ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

Early Identification of Language Disorders Using Natural Language Processing and Machine Learning: Challenges and Emerging Approaches

Jessica M. Lammert,[a] Angela C. Roberts,[b,c] Ken McRae,[d,e] Laura J. Batterink,[d,e] and Blake E. Butler[d,e,f]

[a] Graduate Program in Psychology, University of Western Ontario, London, Canada [b] School of Communication Sciences and Disorders, University of Western Ontario, London, Canada [c] Department of Computer Science, University of Western Ontario, London, Canada [d] Department of Psychology, University of Western Ontario, London, Canada [e] Centre for Brain and Mind, University of Western Ontario, London, Canada [f] National Centre for Audiology, University of Western Ontario, London, Canada

ARTICLE INFO

Article History:
Received July 22, 2024
Accepted October 23, 2024

Editor-in-Chief: Julie A. Washington
Editor: Mahchid Namazi

https://doi.org/10.1044/2024_JSLHR-24-00515

ABSTRACT

Purpose: Recent advances in artificial intelligence provide opportunities to capture and represent complex features of human language in a more automated manner, offering potential means of improving the efficiency of language assessment. This review article presents computerized approaches for the analysis of narrative language and identification of language disorders in children.
Method: We first describe the current barriers to clinicians' use of language sample analysis, narrative language sampling approaches, and the data processing stages that precede analysis. We then present recent studies demonstrating the automated extraction of linguistic features and identification of developmental language disorder using natural language processing and machine learning. We explain how these tools operate and emphasize how the decisions made in construction impact their performance in important ways, especially in the analysis of child language samples. We conclude with a discussion of major challenges in the field with respect to bias, access, and generalizability across settings and applications.
Conclusion: Given the progress that has occurred over the last decade, computer-automated approaches offer a promising opportunity to improve the efficiency and accessibility of language sample analysis and expedite the diagnosis and treatment of language disorders in children.

Identifying language impairment—generally defined by delayed or disordered language processes in the absence of other neurological or physiological conditions—often requires detailed analyses of language samples. These samples comprise meticulously transcribed and detailed notations of linguistic phenomena made by trained raters including educators, clinicians, and researchers. Accordingly, the main barriers to language sample analysis (LSA) include lack of time and low interrater reliability (Stark et al., 2021). Moreover, limited access to practitioners who can provide clinical diagnoses based on LSA outcomes (e.g., speech-language pathologists [SLPs]) due to costs and poor availability in areas where service does not meet community demand (Lim et al., 2017; McAllister et al., 2011) can further drive inequity and discrimination (O'Callaghan et al., 2005; Verdon et al., 2011). Recent advances in artificial intelligence (AI),[1] including machine learning (ML) and natural language processing (NLP), may help alleviate

Correspondence to Blake E. Butler: bbutler9@uwo.ca. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

[1] Colloquially, the terms *artificial intelligence* (AI) and *machine learning* (ML) are often used interchangeably; However, AI describes machines or software capable of intelligent processes (e.g., problem solving, decision making, knowledge representation, and perception), while ML is a subfield of AI concerned with making predictions about unseen data by learning the patterns found in real data. Natural language processing is a subfield of linguistics and AI concerned with the computational modeling of human language.

these barriers by allowing users to analyze language samples automatically or semi-automatically. Some of the major goals of automation include improving the efficiency, accessibility, reliability, and validity of analyses by simplifying diagnostic processes and decision making. In this review, we highlight emerging methods for automated LSA, with a focus on the assessment of children's narrative language skills. The goal of the current article is to help clinicians, educators, and researchers without a formal background in computer science to expand their awareness of automated approaches to LSA, how they work, and how to implement them. We begin with a discussion of typical LSA procedures and the current adoption of AI in clinical practice, followed by a methodological review of the literature on automated language assessment tools. Finally, we discuss the needs, challenges, ethical considerations, and future directions of AI-powered language assessment tools and their applications in health care and related fields.

Language Impairment in Children

Developmental language disorder (DLD) describes language difficulties that cause functional impairments in daily life and a poor prognosis without a clear biomedical cause (D. V. M. Bishop et al., 2017). Up to 8% of children meet the criteria for DLD depending on the population and definition of impairment (Law et al., 2000; Norbury et al., 2016; Tomblin et al., 1997; Weindrich et al., 2000). Language impairments in DLD are not caused by articulation issues; rather, they are associated with problems producing and comprehending the various components of language itself (e.g., phonology, syntax, semantics). These impairments often include issues with expressive and receptive phonology, syntax, semantics, discourse, and verbal memory that persist beyond early childhood (see D. V. M. Bishop et al., 2017, for a review). For example, a child with DLD may have difficulties distinguishing two words based on a single sound, interpreting the meaning of grammatical structures, or may use a more restricted vocabulary compared to their same-age peers.

Accurately diagnosing DLD can be challenging. DLD often co-occurs with other communication and developmental disorders such as autism spectrum disorder and attention-deficit/hyperactivity disorder (D. V. M. Bishop et al., 2017). The processing difficulties associated with DLD are similar to or, in some cases, contribute to co-occurring disorders (R. B. Gillam & Hoffman, 2003). Additionally, the dynamic and interactive nature of language impairments, which can change over time and across contexts, present additional challenges to diagnosis and monitoring (Hansson et al., 2014). DLD presents with a heterogeneous profile (Leonard, 1998), with considerable individual variability in severity, language areas affected, and changes in the nature/severity of deficits over time (Conti-Ramsden & Botting, 1999). Moreover, impairments observed in children with DLD are not limited to language systems but may affect working memory (Montgomery, 2003) and some motor skills (Hill, 2001) as well. Accordingly, approaches to the diagnosis and management of DLD must consider individual differences in the nature of the observed deficits. Clinical evidence shows that early and individualized treatment is effective for communication disorders (McLean & Woods-Cripe, 1997); diagnostic tools that help clearly identify the nature of the specific deficits underlying a child's experience with DLD are fundamental to the development of strategies that address those needs.

Current Assessment Practices

SLPs' Use of LSA

LSA is an exhaustive and ecologically valid method for assessing language impairments with a wide range of clinical and research applications (MacWhinney & Fromm, 2022). The American Speech-Language-Hearing Association (ASHA) recognizes LSA as a comprehensive assessment of language disorders and requires certified SLPs to be proficient in this technique (ASHA, n.d.). Many SLPs report a preference for language sampling approaches relative to standardized norm-referenced tests but cite time constraints, lack of training opportunities, and lack of computing resources as major barriers (Calder et al., 2017; Pavelko et al., 2016; Westerveld & Claessen, 2014).

In a large survey of school-based SLPs, only 67% (N = 1,336) of respondents reported using LSA during the 2012–2013 school year (Pavelko et al., 2016). A more recent small survey reported a substantially higher rate of LSA use (90%, N = 90); however, the authors also noted that respondents demonstrated lack of familiarity with recommended practices (Bawayan & Brown, 2022). LSA is frequently adopted as an informal language assessment measure, with less than one third of school-based SLPs using a specific protocol (Pavelko et al., 2016). Thus, while there appears to be a willingness to adopt LSA, additional education and standardization of practice is required. Indeed, many SLPs report a desire for further training in LSA (84%) and the interpretation of results (83%; Pavelko et al., 2016).

While improved training in administering and interpreting LSA would certainly benefit care providers, it would likely only exacerbate concerns around time constraints. Indeed, a majority of SLPs who do not routinely include LSA in their practice report that the method is too time consuming, regardless of caseload. Fortunately, automated tools are well positioned to support many of

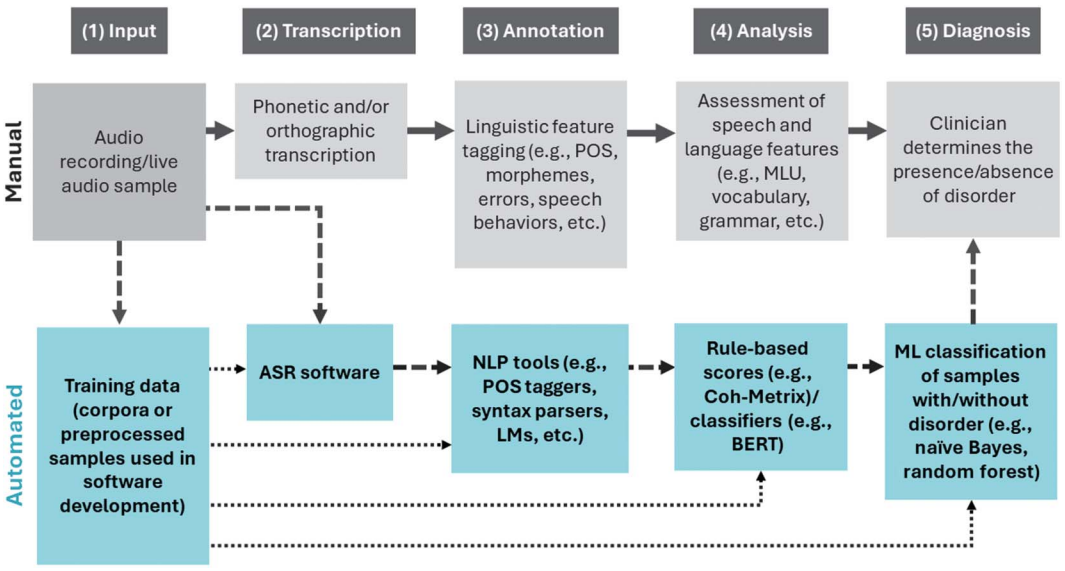
the steps involved in LSA (recording, transcription, feature annotation, measurement, analysis, and diagnosis; see Figure 1). These tools could supplement manual approaches by increasing the efficiency of LSA, thereby reducing barriers to access, and increasing intervention intensity (Baker, 2012). However, despite their accessibility, objectivity, and cost effectiveness, few SLPs (2%–7%) reported using computer-based approaches (McLeod & Baker, 2014; Pavelko et al., 2016). Further research is needed to understand the reasons for SLPs' methods of practice and whether these rates have changed in recent years.

Obtaining Narrative Language Samples

Language samples provide information across several domains, including speech, phonological, semantic, pragmatic, lexical, and syntactic development. In particular, language samples that mirror the style of communication observed in typical social and academic settings provide an excellent opportunity to identify the impact of language impairments in everyday functioning (Costanza-Smith, 2010). Indeed, a large body of research suggests that poor narrative structure can indicate delayed or impaired language in children (S. L. Gillam et al., 2017; Greenhalgh & Strong, 2001; Justice et al., 2006; Kaderavek & Sulzby, 2000; Liles et al., 1995; Merritt & Liles, 1987; Reilly, 2004). For example, children with DLD generally have poorer semantic representations (Sheng & McGregor, 2010) and produce narratives more similar to younger children than to their same-age peers (Botting, 2002). Narrative language sampling (e.g., story retelling) is more naturalistic than traditional language tests that comprise tasks like single utterance or sentence repetition (Westerveld et al., 2004) but is more standardized than spontaneous conversation and play. Moreover, narrative language use is an important and age-appropriate milestone for typically developing school-age children (Common Core State Standards Initiative, 2010; Nippold, 2016) and is suitable for progress monitoring and formative assessments in educational settings (Bailey et al., 2016).

There are several methods that assessors can use to elicit narrative language samples ranging in length and spontaneity. These approaches are often designed to elicit specific linguistic constructs such as tense, verb inflection, speech fluency, and narrative structure. Language sampling procedures—including conversation, interviews, play, spontaneous narratives, narrative retelling, and other forms of structured discourse—effectively capture a broad range of micro- and macrostructural language features (Westerveld, 2011). While task variability may provide some degree of flexibility, it also poses a challenge to scoring assessments and interpreting outcomes. In addition to being true of hand-analyzed samples (Eisenberg et al., 2018; Kapantzoglou et al., 2017), variability is particularly problematic for automated approaches because differences in sampling procedures negatively impact models' generalizability (i.e., the ability to classify previously unseen data) and transfer learning (i.e., the ability to use information derived from one sample to improve performance in another).

Figure 1. Language sample analysis protocol and automated approaches. Overview of a typical clinical language sample analysis protocol and opportunities for computer automation with NLP and ML. Solid arrows denote the manual procedure, while dashed lines denote the automated procedure. Dotted arrows point to processes where training data may be used. ASR = automatic speech recognition; BERT = bidirectional encoder representations from transformers; LM = language model; ML = machine learning; MLU = mean length of utterance; NLP = natural language processing; POS = part-of-speech.



Transcription and Annotation

Manually transcribing and annotating audio- or video-recorded speech samples accounts for a significant portion of a clinician's time spent conducting LSA. Transcribers not only orthographically transcribe the sample but must also segment the sample and annotate language and speech behaviors according to a specific coding system. Two standard annotation systems are the Codes for the Human Analysis of Transcripts (MacWhinney & Fromm, 2022) and Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2015). However, in clinical practice, about half of surveyed SLPs reported favoring self-designed transcription and analysis procedures over packaged systems (Hux et al., 1993; Kemp & Klee, 1997; Pavelko et al., 2016). This was particularly true of more experienced respondents (Pavelko et al., 2016).

There has been no comprehensive survey of SLPs' use of speech-to-text technology, despite the fact that this technology has evolved quickly in recent years and may provide a simple and widely accessible way to reduce this time-intensive step. One current challenge is that many automated speech recognition (ASR) systems (e.g., Google, Microsoft, IBM) are trained on adult speech samples and often produce higher-than-acceptable error rates for children's speech (Hunte et al., 2021; Wu et al., 2019; but see Fox et al., 2021). Children's speech differs from adults, showing age-dependent changes in vocal tract length, decreased fluency and prosody, elongated vowels, and more pitch/volume instability (see Hunte et al., 2021, for a review). An additional challenge is that ASR can be highly sensitive to background noise. The recording location must be quiet to ensure recognition accuracy, which may be challenging for clinicians and researchers testing in school settings (Alim & Rashid, 2018).

While automated transcription captures individual words with some degree of accuracy, clinicians often wish to annotate or "code" specific linguistic structures, acoustic properties, errors, and other nonstandard measures. Identifying clinically relevant measures, such as the omission or overgeneralization of grammatical morphemes, often requires considerable contextual knowledge and discernment by a trained coder. Because transcription and annotation are necessary and time-consuming phases of LSA, whether transcripts are human generated, computer generated, or computer generated and human corrected, significantly impacts accuracy and feasibility.

Toward Automated Assessment

Emerging automated approaches make it easier to extract important features from language data and reduce the time and human effort needed for LSA. Current accuracy levels are such that users cannot completely rely on fully automated end-to-end pipelines (i.e., automation of everything from sample transcription to diagnosis) without human oversight. However, automated measures certainly can supplement other information used in clinical decision making. Moreover, with exponential growth in new patents for language processing AI over the last several years (World Intellectual Property Organization, 2019), the tools available to clinicians and researchers will continue to grow in number and quality.

Given the urgent need to meet clinical demand, the aim of the current article is to provide a better understanding of existing automated methods for narrative language assessment in children. In the next section, we describe automated approaches available at each stage of LSA; assess the quality of the tools, tasks, models, and features included in those techniques; and evaluate evidence supporting their use in identifying DLD. We also discuss several major challenges in the field including generalizability, interpretability, data quality, and bias.

Automating LSA and Applications for the Diagnosis of DLD

Automating Feature Extraction

Manual annotation and scoring of language measures requires considerable training and expertise. Automated methods can save users time transcribing and annotating language samples. NLP tools like part-of-speech (POS) taggers and syntactic parsers can be used to automatically annotate multiple grammatical structures for further analysis. NLP tools are often trained on large corpora to resolve contextual ambiguity and improve prediction accuracy. While not perfect reflections of children's speech, automated POS tagging based on these corpora is expected to be more efficient and accurate than manual coding. Indeed, POS taggers trained on adult language (e.g., TnT tagger; Brants, 2000) have been shown to outperform taggers trained on child speech (e.g., MacWhinney, 2000) for classifying DLD (Gabani et al., 2011), although some manual correction of automated annotation often is required and generally expected for children's speech. Importantly, NLP tools can be used to extract a wide range of micro- and macrolevel features of interest including morphosyntactic features, vocabulary, language productivity measures, general coherence, and narrative structure/quality measures.

Narrative Microstructure

Narrative microstructure refers to a wide range of features, including fluency, lexical diversity, syntactic complexity, and language productivity. As such, several

measures can be used to assess microstructural features, including number of utterances, mean length of utterances, number of unique words used, use of adverbs, presence/absence of elaborated noun phrases, use of conjunctions, syntactic accuracy, and complexity. Automatic extraction of these features typically involves segmenting a transcript at the word or morpheme level, after which automated taggers assign each word in the sentence to a POS such that syntactic parsers can use those tags to generate a syntactic tree. Once annotated, rule-based programs can be used to identify syntactic constructs, vocabulary, and other aspects of language usage.

In one example, Hassanali et al. (2014) used Charniak's (2000) parser and 60 hard-coded rules to automate calculation of the index of productive syntax (IPSyn; Scarborough, 1990), a clinical measure of syntactic development. Their approach reduced the time required to score the IPSyn from 30 min per 100 utterances (manual scoring) to less than 5 min (Hassanali et al., 2014). While initial attempts yielded accuracy scores ranging from 77% to 83% depending on the construct (Altenberg and Roberts, 2016), improvements have increased the accuracy of automated IPSyn programs to above 95% (MacWhinney et al., 2020).

Fox et al. (2022) took a similar approach to scoring the narrative microstructure components of the Monitoring Indicators for Scholarly Language (MISL; S. L. Gillam et al., 2017). Using age-appropriate word banks and the Apache OpenNLP POS tagger, rules were hard-coded for identifying six narrative microstructure elements: coordinating and subordinating conjunctions, metalinguistic and metacognitive verbs, adverbs, and elaborated noun phrases. A comparison between this automated approach and scores assigned by experts yielded very good to near-perfect agreement (i.e., quadratic weighted kappa [QWK] values ranging from 0.74 to 0.89), with the automated approach outperforming nonexpert scorers on four of six measures. This system shows promise in the automated scoring of narrative language samples and is currently being developed into a web-based application for clinical use. Most NLP advances thus far have been in automating the analysis of narrative microstructure, though macrostructural issues are also critical to many language impairments.

Narrative Macrostructure

Narrative macrostructure refers to story content and organization, including measures of semantic content; overall quality; and knowledge of story events, characters, and settings. Automated annotation of narrative macrostructural features is limited, and human scores of narrative quality measures can be highly subjective (Fey et al., 2004). Coh-Metrix uses NLP tools to analyze discourse by extracting features including number of utterances/words, modifiers, intensifiers, and high-level syntactic constructs, Flesch–Kincaid readability tests, conceptual similarity, and many others (McNamara et al., 2014). Including measures derived from Coh-Metrix improves the classification of DLD compared to hand-scored measures alone (Hassanali et al., 2012a, 2012b). That said, Coh-Metrix measures are often subject to the output of third-party parsers, lexicons, and word frequency databases, all of which can impact feature extraction.

An alternative automated approach to macrostructure feature extraction is topic modeling, which provides a numerical representation of the topics in each transcript that roughly correspond to important events (Hassanali et al., 2013). This approach takes advantage of NLP methods like latent semantic analysis or latent Dirichlet analysis, which employ dimensionality reduction for modeling the relationship between terms and concepts in text (Blei et al., 2003). Semantic representations may be derived directly from the sample (Hassanali et al., 2013) or models may be trained on an existing corpus of text and subsequently used to summarize language samples (Bååth et al., 2019). Like microstructural feature modeling, semantic models trained on corpora composed of adult language (e.g., newspaper articles) may use different vocabulary compared to child narrative samples but often produce higher quality representations than those trained on smaller samples of children's speech (Garcia & Sikström, 2014). Indeed, trained topic models have provided insight into semantic development and have been shown to predict DLD based on children's language samples (Bååth et al., 2019; Hassanali et al., 2013).

Jones et al. (2019) compared four NLP methods for scoring the six narrative macrostructural components of the MISL: character, setting, initiating event, plan, action, and consequence (S. L. Gillam et al., 2017). In addition to Coh-Metrix, Jones et al. used language models of varying complexities, including term frequency-inverse document frequency (TF-IDF), global vectors for word representation (GloVe) embeddings, and bidirectional encoder representation from transformer (BERT) embeddings to extract macrostructure features. TF-IDF is a common measure of word importance based on word frequency, while GloVe and BERT are methods of constructing semantic embedding vectors similar to latent semantic analysis and latent Dirichlet analysis. While GloVe maps individual words to vectors, BERT is itself a deep neural network that can take larger sequences (i.e., sentences, whole narratives) and therefore resolve semantic ambiguities based on the surrounding context. The BERT model outperformed all other methods, achieving near perfect agreement (QWK > 0.9) with hand-labeled scores on all elements except "consequence" (QWK = 0.79). Recent advances in NLP show promise in automating the scoring of narrative macrostructure elements; however, models must be fine-tuned

to handle a variety of narrative prompts and language produced by different age groups. Standardized language sampling and open data sharing practices are therefore key to building robust and accurate models. Considerations and ongoing efforts are discussed further in the Challenges in the Clinical Application of ML section.

Using ML to Identify DLD

In ML, a model (e.g., logistic regression, random forest, support vector machine [SVM]) learns the relationship between one or more "features" and some outcome. In the case of NLP, these features commonly comprise standardized test scores or other measures used to assess language skills (C. M. Bishop, 2006), which are associated with an outcome like the presence/absence of DLD. Importantly, such a model—trained on an adequate sample—should be able to classify novel cases. The predicted output may consist of a binary label or probability of belonging to a certain group (i.e., a prediction about the presence/absence of DLD) or, in cases of multiclass classification, a predicted category (e.g., score on an assessment). However, the accuracy of ML models for classifying DLD depends both on the properties of the model itself and the quality of the features used to predict language status. Models typically are trained via supervised learning in which machine predictions are compared to hand-labeled data to create a mapping between the input and expected output. Therefore, diagnostic validity is, in part, dependent on the quality of the hand-labeled data, which serves as the "ground truth" during training, the limitations of which are described further in the Input Quality section below. Once trained, the model should be able to correctly determine the output label of unseen "test" data. In general, there is no "best model," and programmers are expected to experiment with different algorithms and configurations to improve performance for the given use case. For example, while some uses may primarily concern the diagnostic outcome, others may be interested in what predictors led to a given outcome (i.e., feature importance).

ML Algorithms

In the construction of predictive models, multiple ML algorithms often are compared on a single task. To determine which model best classifies children based on the presence/absence of DLD, researchers have submitted a common set of features to logistic regression, LogitBoost, Bayesian network, naïve Bayes, and SVM (Gabani et al., 2011; Hassanali et al., 2012a, 2012b, 2013), as well as neural network and decision tree models (Jones et al., 2019; Oliva et al., 2014; Prud'hommeaux et al., 2011). Across several studies, Bayesian network (Gabani et al., 2011; Hassanali et al., 2012a, 2012b) or naïve Bayes (Hassanali et al., 2013) models showed the best performance. In one study, Bayesian network models trained on micro- and macrostructure features showed an excellent ability to classify DLD from children's narrative language samples ($F1$ = 0.914; Hassanali et al., 2012b). Bayesian classifiers are easier to implement than other classifiers, with fewer parameters to estimate and show good performance with small data sets (IBM, n.d.). Moreover, Bayesian classifiers can handle high-dimensional data such as text, making them appropriate for the classification of language impairments.

Logistic regression, including logistic ridge regression (LRR) models, is also commonly used in classification. In Bååth et al. (2019), an LRR classifier showed 81% accuracy when classifying severe DLD versus typically developing based on latent semantic analysis. However, classification accuracy was considerably lower (68% correct) for less severe cases of DLD, suggesting that this type of model may be less sensitive to mild impairments. One advantage of logistic regression over Bayesian models, however, is that it is easier to interpret the relative importance of features within the trained feature set based on the magnitude of their coefficients (an important consideration if attempting to narrow down the number of measures acquired during diagnosis for example).

Tree-based models are similarly favored for their interpretability, as well as their ability to model complex, nonlinear relationships. Decision trees work by recursively splitting the sample to arrive at a final prediction based on simple decision rules inferred from data features. Decision trees have been shown to be adequate classifiers of DLD versus typically developing children ($F1$ = 79%) but could not discriminate between children with DLD and autistic children ($F1$ = 52%; Prud'hommeaux et al., 2011). Random forest models combine the output of multiple decision trees to arrive at the most likely outcome. They are often more accurate than simple decision trees but are also computationally demanding and more difficult to interpret in part due to the independence of training data, differing results, and quality of the binary splits that form each tree in the ensemble (Schonlau & Zou, 2020).

To summarize, modern ML models can predict language impairment with a high degree of accuracy and agreement with human raters. Some of the strongest classifiers (e.g., Bayesian network, naïve Bayes, BERT) can be challenging to interpret; however, approaches like logistic regression and decision tree modeling can reveal distinct patterns of behavior related to DLD. ML models trained via supervised learning on human scores can reach human-level performance but consequently replicate human errors (see Jones et al., 2019). Performance is dependent on several factors including model complexity, demographic representativeness of the training sample, similarity of

sample elicitation methods, and appropriateness of the features for the sample and classification task.

Training ML Algorithms

Beyond algorithm selection, there are several possible approaches to training ML models via supervised or unsupervised learning. ML models designed to automate LSA have been trained using leave-one-out cross-validation (Bååth et al., 2019; Gabani et al., 2011; Hassanali et al., 2012a, 2012b, 2013; Oliva et al., 2014; Solorio & Liu, 2008), hold-out cross-validation (Jones et al., 2019), or *k*-fold cross-validation (Gabani et al., 2011; Jones et al., 2019).

In leave-one-out cross-validation, each observation is iteratively considered the "validation" or "test" data while the remaining observations serve as the training data. Leave-one-out cross-validation is a common approach, especially for small data sets, but can be computationally demanding for complex models. In hold-out cross-validation, a small subset of the data is withheld from training (usually 10%–20%). The held-out data are then used to validate whether the trained model generalizes to unseen data. Hold-out cross-validation is a popular approach, but there is some concern that unseen data may contain important information that could have improved the performance of the model. Hold-out cross-validation is generally not appropriate for small data sets, where withholding even a small portion of the total data from model training could negatively impact model performance. In *k*-fold cross-validation, data are shuffled randomly and divided into k groups, each of which serves as the test set once and training set $k-1$ times (James et al., 2013). Evaluation metrics can then be summarized and compared or averaged across cross-validation folds for a better understanding of model performance. This approach is purported to produce less biased estimates of model accuracy compared to simpler approaches (Kohavi, 1995) but requires larger data sets. Regardless of the training procedure, the quality and representativeness of the data (discussed further in the Challenges in the Clinical Application of ML section) also impact how well the model generalizes to new samples.

One alternative to supervised learning is cluster analysis—an unsupervised ML task where data are grouped together based on similarity. For example, a clustering analysis approach to DLD would group children based solely on the features described above (e.g., standardized test scores or other measures used to assess language skills), without any associated information about their diagnoses. One such hierarchical clustering analysis based on cognitive features revealed three distinct groups: (a) children with procedural deficits; (b) children with grammatical deficits; and (c) children with processing deficits, with most children with DLD showing procedural deficits (Oliva et al., 2014). This approach is especially useful for identifying how symptoms cluster without a need for preexisting labels (especially when the profile of a disorder is not well understood).

Challenges in the Clinical Application of ML

Generalizability and Transfer Learning

Some elicitation methods and features may be better suited to certain ML approaches than others. To capture a broad range of micro- and macrostructural measures, clinicians use a variety of language sampling procedures including spontaneous conversation, play, narratives, and other forms of structured discourse. Variability among these procedures can lead to problems with generalizability and transfer learning, or the ability to apply ML techniques to novel classification tasks and data.

For example, classifiers that successfully identified DLD when trained on structured narrative language samples were less successful when trained on data from unstructured play sessions and personal narratives (Gabani et al., 2011). An analysis of which features were most important to model accuracy showed that language productivity and speech fluency were important predictors of DLD when the model was trained on narrative samples but were not indicative of DLD in play sessions (Gabani et al., 2011). Even when the same elicitation procedure is used, classifiers may be more accurate at identifying DLD in certain populations over others (e.g., more vs. less severe DLD; Bååth et al., 2019).

Since the goal of many developers is to scale experimental ML approaches into user-facing applications (see Bright et al., 2023; Fox et al., 2022; Scott et al., 2022), it is important to consider how they generalize across language sampling procedures, demographics, and behavior profiles, as this will seriously impact their clinical uptake. Though clinicians can offer a deep understanding of a given condition and potential features of interest, the expertise necessary to discern which NLP and ML approach is best suited to a particular assessment approach is not a core part of clinical training at present and could impact generalizability to clinical cases.

Transparency and Interpretability

Fairness, accountability, and transparency in AI are emerging areas of research concerned with the interpretability and complex social implications of AI. Neural networks and deep learning models such as BERT often operate as a "black box," where the complexity of the model makes it difficult to explain what factors influence predictions. BERT base, for example, consists of 110 million parameters (Devlin et al., 2019). In addition, many of these models are computationally demanding, making

them poorly suited for low-resource environments. However, there are emerging methods of distilling these complex models into simpler, more efficient forms (Sun et al., 2019).

Despite recent growth in the intersections between health care and AI, little is known about how interpretability impacts user trust (Hamm et al., 2023; Woodcock et al., 2021)—a significant potential obstacle to the application of ML in health care settings. Indeed, trust in AI is a value that differs between individuals and groups and may shift over time as technology continues to change (Lukyanenko et al., 2022). As developers and regulators strive to define what constitutes responsible use of AI, several tools are available to implement and evaluate fairness, accountability, and transparency (Bellamy et al., 2019; Huang et al., 2022; Meng et al., 2022; Zhdanov et al., 2022).

A challenge specific to the adoption of computerized LSA is a substantial "research-to-practice gap"—the disconnect between best practices based on empirical research and what is realistically feasible for practitioners (Olswang & Prelock, 2015). Clinicians indicate a strong need for proper training using computerized tools (Pavelko et al., 2016); interpreting the output of ML models and other computerized analysis systems may not be intuitive for nonexperts and new users, which could ultimately lead to frustration and abandonment of automated tools, or of LSA as an assessment approach in general (Klatte et al., 2022).

Input Quality

Sample size has long been one of the biggest influences on bias and generalizability in ML classifiers (Raudys & Jain, 1991). The samples used to train DLD classification models were quite small compared to what is typically needed to resolve complex ML problems (Obermeyer & Emanuel, 2016). That said, larger samples are not always better; the diversity or representativeness of a sample must also be considered. The overreliance on Western, educated, industrialized, rich, and democratic—coined "WEIRD"—samples has rapidly gained attention in the psychological and medical sciences (Henrich et al., 2010). The proliferation of these biases through predictive algorithms can have serious and inequitable consequences on population health (Obermeyer et al., 2019). Beyond age, gender, language, and geographic region, other demographic information such as race, ethnicity, and socioeconomic status (SES) were not generally reported in the literature (Bååth et al., 2019; Fox et al., 2022; Jones et al., 2019; Prud'hommeaux et al., 2011; Solorio & Liu, 2008). Because ML classifiers essentially learn rules from the data provided to them, validating their performance on diverse, independent data sets is an important step in ensuring accurate and equal performance. One concern with inadequate demographic reporting or representation is the potentially massive scale at which data-driven algorithms reflect bias back into our world with limited supervision (Shah, 2018). That is to say, a model trained on data from a specific subset of a diverse population will (mis)apply any underlying biases to its predictions about that broader population (see the Bias section).

Since model accuracy depends on learning critical associations between language features and diagnoses, differences in the way these diagnoses were made and language samples were obtained represent other challenges for models aimed at classifying DLD. Typically, DLD was diagnosed by an SLP (Solorio & Liu, 2008) but may also be determined by researchers based on cutoff measures (Gabani et al., 2011). Moreover, in some cases, language samples are collected years after children received a diagnosis of DLD and began receiving treatment (Gabani et al., 2011).

Manual scoring is often a necessary part of clinical ML model construction, where models designed to classify impairment are trained through supervised learning on hand-scored measures. However, there are several well-known limitations of using hand-scored measures as the gold standard of comparison. For example, interrater reliability can be impacted by myriad factors including fatigue (S. L. Gillam et al., 2017), expertise (Fox et al., 2022), and rater drift—where a scorer's tolerance or understanding of a given concept changes over time (Leckie & Baird, 2011). The variability between ratings used to train a model can have a significant impact on its performance. For example, when Jones et al. (2019) trained a BERT model using a database of nonexpert scores, they found that the model predictions were more closely aligned with nonexpert raters than with experts.

Finally, transcription quality also affects the accuracy and efficiency of automated approaches. ASR has shown to be less accurate for children's speech than for adults. For example, the word error rate obtained from Google ASR was found to be relatively high (25%) for a sample of children compared to other samples reported by Google (5%; Hunte et al., 2021). Many NLP tools are trained on adult language corpora and will frequently require correction by human scorers, especially for children's speech. Transcription accuracy has a considerable influence on output such that few pipelines are capable of fully automating end-to-end analysis reliably. Thus, until this technology develops further, improving the efficiency of LSA may depend on partially automated approaches over full automation.

Bias

Articles 19 and 27 of the Universal Declaration of Human Rights (United Nations, n.d.) support the rights

of people with communication disorders to receive maximal benefit from speech and language research, regardless of race, ethnicity, ability, SES, and so forth. It is thus imperative that ML models for language assessment are developed in a way that is sensitive to algorithmic biases—"instances when the application of an algorithm compounds existing inequities in SES, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems" (Panch et al., 2019). Algorithmic bias can arise at nearly any stage of the clinical ML workflow, resulting in inequitable diagnosis and treatment outcomes across subpopulations (Chen et al., 2023). Outcomes can be skewed by methodological or sociocultural and demographic factors that could create bias toward certain groups or individuals (Akter et al., 2021). Algorithmic bias has important bearings on who stands to benefit from innovations in AI (Panch et al., 2019), and policymakers have only recently addressed it in a legal setting, such as the European Union's Artificial Intelligence Act (European Commission, 2021).

In line with mandates for equitable access (e.g., U.S. Individuals with Disability Education Act, Accessible Canada Act), clinicians are responsible for identifying and using appropriate assessments for patients who are linguistically and culturally diverse (ASHA, 2010). This is critically important, given that the accuracy of ASR (Hannah et al., 2022) and language measures (Overton et al., 2021) vary across language backgrounds, with criterion-referenced language tests showing evidence of bias against populations not represented in the reference (i.e., children from minority ethnic backgrounds and low SES; Campbell et al., 1997). A significant portion of people (approximately 20%) in Canada and the United States primarily speak languages other than English (Statistics Canada, 2018; Zeigler & Camarota, 2019), and about 55% of Canadians are bilingual (Luk, 2017). Moreover, current LSA protocols may potentially disadvantage speakers of nonmainstream English dialects, conflating language *differences* with *disorder*. For example, when the TalkBank model (MacWhinney & Fromm, 2022) was used to assess children who spoke African American English, it was shown that that number of different words—a vocabulary-based measure—could not differentiate children with DLD from typically developing children, suggesting poor sensitivity to language impairment in underrepresented populations (Overton et al., 2021). Thus, tools based on monolingual, English-speaking populations do not represent the linguistic reality of many communities. While there have been some efforts to develop models for other languages (e.g., multilingual BERT; Devlin et al., 2019), significant obstacles remain. For example, transcription and annotation of bilingual speech require the coder to have a high level of proficiency in both languages and familiarity with bilingual language patterns such as code switching. Additionally, ASR systems have shown bias against accented speech, and greater variability and higher error rates for children from diverse linguistic backgrounds (Hannah et al., 2022). While there is a small body of research on automated LSA in bilinguals (e.g., Hassanali et al., 2014), there is little overlap with the identification of DLD.

The goal of ML solutions is to increase the feasibility and accuracy of LSA, but questions remain around which groups will benefit from the application of these models and which groups will be disadvantaged. As distributed representational models, neural networks and other complex AI effectively learn and reproduce human biases (Arseniev-Koehler & Foster, 2022) and can contribute to ongoing prejudice that leads to discriminatory behaviors (Curto et al., 2022). Thus, these time-saving technological advantages may serve to widen an already apparent service gap if these issues remain unaddressed. Measures such as the Gini coefficient (a measure of statistical dispersion intended to quantify inequality; Gini, 1936) and other indices can be used to quantify outcome inequality in ML classifiers, though a large-scale comparison of language analysis tools has yet to be conducted. However, in reaction to the harms and potential risks of rapidly advancing AI technology, regulatory bodies are beginning to set standards for AI in health care to ensure more equitable results. These standards will likely continue to evolve as our understanding of AI and its impact across different segments of the population grows.

Conclusions and Future Directions

LSA is a robust, ecologically valid language assessment tool; however, the processes of data collection, transcription, annotation, and analysis can be time consuming and unreliable. NLP and ML approaches show considerable promise in automating the extraction of language features from a variety of samples and in predicting whether the child who produced that sample is experiencing DLD—all in a fraction of the time of traditional LSA. However, the strengths and weaknesses of automated approaches must be seriously considered before they are put into practice. Many NLP tools are not optimized for child language and require significant correction by hand. The classifiers reviewed above predicted DLD with 40%–90% accuracy depending on the task and sample. BERT and Bayesian classifiers are currently best at identifying language disorder from children's narrative language samples. But in general, model performance depends greatly on tuning the features and tools to best fit the intended purpose.

ML approaches, especially in health care, should be explainable, unbiased, accurate, and trustworthy. Currently, automated LSA models face several challenges associated with algorithmic bias, generalization to different populations, and the interpretation of deep learning models that must be addressed. If uncorrected, these challenges are likely to further contribute to systemic discrimination through the proliferation of learned biases and lack of sensitivity to racialized, disabled, or other underrepresented individuals. Lack of representation is not exclusive to this field (see Henrich et al., 2010), but it does have marked ramifications for the analysis of language, which displays considerable diversity across subpopulations.

SLPs report lack of time and training as major barriers to the feasibility of LSA (Klatte et al., 2022; Pavelko et al., 2016). Additional research is needed to understand the extent to which SLPs currently use AI in clinical practice and to assess their views of whether automated or semi-automated methods hold the potential to save time over fully manual methods in practice. Education aiming to close the research-to-practice gap should acknowledge that it can be a lengthy process for good tools to be validated, released, and widely adopted, while highlighting recent successful prepackaged systems to automate the stages of the LSA process. For example, the SALT (Miller & Iglesias, 2015) and TalkBank (MacWhinney & Fromm, 2022) systems are robust automated LSA programs with built-in databases for comparison. Reference databases range in size from dozens to thousands of samples depending on age, location, and elicitation method. Importantly, users worldwide can submit their own data sets to the TalkBank repository, which includes a diverse range of clinical and nonclinical databases for comparison, and the inclusion of additional data sets is expected to further improve model predictions. SALT, on the other hand, is a proprietary software designed for SLPs in clinical practice that contains bilingual and monolingual Spanish and English language samples of limited ethnic and geographic diversity. The SALT developers offer an extensive training library that makes it much easier for new clinicians to adopt; however, as a proprietary application, SALT's classification accuracies are not widely published. TalkBank, on the other hand, consists of open software that adheres to international standards and responsiveness to user needs, which has made it an excellent resource for those wanting to explore LSA in depth.

Though the current article is limited to narrative language analysis in children, automated discourse analysis is used in several fields including psychology, health care, politics, economics, and more. The limitations of automated LSA are to some degree associated with their intended application; for the identification of DLD in children, these limitations center on the technical specificities, training requirements, and lack of interpretability of existing approaches (Klatte et al., 2022; Lüdtke et al., 2023; Pavelko et al., 2016). Ultimately, developers wishing to integrate AI with health care should follow the "translation continuum" (i.e., the course of translating basic scientific discoveries into clinical applications; Abràmoff et al., 2023) in collaboration with clinical researchers familiar with this process. Tools should be validated on independent samples with different demographics and elicitation methods to ensure their generalizability and fairness. Where possible, approaches should combine features from multiple language domains and produce a detailed output of feature importance, individual scores, and so forth. Future research should work toward eliminating the "black box" associated with AI to increase fairness, accountability, and trust, especially for applications that affect human lives.

Data Availability Statement

Data sharing is not applicable to this article as no data sets were generated or analyzed in preparing the current article.

Acknowledgments

Support for this work was provided by a grant to Blake E. Butler from the Natural Sciences and Engineering Research Council of Canada (05572-2017).

References

Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C., & Maisel, W. H. (2023). Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digital Medicine, 6*(1), Article 170. https://doi.org/10.1038/s41746-023-00913-9

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management, 60,* Article 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Alim, S., & Rashid, N. (2018). Some commonly used speech feature extraction algorithms. In R. Lopez-Ruiz (Ed.), *From natural to artificial intelligence—Algorithms and applications*. IntechOpen. https://doi.org/10.5772/intechopen.80419

Altenberg, E. P., & Roberts, J. A. (2016). Promises and pitfalls of machine scoring of the index of productive syntax. *Clinical Linguistics & Phonetics, 30*(6), 433–448. https://doi.org/10.3109/02699206.2016.1139184

American Speech-Language-Hearing Association. (n.d.). *Spoken language disorders* [Practice portal]. http://www.asha.org/practice-portal/Clinical-Topics/Spoken-Language-Disorders/

American Speech-Language-Hearing Association. (2010). *Roles and responsibilities of speech-language pathologists in schools* [Professional issues statement]. https://www.asha.org/policy/pi2010-00317/

Arseniev-Koehler, A., & Foster, J. G. (2022). Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research, 51*(4), 1484–1539. https://doi.org/10.1177/00491241221122603

Bååth, R., Sikström, S., Kalnak, N., Hansson, K., & Sahlén, B. (2019). Latent semantic analysis discriminates children with developmental language disorder (DLD) from children with typical language development. *Journal of Psycholinguistic Research, 48*(3), 683–697. https://doi.org/10.1007/s10936-018-09625-8

Bailey, A. L., Blackstock-Bernstein, A., Ryan, E., & Pitsoulakis, D. (2016). Data mining with natural language processing and corpus linguistics: Unlocking access to school children's language in diverse contexts to improve instructional and assessment practices. In S. ElAtia, D. Ipperciel, & O. R. Zaïane (Eds.), *Data mining and learning analytics* (pp. 255–275). Wiley. https://doi.org/10.1002/9781118998205.ch15

Baker, E. (2012). Optimal intervention intensity in speech-language pathology: Discoveries, challenges, and unchartered territories. *International Journal of Speech-Language Pathology, 14*(5), 478–485. https://doi.org/10.3109/17549507.2012.717967

Bawayan, R., & Brown, J. A. (2022). Language sample analysis consideration and use: A survey of school-based speech language pathologists. *Clinical Archives of Communication Disorders, 7*(1), 15–28. https://doi.org/10.21849/cacd.2022.00703

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE-2 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3,* 993–1022.

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy, 18*(1), 1–21. https://doi.org/10.1191/0265659002ct224oa

Brants, T. (2000). TnT—A statistical part-of-speech tagger. *Proceedings of the Sixth Conference on Applied Natural Language Processing,* 224–231.

Bright, R., Ashton, E., Mckean, C., & Wren, Y. (2023). The development of a digital story-retell elicitation and analysis tool through citizen science data collection, software development and machine learning. *Frontiers in Psychology, 14,* Article 989499. https://doi.org/10.3389/fpsyg.2023.989499

Calder, S., Stirling, C., Glisson, L., Goerke, A., Kilpatrick, T., Koch, L., Taylor, A., Wells, R., & Claessen, M. (2017). Language sample analysis: A powerful tool in the school setting. *Journal of Clinical Practice in Speech-Language Pathology, 19*(2), 66–71.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research, 40*(3), 519–525. https://doi.org/10.1044/jslhr.4003.519

Charniak, E. (2000). A maximum-entropy-inspired parser. *NAACL 2000: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics,* 132–139.

Chen, Z., Hu, B., Liu, X., Becker, B., Eickhoff, S. B., Miao, K., Gu, X., Tang, Y., Dai, X., Li, C., Leonov, A., Xiao, Z., Feng, Z., Chen, J., & Chuan-Peng, H. (2023). Sampling inequalities affect generalization of neuroimaging-based diagnostic classifiers in psychiatry. *BMC Medicine, 21*(1), Article 241. https://doi.org/10.1186/s12916-023-02941-4

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. http://www.thecorestandards.org/ELA-Literacy/

Conti-Ramsden, G., & Botting, N. (1999). Classification of children with specific language impairment: Longitudinal considerations. *Journal of Speech, Language, and Hearing Research, 42*(5), 1195–1204. https://doi.org/10.1044/jslhr.4205.1195

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education, 17*(1), 9–15. https://doi.org/10.1044/lle17.1.9

Curto, G., Jojoa Acosta, M. F., Comim, F., & Garcia-Zapirain, B. (2022). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & Society, 39*(2), 617–632. https://doi.org/10.1007/s00146-022-01494-z

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT,* 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Eisenberg, S. L., Guo, L.-Y., & Mucchetti, E. (2018). Eliciting the language sample for developmental sentence scoring: A comparison of play with toys and elicited picture description. *American Journal of Speech-Language Pathology, 27*(2), 633–646. https://doi.org/10.1044/2017_AJSLP-16-0161

European Commission. (2021). *Artificial Intelligence Act*. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47*(6), 1301–1318. https://doi.org/10.1044/1092-4388(2004/098)

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription. *Journal of Speech, Language, and Hearing Research, 64*(9), 3533–3548. https://doi.org/10.1044/2021_JSLHR-21-00096

Fox, C. B., Jones, S., Gillam, S. L., Israelsen-Augenstein, M., Schwartz, S., & Gillam, R. B. (2022). Automated progress-monitoring for Literate Language Use in Narrative Assessment (LLUNA). *Frontiers in Psychology, 13,* Article 894478. https://doi.org/10.3389/fpsyg.2022.894478

Gabani, K., Solorio, T., Liu, Y., Hassanali, K., & Dollaghan, C. A. (2011). Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children. *Artificial Intelligence in Medicine, 53*(3), 161–170. https://doi.org/10.1016/j.artmed.2011.08.001

Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences, 67,* 92–96. https://doi.org/10.1016/j.paid.2013.10.001

Gillam, R. B., & Hoffman, L. M. (2003). Information processing in children with specific language impairment. In L. Verhoeven & H. van Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical implications* (pp. 137–157). Erlbaum.

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2017). Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly, 38*(2), 96–106. https://doi.org/10.1177/1525740116651442

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, 208,* 73–79.

Greenhalgh, K. S., & Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools, 32*(2), 114–125. https://doi.org/10.1044/0161-1461(2001/010)

Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets, 33*(1), Article 17. https://doi.org/10.1007/s12525-023-00640-9

Hannah, L., Kim, H., & Jang, E. E. (2022). Investigating the effects of task type and linguistic background on accuracy in automated speech recognition systems: Implications for use in language assessment of young learners. *Language Assessment Quarterly, 19*(3), 289–313. https://doi.org/10.1080/15434303.2022.2038172

Hansson, K., Sandgren, O., & Sahlén, B. (2014). Changing labels for a concept in change. Comment to Bishop, D. Ten questions about terminology for children with unexplained language problems. *International Journal of Language & Communication Disorders, 49,* 407–408.

Hassanali, K., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods, 46*(1), 254–262. https://doi.org/10.3758/s13428-013-0354-x

Hassanali, K., Liu, Y., & Solorio, T. (2012a). Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. *Workshop on Child, Computer, and Interaction,* 7–12.

Hassanali, K., Liu, Y., & Solorio, T. (2012b). Evaluating NLP features for automatic prediction of language impairment using child speech transcripts. *Proceedings of Interspeech 2012,* 1339–1342. https://doi.org/10.21437/Interspeech.2012-321

Hassanali, K., Liu, Y., & Solorio, T. (2013). Using latent Dirichlet allocation for child narrative analysis. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing 2013,* 111–115.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hill, E. L. (2001). Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. *International Journal of Language & Communication Disorders, 36*(2), 149–171. https://doi.org/10.1080/13682820010019874

Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Medical Informatics, 10*(5), Article e36388. https://doi.org/10.2196/36388

Hunte, M. R., McCormick, S., Shah, M., Lau, C., & Jang, E. E. (2021). Investigating the potential of NLP-driven linguistic and acoustic features for predicting human scores of children's oral language proficiency. *Assessment in Education: Principles, Policy & Practice, 28*(4), 477–505. https://doi.org/10.1080/0969594X.2021.1999209

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24*(2), 84–91. https://doi.org/10.1044/0161-1461.2402.84

IBM. (n.d.). *What are naïve Bayes classifiers*? https://www.ibm.com/topics/naive-bayes

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jones, S., Fox, C., Gillam, S., & Gillam, R. B. (2019). An exploration of automated narrative analysis via machine learning. *PLOS ONE, 14*(10), Article e0224634. https://doi.org/10.1371/journal.pone.0224634

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology, 15*(2), 177–191. https://doi.org/10.1044/1058-0360(2006/017)

Kaderavek, J. N., & Sulzby, E. (2000). Narrative production by children with and without specific language impairment: Oral narratives and emergent readings. *Journal of Speech, Language, and Hearing Research, 43*(1), 34–49. https://doi.org/10.1044/jslhr.4301.34

Kapantzoglou, M., Fergadiotis, G., & Restrepo, M. A. (2017). Language sample analysis and elicitation technique effects in bilingual children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 60*(10), 2852–2864. https://doi.org/10.1044/2017_JSLHR-L-16-0335

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*(2), 161–176. https://doi.org/10.1177/026565909701300204

Klatte, I. S., Van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools, 53*(1), 1–16. https://doi.org/10.1044/2021_LSHSS-21-00026

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143).

Law, J., Boyle, J., Harris, F., Harkness, F., & Nye, C. (2000). Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature. *International Journal of Language & Communication Disorders, 35*(2), 165–188. https://doi.org/10.1080/136828200247133

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Leonard, L. B. (1998). *Children with specific language impairment*. MIT Press.

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38*(2), 415–425. https://doi.org/10.1044/jshr.3802.415

Lim, J., McCabe, P., & Purcell, A. (2017). Challenges and solutions in speech-language pathology service delivery across Australia and Canada. *European Journal for Person Centered Healthcare, 5*(1), Article 120. https://doi.org/10.5750/ejpch.v5i1.1244

Lüdtke, U., Bornman, J., De Wet, F., Heid, U., Ostermann, J., Rumberg, L., Van Der Linde, J., & Ehlert, H. (2023). Multidisciplinary perspectives on automatic analysis of children's language samples: Where do we go from here? *Folia Phoniatrica et Logopaedica, 75*(1), 1–12. https://doi.org/10.1159/000527427

Luk, G. (2017). Bilingualism. In B. Hopkins, E. Geangu, & S. Linkenauger (Eds.), *The Cambridge encyclopedia of child development* (2nd ed., pp. 385–391). Cambridge University Press. https://doi.org/10.1017/9781316216491.062

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets, 32*(4), 1993–2020. https://doi.org/10.1007/s12525-022-00605-4

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.

MacWhinney, B., & Fromm, D. (2022). Language sample analysis with TalkBank: An update and review. *Frontiers in Communication, 7,* Article 865498. https://doi.org/10.3389/fcomm.2022.865498

MacWhinney, B., Roberts, J. A., Altenberg, E. P., & Hunter, M. (2020). Improving automatic IPSyn coding. *Language, Speech, and Hearing Services in Schools, 51*(4), 1187–1189. https://doi.org/10.1044/2020_LSHSS-20-00090

McAllister, L., McCormack, J., McLeod, S., & Harrison, L. J. (2011). Expectations and experiences of accessing and participating in services for childhood speech impairment. *International Journal of Speech-Language Pathology, 13*(3), 251–267. https://doi.org/10.3109/17549507.2011.535565

McLean, L. K., & Woods-Cripe, J. W. (1997). The effectiveness of early intervention for children with communication disorders. In M. J. Guralnick (Ed.), *The effectiveness of early intervention*, (pp. 349–428). Brookes.

McLeod, S., & Baker, E. (2014). Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for children with speech sound disorders. *Clinical Linguistics & Phonetics, 28*(7–8), 508–531. https://doi.org/10.3109/02699206.2014.926994

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with CohMetrix* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports, 12*(1), Article 7166. https://doi.org/10.1038/s41598-022-11012-2

Merritt, D. D., & Liles, B. Z. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech and Hearing Research, 30*(4), 539–552. https://doi.org/10.1044/jshr.3004.539

Miller, J. F., & Iglesias, A. (2015). *Systematic Analysis of Language Transcripts* [Computer software]. Salt Software.

Montgomery, J. W. (2003). Working memory and comprehension in children with specific language impairment: What we know so far. *Journal of Communication Disorders, 36*(3), 221–231. https://doi.org/10.1016/S0021-9924(03)00021-2

Nippold, M. A. (2016). *Later language development: School-age children, adolescents, and young adults* (4th ed.). Pro-Ed.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). 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), 1247–1257. https://doi.org/10.1111/jcpp.12573

O'Callaghan, A. M., McAllister, L., & Wilson, L. (2005). Barriers to accessing rural paediatric speech pathology services: Health care consumers' perspectives. *Australian Journal of Rural Health, 13*(3), 162–171. https://doi.org/10.1111/j.1440-1854.2005.00686.x

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Obermeyer, Z., Powers, B., Vogel, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oliva, J., Serrano, J. I., Del Castillo, M. D., & Iglesias, Á. (2014). A methodology for the characterization and diagnosis of cognitive impairments—Application to specific language impairment. *Artificial Intelligence in Medicine, 61*(2), 89–96. https://doi.org/10.1016/j.artmed.2014.04.002

Olswang, L. B., & Prelock, P. A. (2015). Bridging the gap between research and practice: Implementation science. *Journal of Speech, Language, and Hearing Research, 58*(6), S1818–S1826. https://doi.org/10.1044/2015_JSLHR-L-14-0305

Overton, C., Baron, T., Pearson, B. Z., & Ratner, N. B. (2021). Using free computer-assisted language sample analysis to evaluate and set treatment goals for children who speak African American English. *Language, Speech, and Hearing Services in Schools, 52*(1), 31–50. https://doi.org/10.1044/2020_LSHSS-19-00107

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health, 9*(2), Article 010318. https://doi.org/10.7189/jogh.09.020318

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

Prud'hommeaux, E. T., Roark, B., & Black, L. M. (2011). Classification of atypical language in autism. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics,* 88–96.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(3), 252–264. https://doi.org/10.1109/34.75512

Reilly, J. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88*(2), 229–247. https://doi.org/10.1016/S0093-934X(03)00101-9

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*(1), 1–22. https://doi.org/10.1017/S0142716400008262

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata, 20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Scott, A., Gillon, G., McNeill, B., & Kopach, A. (2022). The evolution of an innovative online task to monitor children's oral narrative development. *Frontiers in Psychology, 13,* Article 903124. https://doi.org/10.3389/fpsyg.2022.903124

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), Article 20170362. https://doi.org/10.1098/rsta.2017.0362

Sheng, L., & McGregor, K. K. (2010). Lexical–semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 53*(1), 146–159. https://doi.org/10.1044/1092-4388(2009/08-0160)

Solorio, T., & Liu, Y. (2008). Using language models to identify language impairment in Spanish–English bilingual children. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing 2008* (pp. 116–117). https://doi.org/10.3115/1572306.1572337

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research, 64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Statistics Canada. (2018). *While English and French are still the main language spoken in Canada, the country's linguistic diversity continues to grow*. https://www150.statcan.gc.ca/n1/daily-quotidien/220817/dq220817a-eng.htm

Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for BERT model compression. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 4322–4331). https://doi.org/10.18653/v1/D19-1441

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

United Nations. (n.d.). *Universal declaration of human rights*. https://www.un.org/en/about-us/universal-declaration-of-human-rights

Verdon, S., Wilson, L., Smith-Tamaray, M., & McAllister, L. (2011). An investigation of equity of rural speech-language pathology services for children: A geographic perspective. *International Journal of Speech-Language Pathology, 13*(3), 239–250. https://doi.org/10.3109/17549507.2011.573865

Weindrich, D., Jennen-Steinmetz, C., Laucht, M., Esser, G., & Schmidt, M. H. (2000). Epidemiology and prognosis of specific disorders of language and scholastic skills. *European Child & Adolescent Psychiatry, 9*(3), 186–194. https://doi.org/10.1007/s007870070042

Westerveld, M. (2011). Sampling and analysis of children's spontaneous language: From research to practice. *ACQuiring Knowledge in Speech, Language and Hearing, 13*(2), 63–67.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology, 16*(3), 242–249. https://doi.org/10.3109/17549507.2013.871336

Westerveld, M. F., Gillon, G. T., & Miller, J. F. (2004). Spoken language samples of New Zealand children in conversation and narration. *Advances in Speech Language Pathology, 6*(4), 195–208. https://doi.org/10.1080/14417040400010140

Woodcock, C., Mittelstadt, B., Busbridge, D., & Blank, G. (2021). The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: Experimental study. *Journal of Medical Internet Research, 23*(11), Article e29386. https://doi.org/10.2196/29386

World Intellectual Property Organization. (2019). *The story of AI in patents*. https://www.wipo.int/tech_trends/en/artificial_intelligence/story.html

Wu, F., García-Perera, L. P., Povey, D., & Khudanpur, S. (2019). Advances in automatic speech recognition for child speech using factored time delay neural network. *Proceedings of Interspeech 2019,* 1–5. https://doi.org/10.21437/Interspeech.2019-2980

Zeigler, K., & Camarota, A. (2019). *67.3 million in the United States spoke a foreign language at home in 2018*. Center for Immigration Studies. https://cis.org/Report/673-Million-United-States-Spoke-Foreign-Language-Home-2018#5

Zhdanov, D., Bhattacharjee, S., & Bragin, M. A. (2022). Incorporating FAT and privacy aware AI modeling approaches into business decision making frameworks. *Decision Support Systems, 155,* Article 113715. https://doi.org/10.1016/j.dss.2021.113715