Rule-Based Grammatical Error Detection on Spontaneous Children's Speech

Christopher Gebauer[1], Lars Rumberg[1], Fabian Witt[1], Edith Beaulac[2],
Hanna Ehlert[2], Jörn Ostermann[1]
[1]Institut für Informationsverarbeitung - L3S, Leibniz Universität Hannover,
[2]Institut für Sonderpädagogik, Leibniz Universität Hannover
gebauer@tnt.uni-hannover.de

Abstract: Successful language acquisition is fundamental for the participation in all relevant areas of life. Unfortunately, developmental language disorders (DLD) are the most common developmental disorders during childhood. An early identification of children with DLD is necessary to ensure educational success and to significantly reduce the risk for issues in mental health, social behavior, and skill development in various areas. In this work we explore the suitability of a rule-based grammatical error correction system for written text, to assess grammatical correctness of spontaneous children's speech. We evaluate the tool's capabilities on kidsTALC, a corpus containing dialogues of children with speech language therapists in different elicitation contexts, using the manual and automatically generated transcriptions. A qualitative analysis reveals that disfluencies and a lack of context are the leading causes for misclassifications. Nevertheless, the system achieves a FG_1-score of 0.92 when applied to manual transcriptions.

1 Introduction

Successful language acquisition is a fundamental prerequisite for participation in relevant areas of life, educational success and later career opportunities [1]. With a prevalence of about 7 %, developmental language disorders (DLD) are the most common developmental disorder in childhood [2]. Early identification of children with language support and therapy needs can significantly reduce risks in the areas of skill development, mental health and social behavior [3]. A common area of weakness for children with DLD between the ages of 4 and 8 years is the use of grammatical forms [4]. Therefore, in this work we focus on the aspect of automatic grammatical error detection (GED) within the domain of spontaneous children's speech assessment.

Even on written text GED is a difficult problem as the correct wording and, therefore, the detectable incorrect words depend on the intended meaning of the assessed text [5], which is usually unknown. To address this problem two opposing approaches exist to predict the grammatical proficiency of students (children or second learner). The first group of approaches directly predicts speech assessment scores from human experts, either based on pretrained word embeddings [6] or based on a collection of features like fluency, speaking rate, and part of speech (POS)-tags [7]. The second direction, which we will focus on in this work, predicts the correctness of each word within a given text and enables the derivation of proficiency scores based on the predicted classification. In contrast to direct proficiency scoring the latter method allows for a much higher degree of transparency, as it is possible to trace the location of each grammatical error within the text as well as its error class. Morley *et al.* [8] include the error labels into the arcs of a dependency tree and train a parser model on those labeled data. While the results are promising, high quality data for both, error marks and dependency trees, need to be manually annotated. Additionally, this method has not been adapted to automatically generated transcriptions. He *et al.* [9] trains a large transformer-based system from scratch to predict the grammatical error class based on tokenized text. Bell *et al.* [10] and Lu *et al.* [11] leverage the knowledge of two different pretrained word-embeddings. Both use a bidirectional LSTM network to aggregate the contextual knowledge and predict an error label per word. All these approaches are highly domain-dependent and require finetuning if the data during inference is too distinct [11, 12].

In this work, we predict grammatical errors based on changes derived from a grammatical error correction (GEC) system, similarly to Banno *et al.* [13]. Even tough Hassanali *et al.* [14] demonstrates that rule-based systems are outperformed by feature-based classifiers on GED, we explore the former as a lightweight and transparent baseline for GED on spontaneous children's speech with no necessity of finetuning. To do so, we use LanguageTool[1], a commonly deployed grammar correction tool for written text. We demonstrate the capabilities of LanguageTool to predict grammatical errors on spontaneous children's speech. Further, we qualitatively analyze shortcomings introduced by the domain gap between written text and spontaneous speech. Closing, we investigate the usability of automatic speech recognition (ASR)-based transcriptions as input to the GED-system. Similarly, as proposed by Lu *et al.* [11], we compensate for ASR-based errors by incorporating the ASR-models confidence [15] to discard likely incorrect words.

2 Background

In this section we will explain LanguageTool in detail as well as our approach to GED.

2.1 LanguageTool

Initially, LanguageTool is proposed by Naber [16]. The analyzed text is preprocessed in three parallel pipelines: A POS-tagger, sentence extractor, and phrase extractor. The POS-tags are for each word selected based on two probabilistic, precomputed models. The first takes for each word in the text the probability distribution over the set of possible POS-tags from a look-up table. In the case of out-of-vocabulary words, the most likely POS-tag is predicted based on predefined patters, e. g., a word starting with a capital letter is probably a noun. Next, those probabilities are refined by conditioned probabilities given the neighboring words and its most-likely POS-tags. Afterwards, the POS-tags are greedily selected. The second pipeline extracts sentences from the given text, which is purely based on punctuation in the initial version of LanguageTool. It is further assumed by LanguageTool that the majority of grammatical rule violations do not cross those sentence borders. The last pipeline segments the sentences into grammatically and semantically related phrases. An example is a noun phrase, which contains a determiner, associated adjectives and the noun. Based on the POS-tags, detected word phrases, and the specific word combinations a predefined set of grammatical rules are searched. If a matching rule is found, the text segment is corrected according to the respective rule. Additional rules, independent of the above computed pipelines, are applied as well, e. g., for word repetitions. Furthermore, the information from a statistical n-gram or learned Word2Vec-based [17] language model is integrated. A proprietary extended grammatical rule set and further processing steps are available in the latest software. As no further details on them are publicly available, we will not use them.

[1] `https://languagetool.org/de`

2.2 Grammatical Error Detection

The evaluation data, see Sec. 3.1 for more details, differentiates between lexical, phonological and grammatical errors, but does not contain detailed information on the type of grammatical errors. Therefore, we focus on the binary classification of grammatical (and lexical) errors, i. e., predicting the presence of an error. For brevity, when we refer to grammatical errors or correctness we always include lexical errors if not stated otherwise. To predict grammatical errors we apply LanguageTool, align the original with the corrected text on word-level and classify mismatches in the aligned word pairs as grammatical errors. Grammatical rules that account for recapitalization are always ignored. We further ignore rules for word repetitions and compounding, since those errors are not considered as grammatically relevant in spontaneous speech. The influence will be discussed in Sec. 4. On the automatically generated transcriptions from our ASR-system, see Sec. 3.2, we incorporate confidence scores by treating any word below a given threshold as grammatically correct. The confidence scores are estimated on word-level based on Rumberg *et al.* [15].

3 Experimental Setting

In this section we outline the used speech corpus, namely kidsTALC, the ASR-system, our preprocessing of the transcriptions and, closing, the metrics used to evaluate the GED-system.

3.1 Speech Corpus

We evaluate our GED-system on kidsTALC [18], a corpus of monolingual, typically developing, German-speaking children ranging from 3½-11 years. We extend the test-set by recordings from additional 40 children focusing on the younger age groups of the kidsTALC corpus (AG1: 14; AG2: 23; AG3: 2; AG4: 1). The additional children are almost equally distributed across gender (female: 21; male: 19). All additional children are recorded in the same manner and setting as proposed alongside with the kidsTALC corpus. In total 3 children with speech sound disorders (SSD) and a variety of error patterns that are typical for the respective age are included in the test-set. For all children a manual utterance-level segmentation, transcription, and annotation of error marks are available. The annotations in kidsTALC contain three classes of error marks: Grammatical, lexical and phonological errors. In this work we consider lexical and grammatical errors, further details on the specific error type are not annotated.

To create labels for the automatically generated transcriptions, we align those transcriptions to the manual transcriptions on word-level. We transfer the labels of a manual word, if the aligned, automatically generated word is identical. If the aligned words do not match, the manual labels cannot be correctly transferred. Either those labels are transferred anyway, which results in certain grammatically correct but misaligned words being labeled as grammatically incorrect. If the labels are discarded, by the GED-system correctly identified, grammatically incorrect words, whose labels have been dropped, are penalized. On the other hand, grammatically incorrect words that are not identified as such will not be penalized. We choose to omit those labels to ensure that metrics only relying on the positive class, i. e., words that are labeled as grammatically incorrect, are not skewed by the label transfer but evaluated on a subset of the actual labels. Overall, the omission results in a loss of about half of the manual labels.

3.2 Automatic Speech Recognition

In this work we use a pretrained Wav2Vec 2.0 model [19], specifically a German model from Hugging Face[2]. A small classification head, consisting of two dense layers, is trained on top of the Wav2Vec 2.0 model using the connectionist temporal classification (CTC)-loss [20] and the Adam optimizer [21] with a learning rate of $l_r = 0.0001$. We freeze the feature encoder of Wav2Vec 2.0 and finetune its contextual layer jointly with the classification head. As suggested by Rumberg *et al.* [18], we extend the kidsTALC corpus by Mozilla Common Voice (MCV) [3]. The token set for the orthographic transcript is the full, lower cased, German alphabet and does not include any punctuations.

As decoding scheme we deploy greedy search, i. e., taking the most likely token within each time step, and a beam search algorithm [22]. The word error rate (WER) of the greedy search on the same recordings as used in Sec. 4 is 0.441 averaged across two seeds and two splits. The beam searcher is constrained using a lexicon based on the train-set from kidsTALC as well as MCV. As language model we deploy a 4-gram KenLM [23] trained on the train-set from kidsTALC as well as MCV. We apply a beam size of 500. The beamsearcher achieves a WER of 0.349 averaged across two seeds and two splits.

3.3 Preprocessing

For comparability between the manual and automatically generated transcription, any punctuation, special characters and capitalization are removed from the manual transcriptions. We noticed an improvement in GED performance when punctuations and capitalization are reconstructed, before passing the text to LanguageTool. Therefore, we reconstruct both using a GEC-model from Hugging Face[4] trained on German Wikipedia data only for recapitalization and reconstructing punctuation.

We select context windows from the full transcriptions based on full sentences or a combination of multiple consecutive sentences. The context windows are passed to the GEC- and then to the GED-model at once. To maximize context we keep the utterances from the speech language therapist (SLT) and mask it during evaluation. We constrain the windows to be close to a size of 50 words and shift the windows close to a length of 10 words. This leads to an average window size of 49.7 ± 6.8 and an average hop length of 13.7 ± 9.3. The performance is similar to a fixed sized window with fixed hop length, but more convenient to compare against GED-systems that rely on more context. During experiments, we noticed that the GEC-model above slightly adjusts the processed text besides punctuation and recapitalization (0.51 % of words, compared to 4.47 % by LanguageTool). Therefore, for full stop reconstruction during the selection of context windows a different model from Hugging Face[5] is selected. It is specifically trained on reconstruction of punctuation and keeps the text unchanged. Due to the combination of window size and hop length multiple predictions can be generated per word. Those are aggregated by the *or*-operator, i. e., if in any context window the word is marked as erroneous the word is overall predicted as erroneous.

3.4 Performance Metrics

We evaluate our grammatical error detection using the F_1 gain score [24], which we refer to as FG_1. In contrast to the common F_1-score, the gain score allows to average across different tasks

[2] `https://huggingface.co/facebook/wav2vec2-large-xlsr-53-german`

[3] `https://commonvoice.mozilla.org/en/datasets`, Version 16.1, German

[4] `https://huggingface.co/aiassociates/t5-small-grammar-correction-german`

[5] `https://huggingface.co/oliverguhr/fullstop-punctuation-multilang-large`

Table 1 – F_1 gain score (FG_1) for prediction of grammatically incorrect words on manual transcriptions of kidsTALC. Excluding rules for, e. g., word repetitions and compounding, from LanguageTool's set of rules (ExR) slightly improves the FG_1 score. Recovering punctuations and recapitalize the text (PuRe) leads to a greater improvement. Overall, the combination of both performs best.

Manual	-	ExR	PuRe	ExR+PuRe
FG_1	0.85	0.87	0.91	0.92

by normalizing the score with the proportion of positive labels π:

$$\mathrm{FG}_1 = \frac{\mathrm{F}_1 - \pi}{(1-\pi)\mathrm{F}_1}.$$

This is necessary in our case, as the fraction of positive labels in the automated pipeline varies between seeds due to the mapping of the ground truth labels onto the predicted words. Furthermore, due to the selected method for the transfer of the ground truth labels onto the automatically generated transcriptions, the performance on grammatically incorrect words (positive labels) is especially interesting. Therefore, we use the analog recall gain score, referred to as RecG, for evaluations on the automatically generated transcriptions. Further, we qualitatively discuss general patterns for false positives as well as false negatives and the quality of the manual labels in kidsTALC.

4 Results

In this section we evaluate the proposed GED-system on kidsTALC. First, we predict grammatically erroneous words based on the manual transcription after all punctuations and capitalization are removed. The results are summarized in Tab. 1. Performing GED on the cleaned manual text leads to a FG_1 score of 0.85. Excluding rules for, e. g., word repetitions and compounding, and recovering punctuations as well as capitalization results in an improved FG_1 score of 0.87 and 0.91, respectively. The overall best performance of 0.92 is achieved by a combination of both adaptations.

A qualitative analysis of the results on manual transcriptions reveals different reasons for misclassifications. Starting with false positives, i. e., predicting an erroneous word which is correct, one major cause are disfluencies. Especially, short phrases or exclamations that are incomplete in itself and inserted within an utterance of the child are corrected by LanguageTool. Another source for false positives are transcribed word fractions due to false-tries, neologisms or misspelled words in the transcription. Both these reasons are uncommon or even undesirable for written text and, therefore, correctly marked by LanguageTool. However, they should not be regarded as grammatical errors during analysis of spontaneous speech. Only a small, but noticeable, fraction of the misclassifications originate in missing error marks in the manual transcriptions.

Next, we take a look at false negatives. A major source of misclassifications are pronouns whose correctness is only determinable using context, i. e., the sentence borders need to be crossed to determine what the referred object of the pronoun is. Secondly, tenses of verbs (often the perfect tense), especially for irregular verbs, and the correct usage of auxiliary verbs (*wir sind gegessen* instead of *wir haben gegessen*) are missed by LanguageTool. Next, in certain situations it is not determinable if the current used case of a noun is correct without further context. The determiner *die* in the phrase *hier an die Schnauze, so ganz komisch*, which is a standalone utterance from kidsTALC, is manually marked as grammatically incorrect. Depending on the context this, however, might be grammatically correct as well: *Darf ich hierhin fassen, hier an*

Table 2 – F_1 gain score (FG_1) and recall gain scores (RecG) for predictions of grammatically incorrect words on ASR-based transcriptions from kidsTALC. All scores are averaged across two seeds and two splits. Marking words with low confidence as correct (+Conf.) drastically reduces false positives due to misspelled words, but also reduces the number of true positives.

Automatic	Greedy		Beam	
	-	+Conf.	-	+Conf.
FG_1	0.58	0.94	0.89	0.94
RecG	0.98	0.96	0.98	0.94

die Schnauze?. Interestingly, LanguageTool misses neologisms, irregular composition of nouns and even nonsense words (*unterwasserliges*). Closing, lexical errors are not detectable purely based on the transcriptions. This occurs, if a child uses a wrong but existing name for, e. g., an animal visualized in a picture book. This would only be detectable if corrected by the present SLT verbally or multi-modal information are used.

Next, we evaluate the GED-system using the ASR-based transcriptions. The results are summarized in Tab. 2. The ground truth labels are mapped as described in Sec. 3.1, which results in a loss of half of the labels. All scores are averaged across transcriptions from ASR-systems trained with two different seeds and two different splits. For all settings we recapitalize the text, recover punctuations, and exclude grammatical rules for, e. g., word repetitions and compounding, from LanguageTool's set of grammatical rules as it performed best in the manual setting. In Tab. 2 the poor performance of the GED-system based on the greedy decodings without confidence stands out. This is caused by numerous false positives due to misspellings introduced by the ASR-system. The false positives can be reliably reduced by excluding samples with low confidence, leading to a FG_1 score of 0.94. We select a threshold of 0.95 for the confidence scores without further optimization of this hyperparameter. Due to the constraints to a lexicon the GED-system based on the transcriptions of the beamsearcher does not struggle with high-frequent misspellings and achieves a FG_1 score of 0.89 even without confidence scores. If confidence scores are incorporated, the beamsearcher-based GED-system performs similar to the one using greedy decodings. It needs to be noted that the higher FG_1 scores compared to the manual setting originate in the loss of labels and a reduction of π. While the confidence heavily reduces the false positives, the recall gain scores (RecG) in Tab. 2 demonstrate that true positives are also reduced. Without confidence the GED-system performs similar for the greedy decodings and beamsearcher-based transcriptions in terms of RecG. However, with confidence the RecG scores for both decoding variants drop, the beamsearcher-based GED-system drops slightly more.

5 Conclusion

In this work we investigate the suitability of LanguageTool, a rule-based grammatical error correction system for written text, to assess grammatical correctness of spontaneous children's speech. Further, we investigate the usability of ASR-based transcriptions to overcome the need of tedious, manual transcriptions. We counteract the errors introduced by the ASR-system by excluding words with low confidence. To evaluate our GED-system we use kidsTALC, a corpus containing dialogues of children with speech language therapists in different elicitation contexts. A qualitative analysis reveals that disfluencies and a lack of context are the leading causes for misclassifications. Nevertheless, the system achieves a FG_1-score of 0.92 and 0.94 when applied to manual transcriptions or transcriptions generated using automatic speech recognition systems, respectively.

References

[1] BERENDES, K.: *Sprache als Schlüsselkompetenz für Bildungsprozesse*. *Forum Logopädie*, 38(2), pp. 8–14, 2024.

[2] NORBURY, C. F., D. GOOCH, C. WRAY, G. BAIRD, T. CHARMAN, E. SIMONOFF, G. VAMVAKAS, and A. PICKLES: *The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study*. *Journal of Child Psychology and Psychiatry*, 57(11), pp. 1247–1257, 2016.

[3] KIESE-HIMMEL, C.: *Früherkennung primärer Sprachentwicklungsstörungen – zunehmende Relevanz durch Änderung der Diagnosekriterien?* *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 65(9), pp. 909–916, 2022.

[4] FINESTACK, L. H., E. ANCEL, H. LEE, K. KUCHLER, and M. KORNELIS: *Five Additional Evidence-Based Principles to Facilitate Grammar Development for Children With Developmental Language Disorder*. *American Journal of Speech-Language Pathology*, 33(2), pp. 552–563, 2024.

[5] CHODOROW, M., M. DICKINSON, R. ISRAEL, and J. TETREAULT: *Problems in Evaluating Grammatical Error Detection Systems*. In M. KAY and C. BOITET (eds.), *Proceedings of COLING 2012*, pp. 611–628. The COLING 2012 Organizing Committee, 2012.

[6] WANG, X., K. EVANINI, Y. QIAN, and M. MULHOLLAND: *Automated Scoring of Spontaneous Speech from Young Learners of English Using Transformers*. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 705–712. 2021.

[7] KNILL, K., M. GALES, K. KYRIAKOPOULOS, A. MALININ, A. RAGNI, Y. WANG, and A. CAINES: *Impact of ASR Performance on Free Speaking Language Assessment*. In *Proceedings INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, pp. 1641–1645. ISCA, 2018.

[8] MORLEY, E., A. E. HALLIN, and B. ROARK: *Data Driven Grammatical Error Detection in Transcripts of Children's Speech*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 980–989. Association for Computational Linguistics, 2014.

[9] HE, Z.: *English Grammar Error Detection Using Recurrent Neural Networks*. *Scientific Programming*, 2021(Scientific Programming for Smart Internet of Things), p. 7058723, 2021.

[10] BELL, S., H. YANNAKOUDAKIS, and M. REI: *Context is Key: Grammatical Error Detection with Contextual Word Representations*. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 103–115. Association for Computational Linguistics, 2019.

[11] LU, Y., M. J. GALES, K. M. KNILL, P. MANAKUL, L. WANG, and Y. WANG: *Impact of ASR Performance on Spoken Grammatical Error Detection*. In *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pp. 1876–1880. ISCA, 2019.

[12] KATINSKAIA, A. and R. YANGARBER: *Assessing Grammatical Correctness in Language Learning*. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 135–146. 2021.

[13] BANNÒ, S. and M. MATASSONI: *Back to grammar: Using grammatical error correction to automatically assess L2 speaking proficiency*. *Speech Communication*, 157, p. 103025, 2024.

[14] HASSANALI, K.-N. and Y. LIU: *Measuring Language Development in Early Childhood Education: A Case Study of Grammar Checking in Child Language Transcripts*. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 87–95. 2011.

[15] RUMBERG, L., C. GEBAUER, H. EHLERT, M. WALLBAUM, U. LÜDTKE, and J. OSTERMANN: *Uncertainty Estimation for Connectionist Temporal Classification Based Automatic Speech Recognition*. In *Proceedings INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, pp. 4583–4587. ISCA, 2023.

[16] NABER, D.: *A Rule-Based Style and Grammar Checker*. Diplomarbeit, 2003.

[17] MIKOLOV, T., K. CHEN, G. CORRADO, and J. DEAN: *Efficient Estimation of Word Representations in Vector Space*. In *International Conference on Learning Representations Workshop*. 2013. `1301.3781`.

[18] RUMBERG, L., C. GEBAUER, H. EHLERT, M. WALLBAUM, L. BORNHOLT, J. OSTERMANN, and U. LÜDTKE: *kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech*. In *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pp. 5160–5164. ISCA, 2022.

[19] BAEVSKI, A., H. ZHOU, A. MOHAMED, and M. AULI: *Wav2vec 2.0: A framework for self-supervised learning of speech representations*. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[20] GRAVES, A., S. FERNANDEZ, F. GOMEZ, and J. SCHMIDHUBER: *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*. In *International Conference on Machine Learning (ICML)*, p. 8. 2006.

[21] KINGMA, D. P. and J. BA: *Adam: A Method for Stochastic Optimization*. In *Proceedings of 3rd International Conference for Learning Representations (ICLR)*. 2015. `1412.6980`.

[22] KAHN, J., V. PRATAP, T. LIKHOMANENKO, Q. XU, A. HANNUN, J. CAI, P. TOMASELLO, A. LEE, E. GRAVE, G. AVIDOV, B. STEINER, V. LIPTCHINSKY, G. SYNNAEVE, and R. COLLOBERT: *Flashlight: Enabling Innovation in Tools for Machine Learning*. 2022. `2201.12465`.

[23] HEAFIELD, K.: *KenLM: Faster and Smaller Language Model Queries*. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 690–696. Association for Computational Linguistics, 2011.

[24] FLACH, P. and M. KULL: *Precision-Recall-Gain Curves: PR Analysis Done Right*. In *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.