Automatically assessing oral narratives of Afrikaans and isiXhosa children

Retief Louw[1], Emma Sharratt[1], Febe de Wet[1], Christiaan Jacobs[1], Annelien Smith[2], Herman Kamper[1]

[1]Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]Speech, Language and Hearing Therapy, Stellenbosch University, South Africa

{retieflouw,emsharratt}@gmail.com

Abstract

Developing narrative and comprehension skills in early childhood is critical for later literacy. However, teachers in large preschool classrooms struggle to accurately identify students who require intervention. We present a system for automatically assessing oral narratives of preschool children in Afrikaans and isiXhosa. The system uses automatic speech recognition followed by a machine learning scoring model to predict narrative and comprehension scores. For scoring predicted transcripts, we compare a linear model to a large language model (LLM). The LLM-based system outperforms the linear model in most cases, but the linear system is competitive despite its simplicity. The LLM-based system is comparable to a human expert in flagging children who require intervention. We lay the foundation for automatic oral assessments in classrooms, giving teachers extra capacity to focus on personalised support for children's learning.

Index Terms: child speech processing, language development, early childhood education, computer-aided language learning

1. Introduction

Oral storytelling and comprehension are essential skills for children, and they are strong predictors for later reading proficiency [1–4]. Assessing early oral narrative ability is therefore crucial for understanding language development and identifying delayed literacy [5–8]. However, in many parts of the world, large classroom sizes pose significant challenges for preschool educators. In such contexts, observational assessments, which depend on teacher intuition, are often inaccurate [9, 10]. On the other hand, one-on-one assessments are effective for diagnosing students' language skills [11]. However, the capacity to do individual assessments is often limited by the time, cost, and logistical challenges involved in scaling across large classrooms [12].

Our goal is to develop a system to support preschool teachers in low-resource language settings by identifying children whose oral language skills are below the level expected at their age. The system should facilitate more consistent and efficient evaluations, even in classrooms with large numbers of learners who are too young for written assessments. We focus on two language communities in South Africa, where literacy is a major problem. Only 20% of 10-year-olds can read for meaning [13] and overcrowded preschool classrooms are a norm.

We specifically present the development of an oral narrative assessment system for Afrikaans- and isiXhosa-speaking children between the ages of four and five. A child is shown a series of pictures, prompting them to tell a story. Afterwards, they need to answer a series of comprehension questions. As illustrated in Figure 1, the system takes the child's speech as input. It should then predict as output a range of finer-grained scores, such as narrative cohesion and comprehension, as well as coarser scores, like a binary value for whether intervention is required.

We develop separate Afrikaans and isiXhosa systems. Our proposed methodology consists of an automatic speech recognition (ASR) model that transcribes the child's speech, as shown in Figure 1(b). The text output from the ASR model is then processed (c) by one of two scoring models: a linear model or a large language model (LLM). For the linear method, simple features like word count and narrative length are extracted, serving as input to a regularised model that predicts the scores. For the LLM-based approach, the ASR output is first automatically translated from Afrikaans or isiXhosa to English and then passed on to an LLM, which is prompted using in-context learning to (d) predict scores.

We find that, with a few exceptions, the LLM performs better than the linear model. Even though the error rates of the ASR models are high, the LLM-based approach is able to identify children requiring intervention with an accuracy of more than 80% for Afrikaans and 64% for isiXhosa. For the same task, a speech therapist achieves 80% for Afrikaans and 70% for isiXhosa in a re-assessment test. These imperfect human scores highlight the challenging and subjective nature of the task. We analyse the effects of ASR and automatic translation, and look at differences between Afrikaans and isiXhosa.

Taken together, our findings indicate that early childhood assessment is a complex task, one that even a human professional finds challenging. Nevertheless, we show the potential of an automatic assessment system to assist, complement, and enhance human assessments in real, low-resource settings.
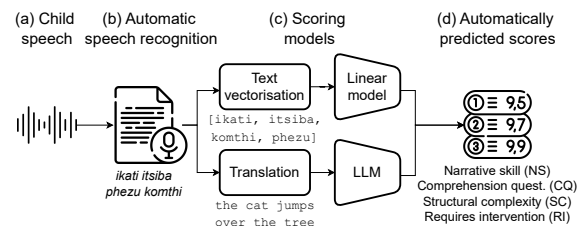

Figure 1: *Our system for automatically scoring narratives told by children aged four to five. An isiXhosa example is shown.*

2. Data

We develop our automatic assessment system using data from a study that involved children aged four to five from lower socioeconomic backgrounds in South Africa [14]. Children were shown a series of six pictures and asked to tell a story in their

Figure 2: *One of the picture sequences used in the MAIN assessment protocol to elicit a narrative from a child.*

home language of either Afrikaans or isiXhosa (Figure 2). The children's speech was recorded using Samsung Galaxy Tab A7 Lite 8.7 tablets and Logitech H111 headsets. The children were evaluated using the multilingual assessment instrument for narratives (MAIN) protocol [15]. This tool is standardised for South African populations, accounting for cultural biases. Assessments were carried out by trained assessors experienced in early childhood development [14].

MAIN assesses three key areas. Narrative skill (NS), ranging from 0 to 16, measures narrative complexity, including inferred goal-directed behaviour (e.g., *the cat wants to catch the butterfly*). Comprehension question (CQ) scores, from 0 to 10, are determined by responses to ten questions assessing understanding of story structure (e.g., *why does the cat jump?*). NS and CQ prediction are regression tasks. Structural complexity (SC) categorises narratives into one of six classes, evaluating how effectively a child conveys goal-directed actions. Higher SC categories indicate complex narratives with goals, attempts, and outcomes (e.g., *he wants to take his ball, he takes it, he plays with it*). SC prediction is a classification task. Children aged four should achieve at least an SC score within the third category [16], so the lowest two SC categories indicate developmental delays. We assign a binary label to these cases that require intervention (RI), and measure performance on this binary classification task.

The data [14] is split into train (roughly 200 children), development (38 children), and test (28 children) sets per language, with no speaker overlap. Each language has approximately 5 hours of active child speech, reflecting its low-resource nature.[1] Differences between Afrikaans and isiXhosa are immediately evident: isiXhosa is morphologically rich with an agglutinative orthography [17], resulting in 3928 unique word types in our data, almost three times the 1140 types in the Afrikaans data.[2]

3. Method

Our automatic narrative assessment system uses an ASR model that feeds into a model that predicts scores, as shown in Figure 1.

3.1. Automatic speech recognition (ASR) models

For each of the languages, we use a corresponding ASR model. These models were developed in previous work, with full details provided in [18]. In short, Whisper-based ASR models [19] are respectively trained on five minutes of transcribed child speech in Afrikaans and isiXhosa. The training data is a subset of the training data described in Section 2. We use this restricted subset for

[1]Only the child's speech is included; the assessor's speech was manually removed. This is an idealised scenario, equivalent to using a perfect diarisation system. Future work will look at a complete system.

[2]isiXhosa and Afrikaans, with 8 million and 7.2 million speakers respectively, are two of South Africa's 12 official languages. isiXhosa is a Southern Bantu language, while Afrikaans is a West Germanic language derived from Dutch. Both languages use the Latin alphabet.

two reasons: (1) this is representative of extreme low-resource settings, and (2) it allows us to see the effect of ASR errors on the downstream score prediction task. The character error rates (CERs) for these models are 28% for Afrikaans and 34% for isiXhosa. isiXhosa exhibits a higher CER than Afrikaans, possibly because Whisper is pretrained on some adult Afrikaans and related languages like Dutch and Flemish, while isiXhosa is not included, nor any language closely related to isiXhosa. We aim to determine the usefulness of these ASR models despite their high error rates.

3.2. Linear model

ASR output is fed into a subsequent scoring model, as shown in Figure 1(c). We consider two options. The first is a linear model. To predict NS and CQ, we use linear regression. We employ logistic regression to predict the SC and RI categories. Linear models are simple, can be efficiently trained with very little data, and are more interpretable than complex models [20].

A range of features are used. We derive speech duration features from the original recordings: duration, mean utterance length, and standard deviation. Using the ASR transcripts, we obtain the total word token and type count as well as counts for individual word types, e.g., there is a feature counting how often *ikati* (cat) occurs. We experimented with lemmatising words, but this did not improve development performance. Our final feature is the Flesch–Kincaid readability score, which quantifies linguistic complexity based on sentence length and syllable count [21,22]. The score was originally designed for English, but can easily be used on other languages and gave benefits in our development experiments.

Other features were also considered in development experiments but did not improve score prediction performance: articulation rate, part-of-speech counts, type-token ratio, sentence complexity, and the Gunning Fog Index [23]. We also explored morphological parsing for isiXhosa, replacing words with morphemes [24], and using n-gram features in both languages, but this did not improve development scores.

While linear models are well-suited for working with limited data, they remain susceptible to overfitting, especially when the number of input features exceeds the number of observations [25], as is the case here. We therefore use lasso regularisation. Lasso adds an L1 penalty $\alpha||\mathbf{w}||_1$ to a linear loss. This shrinks coefficients to zero, thereby selecting relevant features and excluding non-informative ones [26,27]. We tune α on development data. The result is a major reduction in the number of features. For instance, in the Afrikaans model for predicting NS, more than 1140 weights are zeroed leaving only 55 features. We train distinct lasso-regularised linear models for predicting each assessment score in each language.

3.3. Large language model (LLM)

Our second approach uses an LLM with in-context learning to score the children's transcripts, as illustrated in Figure 1(c)-bottom. Despite being multilingual and understanding Afrikaans and isiXhosa, LLMs are primarily trained on English text. To address this, we first translate the Afrikaans and isiXhosa ASR output to English using the Google Translate API. In Section 5 we give results with and without translation.

In-context learning allows pre-trained LLMs to generate task-specific outputs by providing them with examples of ideal input-output pairs. Literature about on in-context learning recommends several techniques of prompt engineering to craft the ideal prompt [28]. The prompt format shown in Figure 3 consis-

Figure 3: *The structure of how we prompt LLMs using in-context learning for scoring oral narratives from children.*

Table 1: *Test set performance of the automated narrative scoring system. We report R^2 for narrative skills (NS) and comprehension (CQ), Cohen's kappa (κ) for structural complexity (SC), and accuracy for children requiring intervention (RI).*

	NS [R^2]		CQ [R^2]		SC [κ]		RI [%]	
	Oracle	ASR	Oracle	ASR	Oracle	ASR	Oracle	ASR
Afrikaans:								
Linear	0.27	0.20	0.28	0.03	0.18	0.12	68	61
LLM	0.67	**0.52**	0.29	**0.20**	0.13	**0.13**	89	**89**
Human	0.82	-	0.79	-	0.25	-	80	-
isiXhosa:								
Linear	0.49	0.32	0.38	**0.10**	0.01	0.19	43	54
LLM	0.58	**0.52**	0.06	0.08	0.39	**0.33**	64	**64**
Human	0.17	-	0.85	-	0.50	-	70	-

tently yielded the best results and is used for all the final results. The prompt includes training examples (each comprising a translation of the ASR output and the three corresponding teacher scores); the prompt instruction; and the unseen translated story.

We conducted development experiments to identify the optimal in-context learning approach. We compared three language models: Gemini-1.5-Flash-8B [29], GPT-4 [30], and Llama-3.3-70B [31]. Gemini-1.5-Flash-8B consistently achieved the best development performance, so it was selected as the LLM for the system. We also investigated several few-shot prompting strategies, such as target-based grouping and balanced sampling [32]. Performance consistently improved with more examples, so we use the entire training set in each prompt for the final system. We also considered adding explanations of how scores are qualitatively determined by the teachers, but this did not improve development performance.

In a study similar to ours, Baumann et al. [33] investigated automatic scoring of German children's MAIN narratives. They fed a German BERT model with the text from human-transcribed narratives. Aggregated embeddings were then given as input to linear models, predicting binary scores for each subcategory of NS. Although they achieved usable results for two NS subcategories, they encountered challenges with others. Our study differs in several regards: we employ an LLM for direct scoring instead of training a classifier, use ASR outputs instead of transcriptions, focus on two low-resource languages instead of a high-resource language, and incorporate additional MAIN scores (CQ, SC) for holistic assessments. Another related study is that of Veeramani et al. [34], who also used a BERT-based score prediction model, but fed it with actual ASR predictions. However, they focused on narratives from English children between the ages of nine and thirteen, while we consider younger children (between four and five) speaking low-resource languages.

4. Experimental setup

For assessing performance, we use R^2 for the narrative (NS) and comprehension (CQ) score regression tasks, Cohen's κ for structural complexity (SC) classification, and accuracy for identifying children requiring intervention (RI).

R^2 quantifies a model's fit by measuring the explained variance in the target variable [35, 36], with 0 for mean prediction and 1 for perfect predictions [35]. Cohen's κ quantifies agreement beyond chance [37] and assesses model alignment with ground truth in machine learning [38]. Ranging from -1 to 1,

κ indicates perfect agreement at 1 and chance-level at 0. Given that SC categories are ordinal (e.g., four is closer to five than to one), we use linear-weighted Cohen's κ, applying a linear penalty for diverging categories [39].

To contextualise results, we compare them with human expert predictions: a speech therapist from the original study [14] manually re-assessed ten samples per language. The resulting R^2, κ, and accuracy scores establish human expert performance for the complex task of child oral narrative assessment, which is inherently subjective [40].

5. Experimental results

Our goal is to determine the most effective automatic system for assessing narratives from preschool children. Table 1 presents the test results for NS, CQ, SC, and RI, comparing the predictive performance of the linear models and LLMs. The *oracle* columns show idealised results when ground truth transcriptions are used in training and evaluating the scoring models—these serve as a reference. The *ASR* columns use automatically produced transcriptions.

LLM vs linear. We start by comparing the linear models to the LLMs for the systems using ASR (oracle results are discussed below). Given its simplicity, the linear-based system performs competitively. But the LLM achieves higher scores in all cases, except for CQ on isiXhosa. As a concrete example, the LLM attains an R^2 of 0.52 when predicting NS in Afrikaans, whereas the linear model scores only 0.20. On Afrikaans, the LLM outperforms the logistic regression model by almost 30% absolute in predicting whether intervention is required (RI), while on isiXhosa, its RI accuracy is 10% better.

Table 2 reports additional recall and F1 scores for RI prediction, supporting our findings. We prioritise recall, as the system's goal is to identify children most in need. The LLM consistently outperforms the linear model across all cases in Table 2, effectively identifying children with learning difficulties. Table 2 also presents the mean absolute error for NS (0 to 16) and CQ (0 to 10) predictions in the original units. The LLM consistently achieves lower error than the linear model for both NS and CQ. Overall, the LLM is superior for automatically identifying children requiring academic support.

Automatic system vs human. We compare our system to a human expert from the original study [14], a speech and language therapist, who re-assessed a subset of the original

Table 2: *Test set results in terms of mean absolute error (MAE) for the narrative skill (NS) and comprehension questions (CQ) scores, and recall and F1 (%) for predicting whether children require intervention (RI) using ASR outputs.*

	MAE		RI	
	NS	CQ	Recall	F1
Afrikaans:				
Linear	1.21	1.91	50	33
LLM	**0.90**	**1.45**	**96**	**94**
Human	0.50	0.90	89	89
isiXhosa:				
Linear	1.34	2.08	60	55
LLM	**0.96**	**1.64**	**75**	**78**
Human	0.80	0.38	100	77

Table 3: *Development set performance comparing LLMs with and without automatic translation to English before scoring.*

	Afrikaans			isiXhosa		
Model	NS $[R^2]$	CQ $[R^2]$	SC $[\kappa]$	NS $[R^2]$	CQ $[R^2]$	SC $[\kappa]$
LLM without transl.	0.18	0.33	0.01	0.17	0.18	0.12
LLM with transl.	**0.23**	**0.41**	**0.08**	**0.30**	**0.19**	**0.13**

data. In Table 1, the human achieves higher R^2 scores for NS in Afrikaans and outperforms the models in both languages when predicting SC and CQ. For isiXhosa RI, the LLM (64%) approaches human performance (70%). Interestingly, the LLM surpasses human scores in two instances: Afrikaans RI (89% vs 80%) and isiXhosa NS R^2 (0.52 vs 0.17). We interpret these superior model scores with caution, as they reflect the limitations and subjectivity in the ground truth data. Human assessment scores varied across all metrics, revealing potential biases and challenges in teacher-generated evaluations, emphasising the complexity of accurately assessing children.

The effect of imperfect ASR. To assess the impact of the ASR model's high CER (Section 2), we compare the oracle and ASR scores in Table 1. Despite relatively high error rates, the expected drop in ASR scores is not substantial. Notably, in some cases, ASR scores exceed oracle scores. For example, when classifying children that should be flagged, the isiXhosa linear model's score of 54% is better than the oracle score of 43%. These results show that even noisy ASR predictions can still be meaningful for downstream assessments.

Afrikaans vs isiXhosa. We find that the LLM generally performs worse on RI for isiXhosa (64%) compared to Afrikaans (89%). This discrepancy cannot be attributed solely to the isiXhosa ASR system's poorer CER, or the translation step, as the oracle RI results are also worse (64% vs 89%). Future work may explore whether the lower isiXhosa performance stems from inherent properties of the data itself, or from limitations in the automatic English-to-isiXhosa translation process. For most of the other metrics, results are similar between the two languages.

The effect of translation. The LLM system only sees English text due to the intermediate translation step. Table 3 shows that translating the text from low-resource languages to English improves the LLM performance regardless of the language or score. The improvement is particularly evident for Afrikaans children, where the κ increases from 0.01 to 0.08 with translation. The Google Translate API's apparent ability to handle noisy ASR predictions, including potentially changing words and fixing errors like stuttering, likely contributes to the improved performance.

We performed these translation experiments on the development set. There is a substantial difference in R^2 and κ between the final row of Table 3 giving development-set scores and the LLM rows in Table 1 giving test-set scores. For example, the LLM achieves an R^2 of 0.41 for CQ on the development set, but only 0.20 on the test set. This difference suggests that the

specific children included in each set also have a considerable influence on the evaluations.

Linear model interpretability. Although not always performing as well as the LLM, the linear model's weights can be inspected. Lasso regularisation pushes most weights to zero and we can examine the largest weights of the remaining features. For predicting NS and CQ in Afrikaans, the largest weights are associated with the features counting words like *gelukkig* (happy), *eet* (eat) and *kwaad* (angry). For isiXhosa, they include *kabuhlungu* (painfully), *ithathe* (take it) and *ilambile* (hungry). For both languages, the largest weights are associated with a combination of emotive, action-related and descriptive word types. In parallel work [41], we explicitly consider how the linear model can be used to gain deeper interpretable insights into the acoustic, lexical and semantic patterns of oral narratives from preschoolers requiring intervention.

In summary, our results show that an LLM-based system outperforms a system that uses a linear model to predict most assessment scores. Giving results in the original units, for predicting the NS score out of 16, our best system achieves a mean absolute error of 0.90 for Afrikaans and 0.96 for isiXhosa, respectively. While a human expert achieves higher scores in most metrics, the LLM shows comparable and even superior performance in the final intervention classification task, suggesting limitations in the human-generated ground truth data. Despite imperfect ASR transcripts, the system maintains performance, with translation further improving LLM results for both languages.

6. Conclusion

We introduced an automated system to assist early childhood educators in assessing children's oral narrative and comprehension skills in two low-resource languages. We specifically combined automatic speech recognition (ASR) with two predictive models. We developed linear models and prompted large language models (LLMs) to predict several development scores from stories told by Afrikaans and isiXhosa children. The LLM-based system aligns more closely with human assessment than the system that uses linear models. Performance was improved by first translating the ASR transcripts before feeding the result to an LLM. In some cases, a human expert was a poor judge when reassessing some of the stories; this revealed the complexity and subjectivity of scoring pre-preschoolers' oral narratives, emphasising the need for further refinement to address the nuanced aspects of automatically assessing young children's spoken language. Nevertheless, we believe that our system will be usable in assisting teachers to identify cases where intervention is most required.

Future work will look at even more approaches for prompting the LLM and end-to-end training methods that go directly from speech to a score.

7. References

[1] D. Hayward, P. Schneider, and R. B. Gillam, "Age and task-related effects on young children's understanding of a complex picture story," *Alberta Journal of Educational Research*, vol. 55, no. 1, p. 54, 2009.

[2] Y.-D. Chiu, "The simple view of reading across development: Prediction of grade 3 reading comprehension from prekindergarten skills," *Remedial and Special Education*, vol. 39, no. 5, pp. 289–303, 2018.

[3] H. N. Hjetland, E. I. Brinchmann, R. Scherer, C. Hulme, and M. Melby-Lervåg, "Preschool pathways to reading comprehension: A systematic meta-analytic review," *Educational Research Review*, vol. 30, no. 3, p. 100323, 2020.

[4] S. Babayiğit, S. Roulstone, and Y. Wren, "Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study," *British Journal of Educational Psychology*, vol. 91, no. 1, pp. 148–168, 2021.

[5] D. K. Dickinson, A. McCabe, L. Anastasopoulos, E. S. Peisner-Feinberg, and M. D. Poe, "The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children." *American Psychological Association*, vol. 95, no. 3, p. 465, 2003.

[6] A. Schick and G. Melzi, "The development of children's oral narratives across contexts," *Early Education*, vol. 22, no. 3, pp. 293–317, 2010.

[7] E. Reese, D. Leyva, A. Sparks, and W. Grolnick, "Maternal elaborative reminiscing increases low-income children's narrative skills relative to dialogic reading," *Early Education and Development*, vol. 21, no. 3, pp. 318–342, 2010.

[8] J. V. Oakhill and K. Cain, "The precursors of reading ability in young readers: Evidence from a four-year longitudinal study," *Scientific Studies of Reading*, vol. 16, no. 2, pp. 91–121, 2012.

[9] M. D. Shermis and J. Burstein, Eds., *Automated Essay Scoring: A Cross-disciplinary Perspective*. Lawrence Erlbaum Associates, 2003.

[10] R. Mozer, L. Miratrix, J. E. Relyea, and J. S. Kim, "Combining human and automated scoring methods in experimental assessments of writing," *Journal of Educational and Behavioral Statistics*, vol. 49, no. 5, pp. 780–816, 2023.

[11] J. Chiquito, "Effects of one-on-one support for English language learners through collaboration, compassion, and engagement," Master's thesis, California State University, Monterey Bay, 2023.

[12] C. E. Snow and T. J. Matthews, "Reading and language in the early grades," *The Future of Children*, vol. 26, no. 2, pp. 57–74, 2016.

[13] K. Roux, S. van Staden, and M. Tshele, "PIRLS 2021: South african preliminary highlights report," Department of Basic Education Pretoria, Tech. Rep., 2023.

[14] A. Smith, "How narrative comprehension and production are intertwined with early learning indicators." Master's thesis, Stellenbosch University, 2023.

[15] N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Välimaa, U. Bohnacker, and J. Walters, "Main: Multilingual assessment instrument for narratives–revised," *ZAS Papers in Linguistics*, vol. 63, p. 20, 2019.

[16] C. E. Westby, "Assessing and remediating text comprehension problems," *Language and Reading Disabilities*, vol. 3, pp. 163–224, 2005.

[17] L. Mzamo, A. Helberg, and S. Bosch, "Introducing XGL-a lexicalised probabilistic graphical lemmatiser for isiXhosa," in *PRASA-RobMech*, 2015.

[18] C. Jacobs, A. Smith, D. Klop, O. Klejch, F. de Wet, and H. Kamper, "Speech recognition for automatically assessing Afrikaans and isiXhosa preschool oral narratives," in *ICASSP*, 2024.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[20] R. M. Dawes, "The robust beauty of improper linear models in decision making," *American Psychologist*, vol. 34, no. 7, pp. 571–582, 1979.

[21] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[22] JP. Kincaid, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Institute for Simulation and Training, Tech. Rep., 1975.

[23] M. Kajzer-Wietrzny and A. Chmiel, "Into B or not into B? the limited impact of interpreting direction on target text fluency and complexity," *SKASE Journal of Translation and Interpretation*, vol. 16, no. 2, pp. 23–43, 2023.

[24] T. Moeng, S. Reay, A. Daniels, and J. Buys, "Canonical and surface morphological segmentation for Nguni languages," in *Southern African Conference for Artificial Intelligence Research*, 2021.

[25] J. Ranstam and J. A. Cook, "LASSO regression," *British Journal of Surgery*, vol. 105, no. 10, pp. 1348–1348, 2018.

[26] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Springer Nature, 2023.

[27] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. Pearson Education, Inc., 2025.

[28] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu *et al.*, "A survey on in-context learning," in *EMNLP*, 2024.

[29] M. Thelwall, "Is Google Gemini better than ChatGPT at evaluating research quality?" *Journal of Data and Information Science*, vol. 10, no. 2, pp. 1–5, 2025.

[30] T. C. McFayden, S. Bristol, O. Putnam, and C. Harrop, "Chat-GPT: Artificial intelligence as a potential tool for parents seeking information about autism," *Cyberpsychology, Behavior, and Social Networking*, vol. 27, no. 2, pp. 135–148, 2024.

[31] M. Ibrahim, "Fine-grained language-based reliability detection in Spanish new with fine-tuned Llama-3 model," in *SEPLN*, 2024.

[32] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *ICLR*, 2023.

[33] T. Baumann, K. Eller, and N. Gagarina, "BERT-based annotation of oral texts elicited via multilingual assessment instrument for narratives," in *ACL*, 2024.

[34] H. Veeramani, N. Balaji Shankar, A. Johnson, and A. Alwan, "Towards automatically assessing children's oral picture description tasks," in *SLaTE*, 2023.

[35] L. J. Saunders, R. A. Russell, and D. P. Crabb, "The coefficient of determination: What determines a useful R^2 statistic?" *Investigative Opthalmology & Visual Science*, vol. 53, no. 11, pp. 6830–6832, 2012.

[36] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Statistical learning," in *An introduction to statistical learning: With applications in Python*. Springer, 2023.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[38] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," *Statistics*, vol. 1050, p. 13, 2020.

[39] K. P. Nelson and D. Edwards, "Measures of agreement between many raters for ordinal classifications," *Statistics in Medicine*, vol. 34, no. 23, pp. 3116–3132, 2015.

[40] L. A. Shepard, "The challenges of assessing young children appropriately," *The Phi Delta Kappan*, vol. 76, no. 3, pp. 206–212, 1994.

[41] E. Sharratt, A. Smith, R. Louw, D. Klop, F. de Wet, and H. Kamper, "Feature-based analysis of oral narratives from Afrikaans and isiXhosa children," in *SLaTE*, 2025.