Feature-based analysis of oral narratives from Afrikaans and isiXhosa children

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

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

emsharratt@gmail.com

Abstract

Oral narrative skills are strong predictors of later literacy development. This study examines the features of oral narratives from children who were identified by experts as requiring intervention. Using simple machine learning methods, we analyse recorded stories from four- and five-year-old Afrikaans- and isiXhosa-speaking children. Consistent with prior research, we identify lexical diversity (unique words) and length-based features (mean utterance length) as indicators of typical development, but features like articulation rate prove less informative. Despite cross-linguistic variation in part-of-speech patterns, the use of specific verbs and auxiliaries associated with goal-directed storytelling is correlated with a reduced likelihood of requiring intervention. Our analysis of two linguistically distinct languages reveals both language-specific and shared predictors of narrative proficiency, with implications for early assessment in multilingual contexts.

Index Terms: oral narratives, child speech processing, automatic assessment, literacy, feature importance, model interpretability

1. Introduction

Even before children learn to read, the ability to tell and understand stories is a key developmental skill. Oral storytelling predicts later reading proficiency [1–3] and narrative assessment can therefore help identify delayed development early on [4–8]. Yet, in many parts of the world large classroom sizes make narrative assessment impossible. Observational assessments, where teachers rely on intuition, are often inaccurate [9–11]. Our goal is to identify features that characterise age-appropriate oral narratives among preschool children in low-resource language settings. These findings aim to inform the development of more effective oral narrative assessment methods to support early identification of language difficulties in pre-literate children.

We focus on Afrikaans- and isiXhosa-speaking children aged four to five from low-income communities in South Africa, where large class sizes and low literacy rates are widespread [12]. We use a dataset where children told stories based on a picture sequence (Figure 1) and answered comprehension questions. Education experts then assessed whether intervention was needed. We do a text-based analysis of the narrative transcripts, using logistic regression to predict which children may need support. In contrast to our parallel work [13], our goal here is not an accurate system for automatically identifying at-risk children, but rather to characterise the types of features that are indicative of risk using an interpretable machine learning approach. Using permutation feature importance (PFI) [14] and exploratory data analysis, we perform quantitative analyses to compare the informativeness of different features. We then compare these quantitative findings to a qualitative analysis based on traditional measures of storytelling development from speech therapy.
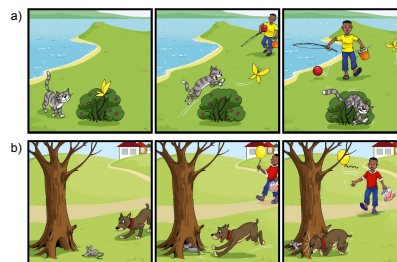

Figure 1: *Three images from the six-picture cat (a) and dog (b) stories used in the MAIN protocol to elicit child narratives.*

We consider a range of features. As in other studies [15, 16], we find that indicative features include those measuring lexical diversity (unique words), productivity (mean utterance length) and syntactic complexity (measured using a readability index). In contrast to other studies [17, 18], we find that speech production features such as articulation rate are of little importance. We also consider grammatical-level indicators (part-of-speech counts) and specific keywords. Although keywords are unique to each language, we find that specific verbs and auxiliaries linked to goal-directed sentences are associated with a lower likelihood of intervention in both Afrikaans and isiXhosa.

In contrast to other studies that look at single languages in isolation [10, 11, 19], our contribution is an analysis of similar oral narratives from two distinct languages. This allows us to identify features that are important to individual communities and also to identify general cross-lingual characteristics of children's oral narratives. We hope that this quantitative study will inform further work on oral narrative assessments.

2. Data

We use data from a speech and language therapy study that involved children aged four to five from lower socioeconomic backgrounds in South Africa [20, 21]. Each child was shown a six-picture sequence and asked to tell a story in their home language of either Afrikaans or isiXhosa.[1] One of two story sequences was used: a cat story or a dog story (Figure 1). Comprehension was subsequently assessed through questions about the story content. The children were evaluated using the multilingual assessment instrument for narratives (MAIN) [22]. This tool was standardised for South African populations, accounting for cultural biases [23]. Assessments were conducted by trained

---

[1]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.

assessors experienced in early childhood development. Based on the combination of stories and comprehension answers, experts finally assigned a binary label for whether intervention is required or not [24]. We call this label *requires intervention* (RI), with a value of 1 if intervention is needed and 0 if not. Children needing additional language support typically struggle to formulate narratives with clear goals, attempts and outcomes, e.g., *She wants it. She jumps. She misses it.*

We divide the data into training (approximately 200 children), development (38 children) and test (28 children) sets for each language, with no speaker overlap between sets. Cat and dog stories appear in all sets. The children's speech was recorded using Samsung Galaxy Tab A7 Lite 8.7 tablets and Logitech H111 headsets. Fieldworkers provided an audio and transcription file for each child. Transcriptions were manually aligned to audio using the Praat toolkit [25]. This resulted in a total of about 5 hours of active child speech for each language. The data shows clear cross-linguistic differences. For example, isiXhosa has an agglutinative orthography, typical of Bantu languages. Its morphological richness leads to a larger vocabulary [26] with over 3,900 unique word types in the isiXhosa data compared to less than 1,150 in the Afrikaans transcriptions.

3. Methodology

We aim to identify which narrative features are most associated with a child requiring intervention. To do this, we train a linear model to predict intervention and use statistical methods to assess feature importance.

3.1. Logistic regression

As our base machine learning model, we use logistic regression to model the combined influence of multiple features, helping to control for potential confounding effects. While linear models are well-suited for working with limited data, they remain susceptible to overfitting, especially in cases such as ours where there are only 200 training examples. We address this in two ways. First, we use `scikit-learn`'s `LIBLINEAR` solver with L2 regularisation [27], recommended for smaller datasets.

Second, instead of training a single model on all features at once, we group specific feature types. Each group is designed to show the relative importance of particular features to one another. For example, we want to see the relative importance of verb count compared to nouns, so these are placed in one group. We consider three groups: verbal language proficiency features, grammatical features and keywords per language. Three distinct logistic regression models are therefore trained.

Because the different features use different scales, we cannot use the weights of the resulting logistic regression model to measure feature importance directly. One remedy would be to normalise the features beforehand, but this resulted in substantially worse development performance. We therefore retain the raw feature values and turn to the following approach.

3.2. Permutation feature importance (PFI)

Permutation feature importance (PFI) quantifies how much each feature contributes to a model's performance by measuring the drop in accuracy when that feature is randomly shuffled. A greater drop indicates greater feature importance [28]. We use `scikit-learn`'s implementation [27], with balanced accuracy as the evaluation metric (Sec. 3.3). First, the model's baseline performance s is computed on the original dataset $\mathcal{D}$. Each feature j is then considered in turn. Column j in $\mathcal{D}$ is shuffled

to produce a corrupted dataset $\tilde{\mathcal{D}}_{r,j}$. The performance $s_{r,j}$ on this dataset is calculated for repetition r and feature j. This is repeated R times. Finally, the importance of feature j is obtained as the average drop in score compared to using the full model: $s - \frac{1}{R} \sum_{r=1}^{R} s_{r,j}$. We use $R = 100$ repetitions.

We visualise feature importance using permutation plots. E.g., in Figure 2, the y-axis lists features and the x-axis measures importance, giving the drop in balanced accuracy across R repetitions when corrupting the feature. While PFI quantifies importance, it does not show whether higher or lower values increase the likelihood of predicting intervention. I.e., a feature may have a large impact on accuracy (high x value) without revealing the direction of influence. To address this, we use colour: red for positive model coefficients (predicts intervention), blue for negative (predicts no intervention). For example, Figure 2a shows *unique words* in blue, meaning more unique words are linked to typical development. In contrast, *articulation rate* is red, suggesting faster speech predicts intervention. However, its low x value also indicates limited importance.

3.3. Evaluation metrics

Children requiring intervention are labelled as positive, RI $= 1$, and those who do not need intervention are labelled as negative, RI $= 0$. A true positive is therefore a correct prediction that intervention is required. We use balanced accuracy as our primary evaluation metric: $\frac{1}{2} \left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}} \right)$. Balanced accuracy accounts for class imbalance. Furthermore, it explicitly incorporates true negative predictions [29] which are not accounted for in metrics like precision, recall and F1. We also report F1 for a full picture.

4. Feature analysis and results

We use three models for our feature analysis, each using a distinct feature group: verbal language proficiency features, grammatical features and frequently occurring keywords. Before we turn to our main question of the relative importance of different features within each group, we report base model performance.

4.1. Base model performance

Table 1 summarises the base systems' accuracy on training and development sets for both languages. We also show the class distribution, e.g., in the Afrikaans training set, positive RI labels make up 36% of the targets. We perform our analyses on the training data. We can do this because our goal is not to evaluate absolute performance but rather to understand what the data reveals about feature relevance. Moreover, because it is a

Table 1: *F1 and balanced accuracy (%) for base models on train and development sets, with true RI proportions per set.*

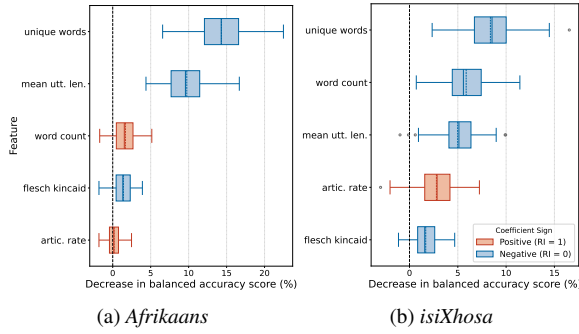| | Afrikaans | | | | isiXhosa | | | |
| | Train (36% RI) | | Dev (32% RI) | | Train (47% RI) | | Dev (53% RI) | |
Model	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
Proficiency	64	71	64	74	69	69	62	61
Grammatical	62	69	52	64	71	71	61	64
Keywords	70	76	55	67	79	79	73	71
All features	77	82	55	67	84	84	61	63

(a) *Afrikaans* (b) *isiXhosa*

Figure 2: *PFI for verbal language proficiency features, showing the drop in balanced accuracy when a given feature is corrupted.*

(a) *Unique words* (b) *Mean utt. length* (c) *Articulation rate*

Figure 3: *Distribution of unique word count, mean utterance length and articulation rate, grouped by RI and language.*

larger set, it yields more robust results. We use the training set's balanced accuracies in Table 1 as the reference scores (Sec. 3.2) for assessing feature importance in the analyses that follow.

4.2. Which verbal proficiency features are important?

Setup: We start by considering the relative importance of verbal language proficiency features. These include the total number of word tokens and types extracted from the normalised transcriptions; mean utterance length (seconds) and articulation rate (characters per second)[2] derived from the recordings; and the Flesch–Kincaid readability score[3] which estimates linguistic complexity based on sentence length and syllable count [30, 31].

Quantitative analysis: As seen in the top bar of Figure 2, unique word count has the highest feature importance for both languages, suggesting a larger vocabulary corresponds with predicting satisfactory development (non-RI). Looking at Figure 2, we see that mean utterance length is the third most important proficiency feature for isiXhosa and the second for Afrikaans. In both languages, longer utterances are linked to typical development. Turning to the word count feature, Figure 2b shows that the number of words produced is a key feature in isiXhosa, with higher counts associated with non-RI classification (blue). Surprisingly, for the Afrikaans model, higher word count is associated with intervention (red). But, its low feature importance suggests that it plays a limited role in classifying whether intervention is required in Afrikaans. Finally, articulation rate is of low importance for both languages, with an unexpected trend that faster speech is associated with requiring intervention.

While PFI shows the importance of each feature in the context of all others, single-feature box plots help visualise how individual features relate to the target variable (RI) in isolation [32]. Figure 3 shows the median, quartiles, outliers and means (dotted lines) for selected features, grouped by intervention status: red (left) for RI and blue (right) for non-RI. The trends in the box plots reinforce the findings of the PFI plots for (a) unique words and (b) mean utterance length, with higher values linked with typical development. These plots also give insight into the surprising articulation rate results (c). Among isiXhosa-speaking children, both RI and non-RI children have similar articulation rates. In contrast, for Afrikaans, non-RI children (blue) have a higher median and maximum articulation rate, but the variance among this group is very large, spanning that of the RI group.
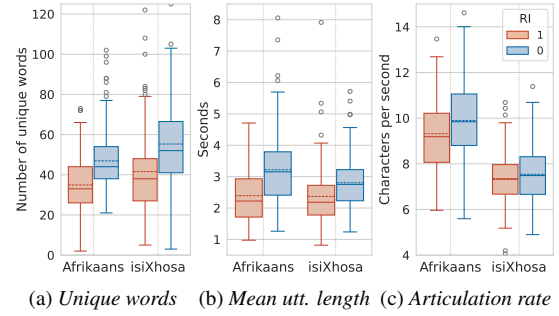
Qualitative analysis: For both languages, unique word count is the strongest predictor of satisfactory development. This aligns with established findings [16], as unique word count measures lexical diversity and reflects a larger vocabulary, which is associated with more complex narratives. Similarly, mean utterance length is an important feature in both models. Utterance length captures syntactic complexity, a microstructural feature related to overall narrative macrostructure. Children requiring intervention typically produce descriptive and action sequences marked by short, simple sentences, often limited to single words or phrases. Consequently, shorter utterances are consistently associated with not meeting age-level expectations. Surprisingly, a higher articulation rate is associated with a higher probability of requiring intervention for both groups. This may suggest that children who produced shorter narratives, often consisting of words, phrases, or simple sentences, spoke at a faster rate. In contrast, more complex narratives, involving more sophisticated syntax, may require a slower speech rate to allow for greater planning and linguistic processing.

4.3. Which grammatical features are important?

Setup: Next we consider the relative importance of grammatical features [33]. E.g., are verbs or nouns more important in determining whether a child requires intervention? Concretely, we derive part-of-speech (POS) counts for both the Afrikaans and isiXhosa data using ctextcore,[4] an open-source Python package that provides corpus analysis tools for ten South African languages using universal POS tags [34]. We select a subset of seven POS tags that apply to both languages. These include counts of verbs, nouns, pronouns, adverbs and adjectives. Additionally, we include auxiliary (helping) verbs and particles. Afrikaans particles include *te* used for introducing an infinitive, the negation particle *nie* and the genitival particle *se*.

Quantitative analysis: Figure 4 presents the PFI analysis while Figure 5 shows box plots for individual features. For isiXhosa, verbs are the most important, as seen in Figure 4b. This is mirrored in Figure 5a, with children who need intervention using fewer verbs. Surprisingly, verbs do not demonstrate much importance in the Afrikaans model; instead, auxiliary verbs stand out in Figure 4a, as is also confirmed in Figure 5b. Although particles are important in both languages in Figure 4, Figure 5c shows that they appear infrequently in the Afrikaans group, with most transcripts containing fewer than three particles. Finally, it is interesting that using more adjectives and adverbs is associated with intervention predictions in both languages.

---

[2]Both languages have regular grapheme-to-phoneme mappings, making characters per second a good proxy for phonemes per second.

[3]In development experiments, we also looked at other readability indices, but Flesch–Kincaid gave the most reliable results.
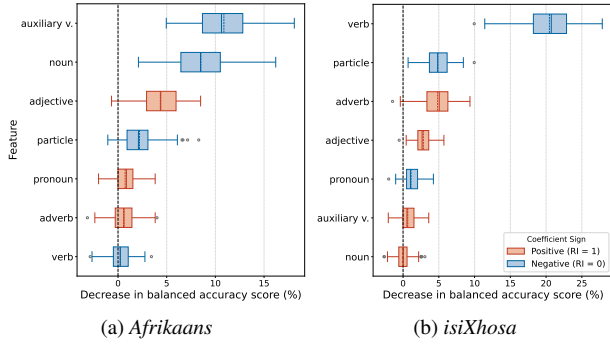
[4]https://pypi.org/project/ctextcore/

(a) *Afrikaans*

(b) *isiXhosa*

Figure 4: *PFI for grammatical features, showing the drop in balanced accuracy when a given feature is corrupted.*

(a) *Afrikaans*

(b) *isiXhosa*

Figure 6: *PFI for specific keywords, showing the drop in balanced accuracy when a given feature is corrupted.*

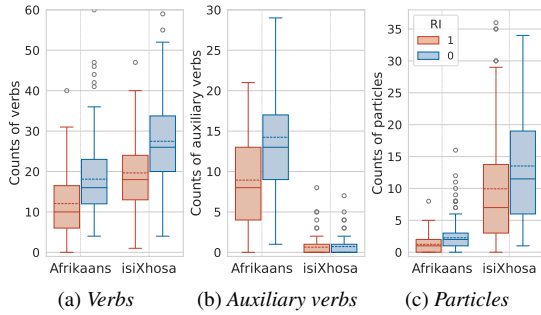(a) *Verbs* (b) *Auxiliary verbs* (c) *Particles*

Figure 5: *Distribution of verbs, auxiliaries and particles in the text, grouped by whether intervention is required and language.*

Qualitative analysis: In general, verbs, adjectives, adverbs, auxiliary verbs and particles contribute to the elaboration of noun and verb phrases, indicating higher syntactic complexity and therefore more advanced language skills (or less need for intervention). The predictive role of verbs and particles in isiXhosa, and of auxiliaries and particles in Afrikaans, is linguistically plausible for identifying non-RI children. Other patterns are more difficult to interpret.

4.4. Are specific keywords important?

Setup: The last model consists of counts of individual keywords, with a control feature to account for differences between the cat and dog stories (Sec. 2). We start with the top twenty most common words from each language. We then use L1 regularisation to select ten keywords from each language. L1 regularisation has the effect of shrinking coefficients to zero, thereby selecting relevant features and excluding non-informative ones [35, 36]. We explored morphological parsing for isiXhosa, replacing words with morphemes [37], and using n-gram features in both languages, but this did not improve development performance.

Quantitative analysis: As shown in Figure 6a, the word *toe* (then) has the highest feature importance in Afrikaans and is associated with the non-RI group (blue). The pronoun *sy* (his), verb *eet* (eat) and auxiliary verb *wil* (want) also contribute meaningfully, with higher counts associated with a non-RI prediction. Among the isiXhosa features, the auxiliary verb *ifuna* (want) and pronouns *yona* (it) and *yakhe* (his) are important features. It is interesting that the Afrikaans word *wil* and the isiXhosa word *ifuna* are important in both cases as they are both translations of the English word *want*. Nouns such as *hond* (dog) in Afrikaans and *ifish* (fish), *inja* (dog) and *ibhaloni* (balloon) in isiXhosa all
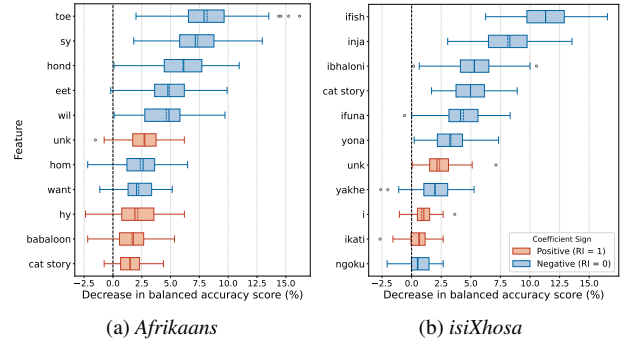
demonstrate importance. These nouns reflect key story elements, and their greater presence may signal longer, more developed narratives. In Afrikaans (Figure 6a), *babaloon* has a positive coefficient (red), so the more it is used the higher the intervention prediction. The word *babaloon* may occur more frequently in the intervention group because it is a mispronunciation of the Afrikaans word *ballon* (balloon). Similarly, we notice that the feature *unk* (unknown), used to label unintelligible speech, is positively associated with intervention for both languages. This aligns with expectations, as unclear speech can be a sign of delays in oral narrative development.

Qualitative analysis: The identification of *toe* as a strong predictor in the Afrikaans model is well-founded; this adverb conveys both temporal progression and causality, which are typical of more advanced narrative structure. Similarly, recognising *wil* and *ifuna* as predictors is reasonable, as these auxiliary verbs are often used in goal-oriented statements that reflect more structurally complex narratives. The pronouns *sy* in Afrikaans and *yakhe* in isiXhosa may contribute as predictors because they elaborate noun phrases, which tend to occur in more syntactically complex constructions. The remaining patterns are less straightforward to interpret qualitatively.

5. Summary and conclusion

We have looked at the relative importance of verbal language proficiency, grammatical and keyword features in oral narratives from four- and five-year-olds. In both our Afrikaans and isiXhosa data, lexical diversity and utterance length distinguished children needing intervention from those who do not. Certain language-specific verbs, auxiliaries and nouns linked to core story elements were indicative of typical development, while unintelligible speech and mispronunciations were associated with requiring intervention. These findings are consistent with prior qualitative linguistic analyses reported in the literature. However, some results challenged expectations: speech production features such as articulation rate and grammatical elements like adjectives and adverbs showed low feature importance. Rather than providing clear answers, these results raise important questions about the role of these features in assessing early narrative skills. Further work by educators and researchers is therefore required to clarify these patterns. In summary, our findings highlight the complexities of assessing early language development, and in particular identifying the salient features linked to the need for language intervention. With continued investigation, we hope these insights can be translated into practical tools to support assessment practices in preschool classrooms.

6. Acknowledgements

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] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] N. Gardner-Neblett and I. Iruka, "Oral narrative skills: Explaining the language-emergent literacy link by race/ethnicity and SES," *Developmental Psychology*, vol. 51, no. 7, p. 889, 2015.

[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, "Bridging human and machine scoring in experimental assessments of writing: Tools, tips, and lessons learned from a field trial in education," in *SREE*, 2021.

[11] ——, "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.

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

[13] R. Louw, E. Sharratt, F. de Wet, C. Jacobs, A. Smith, and H. Kamper, "Automatically assessing oral narratives of Afrikaans and isiXhosa children," in *SLaTE*, 2025.

[14] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[15] P. Uccelli and M. M. Páez, "Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among english and spanish skills," *Language, Speech, and Hearing Services in Schools*, vol. 38, no. 3, pp. 225–236, 2007.

[16] A. Chan, S. Chen, S. Hamdani, B. Tse, and K. Cheng, "Story telling in bilingual Urdu-Cantonese ethnic minority children: Macrostructure and its relation to microstructural linguistic skills," *Frontiers in Psychology*, vol. 14, p. 924056, 2023.

[17] B. J. Amster, "The rate of speech of normal preschool children," Ph.D. dissertation, Temple University, 1984.

[18] T. J. Mahr, J. U. Soriano, P. J. Rathouz, and K. C. Hustada, "Speech development between 30 and 119 months in typical children II: Articulation rate growth curves," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 11, pp. 4057–4070, 2021.

[19] L. Miratrix and R. Ackerman, "Conducting sparse feature selection on arbitrarily long phrases in text corpora with a focus on interpretability," *The ASA Data Science Journal*, vol. 9, no. 6, pp. 435–460, 2016.

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

[21] K. Cain, S. O'Carroll, J. Oakhill, D. Klop, M. Visser, A. Smith, and A. Swart, "Exploring the impact of a story-based teacher training programme on language and early literacy in 4- and 5-year-olds," in *Wordworks*, 2024.

[22] 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.

[23] D. Klop and M. Visser, "Using MAIN in South Africa," *ZAS Papers in Linguistics*, vol. 64, pp. 207–210, 2020.

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

[25] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," https://praat.org/, 2022, version 6.2.09.

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

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] H. Kaneko, "Cross-validated permutation feature importance considering correlation between features," *Analytical Science Advances*, vol. 3, pp. 278–287, 2022.

[29] J.-P. van Zyl and A. P. Engelbrecht, "Analysis of classification metric behaviour under class imbalance," *Egyptian Informatics Journal*, vol. 31, p. 100711, 2025.

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

[31] 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.

[32] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: A simple visual method to interpret data," *Annals of Internal Medicine*, vol. 110, no. 11, pp. 916–921, 1989.

[33] K. Hengeveld, J. Rijkhoff, and A. Siewierska, "Parts-of-speech systems and word order," *Journal of Linguistics*, vol. 40, no. 3, pp. 527–570, 2004.

[34] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *LREC*, 2012.

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

[36] 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.

[37] T. Moeng, S. Reay, A. Daniels, and J. Buys, "Canonical and surface morphological segmentation for Nguni languages," in *SACAIR*, 2021.