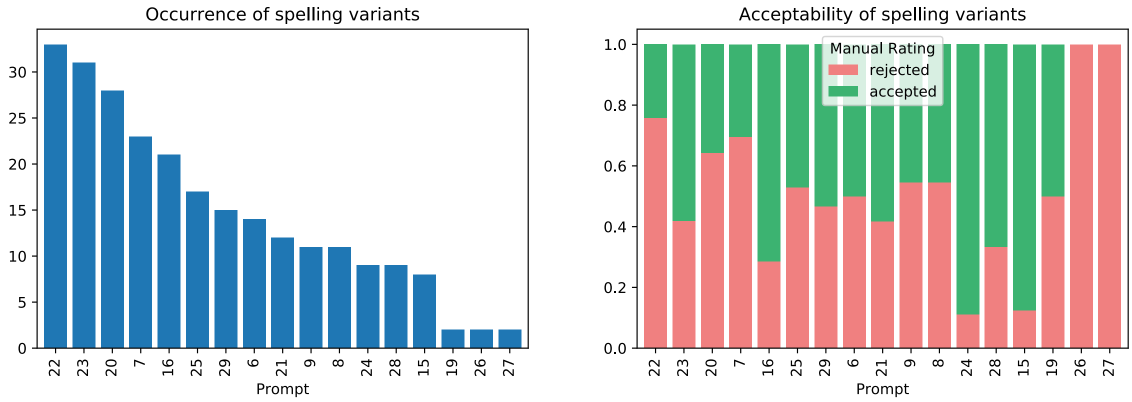


Figure 2: Occurrence (left panel) and acceptance rate (right panel) of spelling variant types per prompt.

4.2 Manual Acceptability Decisions

Test-takers' responses were rated by human experts in a dichotomous format as either *accepted* or *rejected*. Inconsistencies were adjudicated by an additional annotator. Some examples are shown in Table 2. Human raters also need clear criteria to ensure that they mark according to the same standard (Weir, 1993). To achieve this, they were provided with rating guidelines, rater training sessions and standardization meetings.

The rating guidelines consist of general parts, for example that common abbreviations are accepted in an answer, and item-specific parts that contain samples of correct and incorrect answers as well as what is in general expected of a correct response for this item. For example, the guidelines for the target answer *USB-Stick* include the following:

- *USB Stik* is an accepted spelling variant but *USB Tick* and *USP Stick* are not
- *Speicherstik* (*memory stick*) is an accepted synonym with an accepted spelling error (*stik* instead of *stick*)
- *USB Gerät* (*USB device*) is not accepted because it is too general
- *USB* alone similarly does not contain enough information

We compute the inter-annotator agreement of the human experts for the acceptability decision on the same subset as for the annotation if something is a spelling variant. We observe that spelling variants are substantially harder for humans to judge than other answer variants, with a κ value of .60 for spelling variants as opposed to .83 for all other items (see Table 3). Such scoring inconsistencies by human raters despite regular training, annotation guidelines and thorough pre-testing are in line with Buck (2001).

Answer	Accept
Text Entworf	yes
Textentwürf	yes
Testentworf	no
textentw	no
text entworft	no
textintforf	no
Text Einwurf	no

Table 2: Examples of spelling variants and acceptability decisions for the target word *Textentwurf* ('text draft').

	κ	% agreement
all answers	.80	.93
spelling variants	.60	.83
other variants	.83	.94

Table 3: Inter-annotator agreement (Cohen's Kappa) for rating answer variants as acceptable or not.

5 Operationalizing Acceptability Decisions

In the following, we will analyze criteria for the acceptability ratings of spelling variants which could be used by an automatic system. We base our analyses on the set of different spelling variant types. Thereby, we always use the adjudicated labels as the gold standard.

5.1 Surface Distance to Target

The manual scoring guidelines do not prescribe how many errors per word are allowed in order for the answer to count as correct. However, in our sample we can see that the Levenshtein distance

Dist.	SURFACE		STANDARDIZED		PHONEMES	
	#	% acc	#	% acc	#	% acc
0	-	-	1	1.00	20	.85
1	63	.70	147	.66	63	.59
2	72	.60	58	.29	66	.61
3	49	.47	22	.14	36	.31
4	32	.16	13	.15	33	.30
5	17	.24	2	.00	11	.27
≥ 6	15	.07	5	.00	19	.11

Table 4: Frequency and acceptance rate (% acc) of the human raters for all spelling variants with a particular Levenshtein distance (Dist.). The Levenshtein distance is measured on the character level (SURFACE), standardized character level (STANDARDIZED, i.e. ignoring capitalization, hyphens and whitespace) and on the phoneme level (PHONEMES).

between a given answer and the target answer is correlated with the acceptability of the answer. This is detailed in Table 4, column SURFACE. However, despite a trend that words with higher Levenshtein distances are less likely to be accepted, we do not see a threshold above which all answers are rejected or below which all are accepted.

Most frequently, we find a Levenshtein distance of 2 between the given and the target answer. Recall that answers which differ from the target answer only with regard to letter case, hyphenation or splitting a compound in two parts are not included in our spelling variant data set because these deviations by themselves are always acceptable. However, an inspection of the included spelling variants showed that many answers mix capitalization or word-splitting errors with other error types like letter substitutions, e.g. *text entworft* for *Textentwurf*. The Levenshtein distance currently does not take into account that e.g. a capitalization error itself is not as problematic as a different letter substitution. This may blur the actual influence of the Levenshtein distance. Therefore, we standardize the given answers and the target answers by lowercasing, removing hyphens and whitespace and then re-compute the Levenshtein distance.

We can see that a clear majority of standardized answers only has a Levenshtein distance of 1 to the target answer (see Table 4, column STANDARDIZED). Furthermore, there is a clearer trend that the majority of answers with a distance of 1 is accepted while most answers with a higher distance are rejected. Still, an automatic classifier that accepts all answers with a Levenshtein distance ≤ 1 and rejects all other answers would have an accuracy of only 71%. This is clearly above the majority-class baseline of 52% (achieved if all spelling variants are classified as rejected) but far from a sufficiently high accuracy for being used in practice.

5.2 Influence of Keyboard

There are spelling deviations which are intuitively recognized as typos, e.g. *Öffentlivchendienst* for *öffentlichen Dienst*. A typo implies that the test-taker actually knew the word so that it should be marked as correct. As a proxy for whether a spelling variant is actually a typo, we can look whether the substitution or insertion of an erroneous character pertains to a key adjacent to the target key.

Hence, our operationalization of what counts as a typo is as follows: if a standardized answer contains exactly one substitution or one insertion of a character which is adjacent to the target key on a keyboard with QWERTZ, QWERTY, or AZERTY layout, we consider this answer as 'probably only containing a typo'. Using this method, we identified 18 unique typos in the analyzed sample. In 13 of these answers, there are additional deviations in terms of capitalization or the use of whitespace. The human experts scored (only) 12 of the 18 answers as correct, which shows that a spelling variant that is likely a typo is not automatically accepted. The human experts reported that since they cannot know on which type of keyboard a test-taker wrote the answer, they do not explicitly treat (potential) typos differently from other types of errors.

5.3 Phonetic Similarity

In German orthography, most sounds can be represented in more than one way, using different characters or character combinations. For example, a long [aː] can be spelled as <a> (*Tal* 'valley'), <ah> (*Zahl* 'number') or <aa> (*Saal* 'hall'). This means that there can be answers which differ from the target answer in terms of spelling but which are nevertheless pronounced in the same or a very similar way. As with the similarity on the surface level, we can determine the similarity on the pronunciation level by computing the Levenshtein distance between a given answer and the target answer on the phoneme level. We obtained the phoneme representation of each answer from the web service *G2P* of the Bavarian Archive of Speech Signals

Answer		**Target Answer**		**Accept**
Wokshops	*wok shops*	Workshops	*workshops*	yes
Vortag	*previous day*	Vortrag	*presentation*	yes
öffentlichen**dings**	*public thingy*	öffentlichen Dienst	*public service*	no
litersuchen	*liter search*	Literatursuche	*literature search*	no
Testentworf	*test draft*	Textentwurf	*text draft*	no
Text **Einwurf**	*text insertion*	Textentwurf	*text draft*	no
Eigenen**testverwurf**	*own test rejection*	eigenen Textentwurf	*own text draft*	no

Table 5: Examples of real-word spelling variants. Those parts of the word that correspond to another existing word are printed in bold.

(BAS) (Reichel, 2012; Reichel and Kisler, 2014).[6] As one can see from the column PHONEMES in Table 4, most answers with the same pronunciation as the target answer are accepted (85%), but not all. On the other hand, most answers with quite a different pronunciation are rejected, but again there are exceptions. This shows that phonetic similarity alone is not a decisive factor either.

5.4 Similarity to Other Words

In our data sample, we manually identified a total of 34 spelling variants that resulted in other existing words (real-word errors). Most of them occurred only once, resulting in 27 unique variants. Hence, 11% of all spelling variant types are real-word errors. Not all prompts trigger real-word errors to the same degree. For 8 out of the 17 prompts, no real-word error could be found while one of the prompts triggered eight different real-word error types.

Most of the real-word errors are rejected by the human raters – but not all of them: 3 out of the 27 real-word error types were accepted. What is noteworthy is that all of the accepted real-word errors have a Levenshtein distance (on the character level) to the target word of 1. In contrast, the rejected real-word errors have a mean Levenshtein distance of 3.6. Hence, a factor influencing the acceptability of the real-word error seems to be the surface similarity. However, among the rejected answers, there are also four real-word errors with a Levenshtein distance of 1 to the target answer, which shows that there are more complex mechanisms at work. Human experts reported that one factor influencing their decision is whether the meaning of the real-word error would be somewhat plausible yet still incorrect in the context of the given task, and therefore would be confusing in a real-life setting. In contrast, if an answer is far-fetched or consists of a word that is very infrequent, human raters would assume that the error was indeed only an orthographic error and the learner actually meant to write the correct word.

To illustrate this, Table 5 shows some example answers and their acceptability. Most target answers are compound words and the real-word spelling errors mostly only pertain to one part of the word. As a consequence, the error results in a grammatically well-formed answer but often in a non-lexicalized word. In some cases, the meaning of the new compound is far off the meaning of the target answer, e.g. *Workshop* and *Wokshop* (in English, the corresponding words are *workshop* and *wok shop*, i.e. the compound that is a result of the spelling error would have to be written as two words, which is not the case in German). In other cases, the meanings are somewhat close and could lead to a misunderstanding in real communication, e.g. *Textentwurf* ('text draft') and *Testentworf* ('test draft'). It remains to be seen with a larger sample of accepted real-word errors how well this can be operationalized by an automatic scoring system.

5.5 Combination of Features

While all of the criteria presented above play a role for the acceptability decision, we could see that none of these factors alone suffices to differentiate between accepted and rejected answers. In the next step, we examine whether a combination of the features can be used to approximate the human acceptability decisions. We aim for a model that yields interpretable results so that one can identify under which conditions a spelling variant is accepted or rejected.

[6] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Grapheme2Phoneme

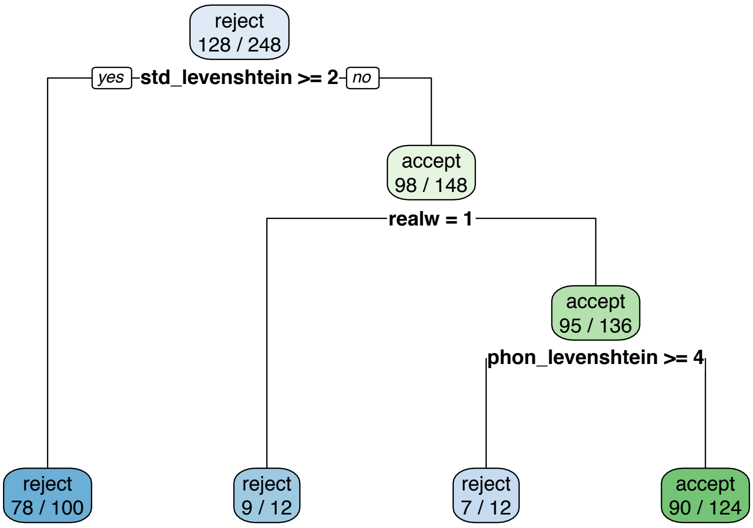


Figure 3: Simple decision tree (pruned) for the acceptability decision.

In order to do so, we train different decision trees on the whole set of spelling error types and the adjudicated gold labels using the R package *rpart* (Therneau and Atkinson, 2019). We then apply the trees to a test set of 127 spelling variant types from 5 new prompts, i.e. a new set of learner and target answers. We use classification accuracy as evaluation metric.

In addition, we apply the trees to the training data set itself in order to get an estimate how consistently the data can be modeled, i.e. whether the features suffice to tell accepted and rejected answers apart or whether there are answers with the same combination of features but different human judgments. The results are shown in Table 6.

Baselines If all instances are classified as rejected, this **majority-class** baseline reaches an accuracy of 52% on the training set. In the test set, the classes are evenly distributed, i.e. the baseline is 50%. Using **character edit-distance** alone as classification criterion, as discussed in Section 5.1, the accuracy rises to 71% on the training set and 73% on the test set.

Simple Trees First, we build a decision tree with default configuration using the features and their operationalizations that were described in the previous sections:

- edit distance on the character level between standardized given answer and standardized target answer, i.e. ignoring letter case, hyphens and whitespace (`std_levenshtein`, numeric)
- edit distance on the phoneme level (`phon_levenshtein`, numeric)
- whether the word is a real-word error (`realw`, binary)
- whether the word probably only contains a typo (`probably_typo`, binary)

This tree is grown with default parameters, which in particular means that it is automatically **pruned**, i.e. not grown to full depth. For a predictive model, this is necessary in order to prevent overfitting on the training data. The resulting tree is shown in Figure 3. In prose, the tree accepts a spelling variant if the edit distance on the character level is < 2, it is not a real-word error and the edit distance on the phoneme level is < 4. The nodes show how many data points fall into the respective class and how many of them are categorized correctly when applied to the training data. In total, the tree reaches an accuracy of 74.2% on the training set and 70.9% on the independent test set. For the test set, this is worse than using character edit-distance alone.

In order to find out whether the features do actually suffice in order to model the data that the tree was trained on, we next grow the tree to **full** depth. The resulting tree has a depth of 8 (compared to the depth of 3 in Figure 3) but still only reaches an accuracy of 76.2% on the training data. This means that there are answers with the same feature set but different acceptability decisions (see discussion in Section 5.6). As one would expect due to overfitting, the full-depth tree performs worse on the test set than the pruned tree.

Advanced Trees One potential limitation of the current feature set is that our version of edit distance is not sensitive to word length. Therefore, we normalize the character edit distance by the number of characters in the target word and also allow for transpositions of characters to count as one edit (`norm_std_damerau_lev`). The other three features remain the same. The default **pruned** tree based on this adapted feature set has a depth of 5 and an accuracy of 75.4% on the training set, which is very similar to the result of the simple tree. See Figure 4 for an illustration of the advanced tree.

However, on the test set, the tree produces much better results than the simple tree with an accuracy of 84.3%. That the result for this tree on the test set is even better than that on the training set indicates that the tree's rules for accepting an answer are indeed transferable to new data sets. In fact, some of the rules even fit the test data better than the training data. For example, 45.6% of the training data and 46.5% of the test data fall into the rightmost leaf node in Figure 4. The answers that fall into this node are predicted to be accepted. In the training data, this decision is correct in 73% of the cases, whereas in the test data, the decision is correct even for 85%.

If we grow the advanced tree to full depth (= depth of 14), the overall accuracy on the training set rises notably, but only to 85.1%. Hence, it still does not reach the adjudicated gold standard but the result is comparable to the human-human agreement of 83%. As we will discuss shortly, the fact that we do not reach 100% accuracy even with this full-grown tree shows that more or different features are needed to tell accepted and rejected answers apart. Since this tree overfits the data, its performance on the test set is much worse than that of the pruned tree, hence it is not suitable for predicting new data points.

5.6 Discussion

We observe that our features do not suffice to perfectly model the acceptability decisions of human raters according to an adjudicated gold standard. There are conflicting cases which cannot be resolved on the basis of the features we currently examine. Some examples are given in Table 7.

	Accuracy	
Method	**Training**	**Test**
majority baseline	.52	.50
char. edit distance	.71	.73
simple pruned tree	.74	.71
simple full tree	.76	.69
advanced pruned tree	.75	.84
advanced full tree	.85	.72
human agreement	.83	-

Table 6: Overview of classification results.

Answer	**Human Accept**
Öffenlichtendienst	yes
offentlischen Dienst	yes
kreatives schrıeben	yes
höffentliche Dienst	no
oofentlichen DIENST	no
offentlichene Dienst	no
krätives schreiben	no

Table 7: Examples of answers (target answers = *öffentlichen Dienst*, *kreatives Schreiben*) that all fall within the same node of the advanced full tree but are rated differently by human raters.

Differences between the accepted and not-accepted cases are subtle and human experts often argue in terms of whether an answer looked "too far off" without being able to specify a general rule supporting their decision. Additional features might be able to distinguish between those cases. However, it may also mean that the human ratings are not fully consistent, which is in line with our observed inter-annotator agreement. In fact, the accuracy of the overfitted tree (85%) is very similar to the human-human agreement on the same data (83%), which we discussed in Section 4.2, hence, we may not expect a system to ever go significantly beyond this value. Therefore, basing the acceptability decision on objectively measurable features instead of individual holistic decisions of human raters could be a way to arrive at more consistent and more explainable results especially in a high stakes test.

6 Conclusion and Future Work

We presented an analysis of the rating of spelling variants in a listening comprehension task from the TestDaF test. We found that spelling variants are more challenging to score for human experts than other types of variants. Furthermore, we ex-

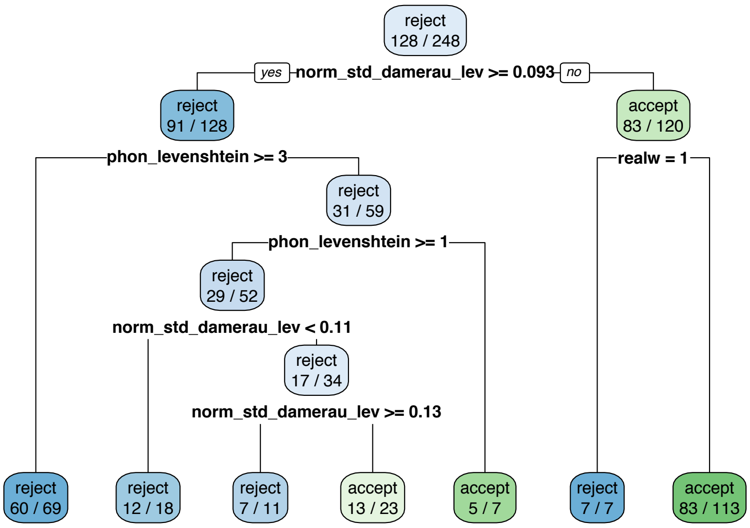


Figure 4: Advanced decision tree (pruned) for the acceptability decision.

plored how the acceptability decision can be operationalized with automatically extractable features such as edit distance and phonetic similarity as a first step towards an automatic scoring system for spelling variants. Their combination in a decision tree reaches a performance similar to human-human agreement, but not exceeding it. This can mean either that human decisions are not fully consistent or that further features are needed to differentiate between cases that currently end up in the same leaf node of the tree.

Options for such additional features include specific error categories as opposed to generic distance-based measures, such as the spelling error categories defined in the Litkey Corpus (Laarmann-Quante et al., 2019). These error categories can be divided into 'systematic' ones (like omitting an <e> that corresponds to an (almost) non-audible [ə]) and 'non-systematic' ones (such as omitting a full vowel). First explorations indicate that 'systematic' errors more likely lead to acceptable spelling variants than 'non-systematic' ones. As another option to obtain more consistent annotations, we plan to explore annotation studies where human raters have access to the automatically extracted features and/or the scoring suggestion learnt by the classifier as a basis for their scoring decision.

Acknowledgments

This research was partly funded by g.a.s.t. (Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V./ *Society for Academic Study Preparation and Test Development*) and conducted in the framework of the Research Cluster D^2L^2 "Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education" of the FernUniversität in Hagen, Germany (https://e.feu.de/english-d2l2). We would like to thank Frederik Wollatz and Lisa Prepens (students from the University of Duisburg-Essen) for their annotation work.

References

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 435–440.

Gary Buck. 2001. *Assessing Listening*. Cambridge Language Assessment. Cambridge University Press.

Luke Harding, John Pill, and Kerry Ryan. 2011. Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly*, 8(2):108–126.

Luke Harding and Kerry Ryan. 2009. Decision making in marking open-ended listening test items: The case of the OET. *SPAAN FELLOW*, 1001:99.

Andrea Horbach, Yuning Ding, and Torsten Zesch. 2017. The Influence of Spelling Error on Content Scoring

Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 45–53, Taipei, Taiwan. AFNLP.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, 4:28.

Gabriele Kecker. 2015. "He listens well who takes notes" – Mit welchen Aufgabenformaten kann Hörverstehen in Vorlesungen an der Hochschule valide getestet werden?". In Jessica Böcker and Anette Stauch, editors, *Konzepte aus der Sprachlehrforschung – Impulse für die Praxis*, pages 511—-526. Peter Lang, Bern, Switzerland.

Ronja Laarmann-Quante, Anna Ehlert, Katrin Ortmann, Doreen Scholz, Carina Betken, Lukas Knichel, Simon Masloch, and Stefanie Dipper. 2019. The Litkey Spelling Error Annotation Scheme: Guidelines for the Annotation of Orthographic Errors in German Texts. *Bochumer Linguistische Arbeitsberichte (BLA)*.

Uwe D. Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *INTERSPEECH*, Portland, Oregon.

Uwe D. Reichel and Thomas Kisler. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In R. Hoffmann, editor, *Elektronische Sprachverarbeitung: Studientexte zur Sprachkommunikation 71*, pages 42–49. TUDpress.

Michael Rost and CN Candlin. 2014. *Listening in language learning*. Routledge.

Terry Therneau and Beth Atkinson. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Cyril J Weir. 1993. *Understanding and developing language tests*. Prentice Hall New York.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 190–200.