Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Word-finding abilities in children with posterior fossa tumors: Insights from lexical diversity in narrative language

Elisa Gottardi
International Doctorate for Experimental Approaches to Language and Brain, University of Groningen
Barbara Höhle
University of Potsdam
Natalie Boll-Avetisyan
University of Potsdam
Ditte Boeg Thomsen
University of Copenhagen
Jonathan Kjær Grønbæk
Rigshospitalet
Aliene Reinders
Center for Language and Cognition (CLCG)
Cheyenne Svaldi
Center for Language and Cognition (CLCG)
Karin Persson
Lund University
René Mathiasen
Rigshospitalet
Andrea Carai
Bambino Gesù Children's Hospital
Bianca Andreozzi
Bambino Gesù Children's Hospital
Angela Mastronuzzi
Bambino Gesù Children's Hospital
Barry Pizer
University of Liverpool
Colin Thorbinson
Alder Hey Children's NHS Foundation Trust
Kristian Aquilina
Great Ormond Street Hospital
Eelco Hoving

Princess Máxima Center

Marianne Juhler

Aarhus University Hospital

Roel Jonkers

Center for Language and Cognition (CLCG)

Vânia de Aguiar

vania.de.aguiar@rug.nl

Center for Language and Cognition (CLCG)

Research Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11554564/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Word-finding abilities in children with posterior fossa tumors: Insights from lexical diversity in narrative language

Elisa Gottardi[1,2,3,4], Barbara Höhle[2], Natalie Boll-Avetisyan[2], Ditte Boeg Thomsen[5], Jonathan Kjær Grønbæk[6], Aliene Reinders[3,4], Cheyenne Svaldi[3,7], Annet Kingma[8], Karin Persson[9], René Mathiasen[6], Christine Dahl[6], Andrea Carai[10], Bianca Andreozzi[11], Angela Mastronuzzi[12,13], Barry Pizer[14,15], Colin Thorbinson[15], Kristian Aquilina[16], Eelco Hoving[17], Marianne Juhler[18], Roel Jonkers[3,4] and Vânia de Aguiar[3,4,19*]

[1]International Doctorate for Experimental Approaches to Language and Brain (IDEALAB), University of Potsdam, Potsdam, Germany; University of Groningen, Groningen, The Netherlands; Macquarie University, Sydney, Australia.

[2]Department of Linguistics, University of Potsdam, Potsdam, Germany.

[3]Center for Language and Cognition Groningen (CLCG), University of Groningen, Groningen, The Netherlands.

[4]Research School of Behavioral and Cognitive Neurosciences (BCN), University of Groningen, Groningen, The Netherlands.

[5]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark.

[6]Department of Paediatric and Adolescent Medicine, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark.

[7]Department of Language and Cognition - Psychology and Language Sciences, University College London, London, UK

[8]Department of Pediatrics/Pediatric Oncology, University Medical Center Groningen, Groningen, The Netherlands.

[9]Department of Health Sciences, Lund University, Lund, Sweden.

[10]Neurosurgery Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy.

[11]Management and Diagnostic Innovations & Clinical Pathways Research Area, Neurorehabilitation and Adapted Physical Activity Day Hospital, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy.

[12]Area of Hematology/Oncology, Cell Therapy, Gene Therapies and Hematopoietic Transplant, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy.

[13]Department of Life Sciences and Public Health, Università Cattolica del Sacro Cuore, Milan, Italy.

[14]University of Liverpool, Liverpool, United Kingdom.

[15]Department of Oncology, Alder Hey Children's NHS Foundation Trust, Liverpool, UK.

[16]Department of Neurosurgery, Great Ormond Street Hospital, London, UK.

[17]Oncological Pediatric Neurosurgery, Princess Maxima Center for Pediatric Oncology, Utrecht, The Netherlands.

[18]Department of Neurosurgery, Aarhus University Hospital, Aarhus, Denmark

[19]Department of Radiation Oncology, University Medical Center Groningen, Groningen, The Netherlands.

*Correspondence address:

Vânia de Aguiar

vania.de.aguiar@rug.nl

Abstract

Word-finding difficulties are reported in children with posterior fossa tumors (PFTs) but have primarily been examined using isolated word-retrieval tasks, such as picture-naming. However, lexical access in everyday communication occurs in context, highlighting the need for more ecologically valid approaches. This study investigated word retrieval in narrative language using lexical diversity (LD) measures in children with PFTs preoperatively and during the acute postoperative phase, to characterize how tumor- and surgery-related changes may affect lexical access in spontaneous language production. We additionally examined whether poorer preoperative performance was linked to the subsequent development of posterior fossa syndrome, focusing on the occurrence of postoperative speech impairment (POSI; i.e., transient mutism or severe speech reduction). Given the typically short language samples and the broad age range of this population, different LD measures were evaluated to assess their developmental sensitivity and the confounding effect of sample length. Two measures emerged as robust indices for short and age-variable samples: moving average type-token ratio and measure of textual LD. Contrary to expectations and prior studies on isolated-word retrieval, no significant differences in LD were observed between patients and controls, either pre- or postoperatively, nor between patients who later developed POSI and those who did not. This discrepancy may reflect differences between isolated and contextual word retrieval, as timed tasks impose greater processing speed demands. Conversely, when given sufficient time, patients appear to perform comparably to controls in narrative contexts.

Keywords
Posterior fossa syndrome, Cerebellar mutism syndrome, Mutism, Infratentorial neoplasms, Preoperative language impairment, Language disorders

Introduction

Posterior fossa tumors (PFTs) are the most common brain tumors during childhood [1,2], involving one or more structures within the posterior fossa, which include the cerebellum, the brainstem (i.e., pons, medulla, and midbrain), cerebello-pontine angle, and the fourth ventricle [3]. Most PFTs are first treated by surgical resection, often followed by adjuvant postoperative treatments, such as radio- and/or chemotherapy. However, both the tumor itself and its treatments, particularly surgical resection, may entail severe and persistent neuropsychological sequelae [4]. More specifically, children with PFTs can present language impairments that impact different language domains (e.g., lexical-semantics, morphosyntax, phonology). These impairments show large interindividual heterogeneity which may persist over time [5].

A possible postoperative complication to PFTs resection in childhood is the development of a complete absence of speech (i.e., mutism) or severe reduction of speech, in which the child's language output is limited to short sentences or single words produced only after intense stimulation [6]. Mutism and reduced speech can be collectively labelled as postoperative speech impairment (POSI) [3]. This is a transient condition with approximately 30% incidence that can arise within days of surgery and can vary in duration but generally resolves within months [3]. POSI is often accompanied by other symptoms such as irritability, ataxia, and hypotonia [6,7], and these symptoms are collectively known as Posterior Fossa Syndrome (PFS) [8,9]. A systematic review [5] suggests that the presence of PFS may be associated with more severe or more frequent postoperative language impairments, spanning lexical-semantic, phonological, verbal learning, and literacy outcomes. The present study further examines potential differences in language impairments between patients with PFS, focusing on the emergence of postoperative mutism or reduced speech, and those with habitual postoperative speech.

Very few studies have assessed children's language abilities *before* PFT surgery [10–12], despite its importance in distinguishing tumor-related from surgery-related impairments in this population, and for accurately interpreting postoperative outcomes [5]. The few studies that included preoperative assessments report that patients may demonstrate speech and language impairments even before tumor resection, including, for example, apraxia of speech, ataxic dysarthria, reduced mean length of utterance, and word-finding difficulties [10,11,13,14]. Interestingly, a study and its follow up [10,13] reported an association between the presence of preoperative language impairments and a higher risk of developing POSI after surgery. Among

the reported language impairments, there were phonological disorders [10,13], impaired lexical retrieval, difficulties in recalling frequently used words, and reduced mean length of utterance [10]. In line with this hypothesis, a recent study [15] presented a comprehensive analysis of preoperative spontaneous language production in children with PFTs, examining multiple linguistic domains (i.e., semantic, lexical, morphosyntactic, and phonological levels, which respectively refer to the knowledge of meanings, word structure, grammar, and speech sounds). They compared language profiles of patients who later developed POSI from those with postoperative habitual speech but found no significant preoperative differences between the two groups. However, before surgery, children who later developed POSI produced more unintelligible speech, suggesting that potential preoperative indicators of later POSI development may be related to speech rather than language impairments. To further investigate the potential association between preoperative language impairments and the subsequent development of POSI, the present study examined preoperative word-finding performance in patients who later developed POSI, patients who retained habitual speech, and healthy controls. Furthermore, because most previous studies lacked preoperative assessments and focused on late postoperative stages, such as the chronic phase or even more than five years after surgery (see [5]), we also investigated word finding during the acute postoperative phase (i.e., one to four weeks after surgery) in patients for whom such data were available.

Word-finding difficulties in children with PFTs

Word-finding ability, which refers to the capacity to quickly and accurately retrieve the appropriate word from one's mental lexicon [11,16], can be impaired in the children with PFTs [5,10–12,17], consistently with evidence implicating the role of the cerebellum in word production [18–20]. Word-finding difficulties are typically evaluated by testing word retrieval in isolation, with picture-naming tasks [21]. Using picture naming data collected within the European study on Cerebellar Mutism Syndrome (CMS) [22], a study examined naming speed and accuracy in 148 children with PFTss before surgery [11]. Results showed that, before surgery, 37% of patients were slower, 24% less accurate, and 16% both slower and less accurate at naming than healthy peers. In a follow-up study based on an expanded dataset that included part of the original sample, authors compared naming accuracy and speed in 184 children with PFTs, tested before and 1 to 4 weeks after surgery [12]. Results showed that, despite 95% of patients performing two standard

deviations slower than age-specific norms after surgery, there was no significant difference in naming performance before and after tumor resection. Word-finding difficulties in children with PFTs have also been reported beyond the European CMS data: a systematic review described lexical-semantic impairments including reduced vocabulary, commonly tested via picture naming tasks, with most studies reporting on postoperative assessments [5].

While picture naming data offers valuable insight into word retrieval in isolation, it does not adequately account for word retrieval in naturalistic settings. One way to address this is through analysis of (semi-)spontaneous language samples, which may be elicited via spontaneous conversations, picture-sequence descriptions, story (re-)telling, or a combination of those (e.g., [23]). This type of task offers high ecological validity, as it more closely resembles everyday communication, with a cognitive demand more similar to those children experience in daily life. Critically, (semi-)spontaneous language tasks deliver similar diagnostic value as formal testing [21,24] but might be preferable in clinical settings, as their administration is less time consuming and they are perceived as less stressful by the patient, especially when they are easily fatigued after undergoing multiple clinical procedures [24]. In addition, while word retrieval in picture naming is at the single-word level, spontaneous language requires additional grammatical and memory processes, which may increase processing demands and potentially affect naming performance [24]. Another factor that may impact naming performance is impaired processing speed, which is often reported in children with PFTs [4]. Such impairment may be particularly evident in picture naming tasks that impose a time pressure (as employed by [11,12]). Together, these findings highlight the complementary nature of picture naming and spontaneous language tasks, underscoring the importance of considering also the latter to gain a comprehensive understanding of word retrieval in children with PFTs, which is the goal of this study.

People with word-finding difficulties may produce more fillers, paraphasias, non-referential terms, and neologisms in their discourse [25]. Additionally, they may also rely on more readily accessible words and reuse the same lexical entries in their communication [26], with a possible preference for general words with broader meaning and usage instead of more specific ones. Such difficulties may hence reduce the speaker's range and word types produced, defined as lexical diversity (LD). While LD reflects access to one's lexicon, it can also be employed as a proxy to evaluate the impact of word-finding difficulties in spontaneous language [25]. For example, a study found that long-term survivors of pediatric cerebellar tumors produced a

significantly higher proportion of general all-purpose verbs (e.g., 'I *did* school' instead of 'I *attended* school') than controls [23]. Moreover, another study reported reduced LD in language samples several years after PFT surgery, as measured by type-token ratio (TTR), which calculates the proportion of unique words (types) to the total number of words (tokens) in a sample [27]. TTR was also used in [15], alongside other measures (e.g., lexical accuracy and word frequency), to investigate lexical processing within a broader assessment of preoperative language impairments in children with PFTs. They reported no significant differences between patients who developed POSI and those who did not. Yet, TTR is not the only measure of LD and presents limitations, such as its sensitivity to sample length. Indeed, LD can be quantified using different approaches, as described in the following paragraph, making it important to determine which measure best characterizes lexical retrieval impairments. To our knowledge, in this population, no studies have examined preoperative LD as a standalone measure, compared different LD approaches to determine which best characterizes lexical retrieval impairments, or investigated whether reduced preoperative LD may serve as an indicator of subsequent POSI development. The present study addresses these gaps by comparing multiple LD measures and assessing whether preoperative reduced LD is linked to subsequent POSI development. In addition, to further characterize lexical retrieval difficulties in children with PFTs, we investigated their manifestation in narrative language during the acute postoperative phase (i.e., one to four weeks after tumor resection).

Lexical Diversity Measures

LD can be measured with several methods, with TTR traditionally being the one used most, also in studies on children with PFTs (see [15,27,28]). However, TTR presents limitations in terms of reliability and ecological validity. For example, it is highly sensitive to sample length: TTR values decrease as sample length increases, leading larger samples to appear less lexically diverse than smaller ones [29–31]. Possible solutions to this are adopting a standardized sample size, such as 300 tokens, which is hard to achieve in children with posterior fossa tumors, or, alternatively, set the shortest sample in a corpus as the standard length and truncate all longer samples accordingly [29]. However, a truncated sample may not fully represent an individual's language abilities, as the arbitrary selection of utterances can influence results. Additionally, if the sample is too short, it may not meet the minimum length required for analysis, risking exclusion from the study [30]. To overcome this challenge and find methods to accurately capture LD in language

samples by children with PFTs, we will include alternative measures of LD less affected by sample length, such as moving average type-token ratio, hypergeometric distribution D, and the measure of textual LD.

Moving average type-token ratio (MATTR; [32]) computes TTRs of sequential, partially overlapping sections of language samples. MATTR selects a segment of *x* tokens and calculates the TTR within that segment. The segment then shifts forward one word at a time (e.g., token 2 to x+1, then 3 to x+2, etc.) until the entire sample is processed. The final score is determined by averaging all computed TTRs [29]. Concerning the hypergeometric distribution D (HD-D), for each word type in a text, it calculates the probability that the type appears at least once in a random 42-token sample[1] drawn without replacement. These probabilities are then summed, giving the expected number of different word types in such a sample, reflected in the HD-D index [34]. As a consequence of this computational procedure, HD-D requires language samples of at least 42 words. Lastly, the measure of textual lexical diversity (MTLD; [34]) is an LD index that measures the average length of sequential word strings that sustain a TTR above a predefined threshold. In a string of text, the TTR value is updated after each incoming word. If words repeat, LD drops. A string is interrupted at the moment TTR values drop below a set threshold (typically 0.72; see [34], for how this value was empirically derived), and at that point the length of the string is calculated. This procedure starts again for the next string of text, and finally the average length across all strings is calculated and corresponds to the MTLD [34].

A LD measure can be considered effective when it captures lexical ability without confounds (e.g., sample length) and shows a clear age-related progression across the developmental period being examined [35]. As mentioned above, previous research reported how TTR is negatively influenced by sample length [29–31]. TTR has also been shown to lack developmental sensitivity [31,35] and be clinically unreliable, as it can misclassify typically developing children and fail to flag children with documented language impairments [31]. In contrast, a study compared TTR, HD-D, MATTR, and MTLD language samples from 4 to 6 year-old children with and without developmental language disorders, showing that MATTR provided the most robust reflection of LD without influence from construct-irrelevant factors such

[1] HD-D was introduced by [33] as a probability-based alternative to voc-D, which estimates lexical diversity using repeated random samples of 35-50 tokens. Because their analyses showed that any sample size within this range yields comparable results [33], they proposed adopting 42 tokens, the midpoint of the range, as the fixed sample size for computing HD-D [34].

as differences in sample length. Though slightly more variable, MTLD performed similarly to MATTR, whereas HD-D and TTR were reported to systematically covary with length [36]. Similar findings were reported by [29] in their study on short discourse samples from people with aphasia, suggesting MATTR and MTLD as reliable measures of LD in this population. Furthermore, adding to MATTR strength, another study showed that MATTR presented no relationship to sample or utterance length, when controlling for age, in either 4- to 9-year-old children with and without developmental language disorders [35]. Concerning the age-growth trajectory, MATTR demonstrated developmental sensitivity in 4- to 9-year-olds [35] and 2- to 6-year-olds [31], though in the latter this measure showed a nonlinear developmental pattern, rising early in development and then leveling off around age four. To the best of our knowledge, no studies investigated the developmental sensitivity of MTLD and HD-D. Moreover, given the wide variation in assessment age among children with PFTs (see [5]), it becomes necessary to examine the age-growth trajectory across a broader developmental span than what has been explored in previous studies. Hence, the present study examined how sample length, utterance length, and age influence different LD measures (i.e., TTR, MATTR, HD-D, and MTLD) in children with PFTs compared to healthy peers, expanding existing evidence in LD research by evaluating these metrics in a previously unexplored population.

Current Research Questions

The present study addresses the following research questions:

1. Which LD measure is most appropriate for language samples from children with PFTs?
2. Do children with PFTs show reduced LD compared to healthy children at the preoperative and acute postoperative stages?
3. Can preoperative LD predict the development of POSI?

To address the first research question, we examined whether LD measures are confounded by sample length, and whether they present developmental sensitivity in relation to age and mean length of utterance in words (MLUw), a reliable measure of language development [37]. For the second research question, once the most suitable LD measure(s) were identified, we compared patients with healthy controls at the pre- and post-operative time points, conducting both group- and individual-case analyses. Finally, for the third research question, we assessed whether children

who later developed POSI already showed lower preoperative LD than those who produced habitual postoperative speech, again considering both group- and individual-case analyses.

Method

Participants

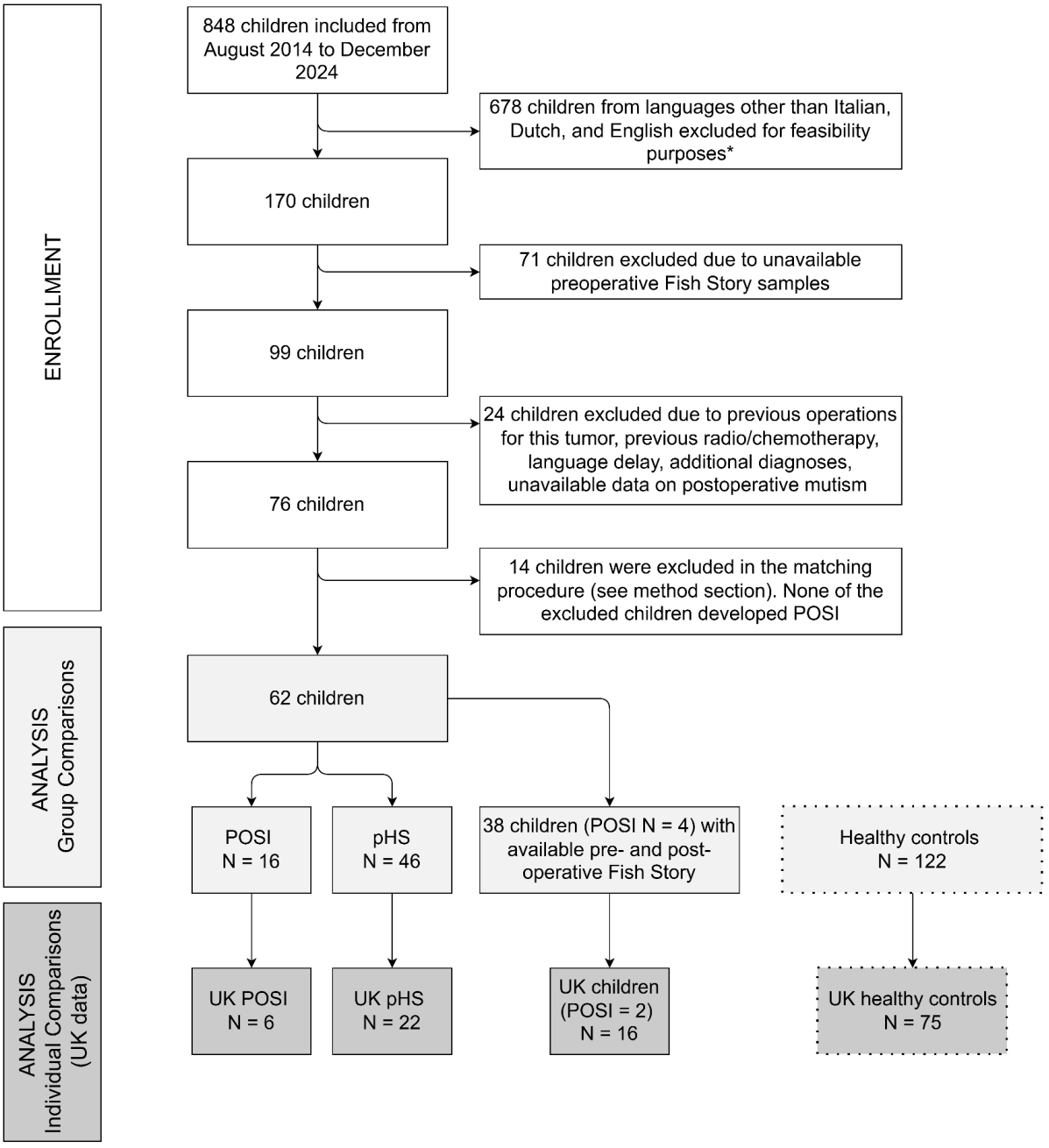
Semi-spontaneous language samples from children with PFTs were obtained from the European Study on Cerebellar Mutism Syndrome (CMS; [22]), a multi-center study collecting clinical and language and speech data on children undergoing PFT surgery across approximately 40 European centers in 13 European countries. For the present study, we included patient data from Dutch, Italian, and British centers. The CMS study was registered on November 24, 2014, with ClinicalTrials.gov (registration number: NCT02300766). Control data are acquired as follows: for the Dutch language, they were collected at the University of Groningen (Netherlands); for Italian, they were collected at the Bambino Gesù hospital (Rome, Italy); for English, they were retrieved from the online repository by Frizelle and colleagues ([38], https://osf.io/fvyqh). Exclusion criteria for the patient group were: (1) absence of a preoperative language sample; (2) history of previous surgery for a PFT; (3) prior radiotherapy and/or chemotherapy; (4) documented neurological, neurodevelopmental, psychiatric, or learning disorders, or a diagnosis/treatment of pre-existing language disorders; and (5) missing information on POSI status at the postoperative assessment conducted 1-4 weeks after surgery. For the control group, children were excluded if they had a history of neurological, neurodevelopmental, psychiatric, or learning disorders, or a diagnosis/treatment of pre-existing language disorders. Participants with reported speech impairments or who received therapy for such speech impairments were not excluded, as long as no underlying language disorder was present, since the focus of the current study is on language. A summary of participant selection can be found in Figure 1. Seventy-six children were eligible according to the inclusion criteria. However, due to the focus of the current study on language, a further selection step was performed to match patients and control groups, as well as patients who developed POSI and those who did not, on the basis of language and sex, to facilitate group comparisons.

The current study aimed to compare pre- and post-operative performance between children with PFTs and healthy controls, both at a group level and at an individual level, given the reported

individual differences in this population (see Svaldi et al., 2024). Participants' age range was from 3 years and 2 months to 17 years and 11 months. The samples included in each analysis are described as follows. For the preoperative group comparison, we included children with POSI, children with habitual speech (pHS), and healthy controls matched for age and sex, as displayed in Table 1. An individual case-to-control analysis was performed on a subset of UK patients for whom we had a sufficiently large pool of eligible and matched controls. Specifically, each patient needed to be compared to five healthy controls, matched for age, sex, and language. Out of the subset of UK patients, twenty-eight of the twenty-nine were included; six of these later developed POSI. One patient was excluded because no sufficient matched controls were available. Out of the eighty-one UK healthy controls available, we included seventy-five for the pre-operative individual case-to-control analysis. The same matched controls were used for multiple patients when they shared the same age and sex.

Figure 1

Participant selection



*Feasibility purposes refer the limited availability of proficient speakers of the patients' languages to transcribe and analyze Fish Story recordings.

Note. *POSI* = patients who later developed postoperative speech impairments (i.e., mutism or reduced speech), *pHS* = patients who did not develop postoperative speech impairments and present postoperative habitual speech, *UK* = English-speaking participants from the United Kingdom

Table 1

Demographics of the preoperative group comparison, divided by group

	N	Age M (SD) years;months	% of males	Tumor types (%)
POSI patients	16 (UK = 6; IT = 7; NL = 3)	9;1 (3;6)	60%	Medulloblastoma (56.3), Pilocytic Astrocytoma (12.5), Ependymoma (6.3), Other types (12.5), n.a. (12.5)
pHS patients	46 (UK = 23; IT = 11; NL = 12)	9;5 (3;5)	63%	Medulloblastoma (17.4), Pilocytic Astrocytoma (58.7), Ependymoma (6.5), Other types (13), n.a. (4.3)
Healthy controls	122 (UK = 81; IT = 26; NL = 15)	8;10 (3;9)	53%	-

Note: *POSI* = patients who later developed postoperative speech impairments (i.e., mutism or reduced speech), *pHS* = patients who did not develop postoperative speech impairments and present postoperative habitual speech, *UK* = English-speaking participants from the United Kingdom, *IT* = Italian-speaking participants from Italy, *NL* = Dutch-speaking participants from the Netherlands, *n.a.* = Not answered.

For the analysis of postoperative language samples, we included thirty-eight patients for whom data were available after surgery (1-4 weeks after surgery). For the group comparison, we made no comparison between patients who developed POSI and those with pHS, as only four POSI patients presented both pre- and post-operative assessments. The included thirty-eight patients (grouping pHS and POSI together) were compared to the same one-hundred twenty-two controls as in the preoperative analysis, as displayed in Table 2. As for the preoperative data, the individual case-to-control analysis was performed only on the English-speaking patients' subset. We included sixteen of the seventeen English-speaking patients with available post-operative data, of whom two developed POSI. One patient was excluded because no sufficient matched controls were available. Out of the eighty-one UK healthy controls available, we included fifty-five for the post-operative individual case-to-control analysis. The same matched controls were used for multiple patients when they shared the same age and sex.

Table 2

Demographics of the post-operative group comparison, divided by group

	N	Age M (SD) years;months	% of males	Tumor types (%)
Patients	38 (4 POSI) (UK = 17; IT = 10; NL = 11)	9;7 (3;7)	66%	Medulloblastoma (18.4), Pilocytic Astrocytoma (57.9), Ependymoma (10.5), Other types (10.5), n.a. (2.6)
Controls	122 (UK = 81; IT = 26; NL = 15)	8;10 (3;9)	53%	-

Note: *Patients* = children with PFTs, *POSI* = patients who developed postoperative speech impairments (i.e., mutism or reduced speech), *UK* = English-speaking participants from the United Kingdom, *IT* = Italian-speaking participants from Italy, *NL* = Dutch-speaking participants from the Netherlands, *NA* = Not answered.

Corpus Collection and Data Processing

All participants narrated the *Fish Story* from the *Expression, Reception, and Recall of Narrative Instrument (ERRNI*; [39]) in their respective language (i.e., Italian, English, or Dutch). The *Fish Story* task consisted of a wordless picture book representing a boy's visit to a pet store. Examiners instructed children to first look through the book in silence and then narrate the story with the pictures as a guide. Examiners were to provide minimal assistance, with allowed general prompts (e.g., "and what happens here?") or occasional forms of encouragement (e.g., "uh uh"). The patients' preoperative language samples were audio recorded by the appointed clinical staff at the respective hospitals (e.g., speech therapists, nurses) in the week preceding the scheduled surgery, and uploaded on the European CMS Study database. For the current study, patients' IDs were pseudo-anonymized. Dutch controls' language samples were collected by PhD and master's students in Linguistics, while Italian samples by speech therapists and neuropsychologists from the partner hospital. The audios were then transcribed via the program ELAN [40] by a native or a highly proficient speaker of the target language and checked by an independent transcriber. The transcriptions of the English controls' data were available in the online repository by [38].

In accordance with the approach adopted by [38], elliptic, prompted, or mimicked (i.e., verbatim repetitions of tester or caregiver) utterances, together with general comments (e.g. "I don't want to do this anymore") were excluded. Similarly, immediate word repetitions without communicative intentions and interjections were omitted. Phonological paraphasias were transcribed in their intended orthographic form based on the transcribers' best judgment.

The different LD measures (i.e., TTR, HD-D, MTLD, MATTR) were then extracted from the preprocessed transcriptions with the R package *koRpus* [41]. The window of MATTR was set at 8 tokens, considering the shortest length reported in the data (9 tokens in total). One Dutch patient who belonged to the POSI group was excluded from the analysis because all the produced utterances were either prompted, elliptic, or mimicked. In addition, HD-D could not be computed for participants whose samples did not meet the minimum length requirement of 42 words. In the preoperative comparison, this affected thirteen participants (age M = 4;6 years, age SD = 9 months, age range = 4;1 to 6;6 years): seven children belonged to the control group, three to the POSI group, and three to the habitual speech group. In the postoperative comparison, this affected nine participants (age M = 4;3 years, age SD = 5 months, age range = 4;1 to 5;3 years): the same seven children belonging to the control groups, one to the POSI group, one to the habitual speech group[2]. By contrast, the other LD measures (i.e., TTR, MATTR, and MTLD) do not impose a minimum sample length and were therefore calculated for all participants.

For the correlation analyses concerning the first research question, the total number of words was extracted from each preprocessed sample, while MLUw was calculated dividing the total number of words by the total number of utterances.

Statistical Analyses

As an initial step, we computed partial correlations between the four LD measures (TTR, HD-D, MTLD, and MATTR) and sample length (i.e., the total number of words), controlling for age. This analysis was intended to assess the potential confounding influence of sample length and to replicate expected patterns. Next, we computed Pearson's correlations between the four LD measures and (i) age and (ii) MLUw. The partial correlation and the Pearson's correlations were

[2] The two patients who did not meet the minimum length requirement of 42 words in the postoperative assessments were also among those patients who did not meet this requirement in the preoperative assessment. Controls were tested only once.

conducted separately for the control group and the preoperative patient group. Finally, we performed Steiger's z tests [42] to compare the magnitude of the obtained Pearson's correlations, in order to characterize the relative developmental sensitivity of the included LD measures. Because Steiger's test requires the compared correlations to be computed from the same sample of participants, participants with missing HD-D values due to samples shorter than the minimum required length (i.e., 42 tokens) were excluded before performing any of the above-mentioned analyses. The excluded participants were 7 controls and 6 patients.

After reinstating the participants who had been excluded from the correlation analyses, we employed group comparisons for all languages and individual-case statistics for English-speaking participants to examine whether individuals with PFT show reduced LD pre- and/or postoperatively when compared with healthy controls. Only the LD measures identified as more appropriate from the abovementioned correlation analyses were considered. For the group comparisons, we fitted general linear models (with LD measure(s) as the dependent variable) to the preoperative language samples and repeated the same analyses for the postoperative language samples. Group was the predictor of interest, whereas age and language were included as covariates. Language was coded using the default treatment (dummy) coding, with English as the reference category. To correct for multiple comparisons, the post hoc Benjamini-Hochberg procedure was used to control the false discovery rate (FDR). Due to the control data availability, the individual case analyses were conducted on a subset of English-speaking patients (preoperative analysis N = 28; postoperative analysis N = 16) and controls (preoperative analysis N = 75; postoperative analysis = 55). In particular, each English-speaking patient's LD score was compared with the mean and standard deviation of five age- and sex-matched English-speaking healthy controls. The two-tailed test of deficit was employed, as described by [43], using the R package *singcar* [44]. A two-tailed test was preferred to investigate patients heterogeneity, which is often reported in this population. The results from the general linear models and individual-case analyses on preoperative data were also used to address the final research question, whether patients who later develop POSI perform worse preoperatively than those who do not. To this end, the *Group* factor was coded using orthogonal Helmert contrasts before fitting the general linear models. Under this coding scheme, the intercept represents the grand mean (i.e., the overall mean across groups). Given the ordering of the factor levels, the first contrast compared the POSI and habitual speech patient groups, whereas the second contrasted the pooled patient groups (POSI

and habitual speech) with the control group. We then further analyzed the individual-case results from the test of deficit, considering possible differences between POSI and habitual speech patients in frequency of findings of high/low LD scores relative to controls. All analyses were performed in RStudio.

Results

Effects of sample length, MLUw, and age on lexical diversity measures

We first analyzed the partial correlations between the four LD measures (TTR, HD-D, MATTR, MTLD) and sample length, controlling for age. The analyses were computed separately for the patient and control groups. Results are summarized in Table 3. Patient and control groups showed consistent results concerning the influence of length on the LD measures: TTR was the only measure with a significant negative correlation with length, after controlling for age, for both patient and control data. All other measures (HD-D, MATTR, MTLD) were not significantly related to sample length.

Table 3

Partial correlations between lexical diversity measures and sample length as number of total words, controlling for age.

	Control group (N = 115)		**Patient group** (N = 55)	
	r	*p*	*r*	*p*
TTR	-0.693	**<.001**	-0.629	**<.001**
HD-D	0.138	0.142	0.243	0.077
MATTR	0.048	0.610	0.230	0.094
MTLD	0.041	0.664	-0.016	0.908

Note: *TTR* = type-token ratio; *HD-D* = hypergeometric distribution D; *MTLD* = measure of textual lexical diversity; *MATTR* = moving average type-token ratio set at 8-word window.

To investigate the developmental sensitivity of the measures, we ran Pearson's correlations between the four LD measures and (i) age and (ii) MLUw, separately for patient and control groups. Results are summarized in Table 4 and presented in Figure 2 and 3. For both the control and patient groups, TTR was not significantly correlated with age but showed a significant

negative correlation with MLUw. In contrast, HD-D, MATTR, and MTLD were all significantly positively correlated with age and MLUw in the control group. In the patient group, however, only MATTR showed significant positive correlations with age and MLUw, whereas HD-D and MTLD were not significantly correlated with either variables.

Table 4

Correlations of lexical diversity measures to age and MLUw

	Correlations with Age				**Correlations with MLUw**			
	Control group		**Patient Group**		**Control group**		**Patient Group**	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
TTR	-0.088	0.349	-0.072	0.602	-0.375	**< .001**	-0.322	**0.017**
HD-D	0.545	**< .001**	0.253	0.063	0.365	**< .001**	0.186	0.174
MATTR	0.476	**< .001**	0.324	**0.016**	0.388	**< .001**	0.434	**< .001**
MTLD	0.516	**< .001**	0.229	0.092	0.309	**< .001**	0.243	0.074

Note: *TTR* = type-token ratio; *HD-D* = hypergeometric distribution D; *MTLD* = measure of textual lexical diversity; *MATTR* = moving average type-token ratio set at 8-word window; MLUw = mean length of utterance in words.

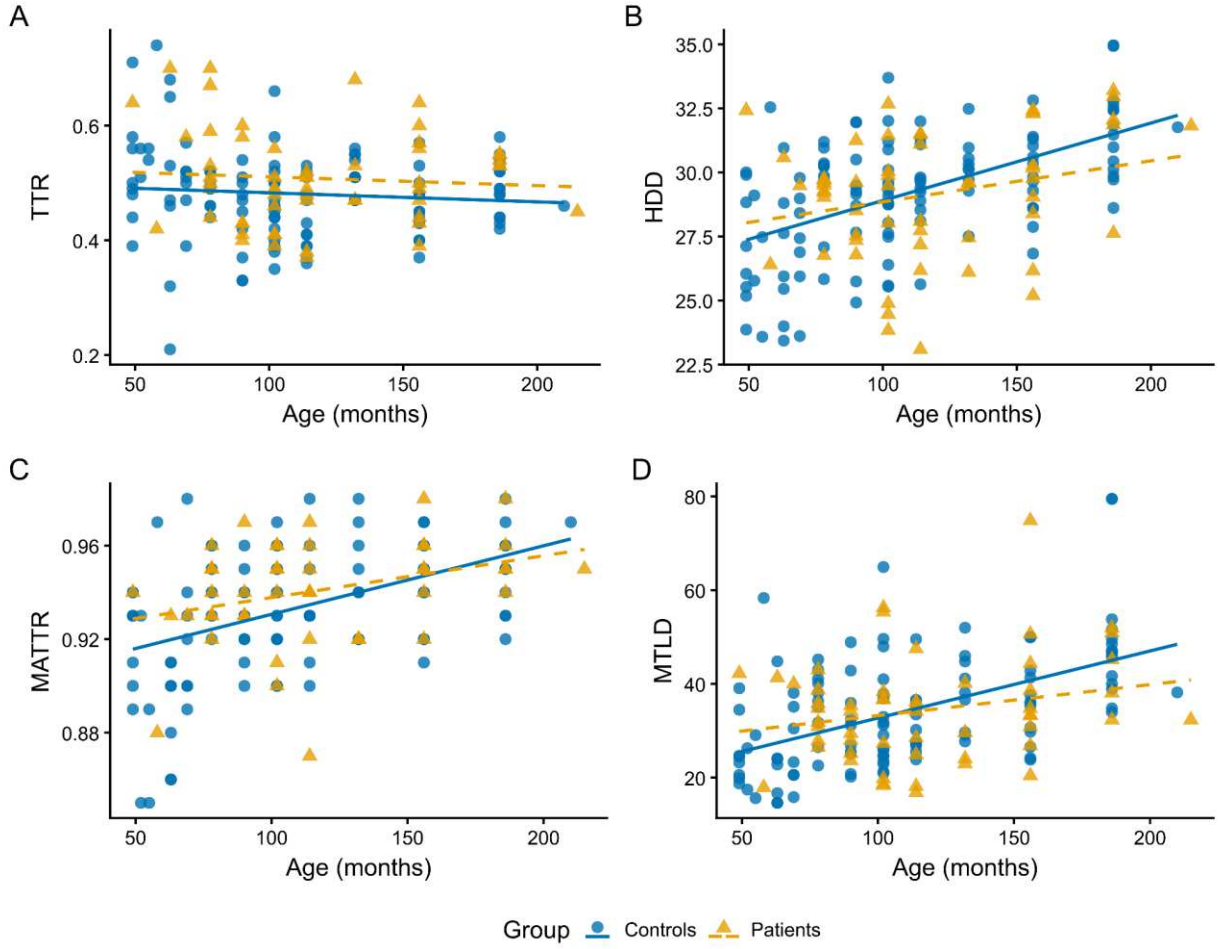
Finally, we compared the magnitude of the correlations between the LD measures and age and MLUw in the control and patient groups using Steiger's z tests. In the control group, the correlations between age and HD-D (Steiger's $z = 6.411$, $p < .001$), MTLD (Steiger's $z = 6.164$, $p < .001$), and MATTR (Steiger's $z = 4.862$, $p < .001$) were all significantly stronger than the correlation between age and TTR. The same pattern was observed in the patient group, where the correlation between age and HD-D (Steiger's $z = 2.311$, $p = 0.021$), MTLD (Steiger's $z = 2.304$, $p = 0.021$), and MATTR (Steiger's $z = 2.150$, $p = 0.032$) were significantly stronger than that between age and TTR, despite the correlations between HD-D and MTLD with age did not reach significance in the patient group. In contrast, the correlations of HD-D, MTLD, and MATTR with age did not differ significantly from one another in either the control or the patient group.

In the control group, the correlations between MLUw and HD-D (Steiger's $z = 7.255$, $p < .001$), MTLD (Steiger's $z = 6.781$, $p < .001$), and MATTR (Steiger's $z = 6.509$, $p < .001$) were all significantly stronger than the correlation between MLUw and TTR. Similarly, in the patient group, the correlations between MLUw and HD-D (Steiger's $z = 3.654$, $p < .001$), MTLD (Steiger's $z = 4.395$, $p < .001$), and MATTR (Steiger's $z = 4.222$, $p < .001$) were also all significantly stronger than that between MLUw and TTR. As for age, the correlations of HD-D, MTLD, and MATTR with MLUw did not differ significantly from one another in either the control or the patient group.

The absence of sample length effects in both groups, together with the stronger associations of HD-D, MTLD, and MATTR with age and MLUw than TTR in the control group, supports the robustness of these measures. In the patient group, MATTR was the only measure significantly correlated with both age and MLUw. Together, these findings suggest that MATTR is the most developmentally sensitive LD measure, followed by MTLD and HD-D. However, because HD-D requires a minimum sample length, only MATTR and MTLD were retained for the subsequent analyses.

Figure 2

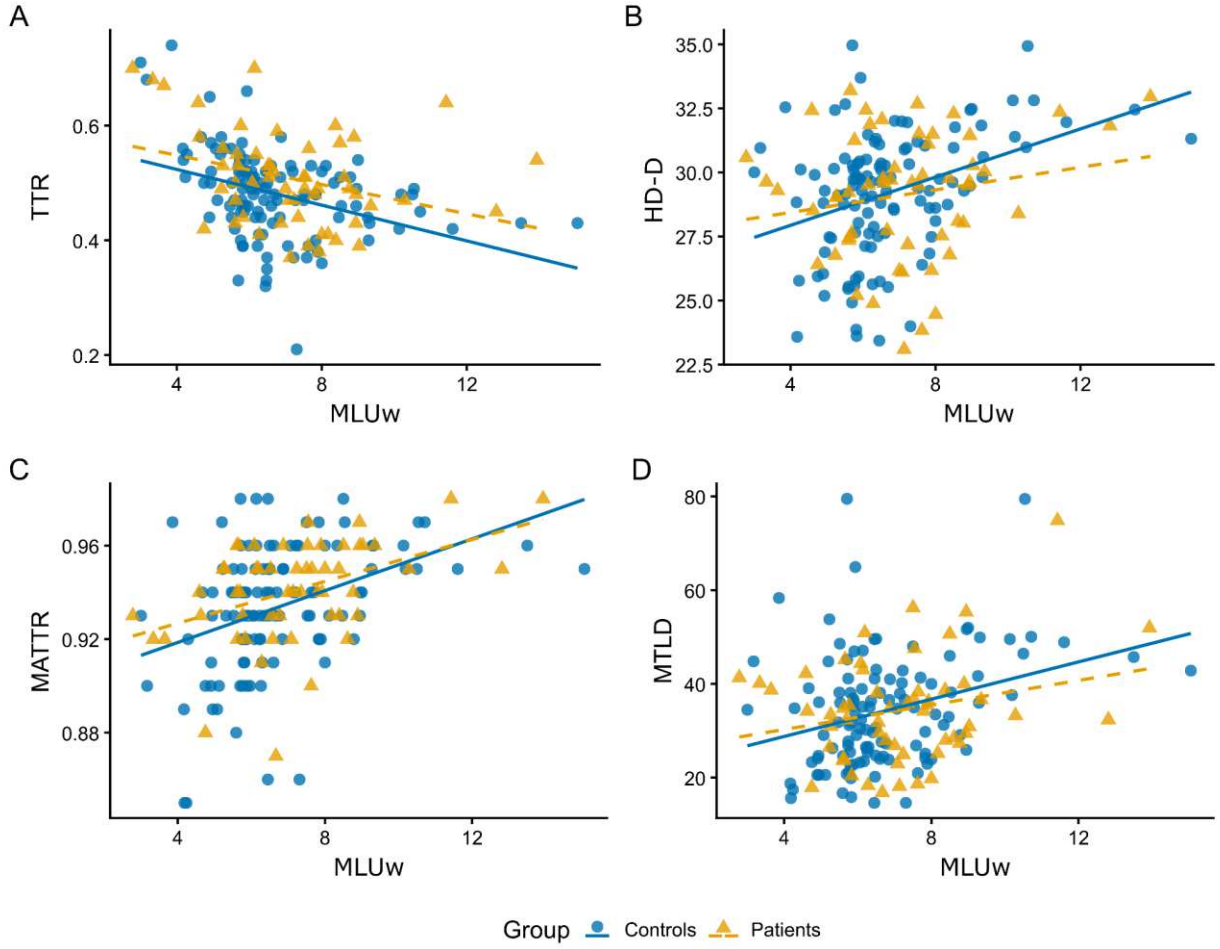
Correlations of lexical diversity measures to age in months for control and patient groups



Note*:* Panel A shows the association between age and TTR (type-token ratio), Panel B between age and HD-D (hypergeometric distribution D), Panel C between age and MATTR (moving average type-token ratio set at 8-word window), and Panel D between age and MTLD (measure of textual lexical diversity). Age was calculated in months. Scatter points represent individual observations for the control and patient groups. Dashed and solid lines represent the fitted linear trends for each group.

Figure 3

Correlations of lexical diversity measures to MLUw for control and patient groups



Note*:* Panel A shows the association between MLU (mean length of utterance in words) and TTR (type-token ratio), Panel B between MLUw and HD-D (hypergeometric distribution D), Panel C between MLUw and MATTR (moving average type-token ratio set at 8-word window), and Panel D between MLUw and MTLD (measure of textual lexical diversity). Scatter points represent individual observations for the control and patient groups. Dashed and solid lines represent the fitted linear trends for each group.

Group Comparisons

Addressing the second research question of this study, results from the linear models showed that, before and after surgery, patients did not significantly differ from the controls for either MATTR or MTLD, the two LD measures deemed more suitable from the correlation analyses. Addressing the third research question of this study, at the preoperative stage, no significant group differences were found between POSI and non-POSI patients. Additionally, Italian participants consistently

showed higher LD than English participants across measures, while no differences were found between English and Dutch participants. Results on preoperative language data are reported in Table 5 and Figure 4, while those on postoperative data are reported in Table 6 and Figure 5.

Both models on preoperative and postoperative data reported a significant effect of age (calculated in months). Results suggest that predicted LD, for both MATTR and MTLD, increases with age for all groups. Nevertheless, as illustrated in Figures 6 and 7, the patient groups show wider confidence intervals around the predicted trajectories than the control group, particularly the POSI group at the preoperative stage. This suggests greater uncertainty in the estimated effects, potentially reflecting smaller sample sizes and/or greater within-group variability. Both models also present a significant effect of language on both MATTR and MTLD.

Table 5

Linear regression results for preoperative group comparisons on MATTR and MTLD, with *p*-values adjusted using the Benjamini-Hochberg procedure to control for false discovery rate (FDR).

	MTLD				**MATTR**			
Predictors	**Estimate**	**SE**	**t**	**FDR *p***	**Estimate**	**SE**	**t**	**FDR *p***
(Intercept)	16.418	2.171	7.563	**<.001**	0.891	0.007	126.985	**<.001**
pHS vs POSI	1.315	1.471	0.894	0.497	0.001	0.005	0.174	0.862
Controls vs Patients (pHS + POSI)	0.926	0.578	1.601	0.167	-0.001	0.002	-0.690	0.536
Age (months)	0.119	0.017	7.038	**<.001**	0.000	0.000	6.118	**<.001**
Italian vs English	12.504	1.776	7.039	**<.001**	0.019	0.006	3.369	**0.002**
Dutch vs English	1.690	2.083	0.812	0.502	0.012	0.007	1.793	0.128

Note: *MTLD* = measure of textual lexical diversity; *MATTR* = moving average type-token ratio set at 8-word window; *POSI* = patients who later developed postoperative speech impairments (i.e., mutism or reduced speech), *pHS* = patients who did not develop postoperative speech impairments and present postoperative habitual speech.

Table 6

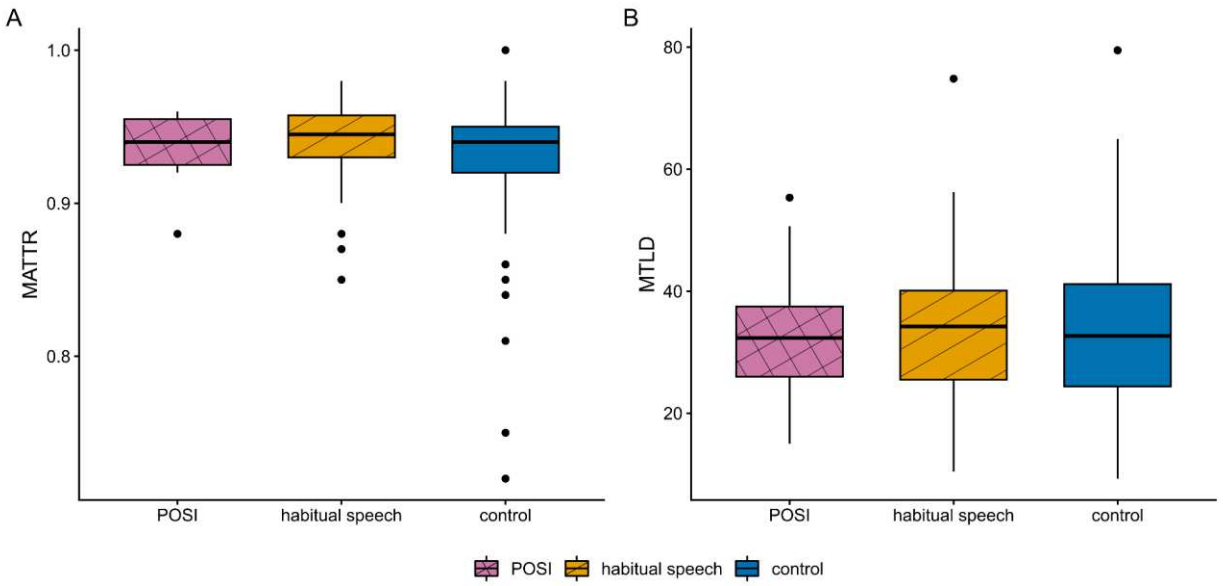
Linear regression results for postoperative group comparisons on MATTR and MTLD, with *p*-values adjusted using the Benjamini-Hochberg procedure to control for false discovery rate (FDR).

	MTLD				**MATTR**			
Predictors	**Estimate**	**SE**	**t**	**FDR *p***	**Estimate**	**SE**	**t**	**FDR *p***
(Intercept)	16.679	2.197	7.590	**<.001**	0.886	0.007	125.631	**<.001**
Patients (pHS + POSI) vs Controls	0.428	1.933	0.221	0.909	0.001	0.006	0.115	0.909
Age (months)	0.135	0.018	7.372	**<.001**	0.000	0.000	5.986	**<.001**
Italian vs English	11.307	1.989	5.683	**<.001**	0.018	0.006	2.826	**0.009**
Dutch vs English	2.411	2.293	1.052	0.368	0.014	0.007	1.969	0.072

Note: *MTLD* = measure of textual lexical diversity; *MATTR* = moving average type-token ratio set at 8-word window; *POSI* = patients who developed postoperative speech impairments (i.e., mutism or reduced speech), *pHS* = patients who did not develop postoperative speech impairments and present postoperative habitual speech

Figure 4

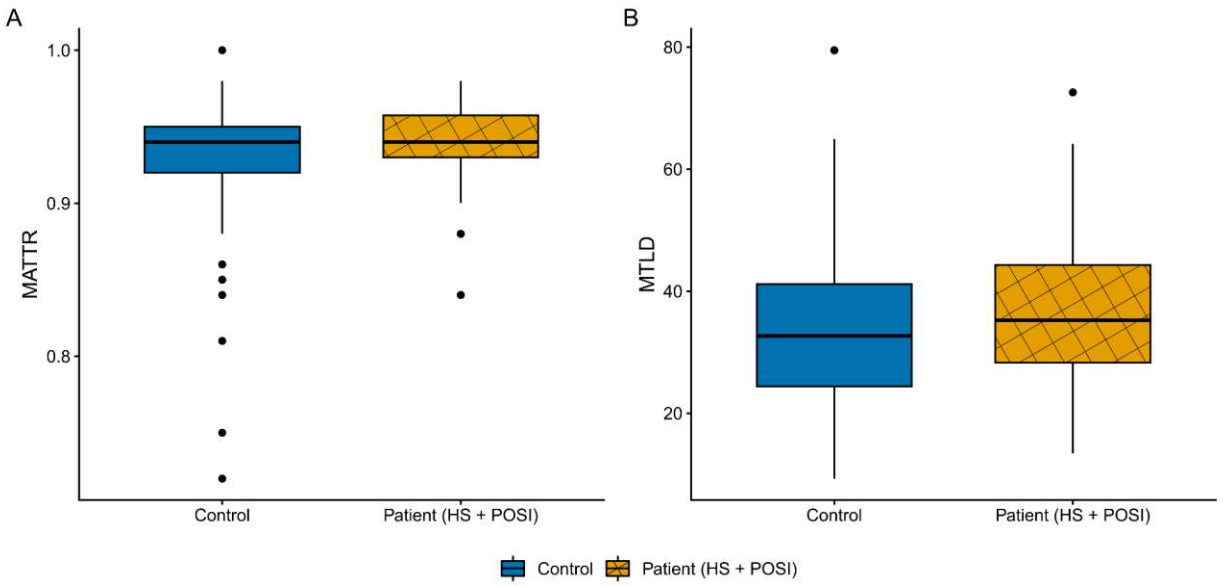
Distribution of preoperative lexical diversity measures across groups



Note: Panel A shows MATTR (moving-average type-token ratio, set at a 8-word window) and Panel B shows MTLD (measure of textual lexical diversity). The x-axis represents the three participant groups: patients who later developed postoperative speech impairments (POSI), patients who did not develop postoperative speech impairments and present postoperative habitual speech, and controls. The y-axis represents the lexical diversity scores. In each boxplot, the horizontal line indicates the median, the box represents the interquartile range, the whiskers indicate the range of non-outlying observations, and points denote outliers.

Figure 5

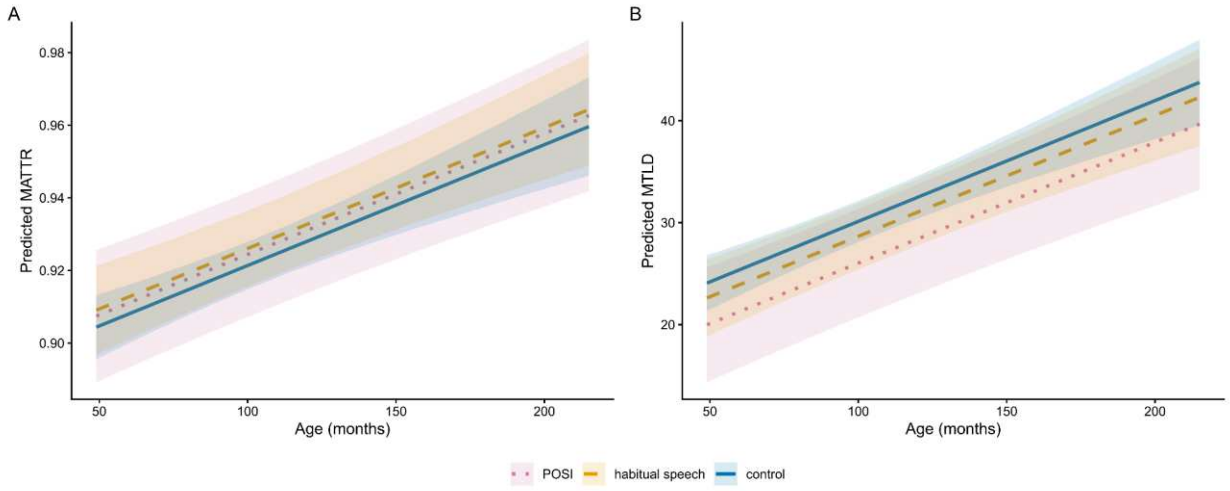
Distribution of postoperative lexical diversity measures across groups



Note: Panel A shows MATTR (moving-average type-token ratio, set at a 8-word window) and Panel B shows MTLD (measure of textual lexical diversity). The x-axis represents the two participant groups: the control group and the patient group, which included those patients who developed postoperative speech impairments (POSI) and those who presented postoperative habitual speech. The y-axis represents the lexical diversity scores. In each boxplot, the horizontal line indicates the median, the box represents the interquartile range, the whiskers indicate the range of non-outlying observations, and points denote outliers.

Figure 6

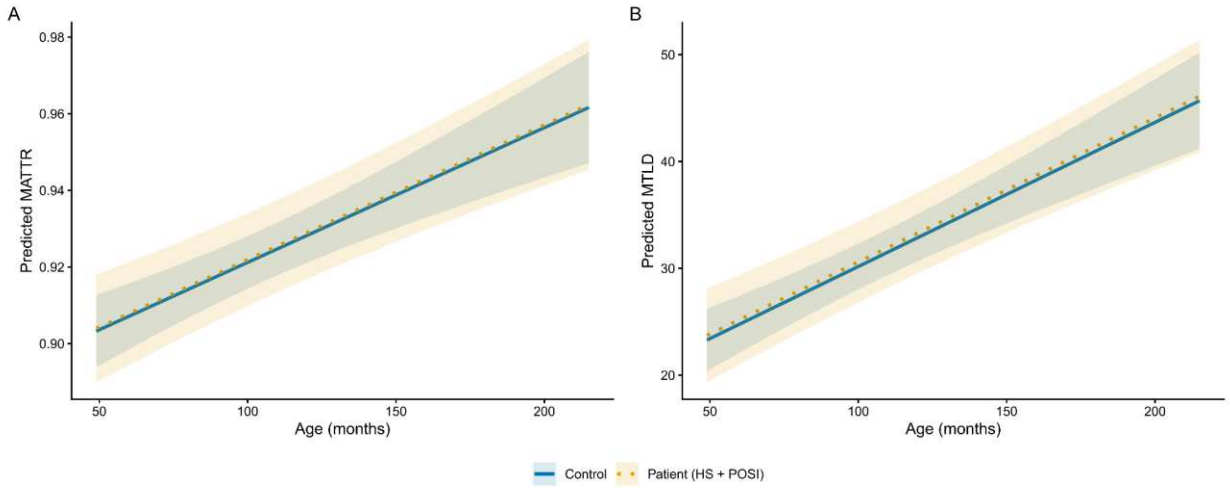
Partial effect of age on preoperative lexical diversity measures across groups



Note: Panel A displays the predicted values of MATTR (moving-average type-token ratio, set at a 8-word window) as a function of age (in months), whereas Panel B displays the predicted values of MTLD (measure of textual lexical diversity). Lines, accompanied by 95% confidence bands, represent model-predicted values of the three participant groups: patients who later developed postoperative speech impairments (POSI), patients who did not develop postoperative speech impairments and present postoperative habitual speech, and controls.

Figure 7

Partial effect of age on postoperative lexical diversity measures across groups



Note: Panel A displays the predicted values of MATTR (moving-average type-token ratio, set at a 8-word window) as a function of age (in months), whereas Panel B displays the predicted values of MTLD (measure of textual lexical diversity). Lines, accompanied by 95% confidence bands, represent model-predicted values of the two participant groups: patients who later developed postoperative speech impairments (POSI), patients who did not develop postoperative speech impairments and present postoperative habitual speech, and controls.

Individual Case Statistics

Addressing the second research question of this study, preoperatively, tests of deficit identified two of the twenty-eight English-speaking patients included as performing significantly worse than controls. Patient P15 showed significantly lower MATTR scores (p = 0.005) and patient P21 showed significantly lower MATTR scores (p = 0.044). Conversely, P17 showed significantly better performance than the control group on MTLD (p = 0.010), respectively. None of these patients subsequently developed POSI.

Addressing the third research question of this study, postoperatively, none of the 16 patients showed significantly lower LD compared to controls. The two patients with significantly reduced preoperative LD either lacked postoperative data (P15) or did not show a significantly reduced MATTR score postoperatively (P21; p = .219). There were no available postoperative data for patient P17. In addition, patients P10 and P18, whose preoperative performance was comparable to controls, presented significantly higher MTLD scores than controls after surgery (P10: p = .011, P18: p = .005). Pre- and postoperative individual results are reported in Table 7, together with additional demographic information about the patients.

Table 7

Individual lexical diversity raw scores and significant differences from control groups

					Preoperative language sample			**Postoperative language sample**		
ID	**Group**	**Tumor type**	**Tumor location**	**Preop. hydrocephalus**	**Raw score**	**Controls M(SD)**	***p* value**	**Raw score**	**Controls M(SD)**	***p* value**
P15	pHS	Pilocytic astrocytoma	Right cerebellar hemisphere, vermis	yes	0.85 (MATTR)	0.942 (0.015) (MATTR)	**.005**	n.a	-	-
P17	pHS	Not answered	Vermis	yes	41.35 (MTLD)	18.6 (4.55) (MTLD)	**.010**	n.a.	-	-
P21	pHS	Pilocytic astrocytoma	Right cerebellar hemisphere	no	0.87 (MATTR)	0.92 (0.01) (MATTR)	**.021**	0.93 (MATTR)	0.92 (0.01) (MATTR)	.219
P10	pHS	Pilocytic astrocytoma	Right and left cerebellar hemisphere, vermis, fourth ventricle	yes	34.20 (MTLD)	29.60 (4.30) (MTLD)	.337	50.49 (MTLD)	29.60 (4.30) (MTLD)	**.011**
P18	pHS	Medulloblastoma	Fourth ventricle	no	27.33 (MTLD)	24.18 (2.09) (MTLD)	.330	37.15 (MTLD)	24.18 (2.09) (MTLD)	**.005**

Note: Patients' ID were pseudo-anonymized. *MTLD* = measure of textual lexical diversity; *MATTR* = moving average type-token ratio set at 8-word window; *POSI* = patients who developed postoperative speech impairments (i.e., mutism or reduced speech), *pHS* = patients who did not develop postoperative speech impairments and present postoperative habitual speech; *n.a.* = postoperative data not available.

Discussion

Word retrieval in context has not been systematically examined either before surgery or during the acute postoperative phase in children with PFTs, despite such an assessment resembling everyday communication more closely and therefore potentially providing a more ecologically valid account of language functioning [24]. Hence, we first assessed the influence of sample length on four lexical diversity (LD) measures (i.e., TTR, HD-D, MTLD, and MATTR) and their developmental sensitivity to age and MLUw to identify the most suitable measure for our population of interest. MTLD and MATTR emerged as the most suitable measures, as they were robust to sample length, developmentally sensitive, and could be calculated from language samples of any length. We then compared patients' preoperative and postoperative LD (considering only MTLD and MATTR scores) to healthy controls, finding no significant results at the group level. Similarly, no evident trend of reduced LD was identified in the individual-case analyses, with two patients performing worse than controls before surgery and none after surgery. Lastly, we further investigated the reported association between preoperative language impairments and the development of POSI after surgery [10,13]. However, we found no evidence of poorer preoperative performance in patients who later developed POSI compared with those who did not, either at the group or individual level. Our findings are discussed in detail in the following sections.

Evaluation of sample length and developmental sensitivity of lexical diversity measures

To determine the most appropriate LD measure for language samples from children with PFTs, we evaluated whether each measure was affected by sample length and whether it was sensitive to developmental changes reflected in age and MLUw. Our results show that, across both healthy controls and patients, TTR was the only LD measure to be affected by sample length after controlling for age, decreasing as the length of the sample increases. This preliminary analysis was conducted to assess the potential confounding influence of sample length and to confirm the expected pattern reported in the literature [29,32–35]. We next investigated whether these LD measures were sensitive to developmental growth by examining their associations with age and MLUw, a reliable indicator of language development [37]. For both patients and controls, TTR was not associated with age, remaining relatively constant across the age range (Figure 2), whereas it was negatively associated with MLUw. These findings are consistent with previous reports of

TTR's limited developmental sensitivity and further support the recommendation to avoid its use in language sample analysis [31,35]. In contrast, among controls, HD-D, MTLD, and MATTR were all positively associated with age and MLUw, increasing in parallel and showing stronger correlations than TTR, with no significant differences among the three measures. Among patients, only MATTR showed significant positive associations with both age and MLUw. However, the stronger correlations observed for HD-D, MTLD, and MATTR relative to TTR in the patient group suggest that these measures may be more developmentally sensitive than TTR, despite only the association for MATTR reaching statistical significance.

Overall, our findings suggest that MATTR is the most suitable measure for comparing LD between children with PFTs and healthy controls, as it is robust to sample length while remaining sensitive to developmental changes in both groups. These results extend previous evidence supporting MATTR as a reliable LD measure in adults with aphasia [29] and children with DLD [35,36], while also supporting its developmental sensitivity in a broader developmental age range [35]. Our findings also support MTLD as a reliable alternative to MATTR. Consistent with previous evidence from aphasia and DLD [29,36], MTLD proved robust to sample length. Importantly, the present study provides the first evidence of its developmental sensitivity, as reflected by the correlations with age and MLUw and the reported age-related effects observed in the regression models. However, these effects were more pronounced in healthy controls than in patients. Lastly, although HD-D showed robustness to sample length and, for control data, developmental sensitivity, its minimum sample-length requirement limits its clinical applicability. Specifically, HD-D requires a minimum length of 42 words per sample, resulting in the exclusion of approximately 20% of patients who later developed POSI (N = 3), 7% of patients with postoperative habitual speech (N = 3), and 6% of controls (N = 7). Notably, those excluded were between 4 and 6 years of age (M = 4;6 years, SD = 0;11 months). Given the challenges of eliciting language samples from children, particularly in clinical populations, and the higher risk of developing POSI with younger age at tumor resection [3], HD-D may be less suitable for this population, although it might be a valuable alternative when longer language samples can be obtained. Taken together, these findings support the use of MATTR and MTLD in the subsequent analyses.

Patients' lexical diversity at the preoperative stage

To investigate whether children with PFTs exhibit reduced LD before surgery, we conducted both group-level comparisons and individual-case analyses relative to healthy controls. These analyses considered only MATTR and MTLD, as these measures showed robustness to sample length and developmental sensitivity, without requiring a minimum number of words to be computed. At the group level, patients who later developed POSI did not differ from those with postoperative habitual speech on either MATTR or MTLD. Similarly, when patients who developed POSI and those with postoperative habitual speech were considered together, they did not differ from healthy controls on either measure. Individual case statistics were then performed on only a subgroup of English-speaking patients (N = 28, 6 POSI), due to data availability. These analyses revealed some degree of individual variability but no consistent trend towards reduced LD in patients: two patients showed lower LD while one patient showed higher LD compared to controls, with no observable trend of POSI patients showing worse performance at the individual level. Based on the two studies that reported a link between the presence of preoperative language impairments (including word-finding difficulties) and a higher risk of developing POSI after surgery [10,13], we expected to observe more severely reduced preoperative LD in patients who later develop POSI compared to those with postoperative habitual speech. However, our findings are consistent with the study that reported no preoperative language differences between patients who later developed POSI and those who did not either at the semantic, lexical, morphosyntactic, and phonological levels [15]. Possibly, risk factors for POSI may lie in the presence of preoperative speech impairments rather than in language impairments, including reduced LD. Nevertheless, further research is required to confirm this interpretation.

The lack of preoperative differences for either MATTR or MTLD between patients and controls in our findings deviates from [11]. Indeed, in their study on picture-naming data, they reported that children with PFTs were either slower, less accurate, or both, than controls before surgery [11]. Given the partial overlap between our samples, as patient data came from the CMS database, we expected to observe similar word retrieval difficulties in narrative language data, reflected in reduced LD compared to controls. This divergence may reflect that, in time-constrained settings such as with the speeded picture naming test, patients might present poorer performances due to impaired processing speed, often reported in cases of cerebellar damage [4]. Alternatively, LD measures may not be sufficiently sensitive to detecting the subtle language

impairments that emerge in more demanding tasks. Additionally, a picture-supported narrative task such as the *Fish Story*, may elicit fewer lexical variation than spontaneous conversations [45]. Furthermore, word retrieval in isolation and word retrieval in context might tap into slightly different aspects of word finding [26]. Despite studies reporting how word finding difficulties are generally present both in isolation and in context, with correlation between picture naming scores and lexical retrieval in connected language [21], dissociations among the two have also been documented, such as in the context of aphasia. For example, people with dynamic aphasia might present impaired word retrieval in context with intact picture naming abilities [21]. In contrast, there are reports of individuals with aphasia failing to retrieve words in isolation, while preserving their ability in context [26]. A proposed explanation is that different variables may impact isolated versus contextual word retrieval. Indeed, lexical retrieval in context requires integration into grammatical structures and coordination with surrounding words, together with the influence of communicative goals and planning. Despite the greater interaction of cognitive processes involved in connected language, speakers can draw on grammatical structure and the meanings of related words to support lexical retrieval. Conversely, in a picture-naming task, isolated lexical retrieval is less flexible, requiring a search for a specific item, often without contextual cues [26]. Additionally, although related, LD measures the variety of words used in a language sample, while word-finding concerns the ability to access and produce a specific word [16]. It is therefore crucial to investigate the presence of lexical retrieval impairment considering both tasks that assess it in isolation and in context, as this combined approach can provide a more comprehensive and informative understanding of the underlying difficulty. When considered alongside previous reports of impaired picture naming, particularly in terms of lexical retrieval speed (see [11]), our findings suggests that children with PFTs may compensate for or mask these difficulties during spontaneous language, such as in narrative tasks, whereas they become more apparent in time-constrained tasks. Indeed, a study reports that children with PFTs might differ in the quality of the words they produce (e.g., concreteness) in their postoperative spontaneous language, suggesting that they may adapt their lexical choices to compensate for underlying language difficulties [28].

Patients' lexical diversity at the postoperative stage

We also examined acute postoperative assessments, in which patients (unfortunately only 38 of the 62 patients, 4 with POSI) were tested between 1 and 4 weeks after surgery. At the group level,

we found no difference for either MATTR or MTLD between patients (considering together patients who developed POSI and those who did not) and healthy controls. Concerning individual case statistics, LD appeared largely preserved in the postoperative phase, with none of the 16 patients included in the analysis showing a significant reduction compared to controls. In contrast, two patients exhibited significantly higher LD than controls: P10 and P18, whose preoperative performance had been comparable to that of controls. However, we expected to observe postoperative reduced LD in patients compared to healthy peers, given the reported reduced diversity in patients' spontaneous language several months or years after surgery [23,27], and the persistence of word-finding difficulties with picture naming task in the acute postoperative phase [12]. As mentioned above, LD measures may not be sufficiently sensitive to detect subtle language impairments. As noted in the preoperative results, even in the weeks after surgery, patients may still be able to use compensatory strategies if given sufficient time. Moreover, given the observable reduction in the number of patients successfully tested immediately after surgery, there may be a potential bias whereby only individuals with relatively preserved language abilities, and, consequently, better lexical retrieval, were able to complete the narrative task. In contrast, patients with more severe impairments may have been unable to complete the task and were therefore excluded from the analysis. Besides, since transient mutism and reduced speech arise within days after surgery and can present variable duration [3], the overlap between the symptom onset window and testing timing explains the reduction of the number of patients who developed POSI, possibly introducing selection bias.

While in picture naming a largely overlapping cohort continued to show below average performance in word retrieval [12], our findings from narrative language provide no overall evidence for (or against) a pre- or post-operative impairment in LD, in line with the results of a recently published study in children with PFTs [45]. Taken together, these findings suggest that the methods vary in their sensitivity, with measures of LD being less sensitive, while perhaps the composition or psycholinguistic profile of the words used in those narratives carrying greater informative value [15,28]. Future studies should compare patients with controls in other properties reflecting the processes of word retrieval in narrative language (e.g., frequency, imageability, age of acquisition), instead of (or in addition to) estimating the diversity of vocabulary. Furthermore, in the acute postoperative phase, outcomes are heterogeneous, with those individuals showing

improvement, maybe reflecting either the impact of the surgical procedure or the potential immediate relief from tumor mass removal.

When interpreting our finding of preserved postoperative LD alongside studies reporting reduced vocabulary range years after surgery [23,27], it is important to consider patients' medical courses and potential contributing factors. The reduced LD observed in patients assessed more than one year after resection may reflect the effects of postoperative adjuvant treatments. Indeed, approximately a few weeks after surgery, patients typically begin radiotherapy and/or chemotherapy, which have been associated with detrimental effects on cognitive functioning [4]. Additionally, disruptions to neural pathways during ongoing brain development, especially at a young age, may lead to new or more severe long-term impairments that were initially unnoticeable. The widening cognitive gap between patients and healthy peers is referred to as "growing into deficit" [46,47]. Accordingly, in children undergoing PFT resection, lexical retrieval difficulties may become more apparent in connected language only during the survivorship phase, potentially reflecting the impact of radio/chemotherapy or "growing into deficit." In contrast, these difficulties may be more readily detected using time-constrained lexical retrieval tasks, both before surgery and in the acute postoperative phase.

Limitations

Despite the common use of LD in language sample analysis, also in the context of individuals with language impairments [29,31], this measure presents some methodological limitations. Indeed, compared to single-word production tasks, in which responses are either correct or incorrect, word retrieval in context allows greater flexibility and, because the speaker's intended target is not always identifiable, it is harder to determine whether an error has occurred [26]. The current study was limited to examining whether patients exhibited reduced LD compared to healthy peers. Therefore, we did not assess the appropriateness of the words they produced. The presence of paraphasias may have inflated the apparent vocabulary richness of patients' samples, resulting in LD measures comparable to those of controls.

The reduced sample length did not allow for separate analysis of LD for nouns and for verbs, which would have been informative given that they can be differentially impaired [27]. Additionally, future research should incorporate supplementary measures of word-finding

difficulties in connected language, such as hesitations, fillers, pauses, and substitutions, with particular attention to the contexts in which they occur.

We also report significant differences in LD across the three languages studied. A plausible explanation is that the R package *koRpus* [41], used to compute the LD measures, operates on unlemmatized texts. Consequently, the consistently higher LD in Italian may partly reflect its greater morphological richness relative to English and Dutch. Thus, most likely, the differences between languages are not a true effect of differences in children's LD but relate to how a word is defined.

Using a picture book to elicit narrative language may have provided additional contextual support through visual cues, potentially facilitating word retrieval. Future studies should complement this method with tasks that do not rely on visual support, such as free recall of a story or personal experience, to elicit more language and minimize the influence of external cues.

Another potential limitation is a post-surgical selection bias, as fewer patients were successfully tested immediately after surgery. Consequently, only individuals with relatively preserved language abilities may have been able to complete the narrative task, while those with more severe impairments were unable to participate and were thus excluded. This issue is particularly relevant for patients who developed POSI, since it typically emerges within days after surgery and varies in duration. The overlap between its onset window and the timing of postoperative assessment likely reduced the number of patients with POSI included in the postoperative analysis, hence limiting their representation in the sample.

Finally, it should be noted that the narrative task itself may have introduced selection bias, as only patients able to complete the task were included, whereas picture-naming tasks might be generally less demanding. This may have favored the inclusion of relatively less impaired patients and should therefore be considered when interpreting the discrepancy between the narrative and isolated word-retrieval findings.

Conclusion

This study provides an important methodological contribution. Our findings indicate that MATTR and MTLD are the most suitable LD measures for short language samples in pediatric clinical populations, being less sensitive to sample length and more developmentally appropriate than TTR. These results support the use of MATTR and MTLD in future research and clinical

assessments. However, we also note that these measures may lack sensitivity, compared to more lexically-specific and time-constrained measurements, such as accuracy and reaction times derived from picture naming. In fact, in our study, both group-level and individual analyses showed preserved LD in children with PFTs, with no differences compared to controls or between patients who later developed POSI and those who did not. The discrepancy with prior picture-naming findings might reflect differences in processing speed demands or that patients might be adopting compensatory strategies, indicating that impairments may emerge more clearly under time pressure than in contextualized speech. Hence, maintaining ecological validity may require moving (semi-)spontaneous language analyses towards measures which reflect the quality of vocabulary, rather than its size or diversity, potentially providing further insights into the added processing costs of compensating for underlying language impairments.

Postoperatively, LD also appeared largely preserved at both group and individual levels. Importantly, the contrast to previous research reporting postoperative impaired LD measures may indicate that later-emerging difficulties reported in the literature may be linked to adjuvant treatments (e.g., radio/chemotherapy), as well as developmental factors like "growing into deficit".

Overall, despite the importance of ecologically valid assessments that better reflect everyday language use, our findings highlight the value of direct language testing, as subtle impairments may otherwise go undetected. They also support complementing these assessments with psycholinguistic analyses of the vocabulary used in connected language. Together, structured and naturalistic assessments can help clinicians better identify not only subtle impairments but also possible patients' points of strength, thereby informing more targeted interventions.

Ethical considerations

The current study used patient data from the European Study of CMS. The Research Ethics Committees of the Capital Region (H-6-2014-002) in Denmark have their approval for the collection of data for this project, and the study was authorized in the UK afterwards.

Author contributions

Elisa Gottardi: Conceptualization, Methodology, Writing - Original draft, Writing - Review and Editing, Data curation, Formal analysis

Barbara Höhle: Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing - Review and Editing

Natalie Boll-Avetisyan: Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing - Review and Editing

Ditte Boeg Thomsen: Conceptualization, Methodology, Supervision, Writing - Review and Editing

Jonathan Kjær Grønbæk: Conceptualization, Methodology, Supervision, Writing - Review and Editing

Aliene Reinders: Data curation

Cheyenne Svaldi: Data curation, Writing - Review and Editing

Annet Kingma: Resources

Karin Persson: Resources, Writing - Review and Editing

René Mathiasen: : Resources

Christine Dahl: Resources

Andrea Carai: Resources

Bianca Andreozzi: Resources

Angela Mastronuzzi: Resources

Barry Pizer: Resources, Writing - Review and Editing

Colin Thorbinson: Resources, Writing - Review and Editing

Kristian Aquilina: Resources, Writing - Review and Editing

Eelco Hoving: Resources

Marianne Juhler: Resources, Writing - Review and Editing

Roel Jonkers: Conceptualization, Writing - Review and Editing

Vânia de Aguiar: Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing - Review and Editing

Funding statement

This study is part of the doctoral research project of E.G. within the International Doctorate in Experimental Approaches to Language and Brain (IDEALAB), funded via the University of Potsdam and the University of Groningen.

This publication is supported by funding awarded to VdA for the project Verb Processing and Verb Learning in Children With Paediatric Posterior Fossa Tumors (with the file number VI.Vidi.201.003) of the research program NWO-Talentprogramma Vidi SGW 2020 financed by the Dutch Research Council (NWO).

The European Study of CMS is supported by the Danish Childhood Cancer Foundation, the Swedish Childhood Cancer Foundation and the Brain Tumour Charity.

DBT and JKG: Inge Lehmann grant for the project The Cerebellum in Language and Speech: Perspectives from Children with Brain Tumors (grant number 10.46540/4302-00027B) from the Independent Research Fund Denmark.

JKG, (European study of the Cerebellar Mutism Syndrome): The Danish Childhood Cancer Foundation, The Swedish Childhood Cancer Foundation, Brain Tumor Charity.

KP The Swedish Childhood Cancer Foundation

Acknowledgements

The European CMS study group is acknowledged for their participation in the data collection.

References

1. Dörner L, Fritsch MJ, Stark AM, Mehdorn HM. Posterior fossa tumors in children: how long does it take to establish the diagnosis? Childs Nerv Syst. 2007;23:887–90. https://doi.org/10.1007/s00381-007-0323-8

2. Fabozzi F, Margoni S, Andreozzi B, Musci MS, Del Baldo G, Boccuto L, et al. Cerebellar mutism syndrome: From pathophysiology to rehabilitation. Front Cell Dev Biol. 2022;10:1082947. https://doi.org/10.3389/fcell.2022.1082947

3. Grønbæk JK, Wibroe M, Toescu S, Frič R, Thomsen BL, Møller LN, et al. Postoperative speech impairment and surgical approach to posterior fossa tumours in children: a prospective European multicentre cohort study. Lancet Child Adolesc Health. 2021;5:814–24. https://doi.org/10.1016/S2352-4642(21)00274-1

4. Cámara S, Fournier MC, Cordero P, Melero J, Robles F, Esteso B, et al. Neuropsychological Profile in Children with Posterior Fossa Tumors with or Without Postoperative Cerebellar Mutism Syndrome (CMS). The Cerebellum. 2020;19:78–88. https://doi.org/10.1007/s12311-019-01088-4

5. Svaldi C, Ntemou E, Jonkers R, Kohnen S, De Aguiar V. Language outcomes in children who underwent surgery for the removal of a posterior fossa tumor: A systematic review. Eur J Paediatr Neurol. 2024;48:129–41. https://doi.org/10.1016/j.ejpn.2023.12.005

6. Gudrunardottir T, Morgan AT, Lux AL, Walker DA, Walsh KS, Wells EM, et al. Consensus paper on post-operative pediatric cerebellar mutism syndrome: the Iceland Delphi results. Childs Nerv Syst. 2016;32:1195–203. https://doi.org/10.1007/s00381-016-3093-3

7. Gudrunardottir T, Sehested A, Juhler M, Grill J, Schmiegelow K. Cerebellar mutism: definitions, classification and grading of symptoms. Childs Nerv Syst. 2011;27:1361–3. https://doi.org/10.1007/s00381-011-1509-7

8. Pollack IF. Posterior Fossa Syndrome. Int Rev Neurobiol. Elsevier; 1997. p. 411–32. https://doi.org/10.1016/S0074-7742(08)60362-1

9. Wickenhauser ME, Khan RB, Raches D, Ashford JM, Robinson GW, Russell KM, et al. Characterizing Posterior Fossa Syndrome: A Survey of Experts. Pediatr Neurol. 2020;104:19–22. https://doi.org/10.1016/j.pediatrneurol.2019.11.007

10. Di Rocco C, Chieffo D, Frassanito P, Caldarelli M, Massimi L, Tamburrini G. Heralding Cerebellar Mutism: Evidence for Pre-surgical Language Impairment as Primary Risk Factor in Posterior Fossa Surgery. The Cerebellum. 2011;10:551–62. https://doi.org/10.1007/s12311-011-0273-2

11. Persson K, Boeg Thomsen D, Fyrberg Å, Castor C, Aasved Hjort M, Andreozzi B, et al. Preoperative word-finding difficulties in children with posterior fossa tumours: a European cross-sectional study. Childs Nerv Syst. 2024;40:87–97. https://doi.org/10.1007/s00381-023-06119-4

12. Persson K, Grønbæk J, Tiberg I, Fyrberg Å, Castor C, Andreozzi B, et al. Postoperative word-finding difficulties in children with posterior fossa tumours: a crosslinguistic European cohort study. Childs Nerv Syst. 2025;41:128. https://doi.org/10.1007/s00381-025-06787-4

13. Bianchi F, Chieffo DPR, Frassanito P, Di Rocco C, Tamburrini G. Cerebellar mutism: the predictive role of preoperative language evaluation. Childs Nerv Syst. 2020;36:1153–7. https://doi.org/10.1007/s00381-019-04252-7

14. Chieffo D, Tamburrini G, Caldarelli M, Di Rocco C. Preoperative neuropsychological and behavioral evaluation of children with thalamic tumors. J Neurosurg Pediatr. 2014;13:507–13. https://doi.org/10.3171/2014.2.PEDS13352

15. Reinders A, Svaldi C, Kingma A, Grønbæk JK, Thomsen DB, Persson K, et al. Analysis of Presurgical Language in Children with Posterior Fossa Tumours Relative to Postoperative Speech Outcomes: Findings from the European CMS Study. The Cerebellum. 2026;25:48. https://doi.org/10.1007/s12311-026-01987-3

16. Messer D, Dockrell JE. Children's Naming and Word-Finding Difficulties: Descriptions and Explanations. J Speech Lang Hear Res. 2006;49:309–24. https://doi.org/10.1044/1092-4388(2006/025)

17. Ahmed R, Reinders A, Svaldi C, Kingma A, Persson K, Thomsen DB, et al. Language Processing in Posterior Fossa Tumour Patients: Psycholinguistic Insights into the Word-Finding Ability. The Cerebellum. 2026;25:84. https://doi.org/10.1007/s12311-026-02026-x

18. King M, Hernandez-Castillo CR, Poldrack RA, Ivry RB, Diedrichsen J. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. Nat Neurosci. 2019;22:1371–8. https://doi.org/10.1038/s41593-019-0436-x

19. LeBel A, D'Mello AM. A seat at the (language) table: incorporating the cerebellum into frameworks for language processing. Curr Opin Behav Sci. 2023;53:101310. https://doi.org/10.1016/j.cobeha.2023.101310

20. Mariën P, Ackermann H, Adamaszek M, Barwood CHS, Beaton A, Desmond J, et al. Consensus Paper: Language and the Cerebellum: an Ongoing Enigma. The Cerebellum [Internet]. 2014 [cited 2026 Aug 1]; https://doi.org/10.1007/s12311-013-0540-5

21. Herbert R, Hickin J, Howard D, Osborne F, Best W. Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? Aphasiology. 2008;22:184–203. https://doi.org/10.1080/02687030701262613

22. Wibroe M, Cappelen J, Castor C, Clausen N, Grillner P, Gudrunardottir T, et al. Cerebellar mutism syndrome in children with brain tumours of the posterior fossa. BMC Cancer. 2017;17:1–7. https://doi.org/10.1186/s12885-017-3416-0

23. Svaldi C, Galli J-I, Paquier P, Keulen S, Van Elp H, Catsman-Berrevoets C, et al. Errors in the Spontaneous Language of Survivors of Pediatric Cerebellar Tumors. The Cerebellum. 2025;24:26. https://doi.org/10.1007/s12311-024-01754-2

24. Rofes A, Talacchi A, Santini B, Pinna G, Nickels L, Bastiaanse R, et al. Language in individuals with left hemisphere tumors: Is spontaneous speech analysis comparable to formal testing? J Clin Exp Neuropsychol. 2018;40:722–32. https://doi.org/10.1080/13803395.2018.1426734

25. Fergadiotis G, Wright HH. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology. 2011;25:1414–30. https://doi.org/10.1080/02687038.2011.603898

26. Kavé G, Goral M. Do age-related word retrieval difficulties appear (or disappear) in connected speech? Aging Neuropsychol Cogn. 2017;24:508–27. https://doi.org/10.1080/13825585.2016.1226249

27. De Smet HJ. Neurolinguistics and the cerebellum: An analysis of speech and language disturbances resulting from acquired cerebellar lesions. [Doctoral dissertation]. Vrije Universiteit Brussel; 2009.

28. Svaldi C, Paquier P, Keulen S, Van Elp H, Catsman-Berrevoets C, Kingma A, et al. Characterising the Long-Term Language Impairments of Children Following Cerebellar Tumour Surgery by Extracting Psycholinguistic Properties from Spontaneous Language. The Cerebellum. 2023;23:523–44. https://doi.org/10.1007/s12311-023-01563-z

29. Fergadiotis G, Wright HH, West TM. Measuring Lexical Diversity in Narrative Discourse of People With Aphasia. Am J Speech Lang Pathol. 2013;22. https://doi.org/10.1044/1058-0360(2013/12-0083)

30. Wright HH, Silverman S, Newhoff M. Measures of lexical diversity in aphasia. Aphasiology. 2003 ;17:443–52. https://doi.org/10.1080/02687030344000166

31. Yang JS, Rosvold C, Bernstein Ratner N. Measurement of Lexical Diversity in Children's Spoken Language: Computational and Conceptual Considerations. Front Psychol. 2022;13:905789. https://doi.org/10.3389/fpsyg.2022.905789

32. Covington MA. MATTR user manual (CASPR research report 2007–05). Atheus, GA.; 2007.

33. McCarthy PM, Jarvis S. vocd: A theoretical and empirical evaluation. Lang Test. 2007;24:459–88. https://doi.org/10.1177/0265532207080767

34. McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behav Res Methods. 2010;42:381–92. https://doi.org/10.3758/BRM.42.2.381

35. Charest M, Skoczylas MJ, Schneider P. Properties of Lexical Diversity in the Narratives of Children With Typical Language Development and Developmental Language Disorder. Am J Speech Lang Pathol. 2020;29:1866–82. https://doi.org/10.1044/2020_AJSLP-19-00176

36. Kapantzoglou M, Fergadiotis G, Auza Buenavides A. Psychometric Evaluation of Lexical Diversity Indices in Spanish Narrative Samples From Children With and Without Developmental Language Disorder. J Speech Lang Hear Res. 2019;62:70–83. https://doi.org/10.1044/2018_JSLHR-L-18-0110

37. Rice ML, Smolik F, Perpich D, Thompson T, Rytting N, Blossom M. Mean Length of Utterance Levels in 6-Month Intervals for Children 3 to 9 Years With and Without Language Impairments. J Speech Lang Hear Res. 2010;53:333–49. https://doi.org/10.1044/1092-4388(2009/08-0183)

38. Frizelle P, Thompson PA, Mcdonald D, Bishop DVM. Growth in syntactic complexity between four years and adulthood: evidence from a narrative task. J Child Lang. 2018;45:1174–97. https://doi.org/10.1017/S0305000918000144

39. Bishop DV. Expression, reception and recall of narrative instrument: ERRNI. Psychological Corporation/Harcourt Assessment; 2004.

40. ELAN. Nijmegen: Max Planck Institute for Psycholinguistics; 2025. https://archive.mpi.nl/tla/elan

41. Michalke M. koRpus: Text Analysis with Emphasis on POS Tagging, Readability, and Lexical Diversity. 2012. p. 0.13-8. https://CRAN.R-project.org/package=koRpus.

42. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull. 1980;87:245–51. https://doi.org/10.1037/0033-2909.87.2.245

43. Crawford JR, Garthwaite PH, Ryan K. Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates. Cortex. 2011;47:1166–78. https://doi.org/10.1016/j.cortex.2011.02.017

44. Rittmo J, McIntosh R. singcar: Comparing single cases to small samples in R. J Open Source Softw. 2021;6:3887. https://doi.org/10.21105/joss.03887

45. Persson K, Boeg Thomsen D, Andersson K, Tiberg I, Castor C, Grønbæk J, et al. Narrative Ability in Swedish Children Treated for Posterior Fossa Tumours: Macro- and Microstructural Performance Before and 1–4 Weeks After Surgery. Int J Lang Commun Disord. 2026;61:e70248. https://doi.org/10.1111/1460-6984.70248

46. Dennis M, Spiegler BJ, Juranek JJ, Bigler ED, Snead OC, Fletcher JM. Age, plasticity, and homeostasis in childhood brain disorders. Neurosci Biobehav Rev. 2013;37:2760–73. https://doi.org/10.1016/j.neubiorev.2013.09.010

47. Kennard MA. Reactions of Monkeys of Various Ages to Partial and Complete Decortication. J Neuropathol Exp Neurol. 1944;3:289. https://doi.org/10.1097/00005072-194407000-00007

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GottardiLexicaldiversityTables.docx